CDMC™

EDMCouncil

Cloud Data Management Capabilities

# Cloud Data Management Capabilities Framework

Best Practices for
Managing Data in the Cloud

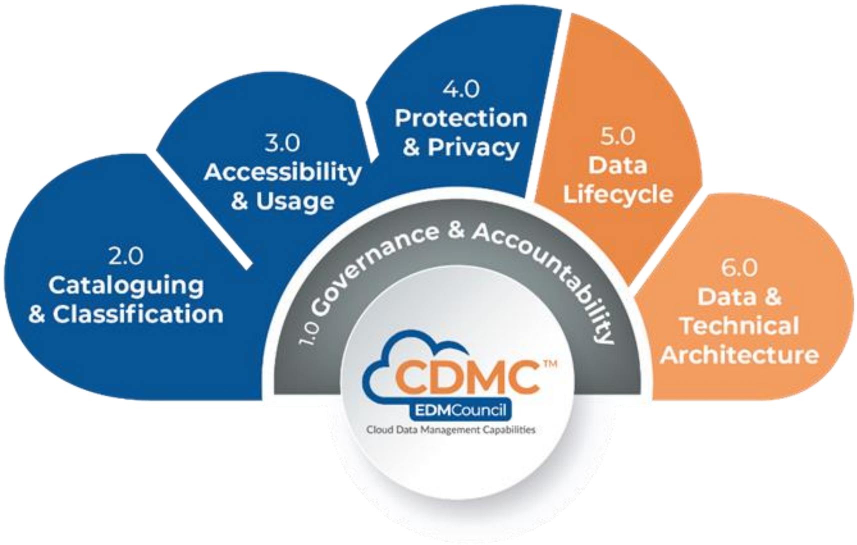
**Version 1.1.1**

September 2021

EDMCouncil®

edmcouncil.org/page/CDMC

## THE CDMC FRAMEWORK


2.0
Cataloguing & Classification
3.0
Accessibility & Usage
4.0
Protection & Privacy
5.0
Data Lifecycle
6.0
Data & Technical Architecture
1.0 Governance & Accountability
CDMC™
EDMCouncil
Cloud Data Management Capabilities

This document is a constituent part of the Cloud Data Management Capabilities (CDMC™) model (“the Model”) and is provided as a free license to any organization registered with EDM Council Inc. (“EDM Council”) as a recipient (“Recipient”) of the document. While this is a Free License available to both members and non-members of the EDM Council, acceptance of the CDMC Terms of Use is required to protect the Recipient’s use of proprietary EDMC property and to notify the Recipient of future updates to the Model.

CDMC™ and all related materials are the sole property of EDM Council Inc. All rights, titles and interests therein are vested in the EDM Council. The Model and related material may be used freely by the Recipient for their own internal purposes. It may only be distributed beyond the Recipient’s organization with prior written authorization of EDM Council. The Model may only be used by the Recipient for commercial purposes or external assessments if the Recipient’s organization has entered into a separate licensing and Authorized Partner Agreement with EDM Council governing the terms for such use.

Please accept these CDMC™ Terms of Use by registering at:

https://app.smartsheet.com/b/form/6e2b0bf4a3024affb98daad174b08483

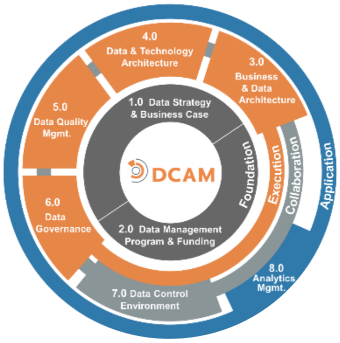
## FOREWORD – JOHN BOTTEGA, EDMC PRESIDENT

When industry identifies a challenge, it's amazing what can be done when talented people collaborate. This is the underlying story of **CDMC – Cloud Data Management Capability Framework**.

The art of data management has evolved. Once thought of as a *behind the scenes* technology function, understanding, curating, protecting and using our information resource is a front and center business, technology and operations function. Data is now the life-blood of our industry and our personal lives. As data professionals, we have a responsibility to ensure information is accurate, timely, trusted, and protected and that it is being put to use effectively and ethically.

It is this goal that has propelled the profession of data management. Chief Data Officers, Heads of Data Quality, Data Governance and Data Architecture are becoming commonplace in our businesses. We now bear the responsibility of curating information from a defensive posture—controlling risk, privacy, safety and security, as well as from an offensive posture—increasing revenue, penetrating new markets, developing new products and services.

To better equip the data professional, the EDM Council developed a data management best practice framework known as **DCAM – Data Management Capability Assessment Framework.**


4.0 Data & Technology Architecture
3.0 Business & Data Architecture
5.0 Data Quality Mgmt.
1.0 Data Strategy & Business Case
DCAM
Foundation
Execution
Collaboration
Application
6.0 Data Governance
2.0 Data Management Program & Funding
8.0 Analytics Mgmt.
7.0 Data Control Environment

DCAM codified data management capabilities, giving the data professional a runbook to build and sustain a data management program. This model prompted several EDMC members to reach out to the Council—to ask us to facilitate an effort to build a similar model for the unique capabilities of data management in the cloud.

With a handful of members, the CDMC – Cloud Data Management Capability Workgroup began. In short order, this effort drew in over 100 companies and over 300 data and business professionals and engineers, including the top cloud service providers, financial institutions, technology companies, and major consultant and advisory firms.

For 16 months, this team worked tirelessly to build a cloud data management framework that would help the industry better manage data in the cloud, better protect data in the cloud, and better enable organizations to realize the benefits of the cloud environment.

CDMC™ EDMCouncil
Cloud Data Management Capabilities

With great appreciation and pride, the EDM Council, working with so many talented people, can now release the CDMC – Cloud Data Management Capability Framework – as a free-license model, to the industry. Thank you to all who contributed – you should feel very proud of the work you did and the benefit this will bring to the industry.

And to those future users of the CDMC Framework, we welcome your feedback. EDMC is committed to keeping this framework current and always on topic. Use the following link to provide your thoughts, ideas and suggestions as we continue to improve and enhance this work: https://forms.monday.com/forms/342ed5577937d03d7cf5ef39a6e72e0a?r=use1

Sincerely,
John Bottega
President, EDM Council

## ACKNOWLEDGEMENTS

We would like to provide special acknowledgement to our CDMC Co-Chairs Oli Bage (LSEG) and Richard Perris (Morgan Stanley) for both their founding inspiration in advocating the CDMC Project to the EDM Council and for their extraordinary CDMC contributions and leadership over the last 18 months. Additionally, special thanks to Morgan Stanley for donating the initial draft of cloud principles that helped jump start the CDMC Project in the early days. Finally, special acknowledgement to our CDMC Project Manager, Jubair Patel (Microsoft formerly with Capco), who with steadfast support from the Capco team, kept the global CDMC project on track and was also an exemplary cloud subject matter contributor.

Over 100 companies have contributed to the production of the CDMC Framework:

- **Cloud Service Providers**: Amazon AWS, Google, IBM and Microsoft
- **Leading financial organizations**, including: Barclays, Citi Bank, Credit Suisse, Deutsche Bank, DTCC, Fannie Mae, Freddie Mac, Goldman Sachs, HSBC, JP Morgan, LSEG, M&G, Morgan Stanley, Societie Generale, Standard Bank, Sterling National Bank, TD Bank and UBS
- **Other major organizations**, including: CPA Canada and Schneider Electric
- **Technology Providers**, including: BigID, Collibra, Informatica, Privitar, Securiti, Solidatus and Snowflake
- **Consultancies and System Integrators**, including: Accenture, Capco, KPMG and Ortecha

EDM Council would like to thank the 300+ individuals who have participated. Those who have provided permission to be named are listed in the following document:

https://edmcouncil.org/resource/resmgr/cdmc_master/CDMC_Framework_Acknowledgeme.pdf

## REVISION HISTORY

| Date | Description | Version |
|---|---|---|
| September 2021 | Initial release of version 1 of full CDMC Framework | V1.1.1 |

## CONTENTS

# INTRODUCTION

## PURPOSE

Digital transformation is fundamentally changing how we do business – personally and professionally. Much of this transformation is taking place in the cloud environment across the globe. Cloud implementations are occurring in all sectors across all industries. There are many benefits of managing and storing data in a cloud environment, including cost savings, flexibility, mobility, improved information security, increased collaboration, and realizing new insights within an organization's *data assets*.

As with any new technology, cloud computing entails many challenges. New cloud implementations face a variety of data, technology and planning difficulties. There remains a lack of consistent industry best practices for applying *data management* capabilities during migrations to and operations in single, multiple and hybrid cloud environments.

Consequently, an organization will likely face cost and complexity risks when adopting cloud computing technologies. Adoption can be especially difficult for regulated entities that must demonstrate precise, consistent data control in both on-premises and cloud environments. *Cloud service providers* (*CSP*s) and technology providers also face complexity as they seek to understand the *data management* priorities of organizations, resulting in challenges to improving their cloud implementations.

The Cloud Data Management Capabilities (CDMC™) Framework defines the best practice capabilities necessary to manage and control data in cloud environments. The creation of this framework represents an important milestone in the global adoption of industry best practices for *data management*. The overall objective is to build trust, confidence and dependability for the adoption of cloud technologies, offering benefits to each of the constituencies within the cloud ecosystem:

- **Cloud Service and Technology Consumers** – provides a structured framework of auditable *processes* and controls, especially for sensitive data.
- ***Cloud service providers*** – provides requirements and controls that can be automated within *CSP* platforms, accelerating adoption and increasing market confidence.
- **Application, Technology and Data Providers** – applies standard, certified CDMC capabilities and controls to services and solutions to ensure a high degree of reliability and operational effectiveness.
- **Consultants and System Integrators** – enables training and assessments, gap analysis, strategy development, and execution services for end clients adopting cloud technologies.
- **Regulators** – provides industry guidance for auditing and validating key cloud environment controls, especially for sensitive data.

CDMC is a best practice assessment and certification framework for managing and controlling data in single, multiple, and hybrid cloud environments. CDMC is used to assess the capabilities of an organization that are necessary to support controlled integration and migration to the cloud environments. The framework focuses and expands on capabilities critical to controlling important and sensitive data and highlights features of contemporary cloud platforms that present opportunities for standardization and automation of *data management* and control.

Though CDMC is a standalone framework, it assumes that an organization already has a strong foundation of *data management* capabilities. A broader set of capabilities is covered in other frameworks such as the *Data Management Capability Assessment Model* (DCAM®) of the EDM Council. Effective *data management* fundamentals, together with the features and capabilities defined in CDMC, will enable an organization to build trustworthy and secure cloud environments—both now and well into the future.

## APPROACH

CDMC was produced by the ***EDM Council CDMC Work Group*** formed in May 2020 with over 300 individual business executives, engineers, technologists and data professionals. The group includes participants from over 100 organizations across the globe, including major *CSP*s, technology service organizations, privacy firms and major consultancy and advisory firms. The objectives of the initiative were to:

- Develop a framework that provides direction and guidance on core *data management* capabilities in cloud *data management* aligned with industry best practices.
- Develop a consistent CDMC scoring *model* for industry organizations to measure maturity and readiness against the cloud *data management* capabilities.
- Collaborate with cloud service and technology providers and industry organizations on a set of priorities for accelerating capabilities for cloud migration and implementations while allowing cloud service and technology providers the opportunity to apply their unique innovations and services to meet these industry requirements.
- Establish methods to continuously improve the CDMC Framework and facilitate training and education on these best practices.

The structure of CDMC and the approach to its creation leveraged the structure and approach of the DCAM® framework, which the EDM Council has maintained since 2014.

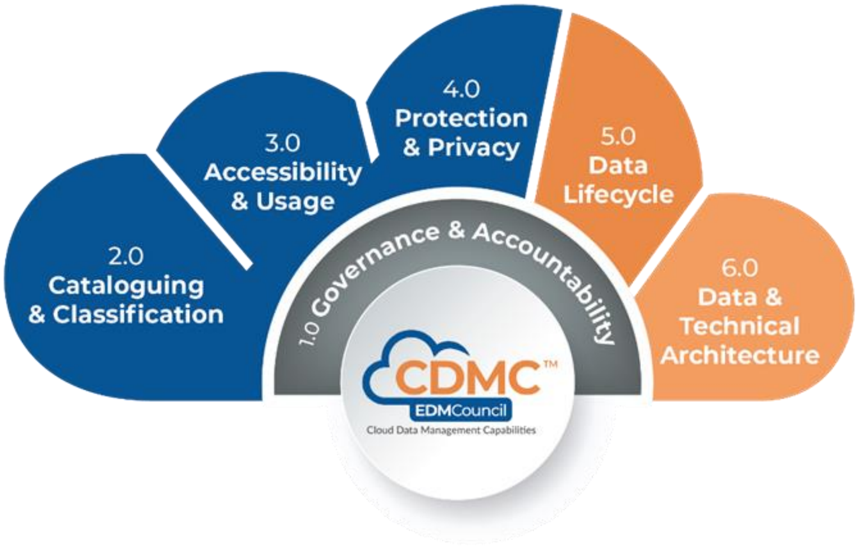
## CDMC – A FRAMEWORK FOR CLOUD DATA MANAGEMENT

Many organizations must establish a broad set of controls to manage data responsibly and comply with applicable regulatory entities. *Standards* and best practices enable an organization to harness the enormous opportunity offered by cloud technologies while avoiding the challenges of developing and adapting home-grown controls and spending time on isolated feature requests between individual companies and *CSP*s.

Controlling data in cloud environments requires a complex set of *data management* capabilities:

- An organization must establish clear accountability, controls and governance for data migrated to or created in cloud environments.
- A critical requirement is always to know what data resides in cloud environments and the sensitivity of each of the *data assets*. Such tracking is essential to automating controls for data access and use. Tracking is also vital to enforcing the controls and maintaining *evidence* for required transparency, security, and protection levels.
- *Data management* controls must be established throughout the *data lifecycle*.
- *Data assets* must be fit-for-purpose and kept to required schedules for retention and archiving.
- As with on-premises *data assets*, the design of the *data architecture* and configuration of supporting technologies are important for ensuring that business objectives are met.

CDMC captures the requirements for these capabilities in six areas. These six Components of the framework include 14 Capabilities and a total of 37 Sub-capabilities. The definition and scope of each component are presented below:



*1.0 Governance & Accountability*

The **Governance & Accountability** component is a set of capabilities that ensure an organization has clear accountability, controls and governance for data migrated to or created in cloud environments. These capabilities provide the foundation of well-governed business cases, effective data ownership, governance of data sourcing and consumption and management of *data sovereignty* and cross-border data movement risks.

This CDMC component helps to:

- Define business cases for managing data in cloud environments, including a value realization framework.
- Ensure that the roles and responsibilities of *data owners* extend to data in cloud environments.
- Ensure that data sourcing is managed with authoritative sources and authorized distributors.
- Exploit opportunities for automation in the cloud environment to support governance of data consumption.
- Improve understanding of the requirements for managing *data sovereignty* and cross-border data movement risks.
- Implement controls for *data sovereignty* and cross-border data movement risk.

*2.0 Cataloguing & Classification*

The **Cataloging & Classification** component is a set of capabilities for creating, maintaining and using *data catalogs* that are both comprehensive and consistent. This component includes *classifications* for *information sensitivity*. These capabilities ensure that data managed in cloud environments is easily discoverable, readily understandable and supports well-controlled, efficient data use and reuse.

This CDMC component helps to:

- Define the scope and granularity of data to be cataloged.
- Define the characteristics of data as *metadata*.
- Catalog the data and the data sources.
- Connect the *metadata* among multiple sources.
- Share *metadata* with authorized users to promote discovery, reuse and access.
- Enable sharing of *metadata* and data discovery across multiple catalogs, platforms and applications.

- Define, apply and use the *information sensitivity* *classifications*.

*3.0 Accessibility & Usage*

The **Accessibility & Usage** component is a set of capabilities to manage, enforce and track *entitlements* and to ensure that data access, use and outcomes of data operations are done in an appropriate and ethical matter.

This CDMC component helps to:

- Express and capture data rights and obligations as *metadata*.
- Ensure that parties respect data rights and obligations over data they are entitled to access.
- Track and report on data access for both regulatory compliance and billing purposes.
- Establish formal organization structures for oversight of data ethics.
- Operationalize ethical access and use of data and ethical outcomes of data decisions.

*4.0 Protection & Privacy*

The **Protection & Privacy** component is a set of capabilities for collecting *evidence* that demonstrates compliance with the organizational *policy* for data sensitivity and protection. The purpose of these capabilities is to ensure that all sensitive data has adequate protection from compromise or loss as required by regulatory, industry and ethical obligations.

This CDMC component helps to ensure that:

- Data loss protection regimes are implemented.
- *Evidence* is gathered to demonstrate the application of required *data security* controls has been accomplished.
- A data privacy framework is defined and approved.
- A data privacy framework is operational.
- *Data obfuscation* techniques are applied to all data types according to *classification* and security *policies*.

*5.0 Data Lifecycle*

The **Data Lifecycle** component is a set of capabilities for defining and applying a *data lifecycle* management framework and ensuring that *data quality* in cloud environments is managed across the *data lifecycle*.

This CDMC component helps to:

- Define, adopt and implement a *data lifecycle* management framework.
- Ensure that data at all stages of the *data lifecycle* is properly managed.
- Define, code, maintain and deploy *data quality rules*.
- Implement *processes* to measure *data quality*, publish metrics and remediate *data quality* issues.

*6.0 Data & Technical Architecture*

The **Data & Technical Architecture** component is a set of capabilities for ensuring that data movement into, out of and within cloud environments is understood and that architectural guidance is provided on key aspects of the design of cloud computing solutions.

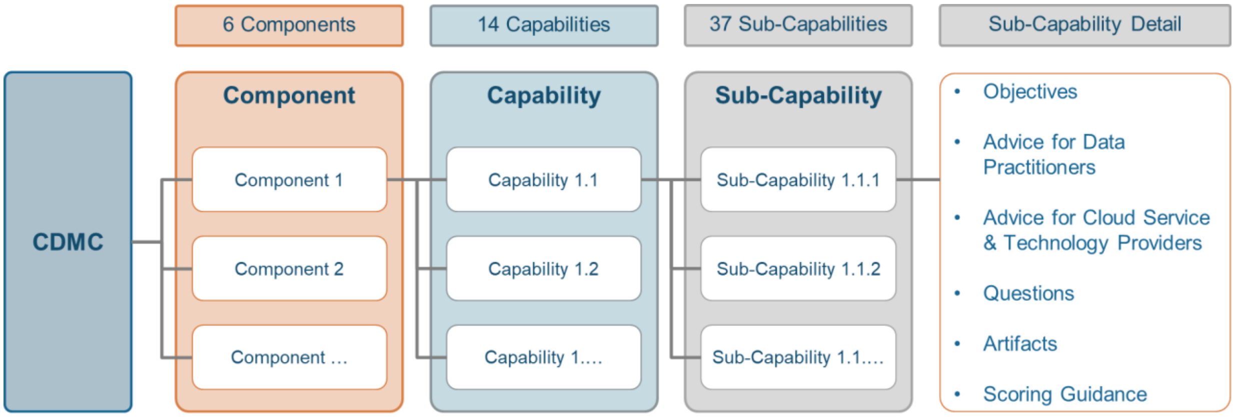
This CDMC component helps to:

- Establish and apply principles for *data availability* and resilience.
- Support business requirements for backup and point-in-time recovery of data.

- Facilitate optimization of the usage and associated costs of cloud services.
- Support data portability and the ability to migrate between *cloud service providers*.
- Automate identifying data *processes* and flows within and between cloud environments while capturing *metadata* to describe data movement as it passes along the data supply chain.
- Identify, track and manage changes to *data lineage*, and provide the ability to explain lineage at a point-in-time.
- Provide tooling to report and visualize lineage such that the outputs are meaningful from a business and technical perspective.

## STRUCTURE OF CDMC

As introduced above, CDMC is organized into six components. Each component is preceded with a definition that describes the components, explains why it is important and explains how it relates to the overall cloud *data management process*. These definitions are written for business and operational executives to understand the cloud *data management* process better. The components are organized into 14 capabilities and 37 sub-capabilities. The capabilities and sub-capabilities are the essences of the CDMC Framework. They define the goals of *data management* at a practical level and establish the operational requirements that are needed for sustainable cloud *data management*. Each sub-capability has a corresponding set of measurement criteria. The measurements are used in an assessment of your cloud *data management* journey.



- **Component** – a group of capabilities that together deliver a foundational tenet of cloud *data management*. A component functions as a reference guide for data practitioners who are accountable for executing the tenet.
    - **Upper Matter** – high-level context for the component—used as a background for understanding the component by data practitioners.
        - **Definition** – formal description of the component—supporting common *data management* understanding and language.
        - **Scope** – a set of statements to establish the guardrails for what is included in the component—used to understand and communicate reasonable boundaries.
        - **Overview** – more detailed context and accounting at a practical level to understand the operational execution required for sustainable cloud *data management*—used as a guide by the respective data practitioners.
        - **Value Proposition** – a set of statements to identify the business value of delivering the cloud *data management* component—used to inform the varied business cases for developing the *data management* initiative.
        - **Core Questions** – high-level but probing inquiries—used to explore the cloud *data management* component.
        - **Core Artifacts** – artifacts required to execute the capability—used to understand deliverables required to support the capability.

- **Capability** – a group of sub-capabilities that together execute tasks and achieve the stated objectives used as a reference tool by the data practitioners accountable for the execution.
    - **Description** – brief aggregate explanation of ***what*** is included in the sub-capabilities required to achieve the capability—used in the assessment process to inform the respondent of the scope of what they are rating.

- **Sub-Capability** – more granular activities required to achieve the capability—used as a reference tool by the data practitioners accountable for the execution.
    - **Description** – a brief explanation of ***what*** is included in the sub-capability—used in the assessment process to inform the respondent of the scope of what they are rating.
    - **Objective** – identified goals or desired outcomes from executing the sub-capability—used as a basis for defining cloud *data management* process design requirements.
    - **Advice for Data Practitioners** – more detailed but casual insight on the best practices of ***how*** to execute the sub-capability with an audit review perspective—used by the data practitioner.
    - **Advice for Cloud Service and Technology Providers** – more detailed but casual insight on how cloud technologies can support the sub-capability—used by cloud service and technology providers.
    - **Questions** – inquiries to direct interrogation of the capability/sub-capability current-state—used by the data practitioner to inform a perspective of the assessment scoring.
    - **Artifacts** – required documents or *evidence* of adherence—used for assessment and audit reference and to link to supporting best practice material—when available.
    - **Scoring Guidance** – insight for defining an assessment score—used when completing an assessment survey.

Each CDMC Component includes references to **Key Controls & Automations,** which are specifications of key controls that must be established at the capability level and highlight opportunities to support the control with automation. These are used as a reference tool by data practitioners accountable for the controls and cloud service and technology providers who support their implementation and automation.

## CDMC USE CASES

Organizations can use CDMC in multiple ways:

- As a well-defined control framework.
- As a tool to assess readiness for migration to and operation in cloud environments.
- As a certification *model* for cloud service and technology consumers.
- As a certification *model* for cloud service and technology providers.

## FRAMEWORK

When an organization adopts the standard CDMC Framework, it introduces a consistent understanding and way of describing cloud *data management*. CDMC is a comprehensive framework—presented as a best practice paradigm—of the capabilities required to manage data in single, multiple and *hybrid* *cloud* environments. It helps accelerate the development of a cloud *data management* initiative and make it operational. The CDMC Framework:

- Provides a common and measurable cloud *data management* framework.
- Establishes common language for the practice of cloud *data management*.
- Translates industry experience and expertise into operational *standards*.
- Documents cloud *data management* capability requirements.
- Proposes *evidence*-based artifacts.

## ASSESSMENT

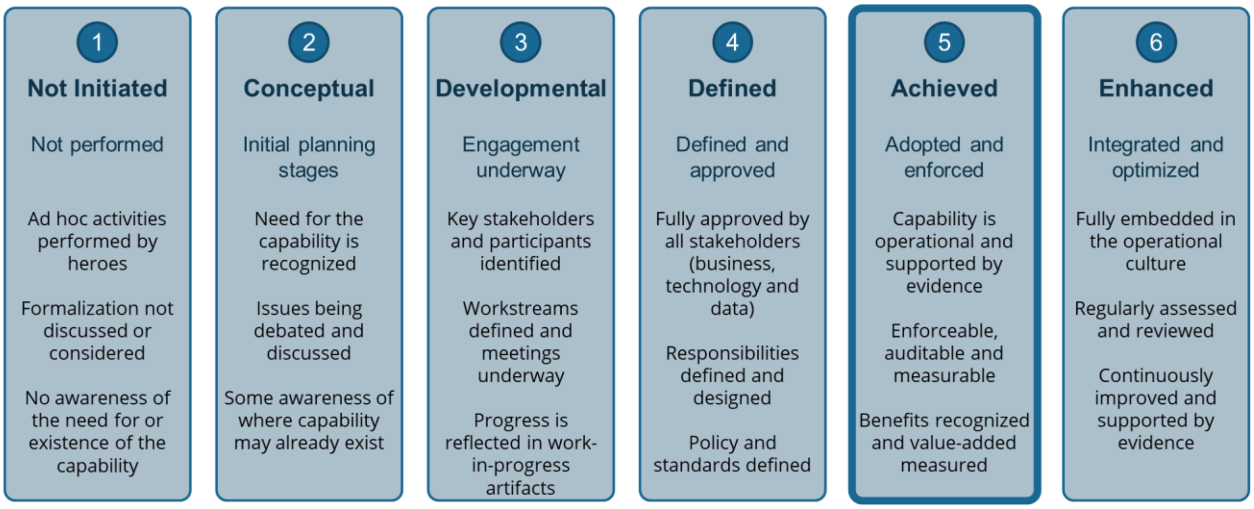
Performing an assessment measures the readiness of an organization to migrate to and operate in cloud environments. The assessment produces results that translate the practice of cloud *data management* into a quantifiable science. The benefits that an organization can gain from assessment outcomes include:

- Baseline measurement of the cloud *data management* capabilities in the organization compared to an industry standard.
- Quantifiable measurement of the organization's progress in establishing the required cloud *data management* capabilities into its operations.
- Identification of cloud *data management* capability gaps to inform a prioritized roadmap for future development and improvement.
- Focused attention to the funding requirements of the cloud *data management* initiative.

Effective use of the CDMC Framework as an assessment tool requires the definition of the assessment objectives and strategy, planning for the management of the assessment and adequate training of the participants to establish a base understanding of the framework. Organizations may either perform a self-assessment or may engage the services of a CDMC Authorized Partner to perform an independent assessment.

*CDMC Scoring Guide*

The CDMC Framework is designed to assess which phase of attainment the organization reaches for each capability. It is not an assessment of the maturity or scope to which the organization has applied the capabilities. The scoring scheme used throughout the framework is as follows:



| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| **Not Initiated** | **Conceptual** | **Developmental** | **Defined** | **Achieved** | **Enhanced** |
| Not performed | Initial planning stages | Engagement underway | Defined and approved | Adopted and enforced | Integrated and optimized |
| Ad hoc activities performed by heroes<br>Formalization not discussed or considered<br>No awareness of the need for or existence of the capability | Need for the capability is recognized<br>Issues being debated and discussed<br>Some awareness of where capability may already exist | Key stakeholders and participants identified<br>Workstreams defined and meetings underway<br>Progress is reflected in work-in-progress artifacts | Fully approved by all stakeholders (business, technology and data)<br>Responsibilities defined and designed<br>Policy and standards defined | Capability is operational and supported by evidence<br>Enforceable, auditable and measurable<br>Benefits recognized and value-added measured | Fully embedded in the operational culture<br>Regularly assessed and reviewed<br>Continuously improved and supported by evidence |

A CDMC assessment must also examine if the key controls have been established. This measurement provides a binary result for each control—the control is either established or not established.

## CERTIFICATION - CONSUMERS

Organizations that achieve all capabilities and establish all key controls can obtain the CDMC Certification. This certification process involves an independent assessment of the achievement of the capabilities and the existence of the controls performed by a CDMC Authorized Partner. If successful, the organization receives a CDMC digital certificate issued by the EDM Council and remains valid for 12 months. This certification is similar to other cloud computing certification programs such as SOC2.

### CERTIFICATION - PROVIDERS

*Cloud service providers* or cloud technology and solution providers can subject their platforms and products to a certification assessment against all or relevant CDMC Key Controls elements to protect sensitive data in cloud environments. An independent CDMC Authorized Partner must perform this certification assessment. Upon successfully completing a certification assessment, the EDM Council will issue a CDMC digital certificate that remains valid for 12 months. This certificate can be commercially represented in the market to indicate that the platform or product supports the respective CDMC Key Controls.

## SUPPORT MATERIALS

Additional materials support the CDMC Framework presented in this document in the following resources.

### CDMC CONTROLS TEST SPECIFICATIONS

Specifications of the CDMC Key Controls tests within the framework form the basis of cloud products and services certification against the framework.

Reference: *CDMC Controls Test Specification Version 1.1* – to be published Q4 2021

### CDMC INFORMATION MODEL

An *ontology* that draws on and combines related open frameworks and *standards* to describe the information required to support cloud *data management*. This *ontology* provides a foundation for the interoperability of *data catalogs* and automation of controls across cloud service and technology providers.

Reference: *CDMC Information Model Version 1.1* – to be published Q4 2021

### DATA MANAGEMENT REQUIREMENTS MODEL

A generic *model* of *data management* requirements with mappings to both CDMC and DCAM capabilities shows the relationship and dependencies CDMC capabilities have on basic *data management* capabilities.

Reference: *Data Management Requirements Model V1.1* – to be published Q4 2021

### TRAINING

The EDM Council and Authorized Partners offer a 2-day training course on the CDMC Framework.

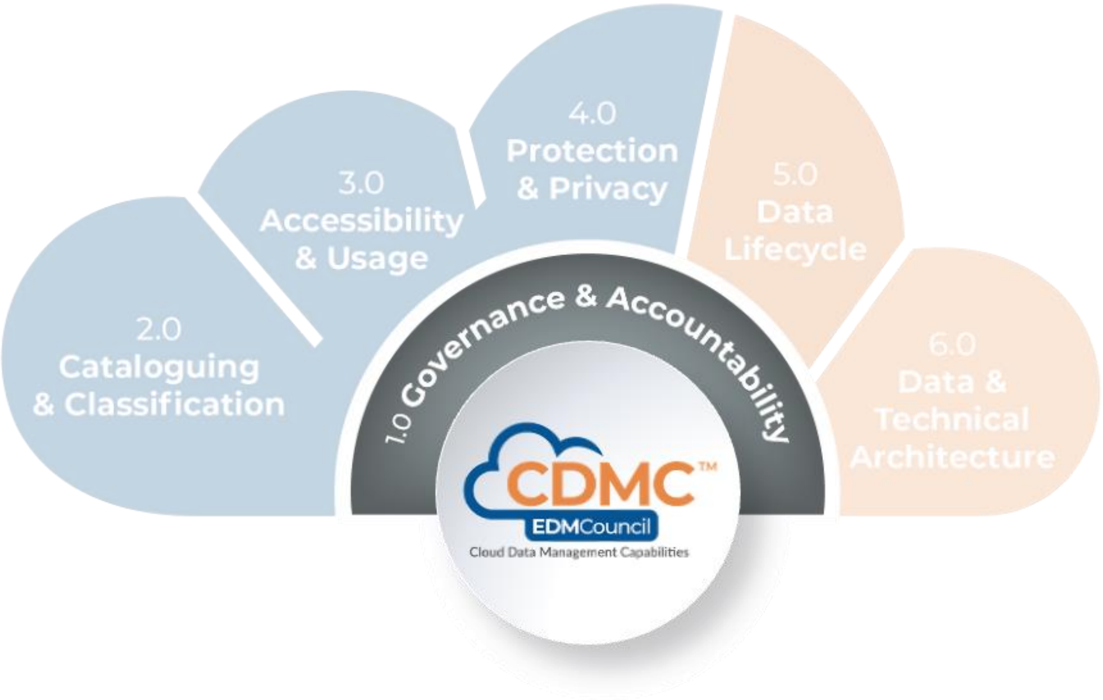
Reference: https://edmcouncil.org/page/CDMCTraining

## BUSINESS GLOSSARY

The EDM Council has developed a *data management* *business glossary* containing approximately 200 data management *term* names and definitions. CDMC v1.1 has applied these terms consistently across the document. Where a *term* is defined in the glossary, the word or phrase is italicized and underlined in the text.

The *business glossary* is available via the following link: https://www.dcamportal.org/glossary/.

# 1.0 Governance & Accountability



# 1.0 GOVERNANCE & ACCOUNTABILITY

## UPPER MATTER

### INTRODUCTION

Governance and accountability are the backbones of the successful management of data in cloud environments. The cloud environment introduces challenges and opportunities for scale, standardization, automation, and the shared responsibility *model*. Consequently, it is important to apply an effective data governance program to data that resides in a cloud environment. All *stakeholders* should have a clear understanding of data controls and accountability for each role. The approach is similar to how data governance, controls and accountability are applied to conventional *data management* in an organization.

### DESCRIPTION

The Governance & Accountability component is a set of capabilities that ensure an organization has clear accountability, controls and governance for data migrated to or created in cloud environments. These capabilities provide the foundation of well-governed business cases, effective data ownership, governance of data sourcing and consumption and management of *data sovereignty* and cross-border data movement risks.

### SCOPE

- Defining business cases for managing data in the cloud, including a value realization framework.
- Ensure the roles and responsibilities of *data owners* extend to data in the cloud.
- Ensure that data sourcing is managed with authoritative sources and authorized distributors.
- Leverage cloud automation opportunities in the governance of data consumption.
- Understand requirements for managing *data sovereignty* and cross-border data movement risks.
- Implement controls for *data sovereignty* and cross-border data movement risk.

### OVERVIEW

Business cases for cloud *data management* must articulate how to manage risk, deliver value, and align with the organization's overall business, data, and cloud computing strategies. The business cases should provide a basis for ensuring there is accountability for the quality of the outcomes.

**Business cases**

Business cases for managing data in a cloud environment need to outline planned activities, dependencies, risks (including plans to mitigate risks, where feasible), timelines, exit strategies, and outcomes based on the use case for that data. The value to be realized as a part of outcomes should link directly to the organization's broader *data management* strategy and cloud strategy. A framework of measures, metrics or key performance indicators must be established to demonstrate progress throughout the cloud *data management* implementation. The framework should include the depth of distinct capabilities matured (such as the number of personas with separate *role-based access controls*) and the coverage across the spectrum of capabilities ( such as the number of users securely accessing the cloud *data management* catalog).

Cloud *data management* business cases must be approved by an appropriate authority and sponsored by accountable *stakeholders*. Successfully managing data in cloud environments requires substantial support from both business and technology *stakeholders* within an organization. The interests of these groups need to be aligned before deployment and consistently represented through deployment.

Cloud *data management* business cases must be enforceable and periodically reviewed by sponsors throughout the deployment. Reviews should compare the original data strategy and cloud strategy that the business case was founded on against the details of interim outcomes and milestones achieved. Acceleration or deceleration of activities within the business case should be considered according to changing the cloud environment and data priorities.

The business cases should outline the key benefits of managing data in the cloud. While cost reduction and risk mitigation benefits are more tangible and easier to project, value-added features are critical to gaining approval from business *stakeholders*. The benefits should be demonstrated regularly to maintain momentum. Examples include:

- Scalability and transparency in managing the products and *analytics* outputs of data science teams.
- Better utilization of *data management* resources by simplifying capacity management.
- Availability of marketplace solutions and accelerators to rapidly mature *data management* capabilities.
- Controlled democratization of data access resulting from centralized storage in the cloud.
- The value from eliminating fixed capital costs and flexibility in the provisioning infrastructure comes at the expense of increased difficulty in forecasting future costs—appropriate mitigating controls should be included in the business case.
- Performing early experimentation and prototyping, enabling the pursuit of *quick wins* at relatively low risk.

There are potential sources of business value realized from managing data in a cloud environment that cannot be easily replicated for data managed on-premises, such as:

- Concentrating *enterprise* *data management* tools, including ease of integration and standardization (storage, compliance, cataloging, *analytics*, security, lineage, sourcing, quality) into fewer providers, reduces architectural variances and complexity.
- Provisioning *data management* infrastructure on variable schedules to account for performance fluctuations.
- Centralizing management of operating expenses to automatically correct for misutilized resources and systemically track the project budget against the plan.
- *Data management* in the cloud can be decoupled from any existing *data management* conducted on-premises, removing the requirement to consider constraints posed by integrating legacy architectures into the *data management* solution in the business case.

**Data ownership**

Data ownership is fundamental to successful data governance, regardless of whether data is on-premises or resides in a cloud environment. Effective data ownership is an enabler for cloud adoption and can drive how an organization leverages new capabilities available in the cloud. It is essential that data ownership is well established and that the responsibilities of *data owners* extend across all environments.

Data ownership is critical to ensuring the appropriate governance of data in the cloud. The *data owner* has overall accountability for the meaning, content, quality, distribution and use of a given *data set*. The *data owner* is supported by other roles such as data stewards, data architects and *metadata managers* in executing this accountability.

An important responsibility of a *data owner* is to ensure that the authoritative sources and authorized distributors of their data are identified and consumption from non-authoritative sources is governed. This importance is accentuated in the cloud, *multi-cloud* and hybrid cloud environments—where there is increased potential of the unnecessary proliferation of copies of data. *Data management* in a cloud environment offers the opportunity to support data sourcing and consumption governance with automation. One example of this is automating denial of data consumption from non-authoritative sources.

*Data owners* also play a role in ensuring *data sovereignty* requirements are understood and addressed in managing risks associated with cross-border data movement. *Data sovereignty* requirements are another area in which cloud computing can increase the potential for a wider geographic footprint of data storage and consumption and offer the opportunity to automate control of *data sovereignty* and cross-border data movement risks.

Characteristics of a *data owner* may include someone who has a good understanding of the meaning and purpose of the data; should be aligned to and familiar with the business areas with which the data is associated; should have a good understanding of the related business *processes* and outputs, and should be aware of *data consumers* to consider the impact of changes to the data.

Data ownership is agnostic to cloud service providers, except when the cloud provider generates the data, such as API or app log files. Ownership is not impacted when data is moved between cloud service providers, and the ownership of the technical data (log files) should not change with each cloud provider. Ownership is the sole responsibility of the organization, not the cloud provider. *Cloud service providers* should deliver the capability to execute data ownership activities for all data objects.

The effects of *cloud service providers* on data ownership include:

- Addressing ownership of new data types, such as log files, that the *cloud service providers* generate.
- Ensuring compliance with *data sovereignty* requirements in environments where data can be easily moved across borders. The *data owner*'s responsibility for establishing *guidelines* and controls for *data sovereignty* increases because of the broad geographic footprint of cloud computing and the abilities of some global data services.
- Understanding the controls available for cloud-managed data and the support available for executing data ownership responsibilities. The design and implementation of controls for cloud-managed data may differ from on-premises controls. *Data owners* should be familiar with the differences to ensure their adequate protection of cloud-managed data. While the controls remain consistent, the implementations of those controls may vary with each cloud service provider.

#### Data sourcing and consumption

Cloud computing provides an opportunity to reinforce requirements for data that is to be consumed from authoritative sources. The ability to expose *metadata* associated with *data assets* enables the discovery of data sources and the enforcement of consumption restrictions. Standardization of data sourcing *processes* that employ *metadata* can support automating authorization of data provisioning and consumption.

Migration of *data assets* into cloud environments or creating new *data assets* in cloud environments can trigger governance workflows that ensure those assets are tagged as authoritative sources, authorized distributors or non-authoritative sources. Similarly, standardization and control of data provisioning and consumption can ensure that the use of data is tracked and that the purpose of the data consumption is recorded.

An organization may want to consider the implementation of cloud data marketplaces supported by automation and driven by discoverable *metadata*:

- Automation can remove the need for a central team to manage data provisioning and access manually for *data producers*.
- Automation can facilitate standardization of the data *entitlement* process, leading to greater transparency for determining cost attributions (apportioning the cost of data sourcing according to variations in consumption) for *data consumers*.
- Automation enhances the transparency of data sources, data usage, provisioning, and organizational accountability for both *data producers* and consumers.

### Data sovereignty & cross-border data movement

*Data sovereignty* and cross-border data movement requirements relate to:

- When data must or must not be stored locally within a particular jurisdiction.

- The storage, transfer or access of data across a border.

These restrictions on data movement across borders are established for various reasons and are generally implemented through privacy, security, bank secrecy, outsourcing or data localization laws, rules, and regulations. The rules also affect how data can be accessed or shared with government authorities and law enforcement across international borders. The increased risk of significant fines and penalties for violating *data sovereignty* and cross-border data movement requirements is causing more organizations to re-evaluate when and how they store, access or transfer data globally. Increasingly strict *data sovereignty* and cross-border data movement requirements must be part of the data strategy for an organization. It is important to document how these requirements will affect business, data storage and *processing* activities.

*Data sovereignty* and cross-border data movement requirements are applicable whether data is stored and processed on-premises or in one or more cloud environments. The use of *cloud service provider (CSP)*s—especially multiple *CSP*s in a global, hybrid cloud environment—increases the complexity of understanding where data (and which data) is being stored, accessed or processed at any given time. This complexity means that organizations should have a framework established by which to understand requirements and ensure compliance. It is also important to extend *data sovereignty* and cross-border data movement reviews and evaluate controls and clearance *processes* with an increasingly larger set of parties to ensure compliance.

*Data sovereignty* and cross-border data movement considerations influence the geographic locations where an organization can locate or process data.

- Organizations need to track and report on the exact jurisdictional location of data to prove compliance with increasingly restrictive requirements.
- Organizations should employ *processes* such as tagging and *classification* to apply jurisdictional rules and mitigate *data sovereignty* and cross-border data movement risk.
- Organizations should mitigate *data sovereignty* and cross-border data movement risk with tools, such as advanced data masking and *encryption* solutions. Refer to *CDMC 4.1 Data is Secured, and Controls are Evidenced.*

Organizations often use multiple cloud service providers, typically with on-premises systems and applications, increasing the complexity of data tracking or the risk of storing, accessing or transferring data in a non-compliant manner. Applications and technology in cloud environments evolve rapidly and change quickly, putting pressure on compliance efforts. Tracking, tagging and automation can make it easier to implement controls around *data sovereignty* requirements. Most cloud service configuration is performed using Infrastructure-as-Code, and this creates a greater opportunity to implement controls at build time and deployment time.

Organizations remain responsible for compliance with *data sovereignty* and cross-border data movement requirements, including:

- Interpreting *data sovereignty* and cross-border data movement rules.
- Checking their applicability to the datasets.
- Implementation of granular data location controls.
- Auditing to determine where data has been stored, accessed or transferred over long periods.
- Reporting on compliance with the *data sovereignty* and cross-border data movement *policies* and *procedures* of the organization.

A *cloud service provider* should provide tooling and support to help the organization implement these requirements. Data practitioners need to understand how the cloud or technology service provider handles data backups, replication, and caching. While the providers are responsible for the functionality, the accountability remains with the organization. Cloud and technology service providers need to provide increased transparency and auditability.

## VALUE PROPOSITION

Organizations that establish strong governance and *data management* controls over data residing in cloud applications have an opportunity to realize all of the benefits of a cloud implementation while managing the associated risks. Data governance and accountability in cloud environments help define effective business case *processes*, identify accountable *stakeholders* and *data owners*, ensure the proper management of data sourcing, and provide proper tracking and control of data movement concerning *data sovereignty* *guidelines*.

Effective data governance and controls help an organization exploit cloud *data management* capabilities to increase the effectiveness of data ownership, improve the ability to track and report on data usage, enforce *policy*, better monitor *data owner* assignment, improve data access controls to authoritative sources and better monitor and control *data sovereignty* requirements. *Data management* in a cloud environment enables an organization to move from systems not built to track data location to new data environments. Data location and types of data can be readily tracked and audited for better compliance.

## CORE QUESTIONS

- Has data governance been established for managing data in cloud environments?
- Have business cases for managing data in the cloud been defined?
- Do cloud *data management* business cases include a value realization framework?
- Are cloud *data management* business cases governed?
- Have the roles and responsibilities of *data owners* been extended to data in the cloud?
- Are *data owners* in place for all cloud data?
- Are all cloud *data assets* identified as authoritative sources, authorized distributors or non-authoritative sources?
- Does the governance of data consumption leverage cloud automation opportunities?
- Are requirements for managing *data sovereignty* and cross-border data movement risks defined?
- Have controls for *data sovereignty* and cross-border data movement risk been implemented?

## CORE ARTIFACTS

- Cloud Data Management Business Cases
- Data Ownership Roles and Responsibilities
- Data Catalog Report – indicating *data owner*
- Register of Authoritative Sources and Authorized Distributors
- Data Sovereignty and Cross-Border Data Movement Requirements Definition
- Data Sovereignty and Cross-Border Data Movement Issues Log

## 1.1 CLOUD DATA MANAGEMENT BUSINESS CASES ARE DEFINED AND GOVERNED

The organization must have clearly defined business cases for the management of data in cloud environments. These must include a framework of measures of the value to be realized. Each business case must be approved by an appropriate authority and sponsored by accountable *stakeholders*.

### 1.1.1 CLOUD DATA MANAGEMENT BUSINESS CASES ARE DEFINED

#### DESCRIPTION

As an organization moves its data and operations to cloud environments, it is important to develop, communicate, cultivate, and support business cases for cloud *data management*. An effective cloud *data management* business case defines the objectives and expected outcomes of the implementation. It is vital to develop an entire cloud business case framework of metrics, measures and key performance indicators to articulate the value of cloud *data management*.

#### OBJECTIVES

- Define a standard process to develop and gain approval for cloud *data management* business cases, justifying what is needed to manage data in the cloud environment.
- Ensure cloud *data management* business cases include measures of the effectiveness of the corresponding cloud *data management* capabilities.
- Document cloud *data management* business cases to include all relevant business problem types for the organization and list the *stakeholder* responsible for each business case.
- Design measures, metrics, or key performance indicators with targets to enable the measurement of progress.
- Ensure cloud *data management* business cases metrics and targets are specific, measurable, achievable, relevant, and time-based.
- Ensure cloud *data management* business cases detail elements of value such as new revenue generated, amount of cost reduction and any mitigated risks.

#### ADVICE FOR DATA PRACTITIONERS

To fully demonstrate the value of *data management* in the cloud, practitioners must develop a value realization framework that includes metrics, measures and key performance indicators for each business case. The framework should include expected outcomes already defined in the organization’s business, data, and cloud strategies.

The precision in the outcome estimates within each business case should be documented. Also, document the risks in failing to achieve the targeted outcomes and explicitly communicate these risks to sponsors and *stakeholders*. The *accuracy* and *data quality* of the metrics must faithfully reflect progress against these business cases. Each business case should quantify each outcome along its respective timeline.

**Cloud data management business case standard**

A business case must include metrics of the effectiveness of all cloud *data management* capabilities that are in use by the organization. Each metric must be specific, measurable, achievable, relevant, and time-based. Metrics must align with the business problems being addressed. Gain approval on targets for each metric and identify *stakeholders* that are to be responsible for achieving the targets. Each metric should have the ability to measure progress. Each element of the value realization framework must be included in the business case:

- Metrics dictionary – a library of measures that align with business outcomes or CDMC capabilities.
- Metrics accountability – document *stakeholder* accountability for each metric.
- Metrics *traceability* – document the correspondence of each business case outcome to the best practices of the organization or industry best practices.

- Outcome projections – document original targets and projections for revenue added, costs reduced, and risks mitigated.
- Assumption *evidence* – document the variables and assumptions (such as discount rates or estimates of regulatory fines) and how each was derived and included in calculations.
- Metrics tracking – document trends (not merely snapshots) accompanied by stated targets with timelines.
- Impact assessment – *evidence* of other cloud *data management* efforts already underway in the organization. Quantify mutually beneficial and any potential detrimental interactions that may result.

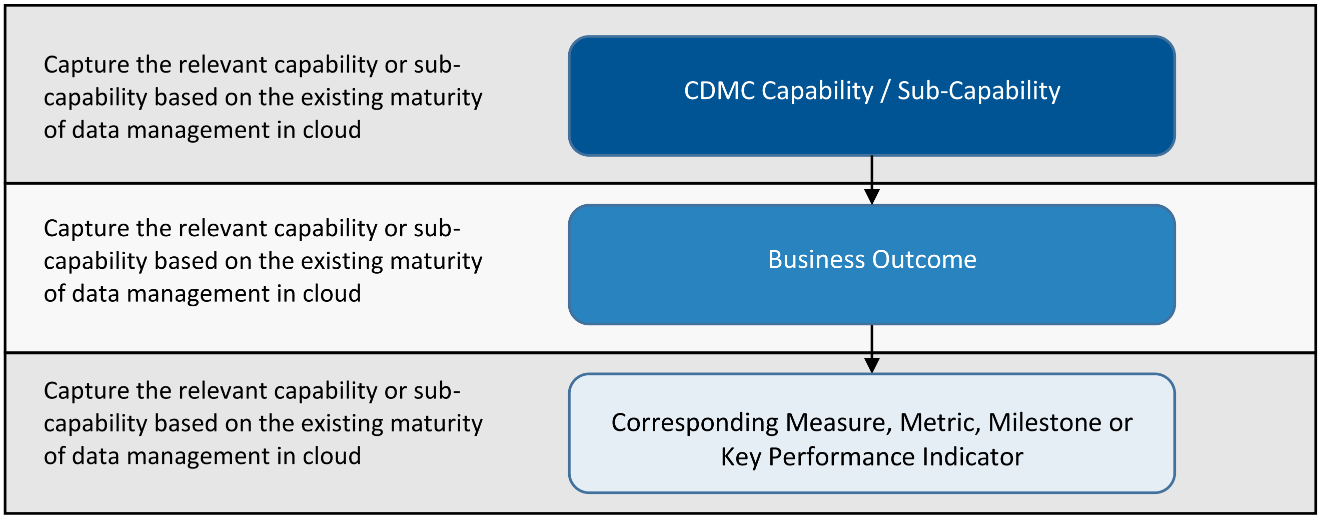
Practitioners should ensure that the implementation team and sponsor are transparent in resource consumption when reporting to *stakeholders*. Establish a baseline or point-of-reference against which to measure tangible value realized in the transition to managing data in the cloud. Create a method for isolating the value resulting from managing data in the cloud from other factors that may also affect revenue added, costs reduced, and risks mitigated. Document how the definition of value generated by managing data in the cloud might need to change as the target state approaches.

Pre-determine the critical junctures at which sunk costs incurred in the implementation phase exceed thresholds that require a review of project scope and progress against any value realized up to each critical juncture. Estimate projected new revenue generated by managing data in a cloud environment and compare that with existing revenue generated from the on-premises environment. Similarly, provide estimates of reduced costs that result from managing data in a cloud environment and compare with the costs from managing data on-premises. One example is the lower storage costs that are typical in a cloud environment.

Discount these estimates using a suitable time value of money. Consider any intermediate costs incurred to complete the transition ( such as temporarily redundant data storage and contracting costs)—separate these one-time costs from any new maintenance costs expected to remain in the future state. Lastly, specify any risks that have been mitigated by managing data in a cloud environment and compare them with similar risks in the on-premises environment. For example, there is typically an increased compliance burden with *GDPR*/CCPA regulation when employing cloud-native tools to track *data lineage*.

*Suggested approach to the identification of success factors and measurements (and constructing the value realization framework). While many organizations will already have adopted frameworks for value realization that can be adapted to suit data management in the cloud, the CDMC has provided one potential approach to providing a structured framework that realizes the objectives outlined above:*

**Example #1: Value Realization Framework - Baseline Business Outcomes and Metrics**



Capture the relevant capability or sub-capability based on the existing maturity of data management in cloud

CDMC Capability / Sub-Capability

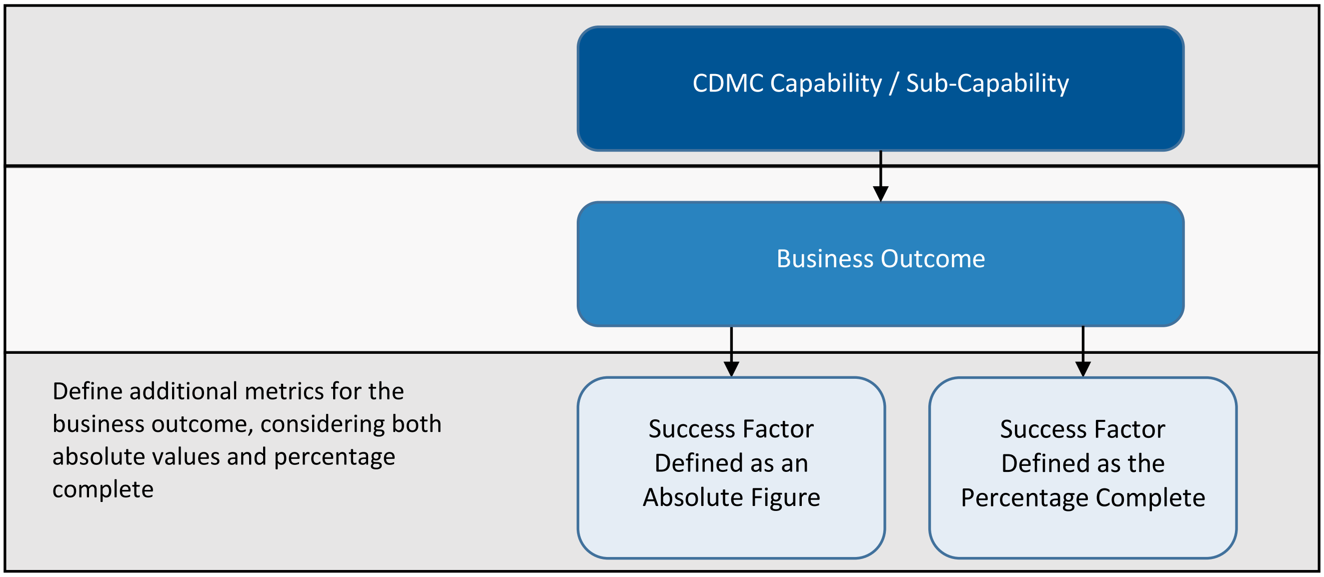
Capture the relevant capability or sub-capability based on the existing maturity of data management in cloud

Business Outcome

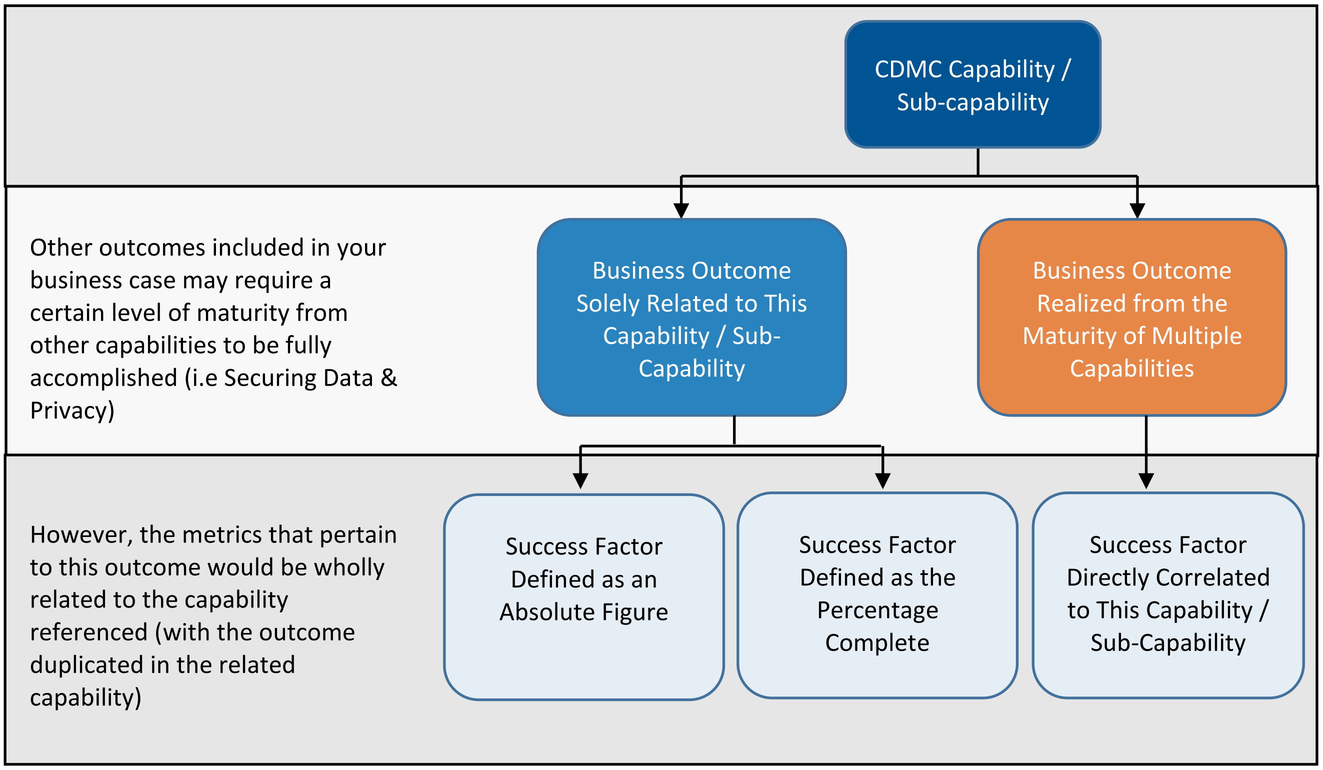
Capture the relevant capability or sub-capability based on the existing maturity of data management in cloud

Corresponding Measure, Metric, Milestone or Key Performance Indicator

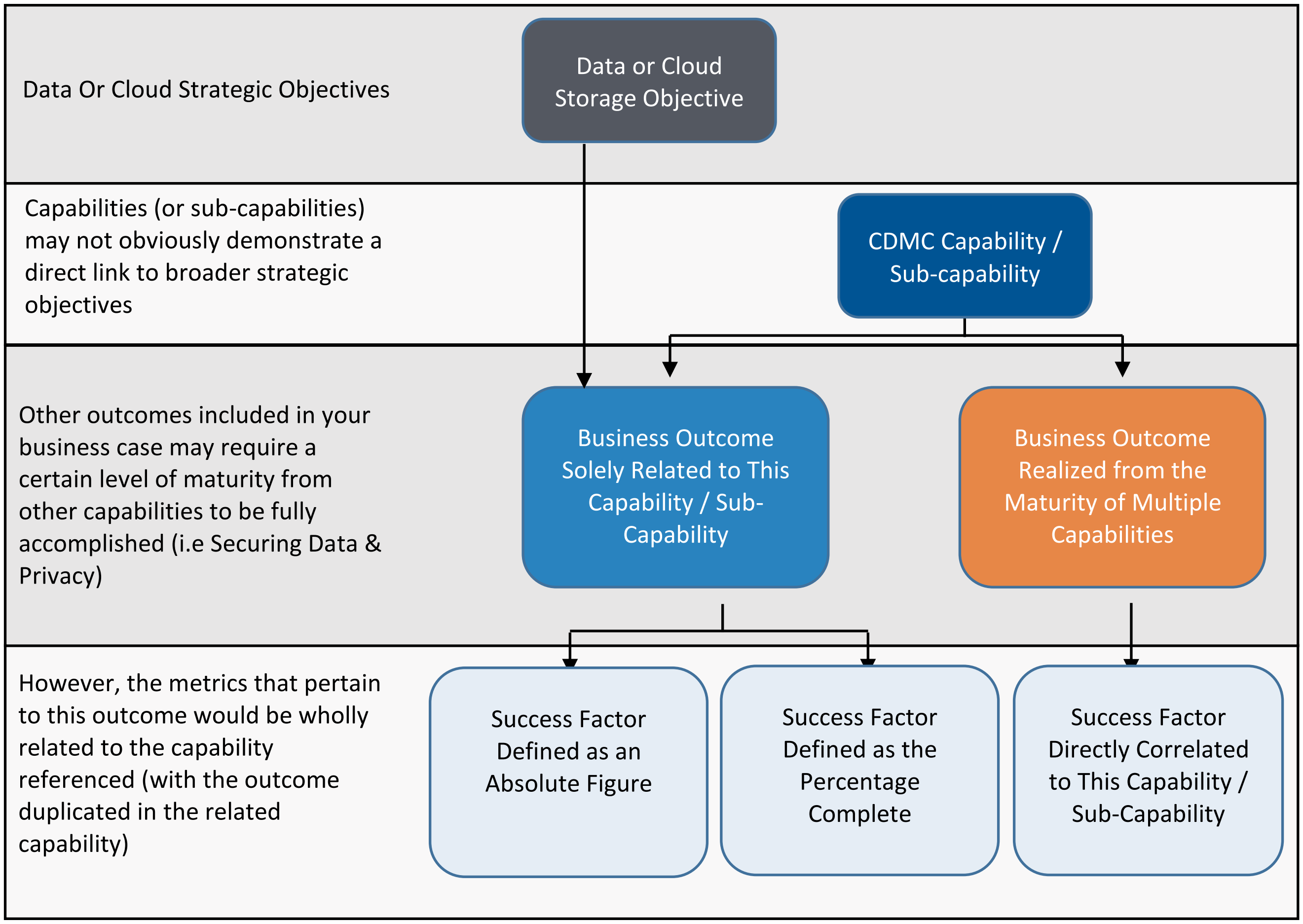
**Example #2: Value Realization Framework - Deeper Realization of an Outcome Already Defined**


CDMC Capability / Sub-Capability
Business Outcome
Define additional metrics for the business outcome, considering both absolute values and percentage complete
Success Factor Defined as an Absolute Figure
Success Factor Defined as the Percentage Complete

**Example #3: Value Realization Framework - Outcomes Realized Through Maturity of Multiple Capabilities**


CDMC Capability / Sub-capability
Other outcomes included in your business case may require a certain level of maturity from other capabilities to be fully accomplished (i.e Securing Data & Privacy)
Business Outcome Solely Related to This Capability / Sub-Capability
Business Outcome Realized from the Maturity of Multiple Capabilities
However, the metrics that pertain to this outcome would be wholly related to the capability referenced (with the outcome duplicated in the related capability)
Success Factor Defined as an Absolute Figure
Success Factor Defined as the Percentage Complete
Success Factor Directly Correlated to This Capability / Sub-Capability

**Example #4: Value realization framework - Outcomes Contributing to The Organization's Broader Data/Cloud Strategy**



Data Or Cloud Strategic Objectives

Capabilities (or sub-capabilities) may not obviously demonstrate a direct link to broader strategic objectives

Other outcomes included in your business case may require a certain level of maturity from other capabilities to be fully accomplished (i.e Securing Data & Privacy)

However, the metrics that pertain to this outcome would be wholly related to the capability referenced (with the outcome duplicated in the related capability)

Data or Cloud Storage Objective
CDMC Capability / Sub-capability
Business Outcome Solely Related to This Capability / Sub-Capability
Business Outcome Realized from the Maturity of Multiple Capabilities
Success Factor Defined as an Absolute Figure
Success Factor Defined as the Percentage Complete
Success Factor Directly Correlated to This Capability / Sub-Capability

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Cloud service and technology providers should understand the *data management* business outcomes organizations are looking to achieve when migrating data to cloud environments. Providers should develop and communicate metrics that organizations can readily employ to optimize *data management* in the cloud environment. Monitoring and tracking capabilities should enable visibility into all costs incurred from managing data in the cloud environment.

In addition, a provider should offer tools and dashboards to automate a broad set of baseline metrics that demonstrate the benefits of managing data through the cloud service. Examples of such metrics include scale of data in cloud, % of data governed, % of data categorized, % of data profiled, % of data with lineage, scale of re-use, % of data measured and the number of access points enabled.

Also, providers should showcase various case studies and benchmarks of quantitative and qualitative outcomes resulting from previous *data management* implementations in the cloud. These examples should include case studies on meeting regulatory requirements that help avoid pitfalls when data is managed appropriately in cloud environments.

Research and develop additional content on avoiding anti-patterns in *data management* design in the cloud that may result in unnecessary costs.

## QUESTIONS

- Is there a standard process to develop and approve cloud *data management* business cases?
- Does each cloud *data management* business case include measures of the effectiveness for the corresponding cloud *data management* capabilities?
- Are cloud *data management* business cases structured to include all relevant business problems being addressed, and does each business case list the *stakeholders* responsible for achieving the targets?
- Have measures, metrics, or key performance indicators been designed with targets to measure progress?
- Are cloud *data management* business cases metrics and targets specific, measurable, achievable, relevant, and time-based?
- Do cloud *data management* business cases detail elements of value such as new revenue generated, amount of cost reduction and risks mitigated?

## ARTIFACTS

- Value Realization Framework – including measures, metrics, or key performance indicators with targets to measure progress
- Cloud Data Management Business Case Standard – including the methodology and framework with standard accountability, assumptions, metrics, *traceability*, outcome projections and monitoring
- Repository of Cloud Data Management Business Cases

## SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal standard cloud *data management* business cases exist. | No formal standard cloud *data management* business cases exist, but the need is recognized, and the development is being discussed | Formal standard cloud *data management* business cases are being developed. | Formal standard cloud *data management* business cases are defined and validated by *stakeholders*. | Formal standard cloud *data management* business cases are defined and adopted by the organization. | The formal standard cloud *data management* business cases are established as part of business-as-usual practice with continuous improvement. |

## 1.1.2 CLOUD DATA MANAGEMENT BUSINESS CASES ARE SYNDICATED AND GOVERNED

### DESCRIPTION

Each cloud *data management* business case must be approved by an appropriate authority and sponsored by accountable *stakeholders*. Successfully managing data in cloud environments requires substantial support from both business and technology *stakeholders* within an organization. The interests of these groups must be aligned early and consistently represented through deployment.

Each cloud *data management* business case must be enforceable and periodically reviewed by sponsors throughout deployment and the cloud *data management* lifecycle. Reviews will ensure that the business cases meet requirements as the organization's objectives evolve and the *stakeholders* change.

### OBJECTIVES

- Ensure cloud *data management* business cases consider the requirements of all key *stakeholders*.
- Obtain approval and support from all key *stakeholders* of cloud *data management* business cases.
- Conduct regular reviews of cloud *data management* business cases.
- Structure and version the cloud *data management* business cases to support an audit.
- Implement governance oversight to ensure that data migrated to, stored in or created in a cloud environment fulfills the requirements of both the cloud *data management* business cases and risk mitigation intentions of the organization.

### ADVICE FOR DATA PRACTITIONERS

Cloud *data management* business cases must account for the priorities of the various *stakeholders*. Some of these priorities are complementary, and some are competing.

An organization should seek to use business cases to balance delivery and execution with risk management and sustainability. The risk-to-benefit appetite of each organization depends on the industry and regulatory environment in which it operates. It is important to consider the risk appetite with full transparency. An organization may choose to address some cloud *data management* considerations. However, the organization must always adhere to legal requirements and address these in business cases.

An organization must conduct sufficient oversight of *data management* controls to ensure a suitable standard for data that will migrate, be stored in, or be created in the cloud. Oversight may occur through automated controls, workflow adjustments, governance reviews, tollgates or other means. Any actions taken should be proportionate to the risk appetite, regulatory environment and size of the organization.

Periodic business case reviews should compare the original business strategy, data strategy and cloud strategy on which the business case was founded against interim outcomes. Decisions on whether to accelerate or delay activities for a specific business case should depend on changes in cloud *data management* priorities.

Key *stakeholders* must approve changes to the business cases with sufficient authority and with appropriate governance. In addition, it is vital to get explicit approval from each of the *stakeholders*.

The table below is a list of potential *stakeholders*, though it is not an exhaustive list. Keep in mind that some organizations may not need each role. The specific roles and responsibilities depend on the business requirements and strategy of each organization. The organization should engage with human resources and vendors to ensure that proper data management and cloud skills are available to support cloud data management and include appropriate funding in the business case. This list of *stakeholders* aims to help data practitioners ensure the major *stakeholder* groups and perspectives have been considered. In addition, plan timeframes are given for each *stakeholder* group.

| Major Stakeholder Group | CDMC Framework Stakeholder Roles | Primary CDM Requirement | Primary CDM Responsibility | Illustrative Planning Horizon | Ongoing Commitment and Review |
|---|---|---|---|---|---|
| ***CDO* and *data management* practitioners** | *Chief Data Officer* / Data Governance Leads / BU data stewards | Accountability for data is well understood and followed across the organization. The approach to cloud *data management* is well documented, suitable and followed. | Setting the vision for sustainable *data management* and high-level requirements (build/run – balancing governance with delivery) | 2-5 years (CDM Vision) | Setting and coordinating the review of the DM strategy, framework and its relation to cloud *data management* |
| | | | | | Review compliance quarterly supplemented with ad hoc reviews |
| **Risk & Finance** | Chief Risk Officer / Chief Financial Officer / Treasury Head | Data is managed in the cloud to the level required by regulatory reporting. Controls are in place to manage risks within appropriate thresholds. | Ensuring data for managing risk and regulatory reporting are sourced correctly, accurately, timely and complete. | 2-3 years (Risk Vision) | Risk management and Regulatory reporting requirements into CDMC Framework |
| | | | | | Informed of any deviations (through quarterly exception reporting, supplemented with ad hoc reports |
| **Business *Data producer* or Consumer** | Business Unit Heads / Operations / *Data management* | As a data consumer: data is consumed following organization data collection principles. Usage is clear, and feedback on *data consumer* requirements is fed back.<br>As a *data producer*: data is delivered to maximize business value and reduce risk and overheads. | Business use cases are clearly defined (which informs value/risk/momentum) and consider the maturity of the organization's cloud *data management* workflows. | 1-year budgeting cycle<br>Conforming to 2-5 years CDM vision | Data usage and business use case reviews (semi-annually) |

| Major Stakeholder Group | CDMC Framework Stakeholder Roles | Primary CDM Requirement | Primary CDM Responsibility | Illustrative Planning Horizon | Ongoing Commitment and Review |
|---|---|---|---|---|---|
| **Technology, Architecture and Transformation** | Chief Information Officer / BU aligned Tech Heads | Cloud *data management* is defined by clear principles and approaches to make it achievable and understandable by tech teams while minimizing administrative overhead. | Ensure data is migrated in a controlled, sustainable and secure. The aim is to ensure sustainable *data management*, proper sourcing, tagging and maintenance. | 1-year budgeting cycle | Cloud *data management* controls implemented for every migration/deployment and development |
| | Chief Architect / CTO / Head of Cloud | Sustainable cloud *data management* architecture is in place and meets interoperability *standards* (multiple cloud environments; suitable range of cloud computing and storage tools). | Data in the cloud is well-organized for sustainability, structured to support organization architecture goals and can scale/adapt with advances in architectural approach. | 2-5 years vision | Annual review of CDM business cases |
| | Cloud Project Teams, Developers, & Engineers | The principles, approach and execution of cloud *data management* concerning deployment and maintenance are well defined, clear and proportionate for an optimal balance of delivery time and risk management. | Inform the cloud *data management* requirements to achieve an optimal balance of delivery time to risk management. Ensure cloud *data management* project teams, developers & engineers adhere to the cloud *data management* approach. | 0.5-2 year delivery horizon | Annual review of CDM business cases with communication of any deviations through quarterly exception reporting supplemented with ad hoc reports |

| Major Stakeholder Group | CDMC Framework Stakeholder Roles | Primary CDM Requirement | Primary CDM Responsibility | Illustrative Planning Horizon | Ongoing Commitment and Review |
|---|---|---|---|---|---|
| **Cybersecurity, Privacy, Legal and Compliance** | Chief Privacy Officer / Head of Cyber / Head of Tech Risk | Privacy, security and technology risks are managed according to risk appetite. Cost is proportionate. Maintenance and controls are robust and sustainable. | Balance cloud *data management* requirements with a specific focus on privacy, security, information lifecycle management and integrity. Continuity controls are well-defined and followed. | 2-3 year | Annual review of CDM business cases with communication of any deviations through quarterly exception reporting supplemented with ad hoc reports |
| | Legal, Compliance & Audit | Cloud *data management* conforms to legal and regulatory interpretation and fulfills organization compliance obligations and *policies*. | Legal rules on data sharing, restriction, and disposition are well-defined, implementable, and communicated to the control owners. | 2-3 year | Annual review of CDM business cases with communication of any deviations through quarterly exception reporting supplemented with ad hoc reports |
| ***Analytics* and Digital Transformation** | Head of Analytics / Data Scientists / Labs / Innovation | Data in the cloud is cataloged, classified and structured to minimize wrangling, responsibly accelerate access, be easy to manipulate, maximizes confidence in the quality and helps achieve value and re-use. | Maximize business value from data managed in the cloud by ensuring data requirements to support analytic use cases are well understood, communicated and maintained. | 2-3 year vision | Annual review of CDM business cases |
| **Cloud Partners** | Cloud Service and Technology Providers | The organization's cloud *data management* requirements, controls, expectations and *stakeholder* landscape are understood. | Ensure the organization is well informed on best practices and avoid common mistakes. A support structure is well-established. Metrics to track sustainability are in place.<br>Maximize business benefit, loyalty and continued growth of cloud *data management*. | N/A | Annual review of the organization's business cases to determine best customer support response |

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Cloud service and technology providers must understand and contribute to the organization's cloud *data management* business cases to help them achieve optimal business outcomes and minimize the risks of cloud *data management*.

Typically, providers have considerable cross-industry experience in helping organizations realize business value from cloud adoption. Understanding and providing input to the business cases benefit the organization from the provider's insight into what has worked well previously. While providers can offer considerable experience in what can work well, it is important that advice remains high-level, non-prescriptive and presented as considerations and challenges to ensure the business case is truly driven and owned by the organization.

*CSP*s should provide appropriate automations to support business cases to control any data migrated, stored or created in the cloud environment to support the organization’s oversight of control mechanisms.

## QUESTIONS

- Have all key *stakeholder* requirements been considered and balanced when constructing the business cases?
- Have all key *stakeholders* approved all business cases, and are they aware of their support roles in the intended outcomes?
- Has the organization set the frequency at which the business cases should be reviewed?
- Has a structure for cloud *data management* business cases been defined that enables them to be audited?
- Does an oversight mechanism exist that is supported by appropriate controls and demonstrates that data created in, stored in or migrated to the cloud conforms to the requirements of the cloud *data management* business cases?

## ARTIFACTS

- Policy, Standard and Procedure – defining and operationalizing the management and governance of cloud data management business cases
- Cloud Data Management Stakeholder Matrix
- Cloud Data Management Business Case Template
- Cloud Data Management Business Case Approval Form
- Cloud Data Management Business Case Governance Forum Charter

## SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal governance of cloud *data management* business cases exists. | No formal governance of cloud *data management* business cases exists, but the need is recognized, and the development is being discussed | The formal governance of cloud *data management* business cases is being developed. | The formal governance of cloud *data management* business cases is defined and validated by *stakeholders*. | The formal governance of cloud *data management* business cases is established and adopted by the organization. | The formal governance of cloud *data management* business cases is established as business-as-usual practice with continuous improvement. |

## 1.2 DATA OWNERSHIP IS ESTABLISHED FOR BOTH MIGRATED AND CLOUD-GENERATED DATA

The roles and responsibilities of *data owners* must be extended to instances of data in cloud environments. Data ownership must be specified for all data, whether migrated to the cloud from the on-premises environment or created in cloud environments.

### 1.2.1 DATA OWNER ROLE AND RESPONSIBILITIES ARE DEFINED

#### DESCRIPTION

Implementing the concept of data ownership requires defining the role and responsibilities of the *data owner* and ensuring the role is applied to data managed in the cloud environment and on-premises.

#### OBJECTIVES

- Define roles and responsibilities of the *data owner* and mandate by the *data management policy*.
- Extend *data owner* responsibilities to data hosted in cloud environments.
- Adapt and extend *data owner* responsibilities to any new data types used by *cloud service providers* (*CSP*s).
- Determine if any *data owner* responsibilities will have more importance concerning data residing in a cloud environment.
- Define cloud technology support requirements for each relevant *data owner* role and responsibility.

#### ADVICE FOR DATA PRACTITIONERS

The *data owner* role must be assigned to a senior business executive to have the necessary authority to perform the role. This required seniority ensures ongoing accountability, even when *personnel* changes occur. *Data management policy* should explicitly ensure that data ownership accountability belongs to the appropriate executive. In most organizations, responsibility for the execution of data ownership tasks will be delegated to supporting roles such as data stewards. Definition of the *data owner* role should extend to and clarify how the execution responsibilities are delegated. This role definition should also be incorporated in and supported by the *data management policy*.

A *data owner* is accountable for the meaning, content, quality, distribution and storage of a given set of data or the contents of a *data domain*. The *data owner* must ensure that all data drawn by its *data consumers* meet fit-for-purpose criteria and align with organizational *standards*. Adopting cloud computing *data management* services can support a *data owner* with automated capabilities that are typically more effective and efficient than conventional systems.

The *data owner* has full responsibility for understanding the quality and scope of the content in a *data domain*. Cloud computing technology typically provides comprehensive, real-time *data catalog* and *data lineage* solutions. Rich *metadata* is available from many of these solutions. This *metadata* enhances the ability of the *data owner* to understand the data landscape and eases the execution of data ownership responsibilities.

Many *data owners* have responsibility for various on-premises applications that rest upon various platforms and legacy technologies. Lack of homogenization and transparency across these *data domains* makes applying granular control across all environments challenging. Many cloud environments can improve standardization of functionality, granular controls standardization and monitoring capabilities.

Cloud environments should provide *standards* for monitoring data and provide summaries for the entire data landscape. *Data owners* will use the monitoring dashboards to drill down to identify various sources of *data quality* and control failures. Such views can extend from *data assets* down to individual *data elements*.

Enhancements in data storage and management homogenization significantly improve the visibility and precision of *data consumer* utilization. Consequently, *data owners* can understand which *data element* controls require prioritization. Better controls improve the ability of the data owner to enforce *data security* and immutability.

A *data owner* should provide transparency about the content, location and consumption of their data. Cloud *data management* can help a data owner manage responsibilities, operate more efficiently, improve transparency and facilitate better systems integration.

Typically, a *data owner* must also solve *data quality* and manage control exceptions. In support of such tasks, the *data owner* should also have the ability to interact with an integrated workflow, direct a course of action or redirect to another *data owner*.

### ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

It is important to recognize that a *data owner* may not have a strong affinity for technology. This understanding is especially true if the *data owner* is from a business, finance, risk, or another background—not Information Technology. Such users should have resources available to navigate and interrogate interactive dashboards and perform some workflow tasks. Any technology competency beyond that expectation should be regarded as optional.

With these expectations in mind, a *cloud service provider* should:

- Provide dashboards, workflow tasks and task execution tracking.
- Provide corresponding training that does not require coding, tedious querying, or any IT knowledge.
- Provide the ability to the *data owner* to execute or manage responsibilities in the *data domain*.
- If necessary, automate any capabilities for the *data owner* to develop and maintain the integration of a *data element* list, definitions, *data quality rules*, controls, *data lineage* and *enterprise* *data model* integration.
- Provide intuitive, non-programmatic interfaces to interact with any automations.
- Provide some ability for *data owners* that may have technical and coding expertise to extend or customize dashboards, workflows and task execution.
- Work with the organization to determine if any *data owner* responsibilities (such as sovereignty) have more importance in managing data in a cloud environment.

### QUESTIONS

- Have *data owner* roles and responsibilities been defined?
- Have *data owner* responsibilities been extended to *data management* capabilities at the *CSP*?
- Does the *data owner's* responsibility include data that is generated by and stored at the *CSP*?
- Does the *data owner's* responsibility include all activities that have higher importance for managing data at the *CSP*?
- Does the *CSP* provide technology to support *data owner* roles and responsibilities?

### ARTIFACTS

- Data Management Policy, Standard and Procedure – defining and operationalizing *data owner* roles and responsibilities

### SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| *Data owner* roles and responsibilities are not defined by *policy*. | *Data owner* roles and responsibilities are not defined by *policy*, but the need is recognized, and the development is being discussed. | *Data owner* roles and responsibilities defined by *policy* are being developed. | *Data owner* roles and responsibilities defined by *policy* are validated by *stakeholders*. | *Data owner* roles and responsibilities defined by *policy* are established and adopted by the organization. | *Data owner* roles and responsibilities defined by *policy* are established as part of business-as-usual practice with continuous improvement. |

## 1.2.2 DATA OWNERSHIP IS ESTABLISHED IN THE CLOUD

### DESCRIPTION

Identifying and assigning ownership for data that resides in a cloud environment should follow the same *guidelines* for on-premises data ownership. Ownership of all *data elements* in any *data domain* within a cloud environment is mandatory and specified by *data management policy* and *standards*.

It is essential to specify data ownership for all data categories.

- **Source data** – data migrated from on-premises data stores or other cloud environments, or data created within the cloud environment such as a *system of record* hosted in the cloud.
- ***Derived data*** – data that uses any existing input data to create new data. Whether generated in a cloud environment or elsewhere, *derived data* will most often consist of data generated from calculators, *models*, metrics, aggregations, return datasets and materialized views.
- **Log data** – data that tracks usage, activities and operations in a cloud environment. The owner of log data is typically the technology *function* that is different from the operational *data owner*. Log files are critical for data privacy, compliance, auditing and organization information barriers.
- ***Third-Party Data*** – data inbound to a cloud environment from an external source, such as public data, open data, client *reference data*, instrument data, and other counterparty data.

### OBJECTIVES

- Ensure that data ownership is consistently assigned and maintained, whether the data resides on-premises or in a cloud environment.
- Gain approval and adopt cloud environment data ownership and accountability *policy*, *standards* and *procedures* that apply consistently across on-premises and cloud environments.
- Establish data ownership before any *data consumer* engages with the data.
- Track data ownership events and changes in each cloud environment according to *data management policy* and *standards*.

### ADVICE FOR DATA PRACTITIONERS

A cloud environment exhibits a shared responsibility *model*. Consequently, data practitioners should work with their cloud and technology providers to establish data ownership for all data and *metadata* within—or exported by—a *data ecosystem*. While some *data management* responsibilities belong to the *cloud service provider* (*CSP*), all data ownership must remain with the organization.

According to the organization's data management policy, managing data ownership in processes that import or add new data into an on-premises or cloud data ecosystem is essential. Develop and maintain an inventory of data to effectively manage data ownership assignments. Sufficiently document and maintain data ownership assignments as *metadata* and conduct periodic review and maintenance routines. It is also important to define ownership for both persistent and temporary data, such as data kept only for the duration of intermediate steps of a calculation.

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Cloud service and technology providers should ensure proper documentation of data ownership assignments for cloud data and *metadata*. This documentation should be created through automated *processes*. This service should support validation, maintenance and auditing of data ownership assignments.

## QUESTIONS

- Is data ownership consistently assigned and maintained across both on-premises and cloud environments?
- Have *policies*, *standards* and *procedures* been defined, verified, sanctioned, published and adopted for cloud and on-premises data ownership assignment?
- Is assignment of data ownership required before data is available for consumption?
- Have technologies been selected that record and track data ownership for all cloud environments?

## ARTIFACTS

- Data Management Policy, Standard and Procedure – defining and operationalizing *data owner* roles and responsibilities
- Process Documentation – inclusive of the required assignment of *data owner* to data in the cloud
- Data Catalog Report
  - Cloud data inventory with *data owner* identification
  - data owner log reflecting assignment and changes over time
- Data Management Tool Stack – inclusive of automated tools to support the required assignment of data ownership in the cloud

## SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| Formal data ownership is not established in the cloud. | Data ownership is not established in the cloud, but the need is recognized, and the development is being discussed. | Data ownership in the cloud is being developed. | Data ownership in the cloud is defined and validated by *stakeholders*. | Data ownership in the cloud is established and adopted by the organization. | Data ownership in the cloud is established as part of business-as-usual practice with continuous improvement. |

## 1.3 DATA SOURCING AND CONSUMPTION ARE GOVERNED AND SUPPORTED BY AUTOMATION

The organization must ensure that data is consumed from authoritative sources or authorized distributors, with data governance that manages the designation of this authority. Cloud platforms must provide automation to enforce consumption from authoritative sources and authorized distributors or highlight consumption from non-authoritative sources.

### 1.3.1 DATA SOURCING IS MANAGED AND AUTHORIZED

#### DESCRIPTION

A data source is an origination point for data that transfers into a primary system. Data sourcing is the act of locating and connecting to a data source, then ingesting data from that source. Data within a cloud environment may originate within that environment, an external cloud environment or on-premises environments. A data source may be one of several in a chain of data sources. An *authoritative data source* is a repository or system designated by a *data management* governing body as the primary or most reliable source for this information.

#### OBJECTIVES

- Formalize a *classification* scheme of *authoritative data sources* and their *provisioning points*.
- Obtain agreement on the usage requirements, system integrations and *provisioning points* for each *authoritative data source*.
- Educate *stakeholders* and *data consumers* about authoritative data sources.
- Establish *procedures* to identify, review and approve new *authoritative data sources* and their *provisioning points*.
- Enable discovery of each *authoritative data source* by authorized *data domains*, capture *metadata* that includes a scope definition.

#### ADVICE FOR DATA PRACTITIONERS

Managing the authorization of data sources is a function of data governance. Authorization and consumption of an *authoritative data source* should be standardized and be applied consistently across all organizational environments—whether in on-premises or cloud environments. Authorization and consumption may differ when comparing data sources that depend on data ingested into the cloud with data generated in the cloud.

A *data management* governing body designates a data source as authoritative when it is a definitive or *standard* source for one or more *data domains*. The use of an *authoritative data source* is typically governed by established *policies* of one or more organizations. The authority to make such a data source available for provisioning and consumption must be clear to all custodians and *data consumers*. To prevent the unauthorized proliferation of valuable data—and to ensure data integrity, validity, and security—it is essential to establish the responsibilities of data source administrators and *data consumers*.

The use of *authoritative data sources* may be constrained to a geography, product, business unit or time period. For an organization that accesses data from *authoritative data sources*, it is vital to establish *processes* supported by *policy*. These *policies* will ensure that *authoritative data sources* exhibit approved *provisioning points* and each data source is identified, approved, utilized for approved application development. Each data source should be periodically reviewed for *accuracy*, compliance and continuing value to the organization.

Data that has been ingested into a cloud environment may originate from other data sources. If necessary, it should be possible to determine that these data sources are authoritative. Data source authorization status and scope should be recorded in a central *data catalog* visible to *stakeholders*.

A common data sourcing use case involves creating a new authoritative cloud environment data source that consolidates data from disparate on-premises and other cloud environments. In such cases, a cloud environment may be created within the existing cloud environment, and such data may not necessarily reside in an authoritative data source. In all cloud environment scenarios, using *authoritative data sources* must be explicitly required by *policy* and approved by the organization.

Any data that resides within a cloud environment or originates from a source external to the cloud environment should be subject to review to determine whether it is authoritative or not. Unless explicitly known at inception, any new data source should be designated as non-authoritative to ensure that a review occurs to confirm that the data source is authoritative. When practicable, automate data ingestion *processes* to send alerts when new data is created and trigger a review when necessary.

Establish and conduct periodic reviews of all data sources. Such reviews should include existing and prospective *authoritative data sources*. These reviews should also consider whether existing *authoritative data sources* continue to satisfy organizational *policies*. Any sources that are no longer compliant should be removed.

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Ensure *data catalogs* provide *metadata* tagging capabilities for identifying the status and scope of *authoritative data sources*. Set the default status of any new data sources to be non-authoritative and prompt *stakeholders* to determine the status of each.

Provide *processes* and controls to align authorized *provisioning points* in the cloud environment with each authoritative data source. Provide data source consumption reports that allow *stakeholder* review of *authoritative data sources*. Provide methods for easily discontinuing *authoritative data source* designation for sources that are no longer viable or complaint.

Offer functionality that automates data source authorization workflows initiated by change events and provides status visibility to all *data consumers* of *authoritative data sources*. Provide methods for verifying, making connections and consuming *authoritative data sources*.

Provide strategic advice for maximizing the value of managing *authoritative data sources* in the cloud environment.

## QUESTIONS

- Has a *classification* system been formalized to approve *authoritative data sources* and their *provisioning points*?
- Has the agreement been obtained on data use requirements, obligations and *provisioning points* for each authoritative data source?
- Are education initiatives in place for *stakeholders* and *data consumers* to create and maintain an understanding of *authoritative data sources*?
- Have *procedures* been established to identify, approve and review new *authoritative data sources* and *provisioning points*?
- Has *metadata* been captured and made available to discover *authoritative data sources* by *data domains*—including the scope of use for the data source?

## ARTIFACTS

- Data Standards – authoritative source methodology overview, data source *classification* scheme, requirements and obligations
- Developer Guide – instructions on how to discover *authoritative data sources*
- Communication Plan – briefing document that describes *authoritative data sources* in use by the organization
- Data Management Procedure – defining and operationalizing data source identification and review
- Data Catalog – directory of all active, *authoritative data sources*

### SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal management and authorization of data sourcing exist. | No formal management and authorization of data sourcing exist, but the need is recognized, and the development is being discussed. | Formal management and authorization of data sourcing are being developed. | Formal management and authorization of data sourcing are defined and validated by *stakeholders*. | Formal management and authorization of data sourcing are established and adopted by the organization. | Formal management and authorization of data sourcing are established as part of business-as-usual practice with continuous improvement. |

## 1.3.2 DATA CONSUMPTION IS GOVERNED AND SUPPORTED BY AUTOMATION

### DESCRIPTION

Data consumption and usage from any environment are largely governed by sourcing from *authoritative data sources*—respecting all applicable legal, ethical and organization *policy* restrictions. Cloud platforms should enforce controls to ensure that data is consumed from *authoritative data sources*. Consuming applications must specify the required data and reference *data catalog* entries, while the cloud platform should automate the access and transfer of data from *authoritative data sources*.

### OBJECTIVES

- Document in *data sharing agreements* all data consumption allowances and restrictions as required by the organization's *policies*.
- Ensure that each data access requests include *metadata* that specifies the intended use of the data.
- Ensure each requested *data element* can be mapped to an authoritative data source.
- Implement reporting to track the use of *authoritative data sources* and govern the use of non-*authoritative data sources*.
- Exploit *metadata* to automate data provisioning and consumption.

### ADVICE FOR DATA PRACTITIONERS

For many organizations, using a cloud platform can change the perception of automated data provisioning from being a best practice to becoming a necessity. Full automation requires rich *metadata* in *data catalogs*, facilitating access requests to authoritative sources and providing access to the data. For example, a *data set* entry in the catalog would include an API specification and location and either an endpoint or information for navigating the virtualization layer.

Automating correct data consumption requires a comprehensive *taxonomy* that specifies conditions for access, use, allowances, and restrictions.

Data access event logging should always be in place—both for auditing and governance purposes. APIs for data provisioning and consumption are a common method for enforcing logging and automating reporting. Tracking the use of non-authoritative sources clarifies the extent of data distribution and is especially important when consuming sensitive data. Refer to *CDMC 3.2 Ethical Access, Use, & Outcomes of Data Are Managed*.

Implementing automated provisioning can be configured to provide additional control by preventing the consumption of non-authoritative sources. Automation also ensures that *data lineage* *metadata* is properly maintained.

Documenting best practices for creating and using provisioning and access APIs is critical to the automation's implementation and support. Exploiting cloud computing capabilities for reporting on data storage, throughput volumes, user access and data usage can provide valuable insights to *data owners*.

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Cloud service and technology providers should deliver capabilities that support the ability of *data owners* to track and control the distribution and consumption of data. *Data owners* should conduct various types of reviews and controls that correspond to the *data classification*.

Providers should furnish APIs to support automating the provisioning and consumption of data. These APIs should integrate with the *data catalogs* to enforce consumption from authorized sources and capture consumption events from non-authoritative sources. APIs should also be available to log provisioning and consumption events at a low level of detail. The logs should be available for audit and reporting purposes. Cloud service and technology providers should supply documentation on best practices for provisioning and access APIs and provide these documents to data practitioners to support implementation and automation.

Providers should offer integrations with workflow functionality for exception reporting and approval of consumption from non-authorized data.

## QUESTIONS

- Are allowances and restrictions regarding data consumption documented in *data sharing agreements* as required by the organization's *policies*?
- Does each data access requests include *metadata* that specifies the intended use of the data?
- Can each requested *data element* be mapped to an authoritative data source?
- Is there reporting to track the use of authoritative sources and govern the use of non-authoritative sources?
- Has *metadata* been exploited to automate data provisioning and consumption?

## ARTIFACTS

- Data Management Policy, Standard and Procedure – defining and operationalizing *data sharing agreements*
- Data Sharing Agreements – including allowances and restrictions captured as *metadata*
- Data Use Taxonomy
- Data Catalog – mapping *data elements* to authoritative sources
- Data Catalog Reporting – with consumption information highlighting the use of authoritative and non-authoritative sources
- API Documentation – detailing integration with *data catalogs* and with guidance to support the implementation of automation

### SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal governance and automated support of data consumption exist. | No formal governance and automated support of data consumption exist, but the need is recognized, and the development is being discussed | Formal governance and automated support of data consumption are being developed. | Formal governance and automated support of data consumption are defined and validated by *stakeholders*. | Formal governance and automated support of data consumption are established and adopted by the organization. | Formal governance and automated support of data consumption are established as part of business-as-usual practice with continuous improvement. |

## 1.4 DATA SOVEREIGNTY AND CROSS-BORDER DATA MOVEMENT ARE MANAGED

The sovereignty of data in cloud environments must be tracked. This information must be used to ensure that the storage and cross-border movement and use of data conform to the relevant jurisdictional requirements.

### 1.4.1 SOVEREIGNTY OF DATA IS TRACKED

### DESCRIPTION

As it becomes easier to allocate resources in the cloud, establishing preventative controls becomes more important. These controls must ensure *data sovereignty* requirements are enforced throughout the *data lifecycle*. Data transfers from one data center to another may result in data movements from one jurisdiction to another. Because of these concerns, organizations need to track and report on *data sovereignty* to demonstrate compliance with various complex requirements. All *data assets*' locations, content, and *data sovereignty* attributes should be clear, accurate, and readily accessible.

### OBJECTIVES

- Establish a *policy* detailing the principles and decision rights for managing a cloud data storage location.
- The *data catalog* defines and captures *metadata* for sovereignty requirements, location, and jurisdiction for cloud *data assets*.
- Ensure *data sovereignty* requirements and restrictions are understood and reflected in *data sharing agreements*.
- Manage the *data sovereignty* impact of any cloud resources or data relocation, change in access or changes to *processes*.
- Establish a process to assess changes in *data sovereignty* regulations and requirements.

### ADVICE FOR DATA PRACTITIONERS

A *data sovereignty policy* defines the types of data and jurisdictions where data is processed and stored. Such *policies* help organizations mitigate legal, technical and business risk issues raised by *data sovereignty*.

Adopting various regulatory frameworks that govern how data is hosted and processed can result in complex compliance environments. Some of the data content, such as their sensitivity, may introduce additional regulations and additional complexity. It is important to have a framework that will provide *guidelines* to cloud service providers on disclosing server locations and providing notice of location changes.

*Data sovereignty* rules can be classified with various taxonomies, One type of *classification* is the source of rules and regulations:

- Governments
- Industry regulators
- Terms of contracts

Alternatively, *data sovereignty* rules can be grouped by the type of action:

- The need to store or not to store data in a particular geographical location.
- The need to store copies of data in a particular geographical location.
- The need to store data using specific security controls, such as *encryption*.
- The need to comply with *processing* rules or accessing the data from within a certain jurisdiction.
- The need to comply with rules about how the data may be used.

There are also various definitions of *personal data* and other types of data in the scope of regulation. Various types of protection are available to accommodate different types of data. In some jurisdictions, there are also specific requirements on access to data. These requirements encompass governmental access, sufficient and timely access to the data for regulators, and *data security* and response obligations. Additionally, there are rules on *data sovereignty* based on the company's origin instead of the local company.

The different types of *data sovereignty* rules are summarized in the following table:

| **Rule type** | **Comment** |
|---|---|
| No-transfer rules | Requirement for data to be kept in specific jurisdictions, including copies made for recovery and infrastructure purposes. |
| Non-personal data restrictions | Restrictions on non-personal data. |
| Outsourcing restrictions | Restrictions on outsourcing of data handling services. |
| Consent restrictions | Data transfers are prohibited unless the individual's explicit consent is given. |
| Infrastructure rules | Requirement for data to be stored and processed by specific methods in the named jurisdictions. |
| Local copy rules | Requirement for a local copy of specific information must be maintained in the country of origin. Typically, this is contained within the database backups. |
| Equivalent *standards* | Allowance for data transfers to a jurisdiction with identical or equivalent data handling rules. |

Having accurate data cataloging and *data classification* is a strong enabler for *data sovereignty* tracking. Capturing *metadata* supports the capturing of *data sovereignty* requirements, outlines the restrictions on the countries to which data may be transmitted, and captures information about the location of each *data element*. *Data sovereignty metadata* is especially important for any data that crosses one or more jurisdictions.

Data practitioners must also understand how their technology providers manage online and offline backups, long-term storage, distribution and temporary data persistence to ensure they meet the *data sovereignty* requirements of the organization. Be sure to explain *data sovereignty* requirements and restrictions in any *data sharing agreement*.

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Cloud service and technology providers should provide services that support *data sovereignty* management and tracking. Most importantly, providing methods for specifying and enforcing rules established by organizational *policy* to meet *data sovereignty* requirements. Also, it is important to provide *metadata* for storage locations and provide the ability for an organization to specify the region(s) in which the data is located. In addition, an organization should have full transparency about the management of online and offline backups, long-term storage, distribution and temporary data persistence.

## QUESTIONS

- Has a cloud data storage location management *policy* been defined?
- Are *metadata* requirements defined to capture sovereignty requirements, location and jurisdiction for *data assets*?
- Are the geographic location and jurisdiction *metadata* captured in the *data catalog*?
- Have *data sovereignty* requirements and restrictions been documented, agreed and incorporated in *data sharing agreements*?
- Is there a process to assess and manage the impact of the relocation of data access or *processing*?
- Is a process in place to review the impact of changes in *data sovereignty* regulations and requirements?

## ARTIFACTS

- Data Management Policy, Standard and Procedure – defining and operationalizing principles and decision rights for cloud data location management
  - for resource reallocation impact
  - for security controls review
  - for impact analysis of *data sovereignty* regulation and requirements changes
- Data Sharing Agreements – including *data sovereignty* requirements and restrictions captured as *metadata*
- Data Catalog Report – showing the geographic location and jurisdiction *metadata*
- Data Sovereignty Requirements Document

## SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal *data sovereignty* tracking exists. | No formal *data sovereignty* tracking exists, but the need is recognized, and the development is being discussed. | Formal *data sovereignty* tracking is being developed. | Formal *data sovereignty* tracking is defined and validated by *stakeholders*. | Formal *data sovereignty* tracking is established and adopted by the organization. | Formal *data sovereignty* tracking is established as part of business-as-usual practice with continuous improvement. |

### 1.4.2 DATA SOVEREIGNTY AND CROSS-BORDER DATA MOVEMENT RISKS ARE MANAGED

#### DESCRIPTION

Organizations must establish controls to manage risks associated with *data sovereignty* and cross-border data movement. When possible, cloud services should be leveraged to automate these controls.

#### OBJECTIVES

- Document *data sovereignty* and cross-border data movement requirements and rules.
- Define and gain approval for a comprehensive set of *data sovereignty* and cross-border data movement controls.
- Specify, design and implement functionality to support enforcement of *data sovereignty* and cross-border data movement controls.
- Ensure *data sovereignty* and cross-border data movement controls are enforced consistently across each of the jurisdictions and environments in the organization.
- Establish ongoing automated monitoring and reporting, providing *evidence* of the effectiveness of *data sovereignty* and cross-border data movement controls.

#### ADVICE FOR DATA PRACTITIONERS

Many organizations seek to modernize their approaches to *data management* and aim to balance control and accessibility. Following the dominant trends in *enterprise* *data sovereignty*, these organizations use tools and techniques to automate cross-border *data management* and unlock data value.

The hurdles that many organizations face during this process are diverse. Perhaps the biggest challenge is the automation of *data classification* and data stewardship duties. Another significant challenge is to find a solution for opening data across organizational boundaries and functional *domains* while maintaining correct data governance principles. In addition, global organizations must comply with regulatory and privacy requirements that vary significantly by jurisdiction.

Faced with large sets of complex regulations, many organizations realize they don’t have a clear view of what data exists in each location and how it is used across geographies and by which external organizations.

*Data management* tools and cloud services can help organizations implement the standardized organizational *policy* for data locations and automated tagging of the new data according to rules and *role-based access controls*. Standardization ensures that *data sovereignty* rules automatically apply to all *data elements*—across various organization systems, departments, and business functions. Automatic tagging for *data sovereignty* and jurisdictional attributes ensures that *data assets* are readily discoverable by authorized users.

Global organizations should use tools and automation frameworks to support the *data sovereignty* requirements of different jurisdictions and establish an automatic auditing and approval process for access to the data from various departments in the organization that need access to the data. Such automation should include approvals for authorized users in specific jurisdictions who are entitled to access specific data.

Wherever *data sovereignty* and cross-border data movement regulations or organizational *policy* stipulate, verify that the organization physically separates the data for each jurisdiction. Also, the organization should consider if *data asset* derivatives that are subject to *data sovereignty* and cross-border data movement rules are also subject to jurisdictional regulations. In some cases, sufficiently abstracted derivatives can be transferred unobstructed throughout global organizations.

When possible, the organization should employ *encryption*, advanced security features, access controls, data masking and obfuscation techniques and backup services available from *cloud service providers* to comply with applicable regulations. Refer to *CDMC 4.1 Data is Secured, and Controls are Evidenced*.

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

The cloud service and technology provider should provide services for organizations to use local cloud offerings in jurisdictions that mandate *data sovereignty* rules. Providers should offer the ability to enforce *data sovereignty* organizational *policies*. Most importantly, the provider should prevent the creation of the resources outside specific locations or data transfers outside specific jurisdictions. For example, it must be possible to develop specific rules that will stop data transfers from the physical location inside jurisdiction and prevent access to the *data assets* by the operations teams from the outside jurisdictions. Data access controls should be configurable to permit or deny creating computational resources outside the jurisdictions that access the data.

In addition, providers should offer services that support organizational *policies* that are enforceable for specific locality rules, including rules for *data-at-rest*, *data-in-motion* and *data-in-use*. Specifying such rules at the organizational level will ensure standardization of the rules between different units and functions.

Providers should make available tools for data discovery that is within the scope of *data sovereignty* rules. These tools should also provide the ability to automatically tag data for the support of other automated *processes* that enforce the *data sovereignty* and cross-border movement of data rules. The organization should have the ability to enforce the locality of the data and configure the location *policy* for each resource.

An organization should have the ability to readily access provider reports on resources in specific jurisdictions or for specific *data sovereignty* data rule types. Reports should be available for the various types of *data assets* in use—including data files, backups and computational resources. In addition, an organization should automatically monitor and report on compliance with *data sovereignty* rules. Monitoring and reporting on the enforcement of the *data sovereignty* rules will highlight any current violations requiring remediation and flag activities causing rules violations.

Together with the provider, the organization must ensure that *data sovereignty* and cross-border data movement rules and restrictions apply to the operational *procedures* of the providers. This requirement includes cases where the provider needs to perform operations on the organization's resources as part of support case investigations or remediation actions.

The provider must provide *evidence* that it is adhering to compliance requirements that pertain to *data sovereignty* and cross-border data movement rules. This requirement includes cases where cloud service and technology providers manage services on the provider platform and cases in which partners provide integration, support or migration services.

## QUESTIONS

- Have *data sovereignty* and cross-border data movement requirements and rules been documented?
- Have *data sovereignty* and cross-border data movement controls been defined and approved?
- Has functionality to support enforcement of *data sovereignty* and cross-border data movement controls been specified, designed and implemented?
- Are *data sovereignty* and cross-border data movement controls enforced consistently across the jurisdictions and environments of the organization?
- Is ongoing automated monitoring, reporting and evidencing of the effectiveness of *data sovereignty* and cross-border data movement controls in place?

## ARTIFACTS

- Data Sovereignty and Cross-Border Data Movement Requirements Document – specifying rules and restrictions
- Data Sovereignty and Cross-Border Data Movement Controls Specification
- Functional Specifications – for automation of enforcement of *data sovereignty* and cross-border data movement controls

- Data Sovereignty and Cross-Border Data Movement Controls Report – showing the extent of implementation and evidencing control effectiveness

### SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal management of *data sovereignty* and Cross-Border Data movement risk exists. | No formal management of *data sovereignty* and Cross-Border Data movement risk exists, but the need is recognized, and the development is being discussed. | Formal management of *data sovereignty* and Cross-Border data movement risks are being developed. | Formal management of *data sovereignty* and Cross-Border data movement risks are defined and validated by *stakeholders*. | Formal management of *data sovereignty* and Cross-Border data movement risks are established and adopted by the organization. | Formal management of *data sovereignty* and Cross-Border data movement risks are established as part of business-as-usual practice with continuous improvement. |

## 1.5 GOVERNANCE & ACCOUNTABILITY – KEY CONTROLS

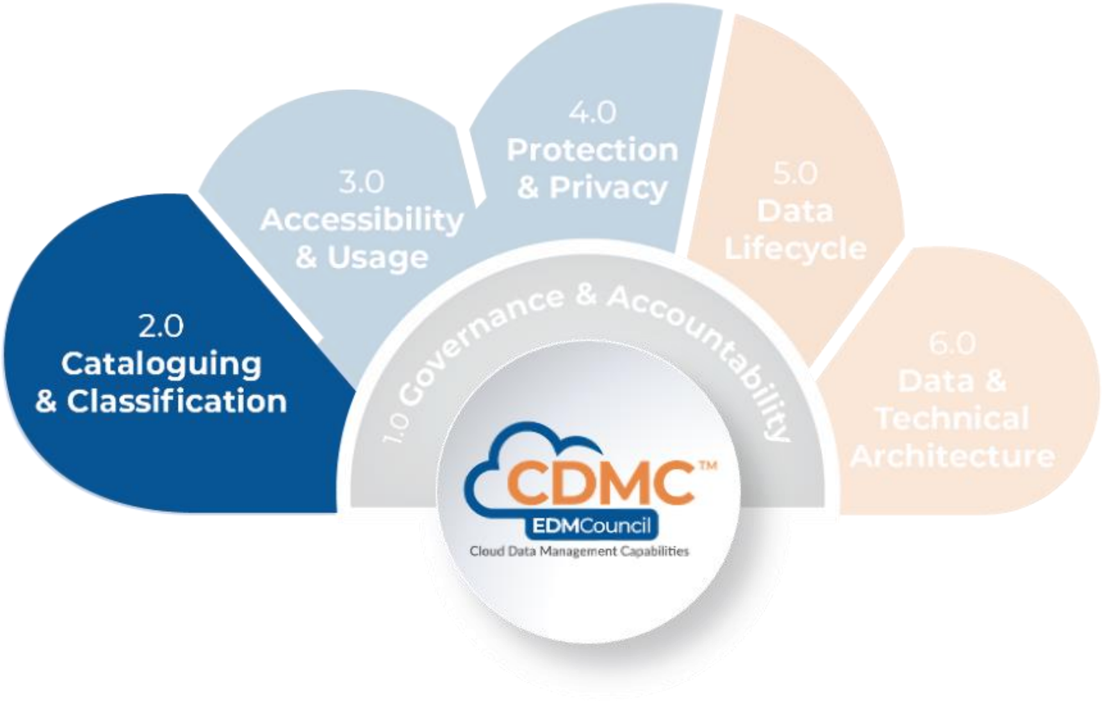
The following Key Controls align with the capabilities in the Governance & Accountability component:

- Control 1 – Data Control Compliance
- Control 2 – Ownership Field
- Control 3 – Authoritative Data Sources and Provisioning Points
- Control 4 – Data Sovereignty and Cross-Border Movement

Each control with associated opportunities for automation is described in *CDMC 7.0 Key Controls & Automations*.

*This page left intentionally blank*

# 2.0 Cataloging & Classification



# 2.0 CATALOGING & CLASSIFICATION

## UPPER MATTER

### INTRODUCTION

Effective cloud data management depends on having full control of all *data assets*. This understanding must include technical characteristics such as formats and data types and contextual information supporting the full CDMC capabilities. These capabilities include business definitions, *classifications*, sourcing, retention, physical location and ownership details. Together, these characteristics comprise the *data catalog*.

### DESCRIPTION

The Data Cataloging & Classification component is a set of capabilities for creating, maintaining and using *data catalogs* that are both comprehensive and consistent. This component includes *classifications* for *information sensitivity*. These capabilities ensure that data managed in cloud environments is easily discoverable, readily understandable and supports well-controlled, efficient data use and reuse.

### SCOPE

- Define the scope and granularity of data to be cataloged.
- Define the characteristics of data as *metadata*.
- Catalog the data and the data sources.
- Connect the *metadata* among multiple sources.
- Share *metadata* with authorized users to promote discovery, reuse and access.
- Enable sharing of *metadata* and data discovery across multiple catalogs, platforms and applications.
- Define, apply and use the *information sensitivity* *classifications*.

### OVERVIEW

Understanding *data assets* in context becomes central when managing those assets through infrastructure controlled by the cloud provider instead of the data organization. Understanding how a cloud provider will control *data assets* is critically important in regulatory requirements such as *data residency* and protection.

An essential task in structuring a *data catalog* is defining an appropriate granularity and breadth of *data assets* to include in the catalog. Care must be taken to assess the overall costs of cataloging driven by legal and regulatory constraints and the amount of data in the cloud. It is essential to compare these costs with the value to the organization.

Comparatively, some challenges of implementing *data catalogs* may be smaller in a cloud computing environment. Consider the following:

- Since a cloud environment tracks all *data assets* that it stores for metering and security purposes, the cloud environment provides a necessary foundation for cataloging that does not exist in many on-premises environments.
- Cloud platforms host relatively few types of data stores. These data types typically offer contemporary methods of integration, such as APIs. Conversely, a typical on-premises data landscape exhibits a much wider variety of data stores, including legacy technologies that may pose intractable challenges to *data catalog* integration.
- Given the near-instant availability of cloud infrastructure, potential time-to-market advantages may be realized by cataloging unstructured data with the aid of natural language *processing*. Other types of data may be easier to find through ontological discovery and exploration.

In maintaining a hybrid cloud environment that spans both on-premises and multiple cloud platforms, an organization will likely need to manage multiple *data catalogs* across multiple data storage technologies. Automation and cross-platform alignment will be critical to the interoperability of these disparate technologies.

*Information sensitivity classification* involves labeling *data elements* according to their business value or risk level. Data presents a 'business risk' if its disclosure, unauthorized use, modification or destruction could impact strategic, compliance, reputational, financial, or operational risk. This labeling is fundamental for security and regulatory compliance in all environments, especially diverse cloud environments with multiple suppliers. With multiple suppliers, data is likely to traverse multiple local or regional jurisdictions in a single workflow.

Implementation of *information sensitivity classification* in a cloud computing environment is essential to realizing these benefits:

- **Availability of new functionality.** Cloud technology provides new functionality, opportunities and approaches for data storage, management, access, cataloging, classification, movement, *processing*, archiving and permanent deletion. Integrating *information sensitivity classification* with this new functionality is vital to support a cohesive and seamless solution.
- **Increased potential for automation.** *Information sensitivity classification* provides a foundation for defining *business rules* that can consistently apply data usage, placement, encryption, distribution, and access across various legal or regulatory requirements.

Understanding the purpose and importance of *Information sensitivity Classification* must be cultivated across an organization through people, *processes* and technologies.

In addition to *information sensitivity classification*, organizations may apply additional *classifications* to support specific *business rules* and precisely manage various data treatments throughout the entire *data lifecycle*. A primary assumption is that all *information sensitivity* and other *classifications* will be captured as *metadata* in the *data catalog*. For an exhaustive list, refer to the *CDMC Information Model*.

*Metadata* within a catalog may contain sensitive information making it important to treat *metadata* itself as a *data asset*. Each *metadata* element should have an *information sensitivity classification* and be controlled by good *data management* practices such as access control and sensitivity tracking.

When it comes time to create that *data catalog*, all cloud *data assets* should be known. Also, each information security and data privacy risk should be known. This knowledge is vital to adopting and supporting the *information sensitivity classification* schemes that will control how to access, protect and manage the data through each stage in the *data lifecycle*.

## VALUE PROPOSITION

Organizations that create, maintain and share comprehensive *data catalogs* gain the ability to maximize controlled reuse of *data assets*.

Organizations that effectively support the *information sensitivity classifications* can benefit from enhancements in transparency and consistent treatment of *data classifications*. Maximum transparency on the precise locations of data storage and data transfer routes will enable automatic *processes* to manage, monitor and enforce the consistent data treatment according to a specific *information sensitivity classification*. Standardizing an *information sensitivity classification* functionality also enables an authorized automatic process to manage, monitor, enforce security and regulatory compliance across multiple jurisdictions. In some instances, *information sensitivity classification* applies to additional *classification metadata*.

## CORE QUESTIONS

- Is there a definition of *data asset* scope and granularity for all data that will be cataloged?

- Is there agreement on a *model* and supporting *standards* for data characteristics to be captured as *metadata*?
- Is there a plan for connecting the *metadata* across multiple *data catalogs*?
- Is the *metadata* available to users and applications to promote discovery and reuse?
- Have *standards* been adopted that facilitate sharing *metadata* and data discovery across catalogs, platforms applications?
- Has an *information sensitivity classification* system been defined, supported in the *data catalogs* and used to control data access and use?

## CORE ARTIFACTS

- Data cataloging strategy and scope
- *Metadata* information *model* and naming standard
- Inventory of platforms and applications to support *data catalog* interoperability
- System interface definitions for machine-readable access to *metadata* in the catalog
- Interchange protocols for controlling the sharing and modification of *metadata* across platforms
- Data Management Policy, Standard and Procedure – defining and operationalizing the *information sensitivity classification* scheme and corresponding *business rules*

## 2.1 DATA CATALOGS ARE IMPLEMENTED, USED AND INTEROPERABLE

*Data catalogs* describe an organization's data as *metadata*, enabling it to be documented, discovered and understood. The data cataloging scope and approach must be defined. Catalogs must be implemented and populated with the *metadata* that describes the data. This *metadata* must be exposed to both users and applications, and *standards* should be defined and adopted to ensure that *metadata* can be exchanged between catalogs on different platforms.

### 2.1.1 DATA CATALOGING IS DEFINED

#### DESCRIPTION

Data cataloging is the process of collecting, organizing, and displaying *metadata* that pertains to *data assets* and is presented in a *data catalog*. Effective *data management* in a cloud computing context depends on a clear understanding of *metadata* that describes the content, source, ownership and other aspects of the *data assets*. *Data catalogs* describe technical information about the *data assets*, such as formats and data types. These catalogs also include contextual information such as classification, ownership, and residency requirements. Business and regulatory requirements define the scope of *data assets* managed by a particular *data catalog* in all data storage environments.

#### OBJECTIVES

- Define the scope of the data to include in the catalog.
- Align each catalog with the business strategy and in consideration of its risk appetite and control framework.
- Define the granularity and types of *data assets* that will be part of the catalog.
- Define the key characteristics of all *data assets*, including relationships among them.
- Define how *metadata* is sourced.
- Provide a catalog definition and scope that will enable *data consumers* to find and understand *data assets* easily.

#### ADVICE FOR DATA PRACTITIONERS

The purpose of *data catalog*ing is to provide a means for fully understanding information about all business *data assets*. The *data catalog* is the repository for identifying, understanding and managing all *data assets*. In addition, the *data catalog* supports ethical, legal, and regulatory compliance issues—for both individuals and *processes*.

Data cataloging involves a level of effort and cost. It is important to develop and implement a data cataloging strategy as part of an overall data strategy. This strategy must include regulatory requirements, business needs, and ethical considerations. It must clearly describe the value to be achieved. While it is important to inventory and document all *data assets*, the granularity of the descriptions and method of contextualization will vary according to the business value.

The business value criteria should define the Key Performance Indicators (KPIs) for revenue generation or cost reduction and the Key Risk Indicators (KRIs) to mitigate risk to the *data assets*. The scope may include any data that the organization requires, including data from internal and external sources.

Examples of *metadata* that *data catalogs* maintain include business terminology, technical metadata such as data types, formats, technical containment, and *data models*. *Metadata* may also include information about data services such as Application Programming Interface (APIs), business *domains*, ownership, licensing, and data movement. In addition, *metadata* may describe data stored in various forms: structured, unstructured and semi-structured.

Always on data cataloging is essential for capturing sufficient *metadata* about each in-scope *data asset*. In addition, *processes* and technologies must be readily available for data specialists to perform data cataloging operations. Also, it is vital to define principles and capabilities for automatically discovering the minimum *metadata* for each in-scope data asset—either at the point of entry or the point of creation. All data definition and *data management* technology must support these principles. Note that a minimum level of data cataloging capability does not impact the availability or security of the contents of the *data catalog*. It may be entirely impracticable for human agents to maintain elements of the catalog. Consequently, automatic *metadata* discovery and capture are important for efficiency and scale.

To develop a flexible, descriptive, and efficient cataloging service, an organization may implement one or more *data catalog* technologies. It is important to ensure compatibility and consistency with on-premise, hybrid and multi-cloud environments when evaluating different offerings.

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

All data within a cloud environment must be inventoried, at least to a level of granularity that will support usage metering. For each data asset, some minimum amount of *evidence* must be available to be included in an inventory. Some *evidence* may be available in some environments, but it may be insufficient to capture the minimum amount of *metadata* for the *data catalog*. It should be feasible to enrich such *evidence* with additional contextual *metadata* or integrate it into a *data catalog* to support business and regulatory requirements. Tools that feature sufficient interoperability will capture minimal levels of *evidence*. Refer to *CDMC 2.1.3 Data catalogs are interoperable across multi and hybrid cloud environments*.

Data sensitivity and ethics are key considerations when dealing with *data assets* managed in a cloud computing environment. For such *data assets*, the *metadata* itself may contain *sensitive personal data* or other *secrets*. In these cases, the *metadata* is itself a data asset, and its treatment must follow *data management policies* like access control and sensitivity tracking. Catalog interoperability becomes a very important requirement for ensuring consistent and secure metadata management across all data catalogs when managing sensitive metadata.

## QUESTIONS

- Has a *data catalog* strategy been defined, published, and communicated to the *stakeholders*?
- Has the *data catalog* strategy been implemented?
- Have data cataloging *policies*, *standards* and *procedures* been defined, verified, sanctioned, and published?
- Do the *policy*, *standards*, and strategy documents identify the scope of *data assets* and *metadata*?
- Does this scope align with the business objectives and strategy?
- Have technologies been chosen to support data cataloging by capturing and maintaining *metadata* for in-scope *data assets*?
- Has *data catalog* governance been aligned with current change-management and data-management *policies*?

## ARTIFACTS

- Data Cataloging Strategy and Scope
- Data Management Policy, Standards and Procedures – defining and operationalizing data cataloging
- Data Catalog Implementation Roadmap
- Data Catalog Architecture Document

### SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal data cataloging exists. | No formal data cataloging exists, but the need is recognized, and the development is being discussed. | Formal data cataloging is being developed. | The formal data cataloging is defined and validated by *stakeholders*. | Formal data cataloging is defined, scoped and used by the organization. | Formal data cataloging is established as part of business-as-usual practice with continuous improvement. |

## 2.1.2 METADATA IS DISCOVERABLE, ENRICHED, MANAGED, AND EXPOSED IN DATA CATALOGS

### DESCRIPTION

A *data catalog* promotes efficient data reuse by describing underlying data with *metadata*. Good *metadata* properly identifies, documents and elaborates on *data elements* available across an entire cloud *data architecture*. Effective *metadata management* and enrichment promote efficient data reuse. Examples include capturing the key *data residency*, classification, ownership, licensing and data protection of cloud *data assets*. Automating metadata harvesting and definition is vital to scale *metadata management* efforts across extensive cloud *data architectures* and on-premises platforms.

### OBJECTIVES

- Define key *metadata management* capabilities in a *data catalog*, so underlying data is readily discoverable, highly organized, well-managed, open to enrichment and easily consumed by humans or by an automatic process.
- Source *metadata* and connect it to other *data assets* within the catalog to assist *data management processes*, improve usability and enhance the data efficacy.
- Correlate *data assets* to identify commonality, minimize duplication and promote contextual discovery.
- Promote the reuse of *data assets* by readily exposing *metadata* through the catalog.
- Ensure quick definition of any mandatory *metadata*, and correlate that *metadata* with underlying *data assets*. *Metadata* should include, for example, semantic meanings for *data elements* that support regulatory compliance.

### ADVICE FOR DATA PRACTITIONERS

Wherever possible, *metadata* should be harvested automatically from as many *data assets* as possible. *Metadata* from internal and external sources should be kept up-to-date and immediately accessible within the *data catalog*. Automation is essential to efficient and effective maintenance of the *data catalog*. However, it is important to understand that automatic discovery may only provide a portion of the necessary *metadata*, which is the case with *technical metadata*. Consequently, methods must be available to create and enrich *metadata* manually. These methods must be governed with appropriate controls and are especially important when it is known that automatic discovery is unavailable or insufficient.

Promoting and implementing data reuse depends on the ability to find relevant data quickly. *Metadata* should be easily searchable and accessible at a suitable level of granularity—either by an individual or process.

The discovery of semantic relationships among various *data assets* and the ability to find related *data assets* promotes data usage. It is important to automate the collection and discovery of relationships among *data assets* and other *metadata* wherever possible. The need to automate is especially true for large, complex systems involving multiple data stores. Also, it should be possible to link and enrich *metadata* manually. To further promote *data asset* reuse, consider collaborative enrichment of *data assets*, such as tagging, commenting, rating, bookmarking, notifications and workflows.

The capture of *metadata* to support *data management* should be practiced through all data life cycle stages. Captured *metadata* includes design changes, implementation and extension of data stores and deployment of *data assets*. The *data catalog* should automatically define, capture, relate, and share *metadata* if it is practical. Examples include:

- **Metrics, KPIs and SLAs and data ownership** – to support *data profiling* and quality management. Refer to *CDMC 5.2.2 Data quality is measured,* and *CDMC 5.2.3 Data quality metrics are reported*.
- ***Classification* of data properties** – to specify sensitivity. Refer to *CDMC 2.2.1 Data classifications are defined*.
- **Capture of provenance information, including data origin and footprint** – to support tracing and authoritative sourcing. Refer to *CDMC 1.3.1 Data sourcing is managed and authorized,* and *CDMC 6.2.1 Multi-environment lineage discovery is automated*.
- **Lifecycle *metadata*** – to manage dataset maturity and enable *records* retention, archival, *disposal policy*. Refer to *CDMC 5.1.1 A data lifecycle management framework is defined*.
- **Usage *metadata*** – to audit access, purpose, sharing and ethical use of data. Refer to *CDMC 3.1 Data Entitlements are Managed, Enforced and Tracked,* and *CDMC 3.2 Ethical Access, Use & Outcomes of Data are Managed*.

**NOTE**: The list above is a summary guide to implementing data cataloging, so it is not exhaustive. Keep in mind that the *data catalog* provides a convergence point for all of this *metadata*.

The *data catalog* should support an information *model* and naming *standards* that satisfy the cataloging requirements of the adopting organization. The catalog should also support interoperability with other *data catalogs* and other cloud and on-premises capabilities. Refer to *CDMC 2.1.3 Data catalogs are interoperable across multi and hybrid cloud environments*.

The *data catalog* should offer the ability to maintain multiple versions of *metadata*, track user actions for auditability and maintain a history of *metadata* to support point-in-time inquiries. Each of these capabilities is important for regulatory and compliance purposes.

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

The *metadata* should be captured alongside the data to ensure it remains up-to-date. Beyond basic technical details, it is important to know the broader *metadata* requirements noted above in ADVICE FOR DATA PRACTITIONERS. Verify that it will be possible to add this additional *metadata* into the *data catalog* or integrate basic metering *metadata*. Interoperability support will make it much easier to integrate. Refer to *CDMC 2.1.3 Data catalogs are interoperable across multi and hybrid cloud environments*.

## QUESTIONS

- Have business and technical users been engaged to define cataloging capabilities, including any ethical concerns?
- Have capabilities been implemented for automatic discovery and enrichment of *metadata*?
- Are relationships actively maintained among the *metadata*, for example, between conceptual terminology and physical *data elements*?
- Are changes in data and *metadata* captured, the changes logged, user actions logged and all critical changes monitored for auditing purposes?

- Are operational Key Performance Indicators (KPIs), metrics and *Service Level Agreements* (SLAs) defined, produced and regularly shared to improve cataloging efficiency and effectiveness?

### ARTIFACTS

- Data Cataloging Principles and Strategy
- Data Catalogs
- Data Catalog Capabilities Implementation Roadmap
- Data Catalog Capabilities Release Notes and Schedule
- Data Catalog Capabilities Communication, Training and Adoption Plan
- Data Catalog Usage Metrics
- Metadata Refresh Log

### SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal *standards* exist for *metadata* discovery, enrichment, management and exposure. | No formal *standards* exist for metadata discovery, enrichment, management and exposure, but the need is recognized, and the development is being discussed. | Formal *standards* for *metadata* discovery, enrichment, management and exposure are being developed. | Formal *standards* for *metadata* discovery, enrichment, management and exposure are defined and validated by *stakeholders*. | Formal *standards* for *metadata* discovery, enrichment, management and exposure are established and adopted by the organization. | Formal *standards* for *metadata* discovery, enrichment, management and exposure are established as part of business-as-usual practice with a continuous improvement routine. |

## 2.1.3 DATA CATALOGS ARE INTEROPERABLE ACROSS MULTI AND HYBRID CLOUD ENVIRONMENTS

### DESCRIPTION

*Data catalog* interoperability is an important capability in *multi-cloud* or *hybrid-cloud* environments. Catalogs should provide the ability to share information across different *cloud service providers*, technology providers and on-premises catalogs.

Enabling *data catalog* interoperability between platforms and applications is achieved by defining structure using:

- Catalog *metadata* naming *standards* and a catalog information *model* for *data sets*.
- Relationships between *data sets*.
- Data services.

Establishing these *standards* and *models* is essential to integration and consistency when sharing or combining catalogs. Maintaining these *standards* and *models* is vital to *metadata* automation, data governance, quality monitoring, *data asset policy* enforcement and compliance capabilities for usage tracking.

Also, *data catalog* interoperability is enhanced by implementing system-level interface *standards* such as APIs that work with interchange protocols that support mutability, mastering and synchronization.

## OBJECTIVES

- Facilitate the sharing and use of common *metadata* across catalogs, platforms and applications for easy access or automatic synchronization.
- Enable a common and consistent understanding of underlying data across multiple platforms and applications.
- Support automatic enforcement of *metadata* *policies* such as data access controls across multiple platforms and applications.
- Support automatic enrichment of *metadata*, including *data quality* analysis and monitoring across multiple platforms and applications.
- Enable discovery of accessible data across multiple catalogs, platforms and applications.

## ADVICE FOR DATA PRACTITIONERS

Automatic discovery and maintenance of *metadata* are necessary to support *always-on* functionality and interoperability between catalogs, applications, and workflows. Refer to *CDMC 2.1.2 Metadata is discoverable, enriched, managed, and exposed in data catalogs*.

It is necessary to create naming *standards* and a *metadata* information *model* based on widely adopted open *standards* to support machine-readability by third-party platforms and applications that drive automation of *data catalog* *metadata*. Automatic synchronization with open *standards* changes ensures that the *data catalog* does not diverge from those *standards*.

The alignment of information *models* requires the definition of standard *metadata* types. These definitions describe how underlying *data assets* are to be defined and described. Such definitions will ensure consistency across multiple catalogs and the applications that use those catalogs. Refer to *CDMC Information Model* for further advice on the *model*. Adopting naming *standards* and a consistent information *model* requires business, data, technology, and regulatory compliance *stakeholders* to ensure a common understanding and consensus. Information *model* alignment supports automation in *metadata* enrichment, knowledge graph exploration, *data lineage*, data marketplaces, recommendation engines, data governance *policy* monitoring and enforcement, *data quality* monitoring, user access controls, usage tracking and compliance tracking and reporting.

After establishing an information *model* and catalog naming *standards*, these will need to be maintained through *policies* requiring the organization to abide by the *standards*. Any adjustments to the naming *standards* or the information *model* should follow the change management requirements defined by the data governance framework. These requirements may include change approvals, impact analysis, controlled implementation and rollout.

The best practices described in *CDMC 1.2 Data ownership is Established for Both Migrated and Cloud-generated Data* apply to *metadata*, such as establishing ownership for *authoritative data sources* and *data sharing agreements* on *metadata* that support multiple catalogs.

*Data catalog* security is essential for interoperability and also alignment with *metadata* *entitlement*, privacy and security *policies*.

While interoperability makes it possible to share *metadata*, the data consumer's responsibility is to manage the ethical sharing of *metadata* across multiple platforms. It is particularly important to limit the sharing of commercially sensitive catalog information between *cloud service providers*. For example, data metrics and usage information from one provider should not be shared with another.

The best practices described in *CDMC 3.0 Data Accessibility and Usage* apply to *metadata* *entitlements*.

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Establishing open naming *standards* and open information *models* is important to ease integration among platforms and reduce adopters' burden.

Establishing a system-level interface that specifies how platforms and applications access *metadata* assets within the *data catalog* is essential to drive automation. Examples of a system-level interface include an API, event-based mechanisms and semantic web traversals. Interchange protocols should also be maintained through *policy* to ensure that shared information is properly managed and risks are minimized for losing the source of truth, such as inconsistent changes in multiple locations.

The *cloud service provider* must supply transparency on the treatment of *metadata* to enable organizations to ensure proper isolation between cloud platform tenants. Therefore, any *metadata* shared through these interoperability mechanisms can be strictly controlled through *entitlement*, privacy and security *policies*.

## QUESTIONS

- Have naming *standards* been established for all *data catalogs*?
- Are catalog naming *standards* consistent and in proper alignment?
- Has an information *model* been created or adopted for catalog *data asset* definitions?
- Are *data catalogs* portable or readable by external applications?
- Is there documentation and tooling for onboarding a new external application to read the current catalogs?
- Is there a *procedure* for onboarding a new catalog that is to be readable by external applications?
- Is there a *procedure* for migrating existing catalogs?
- Are there proper controls to govern *metadata* changes to maintain the source of truth for the *metadata*?

## ARTIFACTS

- Catalog Metadata Naming Conventions
- Metadata Information Model and Naming Standard (Refer to the *CDMC Information Model*)
- Data Catalog Interoperability Technology Tool Stack
- System Interface Definitions – for machine-readable access to *metadata* in the catalog
- Interchange Protocols – for controlling the sharing and modification of *metadata* across platforms

## SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal *standards* exist for *data catalog* interoperability. | No formal *standards* exist for *data catalog* interoperability, but the need is recognized, and the development is being discussed. | Formal *standards* for *data catalog* interoperability are being developed. | Formal *standards* for *data catalog* interoperability are defined and validated by *stakeholders*. | Formal *standards* for *data catalog* interoperability are established and adopted by the organization. | Formal *standards* for *data catalog* interoperability are established as part of business-as-usual practice with continuous improvement. |

## 2.2 DATA CLASSIFICATIONS ARE DEFINED AND USED

From the very moment it is created, data can be both a liability and an asset. Poorly managed data is likely to pose a risk if used inappropriately or unauthorized users access it. Such risks increase in a cloud environment, and many organizations increase their exposure as they move massive amounts of critical data into the cloud.

An *information sensitivity classification* is a scheme for labeling *data elements* according to business risk level or value. Data presents a business risk if its disclosure, unauthorized use, modification or destruction could impact a strategic, compliance, reputational, financial, or operational risk. *Information sensitivity* specifies how to access, treat and manage a *data element* through each stage of its lifecycle. This labeling is essential to security and regulatory compliance in all applications and a growing portion of cloud environments.

### 2.2.1 DATA CLASSIFICATIONS ARE DEFINED

#### DESCRIPTION

The *information sensitivity classification* is defined and approved. *Business rules* that specify how the *classifications* apply to combinations and aggregations of individually classified data are defined and approved.

#### OBJECTIVES

- Define *information sensitivity classifications* within the *data management policy* such that the *classifications* are mutually exclusive and accurately reflect business risk levels and values.
- Define *business rules* to ensure consistent application of order of precedence for *information sensitivity classification*.
- Define *business rules* for classifying combinations of *data elements*. Some combinations of *data elements* will have a collective sensitivity that is greater than the individual *data elements*.
- Define *business rules* for aggregating the *classifications* of individual *data elements* to be held in a repository or moved to an application.
- Define *business rules* for treating unclassified data or setting a default *information sensitivity classification* for individual *data elements*.
- Define principles and *guidelines* that anticipate changes to *data classification* at some point in the *data lifecycle*.

#### ADVICE FOR DATA PRACTITIONERS

The purpose of *information sensitivity classification* is to identify the business risk and value of data. *Information sensitivity classifications* also constrain data accessibility and handling by *data management*, security and downstream business *processes*.

Set a *policy* and *guidelines* that enable fast and intuitive decisions concerning the *classification* of applications, documents, messages and files. Define and use labels and terms that are instantly recognizable and meaningful. Keep the number of different identifiers to a minimum to promote simplicity and consistency. An *information sensitivity classification* can be:

- **Derived according to the content.** Users may be required to identify the content type at the time of creation, or the capability may exist to analyze content to determine or constrain the *classification*.
- **User-driven.** Users may be required to choose the appropriate *classification*.

The *guidelines* should include how to detect and update changes. Potential conflicts between automatic assignment and manual user assignment must reconcile and cascade across all repositories.

Other *data classification* types may be necessary to support additional regulatory compliance requirements, risk reporting and specific business objectives. The assumption is that the use of data will be by *metadata* maintained in the *data catalog,* including the *information sensitivity classification*.

Implementing any *information sensitivity classification* scheme must align with the ethical review of the data access, use, and outcome. Refer to *CDMC 3.2 Ethical Access, Use, & Outcomes of Data Are Managed*.

*Information sensitivity classifications* must be mutually exclusive, where applicable. For example, the same *data element* cannot be simultaneously sensitive and public. There must be rules in place to ensure consistent application of order of precedence for *information sensitivity classification*. The order of precedence determines that the higher level applies when a conflict arises.

Formalize and document *information sensitivity classification* and any accompanying rules for the appropriate level of protection. These additional rules might govern whether an *information sensitivity classification* label is itself to be protected or otherwise obfuscated to protect the sensitivity of the underlying *data elements*.

If required by the *data management policy*, define and implement rules that specify a more stringent *classification*. For example, consider a simple database consisting of three tables, each with a confidential *classification*. Data that is accessed from all three tables simultaneously may be constrained with a highly confidential *classification*. For example, aggregation may render the sensitivity of the repository greater than any of the individual *data elements*. Another example is the three *data elements,* First Name, Last Name and Home Address, may not be sensitive individually. Still, these elements in combination likely identify a specific person and consequently is *Personally Identifiable Information* (*PII*).

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Whether on-premises or in cloud environments, *information sensitivity classification* is important to data and security management. However, a cloud environment may have a more complex control framework.

A suitable cloud environment should provide *metadata* functionality for each distinct data storage entity within that environment. This functionality must permit the assignment of an *information sensitivity classification* value for each *data element* supporting *cross-organization control function policies*.

## QUESTIONS

- Has a unique and precise *information sensitivity classification* scheme been defined and approved?
- Has the *classification* scheme been integrated with the *cross-organization control function policies*?
- Has it been embedded within the culture and aligned with data governance architecture?
- Have *business rules* been established to guide the *classification* of *data element* combinations, and do these rules also guide the inheritance of *classifications* to higher-level repositories and systems?

## ARTIFACTS

- Data Management Policy, Standard and Procedures – defining and requiring the assignment of *information sensitivity classification* schemes and corresponding *business rules*
- Information Sensitivity Scheme Specification

### SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal *data classification* schemes and *business rules* exist. | No formal *data classification* schemes and *business rules* exist, but the need is recognized, and the development is being discussed. | Formal *data classification* schemes and *business rules* are being developed. | Formal *data classification* schemes and *business rules* are defined and validated by *stakeholders*. | Formal *data classification* schemes and *business rules* are established and adopted by the organization. | Formal *data classification* schemes and *business rules* are established as part of business-as-usual practice with continuous improvement. |

## 2.2.2 DATA CLASSIFICATIONS ARE APPLIED AND USED

### DESCRIPTION

An *information sensitivity* label must be assignable to all individual *data elements* and collections and aggregations of *data elements* where possible. An *information sensitivity* label is useful for controlling data access, treatment, and management in each *data lifecycle* stage.

### OBJECTIVES

- Implement *classification* schemes across all on-premises and cloud *data assets*.
- Implement *classification* at the point of creation.
- Support classification with technology that analyzes the content and continually assigns or guides *classification* decisions.
- Provide users the ability to validate any automatic *classification processes*.
- Configure downstream data governance and security solutions to apply *information sensitivity classifications* as the basis for jurisdictional placement, *encryption*, distribution, access and usage.

### ADVICE FOR DATA PRACTITIONERS

Automating the *information sensitivity classification processes* promotes consistency of *classifications*.

*Classification* is always required and must be always on. Best practices in cloud *data management* involve establishing the appropriate *data classifications* before ingesting data into the cloud environment. An organization must define rules for setting default classification for new data elements and any future data changes to achieve this. In addition, rules must be established on how to handle unclassified data. Whenever possible, the application of these rules should be automatic.

### ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Cloud environments should provide the ability for authorized users or services to:

- Inspect the *classification* of a specific *data entity*.
- Assign or modify the *classification* of a specific *data entity*.
- With any specific parent *data entity*, consult the *information sensitivity classification* values list for all child *data entities* (with the option to limit the depth to lower hierarchy levels).

### QUESTIONS

- Is *classification* captured consistently at the point of *data element* creation?
- Is it possible to modify *classification* accordingly when data changes during the *data lifecycle*?
- Are *classifications* before integration of new applications and datasets in the cloud environment evidenced?
- Has technology been implemented to detect and assign data types to improve the quality and consistency of *information sensitivity classification*?
- Is *information sensitivity classification* utilized within all business *processes* and systems?
- Does the *information sensitivity classification* serve as the foundation for access, usage, security at rest and in motion, storage, transport, sharing, archival and data *destruction*?

### ARTIFACTS

- Data Catalog Report – *evidence* of assigned *information sensitivity classification* at all points of creation across the application and data landscape
- *Classification* Recommendation Log – *evidence* produced by technology that analyzes business content to guide users or specify sensitivity *classification* of information
- Data Usage Log – *evidence* that downstream systems and business applications utilize the *information sensitivity classification* scheme as the basis for usage, jurisdictional placement, *encryption*, distribution and access
- Change Management Standard – *evidence* that *classification* definition and application considerations are integral to the approvals in conventional or agile software development lifecycles

### SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal *standards* exist for the application and use of *data classifications*. | No formal *standards* exist for the application and use of data classifications, but the need is recognized, and the development is being discussed. | Formal *standards* for the application and use of *data classifications* are being developed. | Formal *standards* for the application and use of *data classifications* are defined and validated by *stakeholders*. | Formal *standards* for the application and use of *data classifications* are established and adopted by the organization. | Formal *standards* for the application and use of *data classifications* are established as part of business-as-usual practice with continuous improvement. |

## 2.3 CATALOGING & CLASSIFICATION – KEY CONTROLS

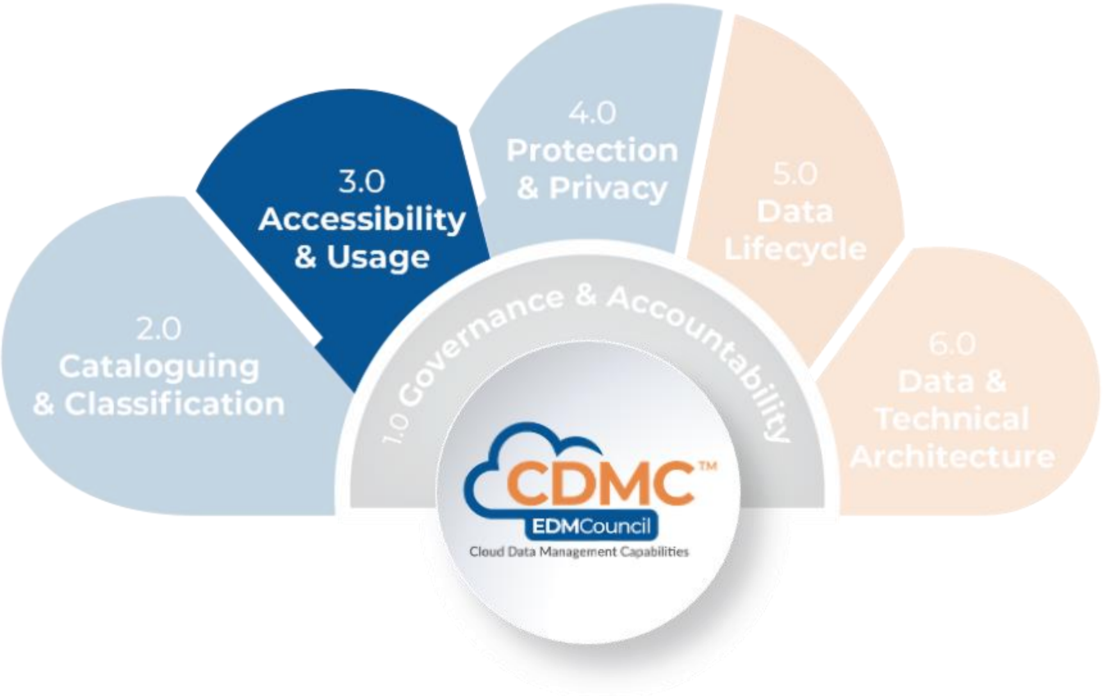
The following Key Controls align with the capabilities in the Cataloging & Classification component:

- Control 5 – Cataloging
- Control 6 – Classification

Each control with associated opportunities for automation is described in *CDMC 7.0 – Key Controls & Automations*.

*This page left intentionally blank*

# 3.0 Accessibility & Usage



# 3.0 ACCESSIBILITY & USAGE

## UPPER MATTER

### INTRODUCTION

Cloud technology offers significant opportunities for organizations to leverage their *data assets* in new ways. Many *cloud service providers* enable machine learning and advanced *analytics* by many more people than on-premises environments. Cloud computing technologies enable the combination of *data sets* in ways that were previously impracticable. When an organization seeks to maximize business value by making data and tools widely available, it is necessary to ensure that its *employees* can only access data to which they have proper *entitlements* for using that data in the organization's service.

### DESCRIPTION

The Accessibility & Usage component is a set of capabilities to manage, enforce and track *entitlements* and to ensure that data access, use and outcomes of data operations are done in an appropriate and ethical matter.

### SCOPE

- Use *metadata* to express and capture the rights and obligations over data.
- Ensure that parties respect the rights and obligations over the data they are entitled to access.
- Track and report on data access for both regulatory compliance and billing purposes.
- Establish formal organization structures for oversight of data ethics.
- Operationalize ethical access and use of data and ethical outcomes of data decisions.

### OVERVIEW

Understanding the data *entitlement* rights and obligations is critical to effective *data management*. The importance increases as more *data assets* move to the cloud. Data practitioners must provide transparency and implement controls for these rights and obligations. The adoption of cloud technology presents an opportunity for data access and use to be managed with newer, better methods. *Metadata*-driven data access control can be standardized and adopted to the extent that is typically not feasible in legacy, on-premises environments.

Data ethics considerations need to be addressed whether data resides on-premises or in one or more cloud environments. The need for focused attention on data ethics in cloud environments stems from the massive increase in volume, access, and use of data and the increased ability to process these data, leveraging various advanced technologies. Data in the cloud is subject to higher rates of use by an ever-increasing number of *stakeholders*. Organizations should develop and depend on data ethics *policies* and controls to ensure that these data are being acquired, accessed, and used in alignment with the organizational values and goals and the expectations of its *customers*.

**Data entitlements**

Typically, access control in legacy, on-premises environments has been addressed by *entitlements* frameworks that are application-oriented. Commonly, cloud computing technologies will blur, or even remove, physical and technical boundaries within and between organizations. Data—and control of data—can pass seamlessly from one party to another. A key aspect of the promise of faster provisioning in a cloud environment is the tighter connections between *data producers* and *data consumers*. Practicably, this promise can only be realized by implementing rights automation to reduce the high transaction friction associated with the licensing and permissions management necessary to provide access to data automatically.

Cloud computing lowers the technical barriers to both data distribution and Data-as-a-Service (DaaS) offerings. The ability of cloud marketplaces to support a greater variety of data, *data producers*, and *data consumers* depends on the implementation of rights automation and access tracking to:

- Offer enforcement at arbitrary levels of granularity, as determined by the data.
- Support more flexible business *models* like usage-based consumption.

Licensing and regulatory challenges are compounded by the need to understand and control how the rights and obligations of aggregated or *derived data* relate to the underlying data's rights and obligations.

These drivers make it increasingly critical that the rights and obligations of data are described by—and move with—the data itself. Though there is additional effort to implement new systems, it is preferable to a collection of disparate, unconnected *entitlements* frameworks.

Cloud computing technologies offer a significant simplification and rationalization of the application and data environment to be controlled through *entitlements* by offering:

- Standardization and commodification of delivery, consumption, and reporting systems.
- A limited number of security and identity *models*, differentiated by cloud service providers rather than by individual industry actors or systems.

*Metadata*-driven automation enables the migration from an application-oriented *entitlements* framework to one focused on the data itself. Migration to one or more *cloud service providers* presents an opportunity to address licensing, regulatory and ethical restrictions through automated rights management.

*Metadata*-driven automation:

- Encourages the representation of rights and obligations as *metadata*.
- Emphasizes the distinction between the expression of rights and their enforcement in the data store.
- Allows the requirements for the control of data to stay with the data as it travels.

Rights automation has dependencies on lineage tracking, *entitlements*, *metadata* catalogs and reporting that can be readily satisfied by standard cloud computing functionality.

**Data ethics**

Data ethics is the study and evaluation of problems pertaining to algorithms, data, and information practices. The purpose of data ethics is to formulate and support morally sound solutions, such as right conduct or right values. Data ethics answers the question: *How should we use and manage data?*

Ethical considerations affect *data management* planning in two ways. First, an organization must determine what structures exist and may be used or expanded to govern data ethics. Such structural analysis should also determine who is accountable for data ethics, who has a stake in how data in the cloud is managed, and the roles and responsibilities of the *stakeholders*. In addition, an organization must decide how to operationalize data ethics, which includes determining the *policies* and *processes* necessary for the needed governance of data ethics.

Ethical considerations for data that resides in a cloud environment should include answering the following questions:

- How are data sourced in the cloud environment? Have both the sources and the methods of sourcing been evaluated against a code of data ethics?
- Do data agreements among *data producers* and *data consumers* explain what data is accessed, how the data will be used, by whom and following data ethics *policies*?
- How might data be processed with machine learning and advanced *analytics*? Given the growth in these methodologies, are the right *policies*, *standards* and *procedures* in place to ensure that ethical outcomes are being assessed and reviewed?

Data ethics considerations are independent of platform or environment. However, cloud computing introduces new opportunities and risks, presenting ethical challenges to existing *policies*, *standards* and *procedures*. For example, data that resides in a cloud environment may be subject to more frequent and distributed access than data stored in an on-premises environment. With increased access comes a greater risk of breach and re-identification of *data subjects*. Required-for-purpose data collection takes on new relevance in a cloud environment since there is a need for greater transparency. Data minimization and performing governance to ensure that the data has been collected and used correctly are essential to *data management* in various *processes*. All of the above must be documented to ensure auditability.

Cloud computing functionality can enable better control over the ethical access, use and outcomes of data, including the cataloging and auditing *analytics* *models* and their outcomes over time. Datasets used for training models may be screened for potential biases more efficiently in the advanced infrastructure of a cloud environment. In addition, an ongoing review of *analytics* *models* and their outcomes may help identify *model* drift—gradual change of *model* behavior driven by changes in the data over time—which may create ethical challenges.

Automated, comprehensive data access, use, and outcomes that are possible in the cloud environment rely on an ability to implement detailed purpose tracking and consent reconciliation. This implementation helps mitigate risk. Cloud computing features can also detect new use cases that may indicate the potential unethical use of the data.

## VALUE PROPOSITION

Organizations that implement *metadata*-driven data access control drive business value by making data readily available for innovative use while minimizing the legal and reputational risk of unauthorized access.

Organizations that achieve a culture of ethical data use and outcomes protect and enhance their business by gaining and maintaining *customers*' trust.

## CORE QUESTIONS

- Are rights and obligations captured as *metadata* in a rules repository?
- Is rights enforcement automated using rights *metadata*?
- Are access and *entitlement* tracking automated?
- Has accountability for data ethics been assigned to a senior executive?
- Does a code of data ethics, data ethics working group and data ethics review committee exist?
- Have data ethics *processes* been operationalized?

## CORE ARTIFACTS

- Rights and Obligations Catalog Report
- Data Entitlement Governance Process Documentation
- Access Logs
- Code of Data Ethics
- Data Ethics Issue Register

## 3.1 DATA ENTITLEMENTS ARE MANAGED, ENFORCED, AND TRACKED

The organization must capture data assets' entitlement rights and obligations as *metadata* and use this information to enforce its *policies* for accessing and using the data. Enforcement of data *entitlements* must be evidenced via automated tracking and reporting.

### 3.1.1 DATA ENTITLEMENT RIGHTS AND OBLIGATIONS ARE CAPTURED AS METADATA

#### DESCRIPTION

Data usage is an *entitlement* expressly granted to authorized users. *Entitlements* are controlled by rights and obligations formalized in licenses, contracts, laws and regulations, *business policies*, codes of data ethics and *data classifications*. The rights and obligations must be captured and expressed in *metadata* as rules. To ensure consistent and reliable usage, people and systems interacting with data must adhere to these rules.

#### OBJECTIVES

- Develop and adopt a *taxonomy* or *ontology* for the expression of rights and obligations.
- Ensure that the requirement to capture and record rights and obligations is supported by *policy*.
- Capture rights and obligations as *metadata* in a rules repository with *traceability* to their source.
- Link rights and obligations to the *data assets* to which they apply in the *data catalog*.
- Ensure rights and obligations are exposed and can be consumed to support automation of their enforcement.

#### ADVICE FOR DATA PRACTITIONERS

There are many benefits of capturing data *entitlement* rights and obligations as *metadata*.

- Enable quick, automatic compliance determinations.
- Reduce the need for third-party audits and the associated risk of fines and other liabilities created by non-compliant data usage.
- Support innovation and the ability to scale.
- Control the security and reputational risks inherent in the data *processing*.
- Simplify change-management tasks when modifying rights and obligations for *data assets*.
- Ensure compliance directives propagate across platforms and applications.
- Ensure compliance directives are understood and recognized by data supply chain *stakeholders* and regulators.
- Reduce the cost of managing data by enhancing rights-management *process* automation.

Non-compliant data usage is risky and costly. Implementing rights management into an existing data governance program can significantly improve monitoring and minimize risk and cost exposures.

A clear understanding of the data *entitlement* rights and obligations is the first step to ensure compliance. It is necessary to acquire knowledge of the licenses, contracts, laws and regulations, *business policies*, codes of data ethics and *classifications* that control data usage. *Refer to CDMC 2.2 Data Classification are Defined and Used.*

Data *entitlement* rights are expressed as rules informed by *cross-organization control functions* and established by *data owners*. The rules define if and when *data consumers* can access the data. For reference, the rules should contain links to the relevant contracts, regulations or document sources.

Clarity and precision are achieved by expressing these rules as sets of:

- Permissions – what actions can be taken, such as *display this data to a trader*.
- Prohibitions – what actions cannot be taken, such as *not sharing this data with a customer*.

- Duties – the actions that must be taken to validate permission, such as a *report on usage*.

It is important to use a consistent vocabulary to specify permissions, prohibitions and duties. Insistence on a limited vocabulary reduces ambiguity and supports human and system interpretation. Free-text entries should be avoided.

*Entitlement* rules and the *data assets* they control are linked by articulating those rules as *metadata* and recorded in the *data catalog*. In this form, the rules allow the *data management* system to trace rights and obligations directly from the data itself. The permission rules function as a semantically rich directive readily acted upon by both people and systems. These directives are readily enforceable to ensure compliance. They also cultivate a common and consistent understanding of rights *metadata* across platforms and applications.

Expressing rights and obligations as *metadata* enable the automatic determination of *entitlement* rights. Refer to *CDMC 3.1.2 Data entitlement rights are enforced,* and *CDMC 3.1.3 Access and entitlement tracking is automated.*

To support compliant data *processing* along the data supply chain and beyond organizational boundaries, articulate rules with an industry-standard Rights Expression Language (for example, the Open Digital Rights Language (ODRL) market data profile).

The life-cycle of rights *metadata* should be managed like all sensitive *metadata*. Access to such *metadata* must be strictly controlled, and all changes must be logged and auditable. Refer to *CDMC 2.1 Data Catalogs are Implemented, Used and Interoperable.*

Cloud data stores enable quick discovery and access to many *data assets*. Understanding the data *entitlement* rights and obligations is critical to effective *data management*. The importance increases as more *data assets* move to the cloud. Data practitioners must provide transparency and implement controls for these rights and obligations.

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Capturing data *entitlement* rights and obligations as machine-readable *metadata* provide broad support for automation and rapid data *processing*. Cloud service and technology providers should offer tools and services that enable *data consumers* to verify compliance before, during and after *processing*. Refer to *CDMC 3.1.2 Data entitlement rights are enforced,* and *CDMC 3.1.3 Access and entitlement tracking is automated.*

Preemptive compliance verification ensures that data use cases are tested for accessing the data inventory targeted for consumption. These verifications could be automated in data marketplaces, improving discovery by ensuring *data consumers* only receive compliant data products. In particular, *entitlement* rights *metadata* is useful in expressing *data consumer* requirements and *data producer* restrictions in a *data sharing agreement*.

As *policies* are applied along the data supply chain, *data consumers* may create stricter versions of the rules to reflect *business policies* and customer relationships. These modified rules should be testable for compliance against the original *policy* requirement. Refer to *CDMC 3.1.2 Data entitlement rights are enforced* and *CDMC 3.1.3 Access and entitlement tracking is automated.*

Both *data producers* and *data consumers* expect all *cloud service providers* to respect their *entitlement* rights *metadata*. When operating from the same *metadata*, compliance determinations must be identical across platforms. *Cloud service providers* should implement common *standards* for expressing *entitlement* rights *metadata*.

## QUESTIONS

- Has a *taxonomy* or *ontology* for the expression of rights and obligations been developed and adopted?
- Is the requirement to capture and record rights and obligations supported by *policy*?
- Are rights and obligations captured as *metadata* in a rules repository?

- Can rights and obligations be traced to their source?
- Are *data assets* in the *data catalog* linked to the rights and obligations that apply to them?
- Can rights and obligations *metadata* be consumed to support automation of their enforcement?

### ARTIFACTS

- Rights and Obligations Taxonomy/Ontology definition *model* or document
- Data Management Policy, Standard and Procedure – defining and operationalizing capturing and recording rights and obligations
- Rights and Obligations Catalog Report – including details on the source of each right and obligation
- Data Catalog Report – tracing links to the rights and obligations that apply to each data asset
- Rights and Obligations API Specification – providing detail on how to access and use rights and obligations information

### SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal capture of data *entitlement* rights and obligations as *metadata* exists. | No formal capture of data *entitlement* rights and obligations as *metadata* exists, but the need is recognized, and the development is being discussed. | The formal capture of data *entitlement* rights and obligations as *metadata* is being developed. | The formal capture of data *entitlement* rights and obligations as *metadata* is defined and agreed to by *stakeholders*. | The formal capture of data *entitlement* rights and obligations as *metadata* is established and adopted by the organization. | The formal capture of data *entitlement* rights and obligations as *metadata* is established as part of business-as-usual practice with continuous improvement. |

## 3.1.2 DATA ENTITLEMENT RIGHTS ARE ENFORCED

### DESCRIPTION

To be effective, an organization must enforce its *policies* for accessing and consuming *data assets* according to *classification* *metadata*, including permissions of users, groups, and applications related to *data assets* *entitlements*. Additionally, the organization must ensure enforcement of transferring *data asset* *entitlements* throughout the *data lifecycle*.

### OBJECTIVES

- Ensure the execution of data access rights by identifying, sharing, implementing consistent enforcement of data *entitlements* as data travels across platforms, applications and environments—throughout the *data lifecycle*.
- Automate *policy*-based rights and permission assignment—both for data access and use.
- Automate rights enforcement according to permission *metadata* that derives from *data catalogs* and *classification* attributes.

## ADVICE FOR DATA PRACTITIONERS

An organization should explore and utilize automated cloud environment capabilities for facilitating and enforcing rights management *policy* as *data flows* into, through and outward from cloud environments. Practitioners should proactively manage and enforce *data asset* permissions between applications and users—throughout the data supply chain and *data lifecycle*.

An organization should progress to fully automatic data rights enforcement over time for all *data assets*. The organization should establish *metadata* that corresponds to each *data asset*. Rights enforcement that uses this *metadata* should support the incorporation of data ownership and transparency. Rights enforcement should also align with data governance programs and be available during *policy* and controls reviews.

Practitioners should ensure that *authentication systems* conform to the business and licensing *policies* of the organization. Consider proactive monitoring for *entitlement* discrepancies, such as varying levels of application of *entitlements* between applications. Also, consider establishing the ability to automate up through the highest level of *entitlement* enforcement to ensure compliance with *policies* and regulations.

Consider monitoring the consistency of each *entitlement* as it travels into cloud environments, various platforms and applications, and the entire data supply chain. Integrate user access controls with *entitlements* *metadata* to accelerate or automate granting, revoking or modifying *entitlements*.

Rights enforcement granularity should correspond to the type of *data asset* and guidance from the *data asset* owner. Commonly, an on-premises solution generally provides for rights enforcement for each application. Rights enforcement granularity should also correspond to rights management *policies*, which should specify the rights for each type of entity (such as database, schema, table, or data element).

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

A cloud service and technology provider should provide the capability to associate access controls with specific *metadata* *entitlement* attributes. The efficient methods that result can be readily automated. A provider should also deliver capability for the organization to an identity management system to support proactive and granular enforcement of licensing, regulatory, ethical concerns for data *entitlements* and data consumption.

In addition, a *cloud service provider (CSP)* should provide the ability to monitor and transfer data *entitlements* as *data flows* into, through and outward from the cloud environment.

## QUESTIONS

- Has the enforcement of data *entitlements* been applied consistently and accurately to data across the *data lifecycle*?
- Have rights and permission assignments been automated—both for data access and use?
- Has rights enforcement been automated using rights *metadata* that derives from *data catalogs* and *classification* attributes?

## ARTIFACTS

- Data Management Policy, Standard and Procedure –defining and operationalizing user and group *entitlements* aligned to the information *classification* scheme
- Access Logs – evidencing the enforcement of access *entitlements*

### SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal data *entitlements* rights enforcement automation exists. | No formal data *entitlements* rights enforcement automation exists, but the need is recognized, and the development is being discussed. | Data *entitlements* rights enforcement automation is being developed | Data *entitlements* rights enforcement automation is defined and validated by *stakeholders*. | Data *entitlements* rights enforcement automation is established and adopted by the organization. | Data *entitlements* rights enforcement automation is established as part of business-as-usual practice with continuous improvement. |

## 3.1.3 ACCESS AND ENTITLEMENT TRACKING IS AUTOMATED

### DESCRIPTION

Data *entitlements* are controlled by rights and obligations formalized in licenses, contracts, laws and regulations, *business policies*, codes of data ethics and *data classifications*. Providing proof of authorized data usage requires *evidence* that demonstrates compliance with such rights and obligations. The *evidence* is especially important when collaborating with multiple internal or external parties.

To ensure compliance and enable scalable automation, it is essential to manage, enforce and track data *entitlements* with *metadata*. Moreover, it is important to record all data access events in a data access event log. Each log entry should include all users, permissions, groups, departments and applications. The level of detail must be sufficient to satisfy reporting, monitoring and compliance requirements.

### OBJECTIVES

- Demonstrate adequate enforcement of *entitlement* rights using *policy* and workflow documentation.
- Record data access and track *data lineage* of *data elements* in the cloud environment.
- Support data sharing compliance, data marketplaces and *data asset* recommendations for *analytics*.
- Automate data access controls aligned with data *entitlement metadata*.
- Establish reporting facilities for *traceability* of data *entitlements* and data access.
- Establish outcome metrics and capture the corresponding measurements for *entitlements* enforcement.

### ADVICE FOR DATA PRACTITIONERS

Incorporating automated data access and data *entitlement* tracking can significantly reduce the ethical, business and regulatory risks of non-compliant data sharing. This tracking system must be built to minimize present and future risks by closely aligning with data privacy, ownership, sourcing, and data ethics policies. This system should provide the ability to identify all abusive patterns of data access and restrict access that is not explicitly authorized. In a cloud implementation, automatic tracking should also consider the initiation and continual evolution of data *entitlements* throughout the entire application lifecycle, user or departmental access to data. For example, in some circumstances, multiple *data elements* taken from independent sources with a lower sensitivity can—when combined—divulge sensitive personal information. In another example, a new permission scheme may enable access that violates the intent of established *policies*.

Explicit, deliberative and continuous tracking is essential for *entitlements* and access *traceability*. Practitioners that operate with only selective or incomplete tracking cannot rely on the sparse audit trails this produces. Such deficiencies will result in multiple *data risks* and security risks. Risks inevitably result in compliance failures and penalties. Organizations should implement a provider- and location-agnostic approach to data *entitlement* and access tracking. The approach should include the recording of all permissions changes. It should also ensure consistency in reporting across all jurisdictions, workflows, users, departments and roles.

A user access management application validates user and group permissions. It is essential to align the user access management application for permissions with access to sensitive data to maximize compliance. Access tracking should be an integral part of *data lineage,* and this lineage should include *traceability* and evolution tracking for user, departmental or role-based permissions. Refer to *CDMC 3.1 Data Entitlements are Managed, Enforced and Tracked,* and *6.2 Data Provenance and Lineage are Understood*.

When providing *evidence* to demonstrate compliant data access, it is essential to have built data *entitlement* and access tracking into operational and compliance reporting. In addition, automation of both tracking and reporting ensures consistency and standardization.

Data *entitlement* and access reporting give *data owners* the ability to examine, assess and proactively manage risk. Reporting also provides a means to assess the value of data, verify data compliance, and consider the necessary ethical treatment of sensitive data in every context.

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

*Entitlement* and access tracking is important to data, *policy* and security management in both on-premises and cloud environments. However, the control framework may be more complex in a cloud environment, containing various hybrid and *multi-cloud* implementations.

Cloud environments must provide functionality for each distinct data storage entity that enables an organization and partners to:

- Manage and track the log of data *entitlement* *classification*.
- Track the evolution throughout any workflow in which sensitive data is accessed across all jurisdictions.
- Require an alert or escalation whenever logging is disabled.
- Report on data access requests in a unified, auditable, historical view, thereby detecting potential violations of organizational *policies* across all cloud locations and for the entire retention period.
- Run historical reports on data *entitlements*, access, and use compliance to support usage-based billing *models*.
- Integrate reporting and tracking with data marketplace environments and other mechanisms for data sharing outside the organization.

## QUESTIONS

- Can enforcement of *entitlement* rights be demonstrated using *policy* and workflow documentation?
- Is data access recorded, and is the *data lineage* of *data elements* tracked in the cloud environment?
- Are data sharing compliance, data marketplaces and *data asset* recommendations for *analytics* supported with automation?
- Have automatic data access controls aligned with data *entitlement* *metadata* been implemented?
- Have reporting facilities for *traceability* of data *entitlements* and data access been implemented?
- Have outcome metrics been established that capture the corresponding measurements for *entitlements* enforcement?

## ARTIFACTS

- Policy, Standard and Procedure – defining and operationalizing data *entitlement* enforcement
- Data Access Event Log – include reporting on data access patterns, data access events and *data lineage*
- Functional Specifications – for automation of access controls and support of data sharing compliance, data marketplaces and data asset recommendations.

- Data Entitlement Traceability Report
- Data Access Report – including entitlements enforcement metrics

### SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal automated access and *entitlement* tracking exist. | No formal automated access and *entitlement* tracking exists, but the need is recognized, and the development is being discussed. | Formal automated access and *entitlement* tracking are being developed. | Formal automated access and *entitlement* tracking are defined and validated by *stakeholders*. | Formal automated access and *entitlement* tracking are established and adopted by the organization. | Formal automated access and *entitlement* tracking are established as part of business-as-usual practice with continuous improvement. |

## 3.2 ETHICAL ACCESS, USE, AND OUTCOMES OF DATA ARE MANAGED

Managing the ethical access and use of data and the ethical outcomes data use requires organization structures to be in place that focuses on data ethics. The organization must establish operational *processes* to report, review and address ethical issues arising from data access, use and outcomes.

### 3.2.1 DATA ETHICS ORGANIZATION STRUCTURES ARE ESTABLISHED

### DESCRIPTION

Managing data ethics for an organization has become a required discipline. Therefore, it is incumbent that organizations establish a formal data ethics oversight function, ensuring the acquisition, access, and use of data is conducted in an ethical manner and that the outcomes of data access and use are being monitored to ensure they fall within acceptable ethical *guidelines*.

Establishing formal organizational structures to support data ethics creates a framework and ensures ethical data access and use accountability.

Formal data ethics oversight includes a governing body, a Code of Data Ethics, and senior executive accountability with defined roles and *processes*. Roles and responsibilities are codified through documentation, training and verification.

### OBJECTIVES

- Assign overall accountability and responsibility for data ethics to a senior executive.
- Create and enact a Code of Data Ethics for the organization as directed by the senior executive.
- Define and implement governance structures for guiding and enforcing adherence to the Code of Data Ethics.
- Identify *stakeholders* and form working groups to implement the Code of Data Ethics.

## ADVICE FOR DATA PRACTITIONERS

To support ethical data access and use, practitioners must identify and document *data subjects*' expectations for how their data is accessed and used and evaluate outcomes of the use. As necessary, establish structures within the organization's governance functions and align those functions with the Code of Data Ethics. Governance must also support *processes* and protections for *personnel* who raise concerns about ethical data access and use.

Collaborative, routine and transparent information practices are essential to ethical *data management* because such practices energize the interdepartmental collaboration necessary to evolve organizational culture from being merely data-driven to being driven by data ethics.

### Beginning a data ethic initiative

Early efforts to cultivate data ethics awareness in an organization typically begin with the formation of a steering committee or working group. The group communicates the principles of data ethics among *stakeholders* through collaborating on activities that examine the importance of data ethics through the lens of the organizational mission and values. Eventually, organizations that agree on a commitment to data ethics develop a formal, chartered data ethics governance that aligns well with the general governance structure of the organization. This governance includes a formal body of diverse *stakeholders* (internal and external) to oversee the ethical acquisition of data, the ethical use of data and the ethical outcomes of data use.

### Realignment of the organization

The working group and the diverse formal body work collaboratively to establish data ethics compliance with the Code of Data Ethics. Such compliance should be mandated by a senior executive and implemented throughout the organization. Many organizations assign overall accountability and responsibility for data ethics to the *Chief Data Officer*. Implementing the Code of Data Ethics with accountability distributed throughout the organization requires data ethics governance structures that align with the overall governance structure of the organization. Examples include an ethics committee that reviews and approves new use cases. These structures allow the senior officer for data ethics to remain accountable for ethical *data management* for the entire organization.

### Ethical outcomes

Ethical outcomes result from data access and use that meet the organization's business needs without infringing on the human dignity of others. Human dignity can be considered what society tolerates—what is generally considered *fair*. Organizations have a moral imperative to interrogate ethical considerations for data used to develop *models* and *analytics*—and the outcomes and effects created by their use. Ignoring such impacts on society is unethical. Governing the ethical outcomes of data access and use requires long-term scenario planning and research into the societal effects of these practices. This accountability is more complex and involved than legal compliance but also links more directly with the values contained within the organization's Code of Data Ethics.

### Going beyond mere legal compliance

The role of legal and compliance in the *data management* initiative is to mitigate legal risk. However, laws and regulations typically lag well behind technological change. Organizations that focus only on legal compliance risks may face costly civil and criminal challenges from data ethics risks. Legal compliance is the minimum. Innovative and responsible organizations strive to be leaders in ethical *data management*.

### Aligning data management practices with a code of ethics

Data often contains one or more meanings. Understanding how meaning changes in different contexts and cultures is vital to ethical *data management*. A Code of Data Ethics describes the values underpinning an organization's *data management*. When structures and *processes* align with the Code of Data Ethics for an organization, decision-making becomes easier since data professionals have clear guidance on ethical use and outcomes expectations.

**Key principles**

A Code of Data Ethics typically includes these general principles:

- Do no harm.
- Interrogate outcomes to mitigate the potential for bias.
- Ensure data use is consistent with the expectations and intentions of its *data subjects*.
- Collect only the data that is necessary for a specific task.
- Provide transparency—*data subjects* have a right to know what data is collected, how it is used and how it is shared.
- Prioritize design practices that promote transparency, clarity, comprehensiveness, explainability, configurability, accountability, and proactive interrogation of training data and outcomes patterns.
- Welcome continuous internal and external ethical review.

A Code of Data Ethics cannot explain the specific expectations or requirements for every situation. Instead, a Code of Data Ethics should provide guidance so all practitioners in an organization understand the values by which they should be making decisions about data.

A key principle of data ethics governance is the empowerment of all *personnel* in the organization to have the ability and the means to raise concerns about data access, use and outcomes. Practitioners must ensure *procedures* exist for addressing such ethical issues. In addition, there should be a way to field ethical concerns from external *stakeholders* as well.

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Regulators hold an organization responsible for any cloud service and technology provider outcomes, so providers should anticipate the organization's periodic testing and reporting to ensure that outcomes align with the data ethics expectations in the third-party agreement. In addition, providers should support the ability of the organization to document both the purposes and outcomes of data use.

## QUESTIONS

- Has overall accountability and responsibility for data ethics been assigned to a senior executive?
- Has a Code of Data Ethics been created and enacted for the organization as directed by the senior *data officer*?
- Have operating governance structures for guiding and enforcing adherence to the Code of Data Ethics been defined and implemented?
- Have *stakeholders* been identified and working groups been formed to operationalize the Code of Data Ethics?

## ARTIFACTS

- Role Definitions Document – demonstrating assignment of data ethics accountability to a senior executive and outlining other roles and responsibilities
- Code of Data Ethics
- Data Ethics Governance Committee Charter – including roles and responsibilities
- Data Ethics Review Committee Charter – demonstrating diverse representation from across the organization

### SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal data ethics organization structures exist. | No formal data ethics organization structures exist, but the need is recognized, and their development is being discussed. | Formal data ethics organization structures are being developed. | Formal data ethics organization structures are defined and validated by *stakeholders*. | Formal data ethics organization structures are established and adopted by the organization. | Formal data ethics organization structures are established as part of business-as-usual practice with continuous improvement. |

## 3.2.2 DATA ETHICS PROCESSES ARE OPERATIONAL

### DESCRIPTION

Operationalizing data ethics in an organization begins with a mandate from senior management, but this mandate must have broad support through specific practices and direct accountability throughout the organization. Establishing the practices and accountability is achieved through defining and operationalizing *policy*, *standards* and *procedures* that execute against the Code of Data Ethics established by the organization.

### OBJECTIVES

- Define and approve data ethics *policies* for the organization.
- Deliver communication and training that reinforces the Code of Data Ethics.
- Establish a *process* for reviewing data acquisition, access, use and outcomes of data decisions against data ethics considerations.
- Establish a *process* for reporting concerns about ethical data acquisition, access, use and outcomes of data decisions.
- Establish a *process* for resolving ethical issues raised concerning data.
- Establish milestones, metrics and measures for quantifying the extent of adherence to the Code of Data Ethics.

### ADVICE FOR DATA PRACTITIONERS

Embedding provenance information into the metadata is one best practice that strongly supports ethical data access and appropriate data use. This approach can significantly enhance data collection and data use transparency. Another important practice is establishing tollgates in *data management* *processes* to verify adherence to the Code of Data Ethics.

It is also important that data practitioners at all levels of the organization can describe and communicate the role and responsibilities of the data ethics senior officer. Practitioners should also understand that accountability for data ethics is a *data consumer* responsibility throughout the organization. Understanding these responsibilities should be reinforced by investing in data ethics training and formalizing roles and responsibilities across the organization.

The organization should consider proactive measures to cultivate trust with its *customers* and partners. It may be necessary to provide individuals with continuous or periodic access to the data held about them beyond jurisdictions with a legal requirement. One proactive measure to reinforce a strong data ethics culture is to enhance the treatment of customer consent by collecting *metadata* for both legal and perceived consent. Other proactive measures include disclosing to *customers* the organization *policy* for data disposition and resolving to collect only necessary data for specific tasks.

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Cloud service and technology providers should anticipate demand for technical solutions that help organizations and partner organizations comply with ethical obligations. Providers should also offer methods by which organizations can define, implement, and audit the purposes of data use to support ethical outcomes.

In addition, providers should have a method for documenting the outcomes of data decisions and measuring against expectations as specified in agreements. For example, proactively verifying authorized data use will strengthen relationships among the various participants in the data supply chain.

Providers should deploy automation to identify when data may be used in a new way and initiate workflows with recommendations for *data owner* review of new use cases.

## QUESTIONS

- Have data ethics *policies* been defined and approved for the entire organization?
- Is there a program for communication and training that reinforces the Code of Data Ethics?
- Have *processes* been established for reviewing data access, data use and outcomes of data decisions against ethical considerations?
- Is there a *process* for reporting concerns about ethical data access, data use and outcomes of data decisions?
- Is there a *process* for resolving ethical issues raised concerning data?
- Have milestones, metrics and measures been established for ensuring the extent of adherence to the Code of Data Ethics?
- Have all of the above been operationalized to 'business as usual' for the organization?

## ARTIFACTS

- Data Management Policy, Standard and Procedure – defining and operationalizing management of data ethics
- Data Ethics Communication Plan
- Data Ethics Training Curriculum and Plan
- Data Ethics Review Process – covering ethical data access, use and outcomes
- Data Ethics Reporting Process
- Data Ethics Issue Remediation Process
- Data Ethics Metrics Report

## SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal data ethics *processes* exist. | No formal data ethics *processes* exist, but the need is recognized, and the development is being discussed. | Formal data ethics *processes* are being developed. | Formal data ethics *processes* are defined and validated by *stakeholders*. | Formal data ethics *processes* are established and adopted by the organization. | Formal data ethics *processes* are established as part of business-as-usual practice with continuous improvement. |

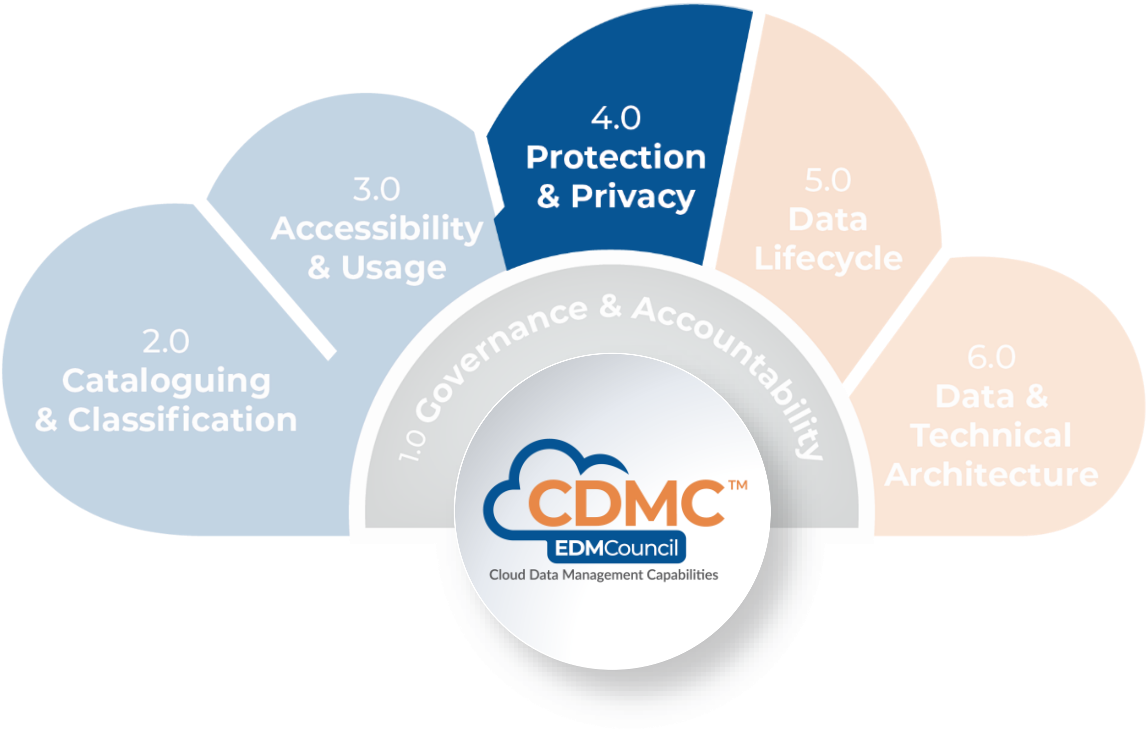
## 3.3 ACCESSIBILITY & USEAGE – KEY CONTROLS

The following Key Controls align with the capabilities in the Data Accessibility & Usage component:

- Control 7 – Entitlements and Access for Sensitive Data
- Control 8 – Data Consumption Purpose

Each control with associated opportunities for automation is described in *CDMC 7.0 – Key Controls & Automations*.

# 4.0 Protection & Privacy



# 4.0 PROTECTION & PRIVACY

## UPPER MATTER

### INTRODUCTION

Protecting the content and the privacy of data in the cloud is a critical requirement in today's cloud environments. Organizations that employ cloud computing technology may be required to comply with multiple jurisdictions' data protection and privacy legislation. The compliance burden can be quite heavy for an organization that is a member of a regulated industry. Teams planning integration with a *cloud service provider* (*CSP*) must exhibit data protection and privacy capabilities that are sufficient to meet both internal *policy* mandates and external regulatory requirements.

### DESCRIPTION

The Protection & Privacy component is a set of capabilities for collecting *evidence* that demonstrates compliance with the organizational *policy* for data sensitivity and protection. The purpose of these capabilities is to ensure that all sensitive data has adequate protection from compromise or loss as required by regulatory, industry and ethical obligations.

### SCOPE

- Implement a Data Loss Protection regime.
- Provide *evidence* that demonstrates the application of required *data security* controls.
- Define and approve a data privacy framework.
- Operationalize the data privacy framework.
- Apply *data obfuscation* techniques to all data types according to *classification* and security *policies*.

### OVERVIEW

Effective and timely management of a large IT infrastructure demands that data protection and privacy *evidence* collection must be reliable, consistent and highly automated. Many organizations sensibly view an external *CSP* environment as posing a higher risk than an internal system and thereby conclude this additional risk necessitates more stringent controls. Additional risk factors do come into play with *hybrid-cloud* solutions and the complexity of feature-variation among multiple *CSP*s.

The numerous challenges of adding complexity and risk to an existing framework of data protection controls can significantly hinder the adoption of attractive technologies. It is vital to identify and implement best practices for ensuring data protection to balance the risks and rewards of integrating with a *CSP*.

Managing sensitive data entails risk. Implementing data protection controls is the most effective approach toward mitigating the universal threats of disclosure, alteration, misuse and repudiation. Effective risk management requires balance. The Data Manager must apply and monitor adequate data protection controls while maintaining ready access to sensitive data for operational and analytical uses.

Organizations should adopt a Zero Trust framework to limit access to specific applications and resources to authorized users. The Zero Trust *model* assumes breach and verifies each request as though it originates from an open network. Zero Trust teaches us never to trust and always verify regardless of where the request originates or what resource it accesses. Every access request is fully authenticated, authorized, and encrypted before granting access. Micro segmentation and least privileged access principles are applied to minimize lateral movement.

Securing sensitive data in a cloud environment requires transferring some responsibility for comprehensive *data security* to the *CSP*. In this shared responsibility *model*, it is vital to ensure the accountability of each participant.

When preparing to transfer *data management* into a cloud environment, each of these steps must be followed:

- Apply adequate levels of data *encryption* to every *CSP* data transmission and data store.
- Demonstrate data protection controls that enforce organization *policies* and privacy *classifications*.
- Ensure controls are fully effective across the entire *data lifecycle*.
- Conform sensitive data access permissions to the principles of need-to-know and access-by-least-privilege while balancing data usability needs across the organization.
- Ensure Data Loss Prevention controls are in place, minimizing the ability to exfiltrate data.

Data protection requirements should be driven by a *data classification* scheme to ensure that the right controls operate correctly, in the right place and at the right time. Refer to *CDMC 2.0 Cataloging & Classification*.

The table gives an example of a simple *information sensitivity* *classification* scheme for a cloud environment.

| ***Data classification*** | **Cloud Environment Encryption** |
|---|---|
| Public | No Requirement |
| Internal Use only | *Data-in-motion*: Encrypt<br>*Data-at-rest*: Service-Based or above<br>SDLC Use: No Requirement |
| Confidential | *Data-in-motion*: Encrypt<br>*Data-at-rest*: Service-Based or above<br>SDLC Use: Protection needed |
| Highly Confidential | *Data-in-motion*: Encrypt<br>*Data-at-rest*: Application Level Encryption<br>SDLC Use: Protection needed |
| Price-sensitive and Secret | *Data-in-motion*: Encrypt<br>*Data-at-rest*: Application Level Encryption<br>SDLC Use: Protection needed |

A data privacy framework consists of the people, *processes*, data and technologies that support business needs, satisfy regulatory obligations, promote trust and deliver appropriate risk-balanced data privacy outcomes. Data privacy encompasses both organizations' and individuals' obligations and rights to manage personal sensitive data. *Data management* practices and controls must be trustworthy, ethical and compliant throughout the entire *data lifecycle*.

An organization that integrates with a cloud environment must consider how the integration should reshape its data privacy framework. An organization implements and operationalizes its data privacy framework through a data privacy program, which typically addresses each of the following:

- Accountability, governance and oversight mechanisms
- Documented *policies*, *procedures* and *processes*
- Documented roles and responsibilities
- Privacy operations and supporting technology
- Training

A cloud environment impacts data privacy in many ways, including:

- **Availability of functionality.** Cloud technology provides functionality, opportunities and approaches for managing data across the entire *data lifecycle*. Any serious consideration of adopting new cloud computing technologies should review and enhance the data privacy framework.
- **Jurisdictional diversity.** Cloud computing functionality and opportunities increase the potential for data to traverse multiple local or regional jurisdictions. Consequently, a data privacy framework must be

flexible and resilient to accommodate many types of legal or regulatory requirements. Refer to *CDMC 1.4 Data Sovereignty and Cross-Border Data Movement are Managed*.

- **Shared responsibility.** Operational use of commercial cloud environments is a shared responsibility *model*. Final responsibility and regulatory accountability remain with the organization that is adopting the cloud technology. Consequently, it is essential that any contract with the *CSP* clearly defines and delegates roles, accountabilities, responsibilities, metrics and measures. To consistently implement its data privacy framework across all operations, the organization must obtain complete clarification on all expectations and responsibilities of the *CSP*.
- **Proliferation of data.** Cloud environments offer low storage costs and easy data movement, so the risk of data proliferation to multiple *data consumers* is much higher. Consequently, the risk of privacy violations or breaches also increases.

## VALUE PROPOSITION

An organization that consistently implements data protection controls will adopt new cloud computing technologies more rapidly and effectively. Also, systems that operate with integral data protection controls are more cost-efficient than retrofitting custom controls.

Applying *information sensitivity classification standards* to integrations with *CSP*s can greatly improve management, monitoring, enforcement and automation of data privacy controls that meet internal, industry and regulatory requirements.

Historically, some organizations have been hesitant to effect *CSP* integrations, primarily because of security concerns. Cloud services have made significant improvements to security and privacy capabilities, integral automation and transparency. These improvements allow effective and efficient privacy risk management across the entire *data lifecycle* through the application of privacy-by-design.

Organizations preparing to integrate cloud computing can access extensive expertise to manage large-scale data repositories in cloud computing environments.

## CORE QUESTIONS

- Has a Data Loss Prevention regime been established?
- Does a documented encryption policy support an approved encryption strategy?
- Can the organization provide *evidence* of *data security* controls?
- Is the data privacy framework updated to manage the impact of cloud adoption and integration?
- Is the internal data privacy framework in operation?
- Is the data privacy framework in operation to cover all *CSP* integrations?
- Have *data obfuscation* techniques been selected, supported and applied?

## CORE ARTIFACTS

- Data Privacy Framework – that reflects requirements for the cloud
- Data Privacy Controls Log – that demonstrates the effectiveness of the controls
- Data Obfuscation and Encryption Strategy
- Data Management Policy, Standard and Procedure – defining and operationalizing data obfuscation and encryption
- Data Loss Prevention Methodology – that includes roles and responsibilities
- Data Security Controls Log – that demonstrates the effectiveness of the controls

## 4.1 DATA IS SECURED, AND CONTROLS ARE EVIDENCED

The organization's *policy* for the *encryption* of data must be extended to cloud environments. They must be enforced for *data-at-rest*, in motion and in use and *evidence* of the implementation of these controls must be captured. Securing data goes beyond *encryption*. Techniques for the obfuscation of sensitive data must be supported and adopted in all environments. A Data Loss Prevention regime must be in place and must cover both on-premises and cloud environments.

### 4.1.1 ENCRYPTION POLICIES ARE DEFINED AND ENFORCED

#### DESCRIPTION

*Data assets* are classifiable by sensitivity level. For each combination of state and sensitivity level, a data *encryption standard* must be enforced by implementing suitable *encryption procedures* available through a *cloud service provider* (*CSP*). Refer to *CDMC 1.2 Data Classifications are Defined and Used*.

#### OBJECTIVES

- Protect sensitive data with *encryption* to mitigate threats, including disclosure, modification, misuse, or attack.
- Protect sensitive data with *encryption* to a level that is acceptable to the organization.
- Protect sensitive data with *encryption* to a level specified by regulatory obligations.
- Consistently apply *encryption* to the extent that it meets or exceeds the risk level and corresponds to the organization's risk appetite and data ethics.
- Consistently apply an *encryption key* management scheme that envelops acceptable risk, the potential for functional loss and operational complexity.

#### ADVICE FOR DATA PRACTITIONERS

Data can exist in one of three states:

- *Data-at-rest*
- *Data-in-motion*
- *Data-in-use*

**Encryption of data-at-rest**

*Data-at-rest* is data that resides in physical storage and is not in transit. This includes data residing in a database, a file or on disk. An organization should encrypt all *data-at-rest* to mitigate the risks of malicious actions such as disclosure, changes to sensitive information or unauthorized access. It is also important to consider applying this type of *encryption* for archived data.

All *CSPs* offer some form of *encryption* for *data-at-rest*, which may be service-based or server-side *encryption*. A *CSP* may permit an organization to manage the *encryption key* lifecycle and thereby control how applications and services use the keys. Also, an organization may choose to generate *encryption keys* and store those keys in a *hardware security module* (HSM) provided by the *CSP*. Another common method is for the organization to import *encryption keys* into the *CSP encryption* solution while retaining backup copies in an on-premises HSM. See the *Encryption Key Management Schemes* section below for more detail on these choices.

**Encryption of data-in-motion**

*Data-in-motion* should be encrypted to ensure that it is accessible only to the intended recipients and entirely impenetrable to any potential interceptor.

Encrypting *data-in-motion* considerations apply to various parts of *data architecture*, including API calls to *CSP* service endpoints, data transfers among *CSP* service components and data movements within applications. The first two considerations are the *CSP*'s responsibility, and the last consideration is the organization's responsibility. The organization must also consider encrypting *data-in-motion* for any data movements between the organization and any third party.

A Transport Layer Security (TLS) protocol should be used for encrypting all *data-in-motion*. For example, as of this writing, NIST SP 800-52 provides specific guidance for selecting and configuring TLS protocol implementations. Consider employing Federal Information *Processing* Standards (FIPS) 140-2 endpoints, if applicable. Such endpoints use a cryptographic library that meets the FIPS 140-2 standard. For financial institutions that manage workloads on behalf of the US government, the use of FIPS 140-2 endpoints may be mandatory to satisfy government compliance requirements.

**Encryption of data-in-use**

*Data-in-use* is data in the process of modification or maintenance. Until recently, it has been necessary for data to be decrypted in memory during *processing*. Privileged users such as DBAs, system administrators and *CSP* operators may access such *plaintext* data in memory. Cyber intruders may illicitly gain access to such data.

As of this writing, preventative *encryption* controls for *data-in-use* are at an early stage of industry development for private and public Software-as-a-Service. Data practitioners should perform risk analysis and evaluate the use of any preventative controls that are part of the emerging confidential computing *model*[1]. Similar analysis should be done for compensating and *detective controls* such as just-in-time (JIT) privileged access, per-access customer authorization of administrative logins into a lockbox and Security Information and Event Management (SIEM) solutions that monitor potential breaches.

**Application-level encryption**

Organization-side *encryption* encrypts sensitive *data elements* before transmission to any storage environment such as a database or cloud storage. Applying this type of *encryption* ensures that sensitive *data elements* will be encrypted before reaching the *CSP*. Because a *CSP* doesn't have access to the organization's *encryption keys*, it cannot decrypt the data. It is important to realize that the inability to decrypt the data may limit, degrade or disable the *CSP* functions for querying the data.

It is possible to combine *application-level encryption* with three other *encryption* types to achieve multiple layers of protection. Refer to *CDMC 4.1.3 Data obfuscation techniques are defined and applied* for other alternatives for protecting application *data elements*.

**Encryption key management schemes**

*Encryption* is useless if the *encryption keys* are not secure. Most *CSP*s offer several different key management solutions to accommodate the requirements of various *data classifications*. Much of the difference among these solutions pertain to the shared management of the *encryption keys*. The table below explains several key management schemes.

[1] For an introduction to confidential computing, see e.g., "The Rise of Confidential Computing" in IEEE Spectrum, June 2020.

| Key Management Responsibility | Key Management Scheme (KMS) |
|---|---|
| ***CSP*** | ***CSP*-managed keys:** Organizations delegate responsibility to the *CSP* for generating, managing and controlling keys throughout the *data lifecycle*. This option is available with most *CSP*s. |
| **Shared option 1** | **Organization-managed *encryption keys*:** The *CSP* key management scheme is used for the entire *encryption*-key lifecycle. The *CSP* and other organization-operated services may be permitted to use keys for *encryption* and *decryption* of organization data. |
| **Shared option 2** | **Organization-supplied *encryption keys* (bring-your-own-key):** The organization operates key management *processes* and infrastructure external to the *CSP*. Organizations upload their *encryption keys* to the *CSP*'s key management scheme. The *CSP* and other organization-operated services may be permitted to use keys for *encryption* and *decryption* of organization data. |
| **Organization** | **Organization-side key management (hold-your-own-key):** An organization's internal key management infrastructure generates its keys. These keys encrypt data before transmitting it to the *CSP*. The *CSP* and other organization-operated services may be permitted to use keys for *encryption* and *decryption* of organization data. |

Irrespective of the key management scheme, a data practitioner should verify that technologies and practices for managing *encryption keys* meet the organization's current *standards*, *guidelines*, and regulatory requirements. *Encryption keys* are sensitive and business-critical. The use of *encryption keys* should be restricted to authorized applications and users. Restrictions should also apply to *processes* that validate access permissions. In particular, the data practitioner should be aware of these relevant technologies and practices:

- Options that employ *role-based access control* (RBAC) and least-privilege access principles to limit access to *encryption keys*.
- Network-level access controls restrict the management of *encryption keys* wherever possible.
- Configuring recovery options, such as soft-delete and *purge* protection, to prevent accidental or malicious key deletion.
- *Encryption key* lifecycle *policy* and *procedures* that include periodic key rotation and immediate (emergency) key rotation.
- A system for retaining data in a storage account that is (a) under organization control, (b) managed with *policy* restrictions and (c) employs configurations that are readily verifiable against the *policy* restrictions.
- Trustworthy log retention and management system for tracking and auditing key usage events such as *encryption* and *decryption* operations.

**Example of an encryption policy**

An organization may have specific risk or impact profiles that require a precise set of *encryption* controls. The controls given in the table provide an example of an *encryption* *policy*. Controls should be customized to match the need and risk appetite of the organization.

| **State** | **Encryption in Transit** | **Encryption at Rest** | **Encryption in Use** | **Application-level encryption** |
|---|---|---|---|---|
| **Sensitivity Level** | *Data in packets on the wire* | *Data in non-volatile memory* | *Data in volatile memory* | *Application data fields* |
| **Critical/Secret** (extreme loss or harm, includes highly sensitive data, payments data) | Required | Required, with organization-managed keys | Considered | Required, with organization-managed keys |
| **Highly Confidential** (material loss or risk) | Required | Required | Not required | Required |
| **Confidential** | Required | Required | Not required | Not required |
| **Internal use only** | Required | Required | Not required | Not required |
| **Public** | Not required | Not required | Not required | Not required |

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Cloud service and technology providers should integrate *application-level encryption* capabilities into *managed services* that store sensitive data. This integration will make it easier for organizations to use *application-level encryption* and derive value and gain insights from the encrypted data. For cases where *application-level encryption* is impossible or impractical (as with some machine learning workloads), the provider should provide built-in mitigating capabilities described in the *Encryption in Use* section above.

Providers should facilitate *evidence* of the existence of *encryption* controls and their operational effectiveness across a broad set of cloud data resources and across large cloud environments where many accounts, regions and data services exist.

*CSP*s should continue to innovate *encryption* and non-encryption controls for protecting all *data-in-use* against unauthorized access. *Data-in-use* includes active data in non-persistent memory such as RAM, CPU caches and CPU registers. *Data-in-use* often contains sensitive data such as digital certificates, *encryption keys*, *personally identifiable information* and intellectual property such as software algorithms and design data. Conventional *encryption* technologies do not protect *data-in-use*.

Cryptographic protection has become a growing concern to businesses, government agencies and other institutions. Threats to *data-in-use* include cold-boot attacks, the connection of malicious hardware devices, *rootkits*, *bootkits* and side channels. Compromising *data-in-use* often exposes encrypted *data-at-rest* and *data-in-motion* as well. For example, an unauthorized user with access to RAM can locate an *encryption key* for *data-at-rest* and access sensitive data.

## QUESTIONS

- Has an *encryption policy* been documented and approved?
- Does the *encryption policy* document accurately portray the risk exposure and the desired level of protection for each category of data in the cloud?

- Do the *encryption* capabilities offered by the *CSP* include options for key management and have these capabilities been documented and assessed?
- Have the organization's security and privacy *stakeholders* reviewed and approved the data *encryption* and *encryption key* management strategies?
- Are monitoring, logging and alerting measures in place to monitor the operational effectiveness of the *encryption* strategy?
- Has a regime been established for reviewing key management practices and technology in use by both the *CSP* and internal staff?

### ARTIFACTS

- Data Encryption Strategy
- Data Management Policy, Standard and Procedure – defining and operationalizing data *encryption*
  - Before installation or upgrade of an application
  - Before migrating data to a *CSP*
- Data Catalog – containing all *classification* information necessary for protecting data
- Security Treatment Plan – containing each application's level of *encryption* and other risk mitigations
- Encryption Strategy Operational Effectiveness Logs
- Key Management Review Procedure – covering review of key management practices and technology in use by both the *CSP* and internal staff

### SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal data *encryption policy* exists. | No formal data *encryption policy* exists, but the need is recognized, and the development is being discussed. | Formal data *encryption policy* is being developed. | Formal data *encryption policy* is defined and validated by *stakeholders*. | Formal data *encryption policy* is established and adopted by the organization. | Formal data *encryption policy* is established as part of business-as-usual practice with continuous improvement. |

## 4.1.2 IMPLEMENTATION OF DATA SECURITY CONTROLS IS EVIDENCED

### DESCRIPTION

*Data security policies* require establishing data protection controls for any *data element* that qualifies for one or more *information sensitivity classifications*. Design and implementation of these controls must be done early in a system or software development project. However, design and implementation are necessary but not sufficient to demonstrate compliance with *policies*. As part of an internal or external audit, it may be necessary to obtain *evidence* of recent application of the controls and the extent to which those controls have been effective.

The sub-capability requires the inclusion of observable and collectible *evidence* that demonstrates the presence of data protection controls. The *evidence* must link directly to *data catalogs* and applicable *information sensitivity classifications*.

*Evidence* should be obtainable from native, local, or third-party applications using Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), or Software-as-a-Service (SaaS) services. Wherever it is practicable, *evidence* collection should be automatic. When the *evidence* reveals exposed sensitive data or *evidence* indicates missing or deficient controls, a resolution plan must be documented to remedy such deficiencies.

### OBJECTIVES

- Define a method for obtaining *evidence* of data protection controls.
- Proactively collect *evidence* that sensitive data is secure and complies with the organization's *data classifications* and data handling *policies*.
- Implement a method of ingesting, storing, *processing* and analyzing *evidence*.
- Collect *evidence* that sensitive data is secure according to regulatory obligations.
- Demonstrate that the organization consistently applies controls for securing data according to risk appetite and data ethics.

### ADVICE FOR DATA PRACTITIONERS

**Know where controls are necessary**

An organization should use *data classifications* to identify the *data elements* that must be secure when stored in a cloud environment. The organization should establish sufficient controls for securing all sensitive data. Activity logs and continuous compliance checks should provide *evidence* of the controls. Refer to *CDMC 2.2 Data Classifications are Defined and Used*.

**Know what controls are necessary**

Some controls are implemented as a default configuration that is broad in scope, such as always-encrypted data storage volumes. However, it may be necessary to identify and enable specific controls wherever precise verification is mandatory. Other examples of specific validation include periodic verification of authorized access to sensitive *data elements* and exhaustive transaction monitoring with a logging facility.

**Observable evidence in custom applications**

All custom applications must comply with data control requirements defined by the organization's privacy and data security policies. Third-party or open-source functions must also accommodate compliance with required data controls. Various architectural patterns and open-source technology frameworks such as Fintech Open Source Foundation (FINOS) are available for adding, managing and observing controls. Also, the organization should consider the best approach for implementing *standards* for controls in *data management* systems.

**Establish data controls in systems and services**

Wherever practicable, data protection controls should be implemented in each system. A *cloud service provider* will typically provide configurations and deployment options to activate controls for the cloud environment or some cloud services. Controls that are implemented for each of the *data management* systems should be readily observable. *Evidence* collection is also easier to implement by engaging with the CSP APIs for accessing system logs and service configurations.

**Infrastructure-as-code (IaC) templates**

Organizations should consider using standardized, automatic and repeatable templates for IaC, which can be quite helpful in implementing appropriate data protection controls to secure data. IaC templates, for example, can automatically activate data *encryption* for *data-at-rest* and *data-in-motion* configurations. In addition, IaC templates can help simplify *evidence* collection by moving the focus toward general segments of the deployment pipeline.

### Collect and report evidence

Custom applications, as well as IaaS, PaaS and SaaS solutions, may have various data protection controls and various mechanisms for gathering *evidence* for those controls. To help document *evidence* that sufficient controls exist, organizations should clearly understand how data is ingested through each data source. This evidence includes logs, IaC artifacts and *CSP* configuration settings.

### Treatment of gaps in evidence collection

In a typical organization, many and varied data protection controls are operating independently in several system contexts. Since gaps may exist in an application or the ability of the *CSP* to collect evidence automatically, additional *evidence* may be necessary to show that all controls are satisfying all *policy* objectives. The organization should plan to resolve these gaps.

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Data practitioners increasingly rely on cloud service providers to manage and store critical data in many contemporary organizations. *CSP* agreements with an organization commonly define a shared burden for enforcing the protection of sensitive data. This shared burden means that a *CSP* has partial responsibility for ensuring data protection and supporting *evidence* collection that demonstrates the security of sensitive data. The *CSP* should offer interfaces, tools, logs and reports that data practitioners can readily access as they collect and exhibit *evidence* for all active data protection controls.

- *CSP*s should ensure that adequate *evidence* for active data protection controls is readily available through APIs that provide access to the *CSP* logs, service configurations and continuous compliance monitoring tools.
- The CSP should provide a simple method to support evidence collection on active data protection controls for applications and services used to store and manage sensitive data.
- The *CSP* should provide simple methods for integrating subsystems that gather and convey *evidence* from *data catalogs* and *classification* systems.
- The *CSP* should support *always-on* controls to secure data for the entire organization.
- The *CSP* should provide a near real-time inventory of available cloud resources, including data stores.
- Across all services, the *CSP* should provide compliance monitoring tools that automatically detect and report any changes in the *data security* configuration. This reporting can greatly simplify audits and reviews.
- The *CSP* should provide a reliable repository of *evidence* that will meet inspection requirements for control functions. This repository should provide simple data extraction and inspection methods and a reliable archiving solution that ensures complete data integrity.

## QUESTIONS

- Has a method for providing *evidence* of controls been defined?
- Are *policies* and design practices in place for any custom applications that have been deployed to the *CSP*?
- Does the *CSP* provide methods for monitoring and ensuring that mandatory controls are active and functioning properly?
- Do *processes* exist for identifying and adjusting any misconfiguration of active controls?
- Is there agreement on how to store *evidence* and compare it to the catalog of active controls?
- Do *processes* exist for identifying and resolving any gaps in the active controls?

## ARTIFACTS

- Data Catalog Report – evidencing execution of required *data classifications*
- Active Controls Log
- Evidence Collection and Review Plan

- Issue Management Report – evidencing capture and resolution of data security defects
- Applications Security Treatment Plan – listing the controls, required *evidence*, and necessary mitigations

### SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal ability to *evidence* the implementation of security controls exists. | No formal ability to *evidence* the implementation of security controls exists, but the need is recognized, and the development is being discussed. | The formal ability to *evidence* the implementation of security controls is being developed. | The formal ability to *evidence* the implementation of security controls has been defined and validated by *stakeholders*. | The formal ability to *evidence* the implementation of security controls is established and adopted by the organization. | The formal ability to *evidence* the implementation of security controls is established as part of business-as-usual practice with continuous improvement. |

## 4.1.3 DATA OBFUSCATION TECHNIQUES ARE DEFINED AND APPLIED

### DESCRIPTION

An organization derives information from all kinds of data to operate and drive business. For organizations that interact with *customers* and other organizations, much of the data is sensitive or proprietary—or both. It is essential to implement security and privacy measures that protect the interest of *data consumers* and custodians of the data. In contemporary computing, a variety of *data obfuscation* techniques are available for protecting data. Techniques should be chosen according to sensitivity classification, business requirements and organizational risk appetites. In addition, it is essential to define *policies* and *standards* that specify the application of obfuscation techniques to datasets with varying sensitivity classifications.

### OBJECTIVES

- Define effective and consistent *data obfuscation* techniques for mitigating *data security* concerns.
- Define the criteria for the appropriate application of various obfuscation techniques.
- Ensure highly secure controls for the reversibility of obfuscation techniques applicable to each *data element*.
- Ensure consistent application of obfuscation techniques to all linked datasets.
- Ensure that *quasi-identifiers* are obfuscated if those identifiers are combinable to reveal the identity of an individual.
- Ensure *traceability* for any obfuscated data, including the ability to track and control the dissemination of such data.

### ADVICE FOR DATA PRACTITIONERS

*Data obfuscation* is the *process* of obscuring, redacting, or transforming all or part of a *data element* to prevent the identification of parties or inappropriate disclosure of private information that may be contained in that data. Typically, this involves substituting placeholder data to represent the actual data. In this approach, data values classified as sensitive are altered so that the original values are no longer available.

In general, there are three types of *data obfuscation*:

1. ***Encryption*** transforms original *plaintext* data into ciphertext, using an *encryption* algorithm and an *encryption key* as input. *Decryption* converts ciphertext back to the original *plaintext* data and requires a separate *decryption* algorithm and a *decryption* key. Refer *to CDMC 2.2 Data Classifications are Defined and Used*.

2. **Data masking** replaces an original value with a character string that results from a data-masking function. Masking may involve substitution, shuffling, or more complex manipulation that obfuscates data while preserving some of the statistical properties of the original *data set* (such as stochastic perturbance). Data-masking functions apply to *data-at-rest* (static data masking) or data-in-transit (dynamic data masking). The original data cannot be exposed by applying any formula to the masked value. Suppose data is masked before ingestion into a data repository. In that case, there is a low risk of exposing sensitive data if that repository is breached (since the contents of the masked elements are fabricated).

3. ***Tokenization*** is the process of substituting a sensitive *data element* with a non-sensitive equivalent known as a token. Such tokens have no extrinsic or exploitable value. The token is a unique reference (identifier) mappable to sensitive data using a highly secure *tokenization* scheme. *Tokenization* typically occurs when creating or importing sensitive data into a system. Tokens require significantly less computational resources to process than either *encryption* or data masking. However, *tokenization* requires a mapping table and high-security measures. *Tokenization* hides sensitive data while substituting comparable data for *processing* and *analytics*. Typically, tokenized data can be processed more quickly, a key advantage in high-performance systems.

Selecting from among various *data obfuscation* techniques must consider the business requirements and desired outcomes. Criteria that would be used to determine applicable obfuscation techniques include:

- Data utility – does the use case require a technique that renders an obfuscated value that retains some measure of *accuracy* or referential integrity?
- Sensitivity *classification* of data.
- Location of the data store.
- Define each of the data perimeters at which specific *data elements* must be obfuscated.
- Define which type of obfuscation (or combination of types) is best applicable to specific data stores or *data elements*.

The table lists a variety of *data obfuscation* techniques and recommendations.

| Category | Technique | Definition | Properties | Best Practices and Appropriate Use |
|---|---|---|---|---|
| *Encryption* | Field-level *encryption* | Replaces the field value with an encrypted value, which derives from the source using a cipher and a key. | Format preservation: Possible (limited) through format-preserving *encryption*.<br>Reversible: Yes | The focus here is on field-level *encryption*. Field-level *encryption* is a less useful alternative to *tokenization* (e.g., organization-side *encryption* may not fit the *processing model* as the system may be too complex to support). |
| Masking | Partial redaction | Partially masking out the field value preserves the general form of the data value to assist in recognizing the data type (such as hiding all but the last four digits of a credit card number). | Format preservation: Possible (limited), some formatting types are preserved, such as value type—but not length.<br>Reversible: No | This technique is suitable when a value section contains information not dependent on the rest of the value, such as zip codes and credit card numbers. |
| | Full redaction | The entire data value is replaced with a single repeated value, such as "XXXXX." | Format preservation: Possible (limited). The length of the value can be shown with a sequence of masking characters.<br>Reversible: No | Using redaction as the *default* supports the principle of data minimization and encourages data users to justify why each field should be retained. |
| | Generalization | Also known as coarsening, this technique is useful in decreasing the data precision or granularity. | Format preservation: Possible (limited).<br>Reversible: No | Generalization can be used to prevent linkage attacks.<br>Examples of generalization include rounding decimal-based coordinates, numeric quantities such as age and zeroing out the last octet of an IP address. |
| | Stochastic Perturbation | Replaces an input with a value that has been perturbed by adding or subtracting a small amount of random zero-mean noise. | Format preservation: Yes<br>Reversible: No | Perturbation aims to protect against identification when an attacker might know a specific value in the dataset. For example, sensitive values in a transaction could be perturbed by any full-unit value in a range, such that an input value of $173 could be perturbed by +/- $7 to generate an output value in the range $166-$180. |
| | Substitution & Shuffling | Replaces an input value or group of values with a value that is taken from a predefined mapping. If the replacement value is from the same *domain* of the masked data, it is called shuffling. | Format preservation: Possible (limited).<br>Reversible: Possible, but becomes increasingly impractical for cases in which the dataset has many unique values. | Substitution allows for general control of the data, but the tradeoff is the substantial effort in configuring the substitution values. An example is the configuration of a mapping in which another unique name replaces every name. This scheme is format-preserving, reversible, and provides referential integrity. However, an increasing number of names also increases the effort to create and maintain the substitution list. |

| Category | Technique | Definition | Properties | Best Practices and Appropriate Use |
|---|---|---|---|---|
| *Tokenization* | | Replaces an input value with a random token that has no extrinsic or exploitable meaning or value. | Format Preservation: Yes<br>Reversible: Yes<br>Linkable: Yes<br>Traceable: Yes | *Tokenization* should be applied as the default for all sensitive values that are not redacted. Consistently tokenized columns retain information about which *records* share the same value. Therefore, it is possible to calculate frequency distributions, perform analysis, and train machine learning *models* on consistently tokenized data with no loss of utility. |

Obfuscation controls should be applied following regulatory obligations, company *standards* and risk appetite. Multiple complementary controls may be applicable to ensure sufficient security in multiple domains. Options include encrypting entire data objects, masking specific sensitive *data elements*, and tokenizing other *data elements*—according to organizational policies that apply to each of the various *data elements*.

Obfuscate the data at the earliest opportunity – preferably when the data is created or imported into the cloud environment. Decide whether it is necessary to preserve referential integrity for some or all of the *output domains*. Establish strict controls for any application or user requests to reverse the obfuscation. Any access request should be logged for audit purposes.

Sensitive data should not move to a lower-grade environment such as QA or Development. If there is an approved business requirement to move the sensitive data, it must be obfuscated at the migration point.

**Tokenization for cloud data storage and exchange**

Consider these best practices for *tokenization* of data that is stored in a cloud environment:

- *Tokenization* is perhaps the best technique for direct identifiers and should be applied near the beginning of the *data lifecycle*.
- Any *tokenization* capability should work across all on-premises and cloud environments in an organization.
- Before implementation of a *tokenization* system, verify compatibility with any application that will depend on that system.
- Access to the *tokenization* mapping table must be secure according to organizational *standards*, and the *output domain* must be sufficiently large to ensure resilience to brute force attacks.
- A systems detokenization should occur only when no viable alternative is available and authorization is explicitly given.
- Use different tokens for different systems to ensure *traceability* and mitigating the risk of sensitive data exposure (that would otherwise occur by linking datasets).

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

To effectively support an organization as it implements *data obfuscation* solutions, *cloud service providers* and technology providers should:

- Offer the organization sufficient transparency for identifying each of the sensitive *data elements* throughout the cloud environment.
- Provide native capabilities and integrations with *data obfuscation* tools that operate across major cloud platforms and on-premises environments.

- Provide functionality that integrates common data cataloging, *information sensitivity classification* and *data obfuscation* solutions.
- Provide the ability to automatically audit data environments to verify compliance with *data obfuscation* requirements that satisfy organizational *standards* and regulatory requirements.

## QUESTIONS

- Have *data classification standards* been established?
- Have the criteria been documented for data *encryption*, data masking and *tokenization*?
- Have obfuscation techniques been selected and applied in alignment with data-usage requirements?
- Does the *tokenization* system prevent reverse translation without access to the mapping content?
- Are applications compatible with the *tokenization* systems (applications that require access to detokenized data)?
- Does the ability exist for controlling reversibility, referential integrity and *traceability* when obfuscating data?
- Have *quasi-identifiers* been obfuscated to protect against linkage attacks?
- Does functionality exist to identify and notify if sensitive data is not obfuscated?

## ARTIFACTS

- Data Management Policy, Standard and Procedure – defining and operationalizing the obfuscation of sensitive data
- Data Management Technology Tool Stack – inclusive of technologies that support the chosen obfuscation techniques
- Obfuscated Data Access Request Log – a record of events and attempts to access obfuscated data

## SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal *data obfuscation* techniques exist. | No formal *data obfuscation* techniques exist, but the need is recognized, and the development is being discussed. | Formal *data obfuscation* techniques are being developed. | Formal *data obfuscation* techniques are defined and validated by *stakeholders*. | Formal *data obfuscation* techniques are established and adopted by the organization. | Formal *data obfuscation* techniques are established as part of business-as-usual practice with continuous improvement. |

## 4.1.4 A DATA LOSS PREVENTION PROGRAM IS ESTABLISHED

## DESCRIPTION

Data Loss Prevention (DLP)—also known as data leak protection—is a strategy to detect and prevent the deliberate or accidental transfer of sensitive data beyond the network and controls of an organization. An effective DLP program includes directive, preventive, and *detective controls* to manage data loss for *data-at-rest*, *data-in-motion*, and *data-in-use*. While DLP software tools are an important element of any DLP program, any DLP system must also address the people and *process* aspects of the organization.

## OBJECTIVES

- Formally establish the DLP strategy and approach within the organization.
- Define and communicate the roles and responsibilities for the DLP program.
- Gain approval and adopt DLP *policy*, *standards* and *procedures* that apply consistently across on-premises and cloud environments.
- Select and implement DLP software tools that align with and support the DLP strategy.
- Develop and deliver DLP awareness initiatives and training.
- Measure and continuously improve the effectiveness of DLP measures.

### ADVICE FOR DATA PRACTITIONERS

A comprehensive DLP strategy must encompass hybrid architectures. Such architectures may include applications that span both cloud and on-premises resources and desktop environments that may be used to access cloud and on-premises resources. To be enduring, develop a DLP program that will scale and encompass an ever-increasing amount and variety of technologies and cloud services. The program must address threats and challenges encountered in cloud environments, such as residency, storage, movement, and data protection.

A DLP strategy must define business requirements that may include the following:

- Prevention of deliberate or accidental disclosure of sensitive data.
- The extent of risk reduction and quantifying the cost of compliance.
- Compliance with contractual obligations relating to any *third-party data*.
- Reduction of reputation and brand risk.
- Protection of intellectual property.

A formal DLP *policy* and supporting *procedures* are fundamental to the establishment of a DLP program. Practitioners should outline acceptable behavior in *policy* documentation and enforce this through defined *procedures*. The program should include supporting incident management and triage capabilities to comply with the principles of zero trust.

Both the DLP program and any DLP software tools should exploit the cataloging and *classification* capabilities (refer to *CDMC 2.0 Cataloging and Classification*) to ensure that the scope of DLP controls are broadly acceptable and *evidence* can be exhibited for each control. The implementation should also exhibit *data security* fundamentals, including *encryption*, obfuscation and access control. Specific DLP control capabilities and *processes* may include:

- Enforcing solutions such as *encryption*, *tokenization* and obfuscation of *data-in-motion*, *data-at-rest*, and *data-in-use*—according to *data classification* and handling requirements.
- Blocking egress traffic from the cloud environment that is excluded from the list of expected *domain* names.
- Implementing the ability to block or monitor the transfer between cloud resources and endpoint devices.
- Monitoring network flow logs for anomalous traffic and connection requests could indicate unauthorized exfiltration of data.
- Analyzing a cloud API system and application logs to identify unexpected and potentially malicious activity.
- Validating continuous monitoring for malicious or unauthorized user behavior is in operation.
- Monitoring of resource configurations to validate compliance against defined *policies* and *standards*.

Practitioners should take advantage of any native DLP capabilities offered by cloud service and technology providers. , Such capabilities include blocking public access to data stores, *encryption*, private connectivity, threat intelligence and detection of anomalies.

The DLP program must provide *evidence* of control coverage and compliance with internal, regulatory, and legal requirements. The program must also ensure that the organization's staff are aware of individual responsibilities and resources available for mitigating DLP.

The effectiveness of the DLP program should be reviewed regularly. Measurements of effectiveness should cover *policy* management, organization coverage, software tool support and automation, communications delivery and training effectiveness.

### ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Cataloging and *classification* capabilities are fundamental in enabling an organization to identify the location of sensitive *data elements* so that DLP controls can be applied. A provider should provide the organization with the ability to distinguish the various instances of cloud applications. This ability enables the organization to implement different controls for different environments—such as production and development.

A provider should offer capabilities that enable the interoperability of DLP software solutions across cloud and on-premises environments. With this capability, an organization can perform DLP tasks for multiple environments from a single console. Interoperability should be supported with the adoption of *policy* language *standards* and DLP rule specification.

Providers should also offer capabilities and standardized outputs to correlate disparate events and help identify and investigate possible DLP events.

DLP risks are lower if organizations can connect to cloud services through private networks—without the need for internet access of public IP addresses. In addition, providers should offer capabilities that permit the organization to identify cloud environments that are not in active use and shut down those environments. Deactivating unused environments also reduces DLP risk.

### QUESTIONS

- Has the DLP strategy and approach for the organization been defined and approved?
- Have the roles and responsibilities for DLP been defined and communicated?
- Have the DLP *policy* and *processes* been defined and implemented in alignment with the DLP strategy?
- Have DLP software tools been selected and implemented in alignment and support of the DLP strategy?
- Have DLP awareness initiatives and training been developed and delivered?
- Is the effectiveness of the DLP program regularly reviewed and enhanced?

### ARTIFACTS

- DLP Strategy – detailing the approach for the organization
- Role Definitions – providing clarity on responsibilities for key roles in the DLP program
- Data Management Policy, Standard and Procedure – defining and operationalizing DLP
- Technology Roadmap – for applications aligned to the DLP strategy
- Communications Plan – specifying the approach to raising awareness of DLP measures and responsibilities
- Training Plan – identifying and implementing required skills for key roles in the DLP program

### SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal DLP program exists. | No formal DLP program exists, but the need is recognized, and the development is being discussed. | A formal DLP program is being developed. | A formal DLP program is defined and validated by *stakeholders*. | A formal DLP program is established and adopted by the organization. | A formal DLP program is established as part of business-as-usual practice with continuous improvement. |

## 4.2 A DATA PRIVACY FRAMEWORK IS DEFINED AND OPERATIONAL

The organization's data privacy framework must be updated to address cloud-specific requirements and considerations. Once defined, *processes* and controls must be established to operationalize the framework.

### 4.2.1 A DATA PRIVACY FRAMEWORK IS DEFINED

### DESCRIPTION

Data privacy encompasses organizations' obligations and requirements and the rights of individuals to manage *personal data* in a trustworthy, ethical, and compliant manner. It is vital to ensure data privacy is enforced throughout the entire *data lifecycle*: when it originates, when processed, and when stored—in both on-premises and cloud environments.

A data privacy framework consists of *policies*, *standards* and *procedures* that collectively ensure the organization meets its business needs, satisfies regulatory obligations, promotes trust, and delivers appropriate, risk-balanced data privacy outcomes. It addresses the people, process, data and technology aspects of these requirements.

### OBJECTIVES

- Ensure the data privacy framework captures cloud-specific requirements and considerations for collecting and *processing* *personal data* throughout the *data lifecycle*.
- Review and refine the roles and responsibilities for data privacy, with considerations for cloud environments.
- Define controls and *processes* that govern data usage within the cloud environment and relate directly to *policy* in the data privacy framework.
- Define *processes* for regular assessment of the design and operating effectiveness of the data privacy framework and supporting controls.
- Align privacy *processes* and supporting technology with the collection and *processing* activities in the cloud environment.

### ADVICE FOR DATA PRACTITIONERS

Data practitioners should develop, extend, and maintain a data privacy framework that accommodates all organization's privacy requirements, themes, and programs. The owner of the data privacy framework should be identified, and its scope should be defined. The framework should include a plan for managing data privacy risk, address data privacy communication and training for the organization, and specify the terms and frequency of *privacy impact assessments*.

### Privacy requirements

Privacy requirements will vary from one organization to another. The requirements vary according to the types of *personal data* collected and processed, the purposes for which *personal data* is collected and processed, and the industry and jurisdictional footprint of the organization.

Privacy requirements can come from various sources, including:

- Privacy and data protection laws, regulations, court rulings, and supervisory authority guidance.
- Society and industry expectations, best practices and norms.
- Internal factors, such as company values and risk tolerance.
- Third-party *stakeholders*, such as *customers*, shareholders, vendors, and business partners.

### Privacy themes

Across all aspects of a data privacy framework, one or more themes can drive privacy requirements. These themes include but are not limited to transparency, choice, individual rights, sharing, data breach notice, retention, and deletion. While global privacy laws and regulations vary, the privacy themes within them are generally consistent. By distilling complex (typically global) privacy requirements into privacy themes, an organization can more easily interpret privacy requirements for effective and efficient privacy programs. A data privacy program must provide transparency for governance, *policies*, notices, roles, operations, monitoring, testing, and reporting to ensure that these are operationalized in an effective and compliant manner.

### Privacy program

An organization implements a data privacy framework through a data privacy program, which primarily includes accountability, governance, and oversight roles and mechanisms. A Chief Privacy Officer—together with a privacy committee—ensures that monitoring, testing, measurement, reporting, escalations and periodic audits occur at the proper intervals.

A data privacy program also generates and maintains documented *policies*, *standards* and *procedures*. All program documentation should articulate privacy requirements in clear language that *stakeholders* in the organization readily understand. This documentation defines all the privacy program operations, including data discovery and mapping, notice drafting, and deployment. Program documents should also specify the collection and implementation of privacy choices and *processing* of *data subject* requests. In addition, program documentation should also include clear explanations of roles and responsibilities, especially who is to be accountable and responsible (RACI) for each privacy program task. Key roles include the Chief Privacy Officer, privacy team, risk, compliance, business, operations, legal and audit.

Privacy tasks are implemented in compliance with *policies* and *standards* through operations and supporting technology—including the operational functions such as data discovery and mapping, *data classification*, notice deployment, consent/preference management, privacy settings in applications and websites, *data subject* requests, data deletion, and data breach notification. Where possible, look for opportunities to simplify and centralize privacy operations, minimizing overlap, redundancy and lengthy notices, multiple privacy-choice delivery channels and *processing* databases, conflicting or unconnected *data subject*-rights *processes*, and data breach notifications. Simplifying and streamlining privacy operations and supporting technology can increase efficiency, reduce costs, ease compliance burden, and mitigate risk.

**Owner of the data privacy framework**

Many roles across an organization will need to consider privacy regularly, and privacy concerns may be substantial for some roles. It is vital to identify an accountable senior executive that leads the definition, development and management of the data privacy framework. This role is often titled the Chief Privacy Officer and should be someone with a high level of data privacy expertise—ideally with certifications. Privacy accountability throughout an organization should be measured with direct oversight from the data privacy program.

An important task in a data privacy program is that the senior privacy executive periodically publish a data privacy report structured around key privacy metrics and corresponding measurements. The executive should present the report to other senior management, thereby communicating the effectiveness and the health of the data privacy program.

**Scope of a data privacy framework**

A data privacy framework must envelop the entire organization, especially all the *personal data* that the organization collects and *processes*—throughout the *data lifecycle*. An organization should consider implementing technology that supports the effective implementation of the data privacy framework through the data privacy program.

It is necessary to establish data discovery and mapping in all areas where *personal data* is collected and processed to identify and capture all *personal data* touchpoints—throughout the *data lifecycle*—across the organization.

**Data privacy risk management**

Data practitioners should adopt and apply a privacy-by-design and a risk-based approach to implementing a data privacy framework. An organization should develop and periodically review its privacy risk appetite. As appropriate, it is important to update the data privacy framework to reflect risk appetite—including privacy requirements, privacy themes, and all other privacy program elements.

The data privacy framework and the privacy program should be assessed against industry *standards* aligned with best-practice guidance. Widely implemented *standards* include the AICPA Privacy Management Framework, SOC 2 Privacy Controls, ISO 27701, Data Protection Management Program (DPMP), HITRUST, and NIST privacy controls.

**Communication and training**

Ensure clarity and consistency for privacy notices in all business units, especially notices about the same individual's personal data. Privacy practices should be clear and consistent to anyone who uses an application, website, or social media space to engage with the organization.

Also, ensure that all relevant departments are proportionally represented in each privacy process—especially for marketing, *data subject* rights, deletions and data breach notices. To properly implement the data privacy framework, it is essential to cultivate strong connections from the Privacy team to the Information Security, *Data management* and *Records* Management teams. In particular, these connections are important to maintain consistency in data privacy definitions and *classifications*.

To cultivate an organizational culture that highly values data privacy, implement suitable training across the organization—especially for management and key roles such as marketing or privacy operations.

**Privacy impact assessment**

Practitioners should work with key *stakeholders* to define and implement an effective *Privacy Impact Assessment process* that includes efficient assessment criteria, tasks, technology and the events that trigger when an assessment is necessary. For example, it may be necessary to conduct a *Privacy Impact Assessment* when a large amount of *sensitive personal data* is processed or *personal data* migrates across jurisdictions.

Also, ensure that _Privacy Impact Assessments_ are designed and implemented with a strong emphasis on data ethics. Refer to *CDMC 3.2 - Ethical Access, Use, & Outcomes of Data Are Managed*. Finally, include cross-border data movement triggers and clearance questions in the _Privacy Impact Assessment_, as appropriate. Refer to *CDMC 1.4 - Data Sovereignty and Cross-Border Data Movement are Managed*.

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Cloud service and technology providers should support flexible _metadata_ tagging of _personal data_ that accommodates multiple jurisdictions with various definitions of _personal data_. In addition, a cloud environment _data catalog_ should readily capture various legal _constructs_, _data classifications_, usage categories and retention _policies_ of its organizations. Practitioners should verify that only authorized users can access such _metadata_.

Technologies should be in place to support integration systems that manage _records_ _processing_. There should be a clear linkage among all _processing_ activities, applications, legal entities and data. Providers must offer support for the consumption and retention of usage, notice, and privacy choice information.

In addition, providers should offer the ability for an organization to comply with _data subject_-rights requests, including access and deletion rights.

## QUESTIONS

- Have roles and responsibilities for Data Privacy been reviewed and refined with considerations for the cloud?
- Does the data privacy framework—including the Privacy Program, _policies_ and _procedures_—capture cloud-specific requirements and considerations for collecting and _processing_ _personal data_ throughout the _data lifecycle_?
- Have _processes_ been defined that regularly assess the design, operating effectiveness and supporting controls of the data privacy framework?
- Have operational privacy _processes_ and supporting technology been aligned to new collection and _processing_ activities within the cloud environment?

## ARTIFACTS

- Data Privacy Framework – reflecting requirements for the cloud environment with a detailed summary of roles and responsibilities and requirements mapped to controls
- Data Management Procedure – for regular reviews and updates of the data privacy framework to ensure the _procedures_ and capabilities address changing requirements and identified shortcomings

## SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal data privacy framework has been defined. | No formal data privacy framework has been defined but the need is recognized, and its development is being discussed. | Definition of the formal data privacy framework is being developed. | The formal data privacy framework is defined and validated by _stakeholders_. | Definition of the formal data privacy framework is established and adopted by the organization. | Definition of the formal data privacy framework is established as part of business-as-usual practice with continuous improvement. |

## 4.2.2 THE DATA PRIVACY FRAMEWORK IS OPERATIONAL

### DESCRIPTION

The organization must show *evidence* that the defined data privacy framework has been defined and is operational as business-as-usual.

Management of *personal data* using protection and privacy controls includes frequent consideration of the *what, where, and why* for this data, the cataloging and classification, and its various uses. Putting a framework into operation should be done according to a set of clear, documented metrics that quantify how the controls ensure compliance with the data privacy framework objectives. All metrics and subsequent operational measures should be readily available to *stakeholders*.

### OBJECTIVES

Each of these objectives must be met to operationalize a best-practice data privacy framework:

- Implement clear data collection and intent-of-use notifications throughout the *data lifecycle*.
- Create, implement and continuously improve the *processes* for *personal data* discovery, *classification* and inventory maintenance.
- Identify and operationalize *Privacy Enhancing Technologies* (PETs) and *data security* controls in all data environments to ensure adequate protection of *personal data*.
- Enhance each PET to improve the automation of preventative data *process* controls that help to mitigate *data subject* risks and privacy risks.
- Implement *processes* that support *data traceability*, *data lineage*, and auditing that produce *evidence* for the usage and provenance of *personal data* according to each data subject's privacy framework and preferences.
- Establish clear *processes*, *procedures* and mechanisms (automated wherever possible) for receiving and responding to requests from *data subjects* and inquiries from regulators regarding the use of *personal data*.
- Define and cultivate a privacy-by-design culture across the organization, in each *data management* platform and *domain* and any reengineering and modernization effort.

### ADVICE FOR DATA PRACTITIONERS

A data privacy framework consists of *policies* driven by the objectives and active controls that underpin the best practices of managing data that must remain private. These *policies* should reflect the risk appetite of the organization that implements the framework.

The complete set of controls that an organization implements in support of its privacy framework should support the objectives given in this sub-capability. Naturally, a specific control may support more than one framework theme.

A control employs one or more technical *constructs* that people, *policies*, *standards* and *procedures* have specified to ensure that data privacy operations comply with the framework. Typically, the complete set of controls covers various functions for collecting and recording, using, maintaining, reporting and sharing *personal data*. Each control should adhere to the known preferences and rights of the *data subject*.

Data practitioners can implement the *policies* in the data privacy framework effectively by concentrating on the following areas.

**Privacy notices and consent management**

Privacy notices must be readily accessible and written to be understood easily by all *data subjects*. Practitioners should be explicit about the various uses of *personal data* and regularly review each use case.

Align privacy consent management *models* to applicable regulatory requirements. Then, ensure that *processes* for capturing consent, distributing intent-of-use notices and demonstrating transparency accommodate any jurisdictional limitations and communicate the nature and extent of compliant behaviors.

Regarding capturing consent for data-use and privacy notices, document all aspects of coordination and the complete set of responsibilities for controllers and processors. Wherever applicable, provide each *data subject* with methods for withdrawing and adjusting consent for *personal data*.

### Classification and cataloging

Effective implementation of the data privacy framework is dependent on the capabilities detailed in *CDMC 2.0 Cataloging and Classification*.

For every use case in each environment, establish risk assessment methods that demonstrate a balance in managing data protection risks and the value of *personal data*. *Data classification* and categorization approaches must take into account the methods of data access on various cloud platforms. Take care to explicitly address multiple *classification* levels while simultaneously anticipating the extent of data proliferation in cloud environments. Any *data classification* capability must support jurisdictional and regulatory hierarchies, layers and intersections of control structures and support complex *personal data* definitions.

Document clear definitions, responsibilities and the coordination necessary for discovering, classifying and categorizing *personal data*, taking into account various capabilities for data origination, transformation, storage and disposition. Adopt technology enhancements for comprehensive management of *personal data*. An example is data analysis systems that identify and classify *data assets*. Discovery tools should continuously accommodate new data types, data structures and data storage environments. These tools should integrate seamlessly together to minimize errors and manual intervention.

### Shared responsibility

Controllers and processors should establish clear *guidelines* for explicitly assignable actions for direct collection and management of *personal data* of all *classifications*. For each data store, process and user type, documentation should exist that outlines the responsibilities for managing and collecting *personal data*. For *personal data* stores and *processes*, define and document the best configuration for *data security* and de-identification. Most *cloud service providers* offer expertise in *data security*, breach detection, mitigation and response.

Controllers and processors must agree on responsibility boundaries for capturing, recording, and reporting each type of data use. Also, each controller and processor should publish the controls each can provide, such as those driven by *policy* or governance, data topological-access controls and manual enforcement.

### Privacy enhancing technologies

*Privacy enhancing technologies* (*PET*s) aim to reduce privacy risks associated with data *processing*. They are sometimes called privacy enhancing techniques or privacy preserving technologies (PPTs). Generally, these technologies protect data by manipulating, replacing, concealing, or perturbing the original data, making it extremely difficult to reidentify. Common techniques include categorization, *tokenization* and *encryption*, data masking and *anonymization*. Refer to *CDMC 4.1 Data is Secured and Controls are Evidenced* for additional information and advice for protecting information using these techniques together with conventional security controls.

A *PET* should accommodate the *policies* in the data privacy framework and provide capabilities for logically and physically organizing both protected and de-identified data and for addressing jurisdictional requirements. *PET*s should support flexible, secure de-identification and *anonymization* capabilities. With conventional *data security* and access controls, a *PET* should protect *personal data* for a wide variety of use cases, each of which may manage data with different risk tolerance levels. Compile requirements and define *processes* for maintaining data privacy through the *PET*s. When practical, automate the transfer of *personal data* in and out of a cloud *data management* platform.

Integrate each *PET* with data privacy management software and tools already in use by the organization. *Data management* *stakeholders* and *processes* should have selective access to the various features of each *PET*, with a clear view of the data protection requirements that are necessary for each specific use case. A new data type, storage type or process may require additional flexibility and extensibility from one or more *PET*s to integrate with native *data management* systems and comply with governance *policies*.

Risk assessments involving *personal data* re-identification should consider exploiting high-volume *processing* efficiencies available in most cloud computing environments. Any personal *data risk* assessment should involve every applicable jurisdiction and organize the assessment according to each jurisdiction. Risk assessments should include outlier analysis, hidden/surrogate identifiers, linkage attacks involving publicly available *personal data* and transactional uniqueness. Such capabilities are typically not possible with conventional on-premises and built-for-purpose systems.

### Data processing

Practitioners should seek to craft each *personal data* capture process precisely and regularly assess each for proper compliance. New and evolving data ingestion, access and manipulation tools will likely require integration with data discovery and cataloging *processes*. These tools should support the careful identification and management of new origination sources of *personal data* collection.

Be alert to the introduction of novel data collection and *processing* techniques, which are part of the value proposition of cloud platforms. These techniques often entail greater complexity to support sufficient *processing* requirements. Novel approaches to *personal data* capture should be well-integrated into *data classifications* and categorizations that comply with *data subject* consent agreements.

Data practitioners should employ data privacy disclosure controls to manage the exchange of *personal data* between cloud environments, jurisdictions and *data domains*. Examples of disclosure controls include minimization, consent and data protection. Also, take care to establish a detailed Record of Processing Activities and ensure each use case has a firm legal basis.

### Data movement and data lineage

Data movements may include minimal decision-making constraints and, consequently, *data lifecycle* management *policies* must account for additional copies of data. Perform analysis to balance lower-friction data movement and storage costs with data collection obligations—especially if some of the data may be virtualized or subject to a *legal hold*.

Tracking and lineage for data movement should support core jurisdictional guidance by providing *evidence* of compliance. Pay special attention to data-sharing use cases, even assuming complete *anonymization* *processes* for data de-identification are in place. Optimally, it should be possible to block or notify on any non-compliant data use.

### Data subject requests

To coordinate *data subject* interface touchpoints and data auditing, insist on simple and timely responses to the *data subject* and regulator requests. Be prepared to accommodate *data subject* requests that originate from various jurisdictions or across multiple jurisdictions.

Coordinate new and evolving data *processing* from various cloud platforms to produce a timely response to *data subjects* and regulatory agencies. Different types of data will require different types of protection from re-identification. It is important to assess the risks of novel and evolving protections carefully. To respond quickly to *personal data*-use requests, consider additional automation as *personal data* proliferates across multiple cloud platforms.

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

To support an organization effectively as it implements its data privacy framework, cloud service and technology providers should consider the following.

### Privacy notices and consent management

Build adequate support for new sources of *personal data* collection and *processing*. Likewise, make corresponding optimize the management of *data subject* consent and notices and make any necessary adjustments to jurisdictional *processing*. Provide documentation that helps organizations define the roles and execute the responsibilities of controllers and processors.

### Classification and cataloging

As detailed in *CDMC 2.0 Cataloging & Classification* build and support the deployment and use of automated *data classification* and categorization schemes that embrace the methods of easy data access and data sharing in cloud platforms. These schemes should support new, complex data types and drive data protection, complex *classification* and data-use hierarchies, lineage capture and controls for managing replication and proliferation. Document responsibilities of both the organization and the *cloud service provider*.

### Shared responsibility

Ensure that documentation emphasizes accountabilities for *personal data* management by controllers and processors—and establish criteria for agreeing to the responsibilities for capturing, recording and reporting on data use. Collaborate with organizations to understand and document *personal data* stores, *processes* and the best *data security* and de-identification configurations.

### Privacy enhancing technologies

Provide privacy-by-design *guidelines* to organizations that support a continuously expanding array of data storage, *processing* and analytical capabilities. Create flexible, consistent, outcome-oriented designs that accommodate complex data and new data types. Support the establishment of cloud computing *standards* for *data security* and data privacy.

Offer the latest technology to support data privacy methods such as de-identification and *anonymization*. Integrate each *PET* with the flexibility to support data protection for a variety of use cases. Innovate to provide options that automate complex jurisdictional *processing* of *personal data*.

### Processing

Provide controls that execute automatically during data transfer operations. Support all organization applications, ensuring the correct implementation of data-use *policies* and monitor for compliance.

## QUESTIONS

- Have data collection and intent of use notifications been implemented throughout the *data lifecycle*?
- Have *processes* been established for *personal data* discovery, *classification* and inventory?
- Have *PET*s been put to use in all organization data environments to protect *personal data*?
- Have *PET*s been enhanced to improve the automation of preventative data process controls that help to mitigate *data subject* risks and privacy risks?
- Is there a process to provide *evidence* for the usage and provenance of *personal data* across the organization?
- Are *processes* in place to receive and respond to *data subject* requests and inquiries from regulators?
- Has the organization embraced a privacy-by-design culture? If so, is there verifiable *evidence* for how the cultural expectations have been communicated, the requirements the organization is expected to meet and how the implementation of those requirements will be confirmed?

### ARTIFACTS

- Data Management Procedures – for the execution of privacy requests and inquiries
- Data Privacy Notification Catalog
- Data Catalog Report – evidencing the discovery and *classification* of *personal data* across all environments
- *PET* Catalog – listing the *PET*s supported in the organization and summarizing their capabilities (including the extent to which they automate preventative controls
- Data Lineage Reports – evidencing the provenance and use of *personal data*
- Privacy-by-design Principles & Guidelines

### SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| The data privacy framework is not operational in cloud environments. | The data privacy framework is not operational in cloud environments, but the need is recognized, and the implementation is being discussed. | Implementation of the data privacy framework in cloud environments is being planned. | Implementation of the data privacy framework in cloud environments has been vaidated by *stakeholders*. | The data privacy framework is operational in cloud environments. | Operation of The data privacy framework is established as part of business-as-usual practice in cloud environments with continuous improvement. |

## 4.3 PROTECTION & PRIVACY – KEY CONTROLS

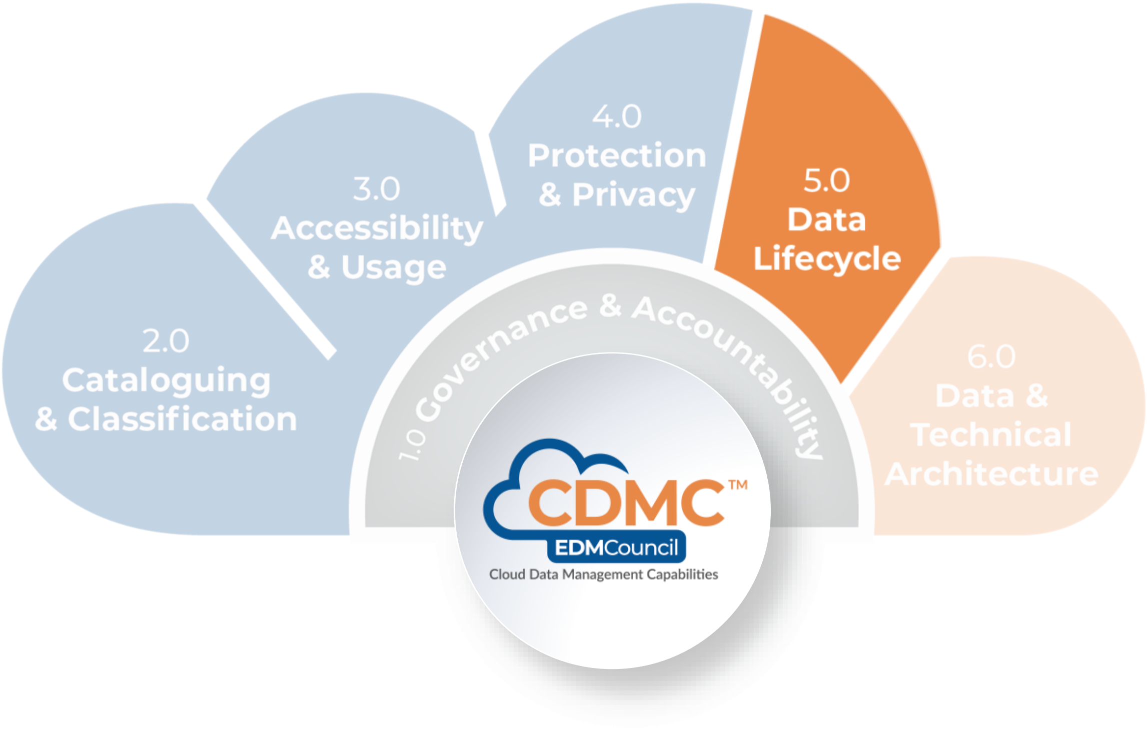
The following Key Controls align with the capabilities in the Protection & Privacy component:

- Control 9 – Security Controls
- Control 10 – Data Protection Impact Assessments

Each control with associated opportunities for automation is described in *CDMC 7.0 – Key Controls & Automations.*

*This page left intentionally blank*

# 5.0 Data Lifecycle


2.0
Cataloguing & Classification
3.0
Accessibility & Usage
4.0
Protection & Privacy
5.0
Data Lifecycle
6.0
Data & Technical Architecture
1.0 Governance & Accountability
CDMC™
EDMCouncil
Cloud Data Management Capabilities

# 5.0 DATA LIFECYCLE

## UPPER MATTER

### INTRODUCTION

A *data lifecycle* describes the sequence of stages data traverses, including creation, usage, consumption, archiving and *destruction*. Data may reside in or move through cloud environments at any of these stages. It may be consumed and used at different stages in different environments. Practitioners must apply proper data management and data management controls across all lifecycle stages to maintain data quality and consistency.

### DESCRIPTION

The Data Lifecycle component is a set of capabilities for defining and applying a *data lifecycle* management framework and ensuring that *data quality* in cloud environments is managed across the *data lifecycle*.

### SCOPE

- Define, adopt and implement a *data lifecycle* management framework.
- Ensure that data at all stages of the *data lifecycle* is properly managed.
- Define, code, maintain and deploy *data quality rules*.
- Implement *processes* to measure, publish and remediate *data quality* issues.

### OVERVIEW

A *data lifecycle* management framework supports effective management of *data assets* in an organization, beginning with creation or acquisition, and continuing through use, maintenance, archiving according to business need, and *disposal*. A well-designed *data lifecycle* management framework will ensure that the most useful and recent data is readily accessible. It can enable storage cost efficiencies as more data becomes obsolete employing automatic migration to various *storage tiers*. A solid framework also includes rules for automatic archiving and *disposal* of data. In addition, data tagging can be used to manage various exceptions to the *data lifecycle*, such as enforcing the retention of data that is subject to *legal holds* or preservation orders.

A *data lifecycle* management framework formalizes the different phases and activities of a *data lifecycle*. Data must be managed consistently throughout the *data lifecycle* regardless of whether the data resides or how it is used in a cloud or on-premises environment. To ensure compliance with legal and regulatory requirements, an organization needs to ensure that data archiving and *destruction* are managed consistently across all environments.

Consistent *data quality* management across the *data lifecycle* is critically important in cloud, *multi-cloud*, and *hybrid-cloud* environments. An effective *data lifecycle* management framework enables the consistent and uniform use of tooling across these environments. For example, it should be possible to execute the same *data quality rule* and generate consistent results regardless of whether the data is at rest or in motion across various environments. The uniform tooling enhances the ability to consistently implement distributed data quality services and rules and integrate outputs into a common repository.

An effective *data lifecycle* management framework also enables transparency and *traceability* of data throughout its lifecycle. Metrics can be established in lineage views of *data flows* across multiple environments, improving the ability to discover the sources of *data quality* issues by exposing the points in *processes* at which *data quality* deterioration is occurring. Cloud-based *data lifecycle* management framework solutions offer the opportunity for a move to nearly instantaneously alerting on *data quality rule* failures, enabling rapid diagnosis of issues, root cause analysis and remediation.

## VALUE PROPOSITION

Establishing an effective *data lifecycle* management framework enables an organization to apply proper *data management* best practices throughout the lifecycle. Data needs to be properly protected and utilized while maintaining data integrity and quality from capture to use. By deploying a *data lifecycle* management framework, an organization can combine the best *data management* practices with the features and functionality of cloud computing to deliver secure and trusted data to their end-users.

An effective *data lifecycle* management framework will:

- Enable better oversight of data through all stages of its lifecycle, ensuring better controls, protection, and appropriate uses of data.
- Enable the use of advanced artificial intelligence and machine learning techniques for detecting *data quality*, data integrity and other issues throughout the *data lifecycle*.
- Enable dynamic sizing of *processing* capacity at all lifecycle stages, providing better on-demand capabilities for high data workloads and avoiding significant capital expenditure on dedicated infrastructure.

Organizations can automate *data lifecycle* management *processes* using *metadata*-driven rules:

- Archiving and *disposal* can be automated by combining retention schedules and *data asset* *metadata* in the *data catalog*.
- *Storage tiers* of *data assets* can be optimized for performance and storage *classifications* such as staging and archiving.
- Information in the *data catalog* can indicate opportunities to reduce data duplication.

## CORE QUESTIONS

- Has a comprehensive *data lifecycle* management framework been defined and approved?
- Has the *data lifecycle* management framework been implemented?
- Is data mapped to an appropriate retention schedule?
- Are *data quality rules* and measurements being managed according to an agreed standard?
- Are *processes* for the design of *data quality* outputs defined?
- Do *data quality issue management* *policy*, *standards* and *procedures* apply across on-premises and cloud environments?

## CORE ARTIFACTS

- Data Lifecycle Management Framework
- Data Management Policy, Standard and Procedure – defining and operationalizing *data lifecycle* management
- Data Retention Schedule Specification
- Data Quality Rules Standard
- Data Quality Measurement Process
- Data Quality Rules Design Process
- Data Management Policy, Standard and Procedure – defining and operationalizing *data quality issue management*

## 5.1 THE DATA LIFECYCLE IS PLANNED AND MANAGED

Effective management of data throughout its lifecycle requires a Data Lifecycle Management framework to be defined and enshrined in *policies*, *standards* and *procedures*. The lifecycle must then be managed for all *data assets*, whether on-premises or in cloud environments.

### 5.1.1 A DATA LIFECYCLE MANAGEMENT FRAMEWORK IS DEFINED

#### DESCRIPTION

Ensuring that data is properly managed throughout its lifecycle is a strategic imperative for any digital organization. A well-designed *data lifecycle* management framework ensures that the most useful and recent data is readily accessible while delivering storage cost-efficiency. Framework design must also include considerations for information security and privacy to ensure compliance with regulatory requirements.

#### OBJECTIVES

- Gain approval on the *taxonomy* of stages of the *data lifecycle* to be adopted by the organization.
- Specify the *metadata* necessary to support automation of *data lifecycle* management and controls.
- Define *policies* for storage tiering as data progresses through the stages of the lifecycle.
- Define a *policy* and *standards* for data placement, retention and *disposal*.
- Ensure the retention and *disposal* *policy* addresses lifecycle exceptions.
- Define *standards* for the secure *disposal* of data from storage media such that data is not recoverable by any reasonable forensic means.

#### ADVICE FOR DATA PRACTITIONERS

The *data lifecycle* management framework must be defined and documented in *policies*, *standards* and *procedures* with the approval of all key *stakeholders*. *Data management* *policies*, *standards* and *procedures* must support the *data lifecycle* management framework for data hosted on-premises and in cloud environments.

The *data lifecycle* management framework should address the various requirements that pertain to *data domains*, data sensitivity, legal ownership and location. *Metadata* for each dimension should be captured in the *data catalog* (refer to *CDMC 2.1 Data Catalogs are Implemented, Used, and Interoperable*). This *metadata* can support the automation of controls that enforce the *policies* and ensure compliance with applicable laws and regulations. Legal ownership and *data sovereignty* requirements will influence how backup, archiving, access, retrieval and *disposal* are designed, supported and implemented.

Cloud environments offer different *storage tier*s and *policy*-driven placement, presenting cost-saving and automation opportunities. Storage-tiering *policies* and rules will typically be based on *metadata* such as age, last modified date, last accessed date, lifecycle status and *data domain*. Such *policies* can deliver cost savings, but the *policies* should also ensure the satisfaction of business requirements such as availability, resiliency, speed of access and retrieval and retention (in alignment with the master retention schedule of the organization. *Policies* should also maintain compliance with applicable laws and regulations.

The *data lifecycle* management framework should address any deviation from a typical lifecycle that may be in practice by the organization. Any departmental exceptions that need to be addressed in the *policy* and *standards* should consider the required response to events. Examples of such events include e-discovery requests, *legal hold* instructions and right-to-be-forgotten requests.

#### ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Since various organizations are likely to define different stages in their *data lifecycle*, cloud service providers should offer the flexibility for the organization to specifically define its *data lifecycle* stages and choose technology services that will adequately support its particular requirements. The organization will want the *data lifecycle* to be managed consistently across all environments. *Policy*-driven data placement rules should operate effectively across multiple cloud environments.

### QUESTIONS

- Has organization approval been achieved on the *data lifecycle* stages *taxonomy*?
- Has the necessary *metadata* been specified to support automation of lifecycle management and controls?
- Have *policy* and *standards* been defined for the use of storage tiering as data progresses through its lifecycle?
- Have *policies* and *standards* been defined for data placement, retention and *disposal* to ensure that data is stored, accessed, archived and disposed of in compliance with applicable rules and regulations?
- Do the *policy* and *standards* address lifecycle exceptions?
- Have *standards* for the secure *disposal* of data been defined?

### ARTIFACTS

- Data Management Policy, Standards and Procedures – defining and operationalizing *data lifecycle* management, including specification of a standard *taxonomy* of lifecycle stages and addressing the use of storage-tiering
- Data Management Policy, Standards and Procedures – defining and operationalizing data placement, retention and *disposal*, addressing compliance with applicable rules and regulations and including coverage of lifecycle exceptions and secure *disposal* of data
- Data Catalog Report – evidencing the *metadata* required for *data lifecycle* management

### SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal Data Lifecycle Management framework exists. | No formal Data Lifecycle Management framework exists, but the need is recognized, and the development is being discussed. | A formal Data Lifecycle Management framework is being developed. | A formal Data Lifecycle Management framework is defined and validated by *stakeholders*. | A formal Data Lifecycle Management framework is established and adopted by the organization. | A formal Data Lifecycle Management framework is established as part of business-as-usual practice with continuous improvement. |

## 5.1.2 THE DATA LIFECYCLE IS IMPLEMENTED AND MANAGED

### DESCRIPTION

All *data assets* must be managed throughout the entire *data management* lifecycle for data on-premises or in a cloud environment. Managing data in a cloud environment offers opportunities for *metadata*-driven automation of the *data lifecycle* management *processes*— especially for data retention, archiving, *disposal* and *destruction*.

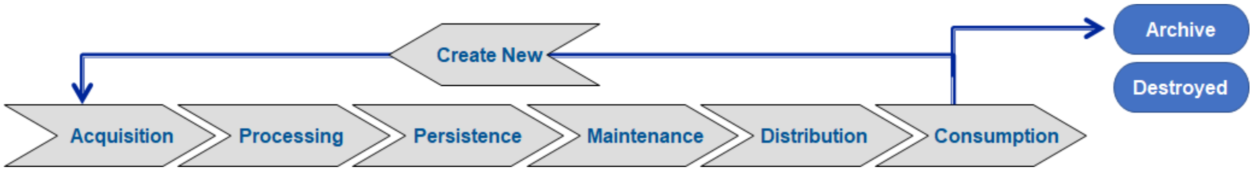
## OBJECTIVES

- Implement *data lifecycle* management *processes*, *procedures*, controls, and roles and responsibilities that cover both on-premises and cloud environments.
- Ensure that all *data assets* are mapped to an appropriate retention schedule and the applicable stages in the *data lifecycle*.
- Ensure the minimum *metadata* required for *data lifecycle* management is collected to comply with applicable laws, regulations, contracts and internal *policies*.
- Implement and demonstrate the effectiveness of controls to respond to e-discovery requests in a timely manner.
- Implement *metadata*-driven systems and *processes* in which data is disposed of, removed from operational use or destroyed (not recoverable by any forensic means).

## ADVICE FOR DATA PRACTITIONERS

The *data lifecycle* management framework describes the typical phases through which data moves in an organization. This movement begins with creation or acquisition and continues through *processing*, maintenance, archiving according to business need and *disposal*. The framework also describes the different applications and use-cases for *data lifecycle* management, including *records* management.

The figure below depicts a typical *data lifecycle*.



Generally, data *processing* at each phase involves:

1. **Creation / Acquisition** – new data is proposed, created or received by an organization.
2. ***Processing*** – data is extracted from internal and external sources. *Data quality rules* and *standards* are applied, and appropriate data remediation is put into effect.
3. **Persistence** – data is cataloged with *metadata*, described using a standard dictionary or *taxonomy*, and mapped to an appropriate retention schedule.
4. **Maintenance** – data is maintained according to defined *processes* related to defined rules and dimensions such as quality, *timeliness*, *accuracy*.
5. **Distribution** – data is distributed according to defined methods and services for the controlled data access (data-as-a-service).
6. **Consumption** – data is accessible and retrievable in a secure and timely manner— according to business requirements. Data may be moved to alternative *storage tiers* to reduce cost and increase operational capacity as data becomes obsolete.
7. **Archiving** – data retained for legal and regulatory purposes may be moved to an archive environment. The aim is to reduce costs while maintaining compliance with access and retrieval requirements.
8. ***Disposal*** – data is deleted entirely and removed from operational access. While such data may be technically recoverable, it is no longer accessible to users or *data consumers*.
9. ***Destruction*** – data is permanently destroyed such that it is no longer recoverable by any reasonable forensic means.

Data governance occurs throughout the entire lifecycle, and the governance is typically codified in data *policies* and *standards*. At any stage of the lifecycle, data may be of specific interest to regulators and litigators. Consequently, the *data lifecycle* management framework, *processes* and systems must provide reliable methods for cataloging and protecting data from deletion until all *legal holds* or preservation orders are removed.

An organization should define rules, *processes* and controls to efficiently manage multiple document versions—both structured and unstructured documents. Rules should exist to ensure that earlier versions of data are treated with the same protections as the current version. When practicable, earlier versions of documents that do not need to be retained should be automatically deleted.

Data retention requirements vary according to data type. The *taxonomy* for data types and the corresponding retention schedules should be defined for the entire organization and consistently applied to all divisions and departments. Data practitioners should establish *processes* to review and validate data cataloging (with relevant *metadata*), searching, access and retrieval practices. The aim is to ensure that all data, including archived data, is readily locatable and accessible. Archived data should be anonymized and available in a format that renders the data accurately.

Practitioners should develop and implement an archiving solution that meets the organization's requirements while remaining compliant with applicable laws and regulations. An example is a region-specific requirement such as Write Once Read Many (SEC Rule 17a-4). If data must remain available after its retention period, practitioners should ensure all actions are taken to protect against the inappropriate or incorrect use of the data. For example, if data needs to be retained beyond its retention period for *analytics* purposes, it must not be possible to link it to an individual. Refer to *CDMC 4.1 Data is Secured and Controls are Evidenced*.

Practitioners should implement automatic, *policy*-driven tiering that aligns with the *data lifecycle* while satisfying the organization's data requirements and complying with applicable laws and regulations. Seek to reduce costs by optimizing the use of *storage tiers*. Tiers vary by many factors, including cost, location, availability, resiliency, speed of access and retrieval, and minimum storage durations.

*Metadata* may be required about location and legal ownership and establishing/enforcing clear access and transfer rules (backup, archiving, access, retrieval, *legal hold* and *disposal* decisions will be sensitive to legal ownership and location). *Classification* of data may influence storage decisions and controls of data at various stages of the *data lifecycle*.

Consider segregating newer *data assets* from older *data assets*—which may not have enough *metadata* for identification. Older *data assets* are often kept beyond required retention periods. It is best to develop a risk-based methodology that makes *disposal* decisions with the best available data. Practitioners should avoid migrating *data assets* to cloud storage if that data lacks the minimum *metadata*. Before *data assets* are moved to a cloud environment, they should be reviewed and classified. The data should be disposed of if it is over-retained.

The training curriculum for the organization should include cloud environment user training. All cloud computing roles should be identified, and this *personnel* needs to understand governance, general architecture, and the *procedures* for *disposal* or amendment of holds and retention schedules.

### ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Generally, the organization owns the data, while the provider is responsible for adhering to regulatory requests. The provider's responsibility is typically done on a best-endeavors basis, subject to *encryption*/ technical access. In particular, the provider must provide the necessary mechanisms and functionality to manage data throughout the entire lifecycle. Business engagement agreements typically contain contract terms of shared responsibility for the custody of data in the context of any necessary legal compliance actions.

The provider should provide the methods for associating the minimum set of mandatory *metadata* with data and *records*. *Policy*-driven rules must directly relate to *metadata* such as age, last modified date, last accessed data, lifecycle status, and other *metadata* stored in the *data catalog*. Providers should offer flexible *metadata* APIs to access *metadata* associated with data retention schedules and exceptions to those schedules driven by business events, such as the need to apply *legal holds*.

The provider must ensure that a workflow for *disposal* or amendment of *legal holds* and retention schedules is transparent and granular enough to provide an audit trail for a decision. Functionality should also enable an organization to receive alerts for key events like the end of a customer relationship or application *legal hold*. Such events may affect the retention period, lifecycle status, or data tiering. In addition, the provider should consider providing auditable proof of data movement, retention, and *disposal* decisions.

The tiering of data and *records* should be *policy*-driven and automated, and providers should offer services that support tiering. Many providers do offer automatic archiving, and *disposal* can significantly reduce effort and costs. Another common offering is automatic switching to larger *storage tiers*, which can help control costs as an organization increases its data volumes.

Providers should ensure that any data *destruction* task is complete, such that the data is no longer recoverable by any reasonable forensic means. Note that this is distinct from data disposition, in which data has moved into the retention staging and is sent for archiving.

Each *cloud service provider* should provide training that explains the services that support automatic data retention and *disposal* and the functionality for managing exceptions to retention schedules.

## QUESTIONS

- Do *data lifecycle* management *processes*, *procedures*, controls, roles and responsibilities cover both on-premises and cloud environments?
- Is the lifecycle stage of each *data asset* recorded and maintained?
- Has each *data asset* been mapped to an appropriate retention schedule?
- Has the *metadata* required by *data lifecycle* management been collected?
- Can e-discovery requests be responded to in a timely manner?
- Is data archiving and *destruction* automated and driven by *metadata*?

## ARTIFACTS

- Data Lifecycle Management Procedures and Controls – with defined roles and responsibilities, applying to both on-premises and cloud environments
- Data Catalog Report – demonstrating recording of *data lifecycle* stage of *data assets* and capture of required DLM *metadata*, and mapping *data assets* to retention schedules
- e-Discovery Request Logs – demonstrating timeliness of response
- Data Archiving / Destruction Logs – demonstrating *metadata*-driven execution

### SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| The *data lifecycle* is not formally implemented and managed. | The *data lifecycle* is not implemented and managed formally, but the need is recognized, and the development is being discussed. | Formal implementation and management of the *data lifecycle* are being planned. | Formal implementation and management of the *data lifecycle* are defined and validated by *stakeholders*. | Formal management of the *data lifecycle* is established and adopted by the organization. | Formal management of the *data lifecycle* is established as part of business-as-usual practice with continuous improvement. |

## 5.2 DATA QUALITY IS MANAGED

Management of *data quality* starts with the management of *data quality rules*. These rules must then be deployed and executed to operationalize *data quality* measurement. *Data quality* metrics that result from this measurement must be reported and made available to owners, *data producers* and *data consumers*. *Processes* must be established to manage the reporting, tracking and resolution of *data quality* issues that are identified.

### 5.2.1 DATA QUALITY RULES ARE MANAGED

### DESCRIPTION

*Data quality rules* management includes rules governance to control how rules are put into place, rules lifecycle management to handle creation, maintenance and retirement of *data quality rules*, and rules change management and auditability to ensure that rule-based decisions can be properly understood retroactively.

### OBJECTIVES

- Establish and enforce *standards* for *data quality rules*.
- Define *processes*, roles and responsibilities for the creation, review and deployment of *data quality rules*.
- Specify and approve the lifecycle states for *data quality rules*.
- Define *processes* for the management of transitions between lifecycle states.
- Implement regular reviews of *data quality rules*.
- Ensure that changes to *data quality rules* can be audited.

### ADVICE FOR DATA PRACTITIONERS

*Data quality rules* are the cornerstone of Data Quality Management, formalizing the requirements against which data quality will be assessed. As the access to and use of data increases in cloud environments, so will the number of *stakeholders* involved in defining and executing *data quality rules*. The additional stakeholder involvement increases the importance of effective rules management to ensure they can be applied consistently across multiple clouds and on-premises environments.

*Data quality rules* management covers rules definition *standards*, rules governance, rules lifecycle management and rules change management and auditability:

- Rules definition *standards* underpin consistency in the *data quality rule* definition, for example, by standardizing the categorization or rules according to core dimensions of *data quality*.
- Rules governance defines the *processes*, roles and responsibilities for how rules are be created, reviewed and deployed.
- Rules lifecycle management defines the states that rules go through and how the transitions between those states are managed.
- Rules change management and auditability define how and when rules need to be changed and how those changes are tracked for later auditing.

*Data quality rules* encapsulate the expectations of data *stakeholders*. The specification of *standards* for *data quality rules* supported by *policy* for their adoption will provide consistency across the many *data quality rules* in an organization. The *standard* should ensure that *data quality rules* are easy to understand and explain and that they can be implemented. It should specify how rules should link to *data assets* and the definitions in the *data catalog*. It should also specify where the rules will be cataloged and how they will be cross-referenced to the applicable data.

Organizations should adopt a standard set of types or dimensions of *data quality rules*. For example, the EDM Council *Data Management Business Glossary* includes definitions of seven *data quality dimensions*. The relevance of each dimension should be considered for each *data element* for which rules are being specified. The appropriate number of *data quality rules* may depend on the business criticality of the asset. Data with higher business criticality will require greater coverage of rule dimensions.

Validation of a rule's adherence to the *data quality rule standard* is one aspect of rules governance. Automation of this validation should be considered. Responsibilities for creating rules and for their subsequent review, approval and deployment must be clearly defined. The *processes* for these activities should be standardized. However, organizations should consider applying different levels of governance depending on the business criticality of the data.

*Data quality rules* should be reviewed periodically to ensure they remain relevant. Reviews may result in decisions to update or decommission rules. The decommissioning must follow the change management *processes*.

As the volume and application of *data quality rules* increases in an organization, the need for clarity on the status of any particular rule becomes increasingly important. Lifecycle states should be defined that indicate a rule's progression from drafting, through approval to implementation and potentially to retirement. The rule governance processes and responsibilities should reference transitions between states. The organization should consider requirements to track and manage the state of groups of rules.

The results of executing *data quality rules* will be used to drive business decisions, particularly whether data is fit-for-purpose. The *data quality rules* should be treated as assets with version history maintained to enable auditability from decisions back to the rules on which they were founded. Rule creation and change management should encompass rule description, version control, change approval and deployment process, and should align with the organization's change management *standards*.

The *data quality rule standard* itself should be included in the scope of governance and change management.

### ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Any *data quality* tool should prove the ability to define and manage rules across multiple environments, whether cloud or on-premises.

*Data quality* cloud service and technology providers can facilitate consistent definition and implementation of *data quality rules* by providing access to *metadata* on all rules stored in or implemented by their products or services.

*Data quality* cloud service and technology providers should offer functionality such as workflow integration and feedback capture to support an organization's *data quality rule* governance, lifecycle management and change management *processes*.

*Data quality* cloud service and technology providers should enable automated validation of *data quality rules* against an organization's *standards* for those rules.

### QUESTIONS

- Has a *standard* for *data quality rules* been defined?
- Does the *standard* include a categorization scheme for *data quality dimensions*?
- Have *processes*, roles, and responsibilities been defined to create, review, and deploy *data quality rules*?
- Have the lifecycle states for *data quality rules* been defined and approved?
- Do standard *processes* exist for the management of transitions of *data quality rules* between lifecycle states?
- Have regular reviews of *data quality rules* been implemented?
- Are changes to *data quality rules* recorded and auditable?

### ARTIFACTS

- Data Quality Rule Standard – including a categorization scheme for *data quality dimensions*
- Data Quality Rule Governance Procedures – with defined roles and responsibilities for the creation, review and deployment of *data quality rules*
- Data Quality Rule Lifecycle Management Processes – referencing standard lifecycle states and addressing the management of transitions of *data quality rules* between lifecycle states
- Data Quality Rule Status Report – generated from rules repository and including date of the last review
- Data Quality Rule Change Management Log

### SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal *data quality rules* management exists. | No formal *data quality rules* management exists, but the need is recognized, and the development is being discussed. | Formal *data quality rules* management is being developed. | Formal *data quality rules* management has been defined and validated by *stakeholders*. | Formal *data quality rules* management is established and adopted by the organization. | Formal *data quality rules* management is established as part of business-as-usual practice with continuous improvement. |

## 5.2.2 DATA QUALITY IS MEASURED

### DESCRIPTION

*Data quality* measurement is the ability to capture metrics generated by executing the *data quality rules* established in *CDMC 5.2.1 Data quality rules are managed*.

### OBJECTIVES

- Define standard *processes* for *data quality* measurement that provide consistency across cloud and on-premises environments.

- Generate *data quality* metrics that align and integrate with relevant *metadata*. The metrics and corresponding measures should be transparent, traceable and auditable.
- Execute *data quality* *processes* in a timely, accurate and consistent manner throughout the *data lifecycle*.
- Implement regular reviews of the scalability and efficiency of *data quality* measurement *processing*.
- Seek guidance from *data owners* to clearly define and communicate *data quality* roles and responsibilities to appropriate *stakeholders*.

## ADVICE FOR DATA PRACTITIONERS

Throughout many industries, there continues to be an increase in the volume of cloud data storage, the number of *data quality rules* that interact with this data and the number of *data consumers* accessing this data. It is critically important that all *stakeholders* interested in cloud data storage receive accurate and timely *data quality* assessments within their cloud environments. Critically, these assessments depend upon the establishment and operationalization of a *data quality* measurement program.

Data exists primarily in two states—*data-at-rest* and *data-in-motion*. This sub-capability advocates frequent capture of *data quality* measurements—both from *data-at-rest* and *data-in-motion*.

### Examining data-at-rest

Practitioners should periodically examine *data-at-rest* to ensure that any data changes data are compliant with all applicable *data quality rules* of the organization. Perform *data-at-rest* analysis and measurement in a non-blocking way, such that operational dependencies on the data are not compromised. When the analysis is complete, perform routine data remediation following *data quality rules* and established metrics. A careful approach will ensure *data-at-rest* is of the highest quality. *Refer to CDMC 5.2.4 Data quality issues are managed*.

### Examining data-in-motion

*Data quality* measurements for *data-in-motion* typically run within a data production process and may sometimes be performed in a blocking way. *Data quality* measurement outputs may be intentionally configured to prevent recent data products from being published. It is important to realize that the tight coupling between data production and measurement *processes* may limit the flexibility and scalability of *data quality* controls.

For either published *data-at-rest* and *data-in-motion*, end-users should retain the option of either consuming or refusing to use the published data acquired. The choice of a user would depend on specific *data quality* limits and threshold requirements.

Take care to explicitly define all the *data quality* control points at which measurements will be taken. The *data producer* and *data consumer* are both accountable for ensuring *data quality*. Place control points near the data source and data consumption to address these tradeoffs:

- **efficiency** in identifying issues and reducing the negative consequences of data that is not fit-for-purpose (achieved by early measurement)
- ***accuracy*** of the measurement outputs to provide value for *data consumers*—typically achieved by measurements downstream in the *data processing pipeline*
- **data latency** caused by the additional *processing* time necessary for data measurement in the synchronous mode (early measurement is likely to hold up more data when failures are detected)

### Measuring other types of data

When capturing *data quality* measurements, consider all forms of data that require monitoring—including semi-structured and unstructured data. Also, take care to perform *data quality* monitoring that is most applicable to the data under examination.

*Operational metadata* for *data quality* measurement should support:

- Tracking the comprehensiveness of measurement coverage for *data assets* against established *standards* and *policies*.
- Monitoring the *data quality* measurement process and costs.
- Visibility into the operational status of *data quality* measurement (examples of status include *not started, initiated, in progress* and *comprehensive*)
- *Traceability* from *data quality rules* and *data quality* measurements through to *data quality* outputs.

**Balancing centralization and federation in measuring data quality**

A best-practice *data quality* measurement *model* balances both centralized and federated *data quality*. A central team can provide a *data quality* measurement service to all *data domains*. Measurement *standards*, tooling and outputs would be provided centrally. Each *data domain* supplies *data quality rules,* exposes data for measurement, and performs actions according to measurement outcomes. This approach benefits from higher consistency in execution and adherence to *standards*, less complexity and lower effort for each *domain*. The tradeoff with this approach is that it provides less flexibility and control for the *data domains*.

A central team provides *data quality* measurement *standards* and perhaps some tools with the federated *data quality model*. *Data quality* measurements and capture of outputs occur locally—in each *data domain*. This approach may benefit from higher flexibility and level of control for *domains*, resulting in more overall effort and some risk of divergence among the various *data domains*.

**Other measurement considerations**

For environments that exhibit frequently changing *data quality rules*, measurement should align with the existing governance *processes*. This alignment promotes consistency, *traceability* and enables *data quality* measurement to benefit from the standard capabilities provided by governance mechanisms.

In addition, practitioners should seek to monitor and optimize *data quality* measurement proactively. Since measurements will need to scale together with an ever-increasing volume of *data assets*, it is important to ensure that cost and efficiency targets related to *data quality* measurement remain within acceptable limits.

These are some methods that support such monitoring and optimization:

- Establishing *operational metadata* and *Service Level Agreements* around *data quality* measurements.
- Performing *data quality* measurements incrementally, wherever possible.
- Ensure the *data quality* infrastructure can scale as *data assets* increase in volume.
- Establish SLOs and a production support framework for *data quality* measurement capabilities.

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Cloud platforms and *data quality* measurement tooling should allow organizations to maintain consistency by deploying *data quality* measurement *processes* in heterogeneous and *multi-cloud* environments. Measurement tooling should exploit elastic cloud infrastructure to scale measurement *processes* as data volumes increase. Cloud data platforms should provide an option to run *data quality* measurement *processes* without extracting any of the data to support efficiency and *data security*.

Cloud platforms should provide common interfaces for capturing and storing *operational metadata* that supports *data quality* measurement and enables a broad range of *data consumers* to access this *metadata*. Cloud data platforms typically provide cost-efficient and execution-efficient data validation capabilities by supporting easy identification of new and modified data.

### QUESTIONS

- Have standard *processes* for *data quality* measurement been defined that provide consistency across cloud and on-premises environments?
- Have *standards* been defined and adopted for the design of *data quality* measurement control points and *processing*?
- Have standard *processes* for *data quality* measurement been implemented, such that these *processes* execute in a timely, accurate and consistent manner throughout the *data lifecycle*?
- Have regular reviews of the scalability and efficiency of *data quality* measurement *processing* been implemented?
- Have *data quality* roles and responsibilities been communicated to the appropriate *stakeholders*?

### ARTIFACTS

- Data Quality Process Documentation – providing consistency across cloud and on-premises environments
- Data Quality Measurement Standard – covering the design of *data quality* measurement control points and *processing*
- Data Quality Measurement Review Report – assesses and provides recommendations on the scalability and efficiency of *data quality* measurement *processing*
- Data Quality Measurement Operating Model – covering implementation, ongoing support and alignment with data ownership
- Data Quality Measurement Review Report – exhibits the consumption of *data catalog* *metadata* to demonstrate coverage and comprehensiveness of *data quality* measurement

### SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal *data quality* measurement exists. | No formal *data quality* measurement exists, but the need is recognized, and development is being discussed. | Formal *data quality* measurement is being developed. | Formal *data quality* measurement is defined and validated by *stakeholders*. | Formal *data quality* measurement is established and adopted by the organization. | Formal *data quality* measurement is established as part of business-as-usual practice with continuous improvement. |

## 5.2.3 DATA QUALITY METRICS ARE REPORTED

### DESCRIPTION

*Data quality* metrics result from the application of *data quality rules*. Reporting on *data quality* metrics disseminates information to data stewards, *data consumers*, *data producers*, data governance teams and other business *stakeholders* interested in a particular *data domain*, category or feed. Such reports may take the form of dashboards, scorecards, interactive reports or system alerts.

### OBJECTIVES

- Formalize the *processes* for the design and approval of *data quality* metrics reports that take data sensitivity and *data consumer* needs into account.
- Establish guidance for the design of relevant and actionable *data quality* metrics reports.

- Produce *data quality* metrics reports that combine measurements from all *data quality rules*, *data assets* and control points.
- Ensure that *data quality* metrics reports can combine time-series *data quality* measurements supporting the identification and analysis of trends.
- Ensure *data quality* metrics reports are accessible from the *data catalog* to support the *data quality* management process.

## ADVICE FOR DATA PRACTITIONERS

In *data management*, *data quality* metrics reporting serves two main purposes. Complete identification of good data builds trust and confidence in the data. In addition, identifying defective data informs *stakeholders* of the need to assess the impact of *data quality* issues, driving follow-up activities to investigate, prioritize and remediate the issues.

The need for *data quality* measurement is even more important in a cloud environment since many of the restrictions that are imposed on on-premises systems do not apply in a cloud environment. Many cloud computing environments provide various standard and comprehensive abilities for measuring *data quality* and immediate alerting of critical issues. Therefore, it is important to design *data quality* metrics reports being relevant and actionable for all intended recipients.

Many types of *stakeholders* will make use of *data quality* metrics reports. However, the importance of ensuring the viability and accountability of the data that corresponds to *data quality* metrics demands that the *data owner* is the accountable recipient of the metric reporting. The *data owner* will solicit input from the other data *stakeholders* and initiate an issues management process after assessing the impact of the defective data. Refer to *CDMC 5.2.4 Data quality issues are managed*.

**Designing data quality metrics reports**

When designing *data quality* metric reports, there are many considerations. *Data quality* metrics reporting should enable *stakeholders* to make informed decisions on whether the data is fit-for-purpose. *Data owners* and *stakeholders* must consider whether the data is fit-for-purpose when defining *data quality* metrics and deciding which will be reported. For example, some use cases may require high-quality data (such as customer billing and credit-risk modeling). In contrast, other use cases may tolerate data omissions or errors (such as marketing communications). A common approach for convenient aggregation of various use cases is to use a *data quality* scorecard. It should be possible to view individual metrics, aggregated metrics and an overall metric for each *data domain*. *Data quality* metrics should be available in the *data catalog*.

Information in *data quality* metrics reports should be actionable and informative. Provide summaries of *data quality* issues and the status—open, pending investigation or closed/resolved. Avoid excessive amounts of extraneous information. Highlight *data quality* defects that correspond to standard *data quality dimensions* such as *completeness*, *conformity* and valid values. Provide audit information and any necessary technical metrics such as the number of *records* transferred any data transmission failures.

To facilitate issue management, consider implementing visualizations that provide drill-down capability into the details of defective data. Data metrics can be captured and presented at a *data element* level, across elements and *data sets* but should only include relevant metrics for those who access the reports. Use visualization techniques such as trends, summaries, ranges and colored ratings to help users locate and understand the information. Also, organize *data quality* metrics using groups, aggregations, categories, business units, departments, geographies, and product lines.

Where sensible, implement trend analysis to indicate changes to *data quality* and support issue management and resolution. Time-series measurements can show progress in resolving *data quality* issues, so it is important to choose sensible tracking periods (daily, weekly, and monthly).

To see the impact of real-time adjustments, users must have the ability to refresh *data quality* metrics in the dashboards and reports. Provide the ability for each *stakeholder* to subscribe to specific categories of *data quality* metrics. Consider integrations with alerting and issue management systems to communicate *data quality* issues as they arise. Also, consider progress reporting as *data quality* issues are resolved.

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Comprehensive *data quality* metrics reporting depends upon collating information from multiple sources. Cloud service and technology providers must offer open APIs that provide the ability to centralize *data quality* measurements. Providers should provide standard operational reports, including data volume and transmission details, failure information, and data sampling statistics. Data-pull and data-push subscriptions should also be available to organizations to satisfy different *data consumer* requirements. In addition, some organizations depend upon frequent information updates, so it is important to have the ability to update *data quality* metrics reporting as often as possible.

Provide standard formats for both *metadata* and *data quality* metrics. Employ standard *data models* for information exchange and integration to support ease-of-use and combining metrics for reports. Also, provide readily understood visualizations and interactive dashboards to improve data quality metrics reporting effectiveness for business users and data *stakeholders*.

## QUESTIONS

- Have the *processes* for design and approval of *data quality* metrics reporting been formalized, and do they account for *information sensitivity* *classifications* and *data consumer* needs?
- Is guidance available on the design of relevant and actionable *data quality* metric reports?
- Are *data quality* metrics reports available, and do they present results from all *data quality rules*, *data assets* and control points?
- Do the *data quality* metrics reports provide flexibility in combining time-series *data quality* measurements?
- Are the *data quality* metrics reports accessible for the support of the *data quality* management process?

## ARTIFACTS

- Data Quality Metrics Reporting Process Document
- Data Quality Metrics Reports Design Guidance Document
- Data Quality Metrics Report Catalog – including a description and location of each report along with descriptions of time-series *data quality* measurements in each applicable report

## SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal *data quality* metrics reporting exists. | No formal *data quality* metrics reporting process exists, but the need is recognized, and the development is being discussed. | Formal *data quality* metrics reporting process is being developed. | Formal *data quality* metrics reporting is defined and validated by *stakeholders*. | Formal *data quality* metrics reporting is established and adopted by the organization. | Formal *data quality* metrics reporting process is established as part of business-as-usual practice with continuous improvement. |

## 5.2.4 DATA QUALITY ISSUES ARE MANAGED

### DESCRIPTION

*Data quality issue management* entails identifying, categorizing, handling, and reporting data quality issues arising from manual or automatic *data quality* measurements. An organization's *data quality issue management policy*, *standards*, and *processes* must have consistent application in all on-premises and cloud environments.

### OBJECTIVES

- Gain approval and adopt *data quality issue management policy*, *standards* and *procedures* that apply consistently across on-premises and cloud environments.
- Provide integrated *data quality* issue reporting for all on-premises and cloud environments.
- Ensure *data quality* issues link directly to specific *data assets* and the relevant *metadata* in the *data catalog*.
- Establish metrics that provide *evidence* for sufficient coverage and effectiveness of *data quality issue management*.

### ADVICE FOR DATA PRACTITIONERS

Increasing the degree of automation for a *data quality issue management* process typically drives an increase in efficiency. Since *data quality issue management* impacts much of an organization, seek to standardize and automate wherever practicable. Data practitioners should integrate the *data quality issue management processes* into the organization-wide issue management *processes*. The visibility and routines of organization-wide issue management activities will attract valuable *stakeholder* attention and marshal resources to urgent *data quality* tasks. In addition, practitioners should establish transparent issue management workflows that are visible to the entire organization.

Practitioners can use the values of the *data quality* measurements to shape a risk-based prioritization of any *data quality* issue. Refer to *CDMC 5.2.3 Data quality metrics are reported*. Practitioners should develop closure criteria for each auditable *data quality* issue to ensure a complete response that demonstrates how each issue is identified, how the ownership was allocated, and how the issue was prioritized, resolved, mitigated or accepted with documented risk.

Many *data quality* issues are manageable as part of a larger problem. Practitioners should establish *processes* for automating the identification of common root causes across multiple issues. These *processes* will support the ability to prioritize more effectively and allocate resources to the highest priority issues.

It is good practice to annotate *data assets* with descriptive *metadata* upstream in a *data processing pipeline*. Such *metadata* would minimally include *reference data*, data inventory and lineage. *Data quality* issues can be automatically attributed to relevant *data owners*, *data producers*, and *data consumers* with some additional configuration. Practitioners will benefit substantially by examining the *data catalog* to identify the responsible *data owner* and assigning accountability for specific *data quality* issues. This step minimizes manual triage and supports issue ownership allocation against the root cause rather than at the point of discovery.

Data practitioners should also generate reports that compile the *stakeholder* accountability matrix of the organization. This matrix should include the extent of accountability for business process owners, technology platform owners, data stewards and *data owners*.

### ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

A key driver of complexity in *data quality issue management* is the potential variety of participants involved at different steps in the process. Some of these roles are the issue identifier, the *data owner*, business owners, data architects, IT *personnel* and various organizational *stakeholders* affected by *data quality* issues. Other participants include members of the project management office, who will need to allocate funding for issue resolution.

Cloud service and technology providers (*CSPs*) should provide abilities for issue ownership identification and assignment capability. In addition, *CSPs* should provide workflow automation that facilitates active monitoring and alerts for *data quality issue management*.

*CSPs* should provide capabilities for integrating *metadata* for *data quality* issues into the general *data domain* management feature set. A cloud computing facility for managing such *metadata* provides enhanced reporting for managing high-criticality issues. Such a facility supports risk-based prioritization of a *data quality* issue based on the impact on relevant business *processes*. This prioritization method increases confidence in the integrity of those *processes*.

To provide a broad foundation for reporting on the impact of a *data quality* issue, *CSPs* should ensure that *metadata* for *data quality* issues is linkable to other *metadata* in the *data catalog*. For example, a data practitioner may want to identify all downstream process impacts that might result from a *data quality* issue. This impact identification is most easily made by enumerating all the impacted platforms, data structures, *data elements* and business *processes*. Since this requires viewing the entire *data lineage* and relevant *metadata*, *CSPs* should provide this ability for all cloud-based data.

## QUESTIONS

- Have *data quality issue management policy*, *standard* and *procedures* been approved and adopted, and have they been applied across on-premises and cloud environments?
- Is *data quality* issue reporting integrated across on-premises and cloud environments?
- Do all *data quality* issues link to *metadata* in the *data catalog*?
- Are metrics in place that provide *evidence* for sufficient coverage and effectiveness of *data quality issue management*?

## ARTIFACTS

- Data Management Policy, Standards and Process Documents – defining and operationalizing *data quality issue management* that covers both on-premises and cloud environments
- Data Quality Issue Reports – presenting the integration of issues from both on-premises and cloud environments
- Data Catalog – including *metadata* that links *data quality* issues to data in the catalog
- Data Quality Issue Management Metrics Report – that provides *evidence* of issue management coverage and effectiveness

## SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal *data quality issue management* exists. | No formal *data quality issue management* exists, but the need is recognized, and the development is being discussed. | Formal *data quality issue management* is being developed. | Formal *data quality issue management* is defined and validated by *stakeholders*. | Formal *data quality issue management* is established and adopted by the organization. | Formal *data quality issue management* is established as part of business-as-usual practice with continuous improvement. |

## 5.3 DATA LIFECYCLE – KEY CONTROLS

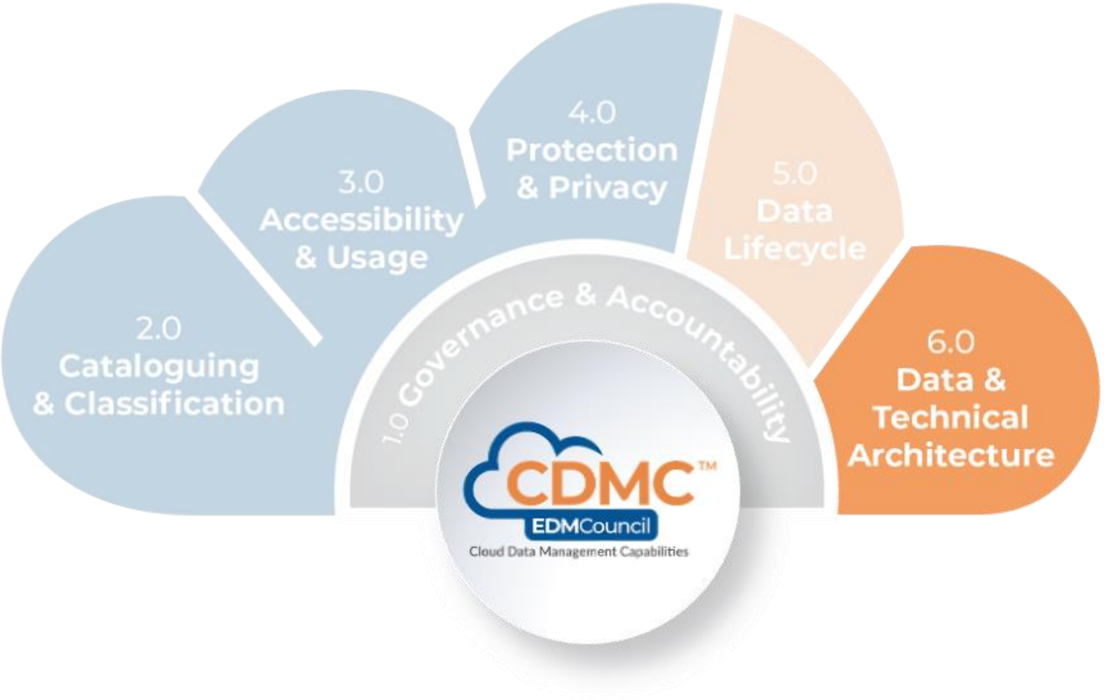
The following Key Controls align with the capabilities in the Data Lifecycle component:

- Control 11 – Data Quality Metrics
- Control 12 – Data Retention, Archiving and Purging
-

Each control with associated opportunities for automation is described in *CDMC 7.0 – Key Controls & Automations*.

*This page left intentionally blank*

# 6.0 Data & Technical Architecture



# 6.0 DATA & TECHNICAL ARCHITECTURE

## UPPER MATTER

### INTRODUCTION

Data and technical architecture address architectural issues unique to cloud computing and affect how data is managed in a cloud environment. With most *cloud service providers*, there are many options for how business solutions can be designed and implemented in cloud environments, using a variety of cloud services and in how these services are configured and consumed. Developing and adopting specific architecture patterns and *guidelines* can provide a foundation for best practice *data management* in cloud environments.

### DEFINITION

The Data & Technical Architecture component is a set of capabilities for ensuring that data movement into, out of and within cloud environments is understood and that architectural guidance is provided on key aspects of the design of cloud computing solutions.

### SCOPE

- Establish and apply principles for *data availability* and resilience.
- Support business requirements for backup and point-in-time recovery of data.
- Facilitate optimization of the usage and associated costs of cloud services.
- Support data portability and the ability to migrate data between *cloud service providers*.
- Automate identifying data *processes* and flows within and between cloud environments, capturing *metadata* to describe data movement as it traverses the *data lifecycle*.
- Identify, track and manage changes to *data lineage*, establishing the ability to explain lineage at any point-in-time.
- Provide tools that meaningfully report and visualize lineage—from both a business perspective and a technical perspective.

### OVERVIEW

Cloud computing introduces capabilities that an organization should include in its *data management* and architecture best practices. These capabilities allow an organization to adopt leading-edge approaches to *data management*, such as Data-as-a-Service (DaaS) or data fabrics. However, many organizations may find it best to begin more simply and seek or develop guidance on various aspects of data storage such as speed (*storage tier*), type (including data stores, object stores and file stores) and geographic location. In any solution design, it will be necessary to balance cost and functionality considerations with consumption-based pricing (based on the volume of data stored and the volume of data ingested or egressing), providing more flexibility than is typically the case in on-premises environments.

A *cloud service provider* uses APIs for managing data services, typically available for manual and automatic use. Many *cloud service providers* offer an array of APIs for computation, storage management, *data management*, scaling, monitoring and reporting capabilities—often well beyond what is typically available in on-premises environments.

There are different types of data recovery options available within most cloud environments. The option chosen will determine the speed at which any recovery can be accomplished. Availability zones, block-based storage replication and other options can enable the organization to exploit various techniques to facilitate recovery based on the criticality of specific data access and application requirements. Dynamic scalability is a key feature of cloud computing and can be used in various ways to enhance *data resiliency* and availability by separating computational and storage functionality. *Multi-cloud* environments extend that scalability further. An organization will need to develop *guidelines* that aim to improve cost efficiency and address the cost of data movement between cloud environments while supporting business data needs for resiliency and availability.

Compared with on-premises environments, tools and services within cloud environments typically enable better discovery of detailed *data lineage* and provide more detailed, accurate, and up-to-date lineage tracking than on-premise. Consequently, a much greater degree of *data lineage* detail is available within a cloud environment and enables:

- Validation of data sources.
- Analysis of the impact of change and improved root-cause analysis of *data quality* issues.
- Detection of duplicate or conflicting *data transformations* and derivations.
- Detection and assessment of data replication and redundancy.

Migration to cloud computing is an opportunity for an organization to rationalize its *data ecosystem* and simplify its *data lineage*. Data movements within and between environments will expand. Cloud computing greatly enhances the ability of an organization to detect and record these movements automatically. The immediate scalability of *processing* power in a cloud environment enables a level of detail to be captured in *data lineage* that is rarely feasible within on-premises environments. Automated monitoring reduces the effort necessary to maintain this lineage data.

The capture of *data lineage* is critical to controlling data in a cloud environment. Understanding the actual source of data and the movement of data from source to consumption provides confidence in the data that is put to use in business *processes* and *analytics*. It underpins regulatory compliance, impact analysis, quality troubleshooting and detecting any data duplications. Cloud technology offers organizations significant potential to automate many aspects of *data lineage* discovery and management.

While *data lineage* tracking is more readily performed in the cloud, it is important to note that when *data flows* are moved from an on-premises environment without re-engineering, the lineage may not be discoverable by monitoring services in the cloud environment. In such cases, relevant *metadata* must be loaded to the cloud data store with the associated data.

## VALUE PROPOSITION

Organizations that establish best practice architecture patterns and *guidelines* for the adoption of cloud capabilities can confidently maximize the ability to realize value from those capabilities:

- *Data management* best practices can be engineered into cloud solutions.
- The compute and storage scale of the cloud offers great global availability and resiliency for increased data accessibility and recovery. Scalability also offers greater cost efficiencies.
- Multiple cloud environments reduce the perceived business risk of data access.
- Choices of technical options such as storage tiering, availability zones and replication can match business needs.
- Cost visibility and control can be built into the solution design.

Adopting best practices and *standards* for data portability provides a basis for exiting or changing cloud services in response to commercial or regulatory drivers.

Organizations that take advantage of the enhanced capabilities for managing *data lineage* within cloud environments can reduce the costs associated with *data lineage* and create opportunities to realize business value from well-understood *data lineage*:

- Costly and error-prone manual activities can be eliminated.
- Detailed point-in-time information can be produced with ease to satisfy regulatory audits.

- The *analytics* environment can be simplified by automated monitoring of the *data ecosystem*. The automation minimizes both the need for the analysts to manually research the data sources and the risk that unwanted changes to the provenance of their data will go undetected.

## CORE QUESTIONS

- Has architectural guidance for the design of backup approaches been provided?
- Has architecture guidance and patterns been provided for data *processing*, use, storage and movement?
- Are there architectural *standards* on how solutions should provide data transfer and *processing* to another provider?
- Have architecture patterns been selected and implemented to support business requirements for availability and resilience?
- Have *policies* and *standards* been established for backup strategies, planning, implementation and testing?
- Is lineage automatically discovered and recorded across all in-scope environments?
- Have *policies* and *procedures* for lineage change management been defined?
- Can lineage be reported for any historic point-in-time?
- Have lineage reporting and visualization requirements been documented and approved?

## CORE ARTIFACTS

- Architecture Patterns – addressing backups, data *processing*, use, storage and movement, availability and resilience, and data portability
- Data Management Policy, Standard and Procedures
    - Defining and operationalizing data backup
    - Defining and operationalizing *data lineage* change management
- Cloud Provider Exit Plan
- Data Lineage Reports
- Data Lineage Reporting and Visualization Requirements

## 6.1 TECHNICAL DESIGN PRINCIPLES ARE ESTABLISHED AND APPLIED

Technical design principles must be established to facilitate the optimization of cloud use and cost efficiency. They must guide the implementation of solutions that meet availability, resilience, back-ups, and point-in-time recovery requirements. The ability to exit cloud services must be planned and tested, facilitated by data portability between cloud and on-premises environments.

### 6.1.1 OPTIMIZATION OF CLOUD USE AND COST EFFICIENCY IS FACILITATED

#### DESCRIPTION

The placement, storage and use of data in a cloud environment offer an organization greater flexibility and capability, which is not typically available in an on-premises environment. However, all activities associated with a cloud environment, including data *processing*, use, storage and movement, will incur costs. In addition, many decisions must be made for architecture, design and implementation solutions—and these must follow best practices. Simultaneously, design principles and *guidelines* must be migrated, revised and established to consider all necessary trade-offs between functionality, use and maximizing cost efficiency.

#### OBJECTIVES

- Establish architecture *guidelines* and patterns for data *processing*, use, storage and movement—emphasizing automation to drive standardization.
- Provide *guidelines* on solution design that optimize functionality and cost while sufficiently addressing constraints such as security, integrity and availability.
- Identify and gain approval on the cost drivers that must be addressed in cloud-based solution business requirements, including data retention, availability, and sovereignty.
- Define and capture usage and cost transparency metrics that adequately support management decision-making and ongoing oversight.

#### ADVICE FOR DATA PRACTITIONERS

Architecture guidance and patterns should be used to capture and formalize best practices for designing cloud solutions. There are many considerations. It is important to begin by matching requirements driven by the sensitivity of the data with the cloud provider features, balancing functionality and cost. Practitioners may need to lead an effort to make decisions on single-cloud, *multi-cloud* and *hybrid-cloud* designs; the location of data stores and *processing*.

Architectural guidance should encourage the decoupling of compute and storage, enabling the ability to scale each independently, supporting both a cost-effective and high-performing solution. It may be necessary to create operational duplicate data stores to meet availability requirements. Wherever practicable, automate and standardize data movements outbound, inbound and within cloud environments.

In addition, *data lifecycle* management should be driven by organizational *policy*. Also, use compressed formats to reduce data storage and transfers costs and avoid unintentional or unnecessary *processing* and data movement.

Practitioners should verify that cloud service and technology providers can support the architecture patterns and provide the major cost contributors and rate information related to any guidance provided. While the selection and management of service providers are beyond the scope of this sub-capability, the ability of providers to optimize usage, efficiency and costs should be factored in the selection of new providers.

Practitioners should also consider how cost optimization will be assessed and managed. Information on actual costs incurred should be used to justify employing existing systems. Implemented designs must continue to be cost-effective and demonstrate ongoing potential for optimization. Measurements to inform this process will depend on well-defined metrics, effective cost management and an effective internal chargeback process that allocates specific costs to the business solutions that incur them.

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Cloud service and technology providers should automate data collection and report on cost drivers, including data replication, *storage tier*, retention period and *destruction* deadlines. Providers should also offer comprehensive reporting on resource utilization, billing costs for storage, usage, access, and data movements. The organization should be able to act on alerts that trigger on pre-defined thresholds. Examples include:

- alerting on a request to move data from a lower to a higher cost *storage tier*.
- alerting when an attempt is made to delete data with an incorrect data age.

Such abilities enable an organization to proactively monitor and manage its cloud environment(s) from risk management, cost and business *function* perspectives.

In addition, providers should give the organization the ability to readily access logging services that give full visibility into all data activity and movements. Ideally, log access APIs should be available with different levels of user access. This API functionality should include uploading, extracting, and accessing the log data. These abilities give the organization the benefits of detail monitoring, analysis and gaining insights on minimizing costs.

## QUESTIONS

- Have architecture *guidelines* and patterns been established for data *processing*, use, storage and movement?
- Do architecture *guidelines* emphasize automation to drive the adoption of standard patterns?
- Are cost drivers identified and approved that must be addressed in business requirements for cloud-based solutions?
- Have *guidelines* been established for solution design that optimizes functionality and cost and sufficiently address security, integrity and availability constraints?
- Are usage and cost transparency metrics defined and captured that adequately support management decision-making and ongoing oversight?

## ARTIFACTS

- Cloud Architecture Requirements and Guidelines – including advice on automation and adoption of standard patterns
- Cloud Architecture Patterns – including approved designs for data *processing*, use, storage and movement
- Business Requirements Template – including requirements that affect costs
- Solution Design Guidelines – including guidance on optimizing functionality, cost and constraints such as security, integrity and availability
- Cloud Use and Cost Reports – including metrics defined and captured to support management decision-making and ongoing oversight

### SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| Optimization of cloud use and cost efficiency is not facilitated | The need for optimization of cloud use and cost efficiency to be facilitated has been identified, and the facilitation is being discussed | Facilitation of cloud use and cost efficiency optimization is being developed | Facilitation of optimization of cloud use and cost facility is defined and validated by *stakeholders*. | Facilitation of optimization of cloud use and cost efficiency is established and adopted by the organization. | Facilitation of optimization of cloud use and cost efficiency is established as part of business-as-usual practice with continuous improvement. |

## 6.1.2 PRINCIPLES FOR DATA AVAILABILITY AND RESILIENCE ARE ESTABLISHED AND APPLIED

### DESCRIPTION

The business requirements for *data availability* and resiliency in a cloud environment must be documented and approved. These requirements are applied to the cloud architecture design to employ cloud capabilities that support availability, accessibility, replication, and resiliency across the entire architecture.

### OBJECTIVES

- Define and gain approval of business requirements for *data availability* and resilience.
- Provide availability and resiliency *guidelines* for selecting storage and access options available from the *cloud service provider*.
- Provide *guidelines* on employing and configuring availability zones to meet requirements for resilience and high availability.
- Ensure each data resource has a corresponding are tagged with their availability and resilience *service level agreement* (SLA) and *service level objective* (SLO) and in the *data catalog*.
- Develop architecture patterns for providing data consistency, availability, and partition tolerance.
- Adopt appropriate architecture patterns in line with business requirements for *data availability* and resilience.

### ADVICE FOR DATA PRACTITIONERS

Typically, an organization relies on data as the cornerstone of its business. Solid cloud architecture and design can significantly improve the outlook for achieving the best possible *data availability* and resiliency.

A key principle in establishing *data availability* and *data resiliency* for *data management* in a cloud environment is to establish controls that ensure data is available only to authorized users in a controlled manner (adhering to data protection and privacy) to satisfy business requirements. Another key principle in achieving high availability and resiliency is to employ cloud storage capabilities such as storage options, availability zones and optimizing for area considerations. One more key principle is selecting and configuring a cloud storage architecture that meets the organization's various availability and resiliency requirements of all relevant user types.

Fundamentally, the cloud environment architecture pattern must balance data duplication for availability with the costs and consistency implications of that duplication. The architectural pattern should address features such as repeatable results for data loading and provide a restart capability from the last successful point during *processing*. Architects should be aware of the trade-offs and choices implied by the CAP Theorem, also known as Brewer's

Theorem, which states that only two of the three properties of consistency, availability and partition tolerance can be guaranteed.

It will be necessary to analyze and re-architecture business applications developed for on-premises to fully utilize cloud services' availability and resilience capabilities. Architecture blueprints and patterns for cloud-native applications will guide this re-architecture.

All *stakeholders* responsible for implementation in the cloud environment must establish a system of controls to monitor the SLAs and SLOs for the environment. Contractual agreements with *cloud service providers* should include SLAs for *data availability*. The SLO is a component of an SLA and enables measuring the service provider's reliability at guaranteed thresholds defined by the SLA. For availability and resilience, the SLO provides a quantitative document for defining the level of service the organization can expect through metrics such as up-time and network throughput.

As stated in the Upper Matter for this component, data practitioners should take advantage of cloud service providers' architecture training and education resources.

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Cloud and technology service providers should provide architectural blueprints to data practitioners, giving detailed information on the various storage types and capabilities that support referencing *data availability* and resiliency. Providers should communicate guidance on how the cost of common or relevant storage and workloads must be considered when selecting an approach that satisfies the business requirements. Refer to *CDMC 6.1.1 Optimization of cloud use and cost efficiency is facilitated.*

Providers should also describe how availability zones use replication to support high availability in a cloud implementation. APIs should be available to enable the automation of *data availability* and resiliency. Lastly, providers should offer playbooks for assessing the suitability of how each architectural pattern could be implemented.

## QUESTIONS

- Are there approved business requirements for availability and resilience?
- Have *guidelines* for the selection of storage and access options been documented?
- Has guidance on the use of availability zones been provided?
- Have patterns for consistency, availability and partition tolerance trade-offs been developed?
- Does each data resource have a corresponding availability and resilience *service level agreement* (SLA) and *service level objective* (SLO), and is each document accessible in the *data catalog*?
- Have architecture patterns been selected and implemented to support business requirements for availability and resilience?

## ARTIFACTS

- Business Requirements Document – including requirements for *data availability* and resilience
- Service Level Agreement – including SLOs for *data availability* and resilience
- Architecture Standards, Patterns and Guidelines – covering the selection of storage and access options, use of availability zones and consistency, availability and partition tolerance trade-offs
- Data Catalog – including cloud service and technology provider SLA / SLO tags

## SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal principles for *data availability* and resilience exist. | No formal principles for *data availability* and resilience exist, but the need is recognized, and the development is being discussed. | Formal principles for *data availability* and resilience are being developed. | Formal principles for *data availability* and resilience are defined and validated by stakeholders. | Formal principles for *data availability* and resilience are established and adopted by the organization. | Formal principles for *data availability* and resilience are established as part of business-as-usual practice with continuous improvement. |

## 6.1.3 BACKUPS AND POINT-IN-TIME RECOVERY ARE SUPPORTED

### DESCRIPTION

The ability to perform data backups is critical for disaster recovery and, more general point-in-time data recovery. Backups are copies of data typically stored in a location that is different from the primary data store. Data retrieved from a recent backup is used to restore data and system configuration in a disaster recovery context. In addition, data from an older backup can be used to restore data that had been in use at an earlier time.

### OBJECTIVES

- Gain approval and adopt backup and recovery strategy, planning, implementation and testing *policy*, *standards* and *procedures.*.
- Ensure backup and recovery capabilities support disaster recovery and point-in-time data recovery.
- Ensure backup and recovery *standards* specify isolation and *data residency* requirements.
- Ensure backup and recovery *standards* specify security requirements that align with the *data asset classifications* in the *data catalog*.
- Provide architectural guidance for designing backup and recovery approaches.
- Ensure that the backup and recovery plans reflect the *standards* and guidance.

### ADVICE FOR DATA PRACTITIONERS

**Cloud operational backups**

Data backups are special copies of the contents of data stores. Backups are stored in a different location from the original data. When necessary, data is taken from a backup copy to restore data to a correct state. Backups must be secure and provide reliable recovery mechanisms to ensure that a logical recovery can occur when needed. A solid backup plan ensures that data is readily recoverable with minimal data loss.

Above all, a backup and recovery plan must outline how to recover data quickly and completely. After compiling a recovery plan, it is essential to test it immediately and at regular intervals. When designing a backup plan, it is also important to specify retention periods, backup file capacity requirements and the method for disposing of unnecessary backup files.

Using proprietary formats for data backups may be problematic when attempting to restore data if there is a failure in the proprietary system.

### Data archiving

A data archive is a copy of data that is put into long-term storage. The original data may or may not be deleted from the source system after the archive copy is made and stored, though it is common for the archive to be the only copy of the data.

Archiving is different from an operational backup and is typically done to support regulations and legal requirements. Many archiving solutions use simple, generic methods for storing copies of data. The archive copies are independent of the archiving system and the primary data storage system

An archive may have multiple purposes. By maintaining data archives, an organization can maintain an extensive permanent record of historical data. Commonly, a data archive directly supports information retention requirements for an organization. If a dispute or inquiry arises about a business practice, contract, financial transaction, or *employee*, the *records* about that subject can be obtained from the archive. Refer to *CDMC 5.1 The Data Lifecycle is Planned and Managed*.

### Backup scope

A basic assumption of an effective backup and recovery plan is that all data becomes inaccessible in the operational environment. The data itself is typically only one aspect of a system that requires recovery. The plan must account for the infrastructure configuration, environment variables, utility scripts and source code and other relevant subsystems that integrate with the main application.

Also, it may be necessary to consider:

- Machine learning and other time-variable modeling are not easily reconstructed from source code, the data, or the *model* outputs. Therefore, it is essential to make provisions to protect the training *data sets*.
- If *encryption keys* protect any data, the loss of these keys will render the data unrecoverable. All data *encryption keys* should be part of a backup and recovery plan.

### Backup resiliency

A backup and recovery process must be resilient against a multitude of possible issues to be effective. Regulators routinely require minimum data backup protections. The *air gap* is a common implementation for satisfying regulations for protecting data backups.

The OCC asks GSIPs to "*Logically segment critical network components and services (e.g., core processing, transaction data, account data, and backups) and, where appropriate, physically air gap critical or highly sensitive elements of the network environment."* They go on to highlight backups. "s*ecurely store system and data backups offsite at separate geographic locations and maintain offline or in a manner that provides for physical or logical segregation from production systems."*[2]

The FFIEC (a coalition of Fed RB, FDIC, NCUA, OCC and CFPB) defines an air gap as:

*"An air gap is a security measure that isolates a secure network from unsecured networks physically, electrically, and electromagnetically."*

*"In accordance with regulatory requirements and FFIEC guidance, financial institutions should consider taking the following steps. Protections such as logical network segmentation, hard backups, air gapping, maintaining an inventory of authorized devices and software, physical segmentation of critical systems, and other controls may mitigate the impact of a cyber attack involving destructive malware…."*[3]

[2] https://www.fdic.gov/news/financial-institution-letters/2020/fil20003a.pdf

[3] https://www.ffiec.gov/press/PDF/2121759_FINAL_FFIEC%20Malware.pdf

### Backup resiliency in the cloud

Historically, many on-premises backup *procedures* involved taking a data backup, storing it in another onsite location and duplicating a copy of the backup to a storage medium that would be stored offsite. This *procedure* and movement to segregated storage would meet the air gap's physical, network and electromagnetic isolation requirement. The air gap requirement intends to provide backup isolation and have no single point of failure between the primary data store and backup storage.

When employing cloud computing solutions to support a backup and recovery plan, data practitioners should be aware of the technology options available from the *cloud service provider* (*CSP*) to ensure that physical, network, electrical and electromagnetic isolation requirements are met.

It is the responsibility of the data practitioner to understand, configure and verify that *CSP* solutions meet the requirements in the backup and recovery plan. The data practitioner should provide testing *evidence* that the backup and recovery plan is readily executable through chosen *CSP* technologies.

The data practitioner must carefully examine the isolation capabilities of the *CSP*, and the *CSP* must provide *evidence* that its technologies provide backup isolation that meets organization requirements.

There are many other techniques available to the data practitioner to help satisfy air gap requirements.

- **Network isolation** – Using a separate Virtual Private Cloud (VPC) to isolate operational and public-facing components from backup environments.
- **Logical separation** – Implementing security and permission schemes with the application environment to support the division of duties and prevent operations from adversely impacting the backup and configuration areas.
- **Physical redundancy** – Backup replication can be done in the local region or in other availability zones to mitigate localization risks. When practicable, it is best to protect backups from electronic, electromagnetic, and physical risks. These protections should be done with consideration for any residency, sovereignty, or localization requirements.
- **Immutable storage** – Write-Once-Read-Many (WORM) storage devices are useful in mitigating the risks of corruption, deletion, unauthorized modification or unintended alteration of data. WORM storage and other similar storage offerings can address part of the air gap requirement since such technologies are highly impervious to overwrites or deletions.
- **Security and *encryption* of backups** – Since much of the data in the operating system must be protected, the backup files and environments must be secure. Though there are performance tradeoffs, it is important to consider the *encryption* of all backup files.

### Backup gold copy

One approach that ensures the backup gold copy status configures backup and recovery to employ the cloud computing environment to write backup files to immutable cloud storage in a secure, segregated network with replication over physical zones. The segregated network would ensure isolation, immutable storage would prevent file corruption, and the redundancy provided by the *CSP* would protect the backup from electromagnetic or physical risks. Also, backup files can be encrypted and stored with exclusive access rights.

Typically, cloud storage is highly redundant. Many providers offer three or more availability zones, 99.99% availability and extreme durability (> 99.999% recoverability). Using a *CSP* for backup storage is a strong mitigator of the risks of site failure and localization. Establishing network isolation and implementing highly restrictive access controls prevent accidental corruption and negative effects from malicious software or bad actors.

### Planning for point-in-time recovery

Point-in-time recovery gives administrators or users the ability to restore data from a backup. The contents of the *operational data* will be identical to a specific point in the past. Examples include an accidental drop of a database table, an unintentionally committed update, or a process that maliciously corrupts data in files or systems.

Point-in-time recovery can employ cloud storage replication *processes* that simplify backup-and-restore *processes*. Point-in-time recovery can also exploit cloud capabilities such as availability zones or multi-tier storage.

These are common use cases for point-in-time recovery:

- **Transaction failure** – If a transaction, system write or save fails before completion, the system may not be successful in restoring all data to the correct values, and inconsistent data may result. A point-in-time recovery would be a suitable remedy in such cases.
- **Rogue or malicious process** – If an unauthorized update, deletion or change results in corrupt data, point-in-time recovery is suitable for restoring a system, subsystem, table, or file to the state before the corruption.
- **System failure** – To recover from a failure that may have corrupted data at the system level, such as a software release gone bad. Point-in-time recovery is also effective for restoring data that has become broadly corrupt due to a system failure or software update.
- **Media failure** – A hardware failure is very similar to a system failure, and the recovery plan is nearly identical. A failure is less likely to occur in a highly redundant cloud computing environment.
- **Point-in-time for disaster recovery** – For many organizations, point-in-time recovery is also used for disaster recovery planning. A common approach is to activate a standby site when a system with no high-availability capability needs to be brought online following a failure.

**Planning for recovery point and recovery time objectives (RPO & RTO)**

It is good practice to design a backup and recovery plan that accommodates various system criticality. A plan for point-in-time recovery should primarily be driven by a recovery point objective and recovery time objective. In a large organization, practitioners can create patterns or blueprints for systems that share similar RPOs and RTOs.

A recovery point objective is a specific volume of data that an organization identifies as an acceptable loss in a disaster. Replication can provide close to a real-time recovery point and replicate all changes to another location. Systems may not warrant such a strategy. Business and IT demands should shape the recovery point objective and associated plan.

A recovery time objective is equally important as the recovery point objective since it provides the business requirement for a tolerable outage duration. All relevant systems must be operational before the recovery time objective duration elapses. The amount of time to recover a system depends on the recovery method, frequency of backup checkpoints and the volume of data to recover. Recovery time may lengthen with the consumption of storage capacity and decreasing network capacity.

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Cloud service and technology providers (*CSP*s) should ensure that backup and restore utilities are accessible through a console, command line, and API. Backup and restore services should be available through these methods for each type of storage (block, object, and database).

The *CSP* should explain each of its available *storage tiers*, including media used and costs. In addition, an organization should be aware of storage accessible in one or more availability zones and regions depending on the organization's needs. If the *CSP* provides it, the organization should know what options are available for write-once-read-many (WORM) storage and the features that are available with this storage type.

At a minimum, a *CSP* should provide the ability to encrypt backups and backup files, the ability to create and track multiple versions of backups, the ability to provide visibility and tools for application stack versions and underlying services to aid in the restoration of versions, and the ability to isolate backups from operational systems using virtual networking. Where applicable, present the options that are available for self-managed and *managed services* for backups.

*Evidence* that these solutions have been implemented should be readily available and presented in reports that include the network, region and other attributes that show isolation characteristics. Ideally, this *evidence* should be consistent and accessible for each of the backup solutions.

## QUESTIONS

- Are there *policies* and *standards* for backup strategies, planning, implementation and testing?
- Do backup and recovery capabilities support disaster recovery and for point-in-time data recovery?
- Do the backup and recovery *standards* specify isolation requirements?
- Do the backup and recovery *standards* specify security requirements that align with the *data asset classifications* in the *data catalog*?
- Is architectural guidance available for designing backup and recovery approaches?
- Do the backup and recovery plans reflect the *standards* and guidance?

## ARTIFACTS

- Data Management Policy, Standards and Procedures – defining and operationalizing data backup
  - Covering backup strategies, planning, implementation and testing
  - Specifying isolation and *data residency* requirements
  - Specifying security requirements that align with the *data asset classifications* in the *data catalog*
- Backup Architecture and Design Guidance – covering both disaster recovery and point-in-time recovery
- Backup and Recovery Plans – reflecting the *standards* and guidance

## SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal support of backups and point-in-time recovery exists. | No formal support of backups and point-in-time recovery exists, but the need is recognized, and the development is being discussed. | Formal support of backups and point-in-time recovery is being developed. | Formal support of backups and point-in-time recovery is defined and validated by stakeholders. | Formal support of backups and point-in-time recovery is established and adopted by the organization. | Formal support of back-ups and point-in-time recovery is established as part of business-as-usual practice with continuous improvement. |

## 6.1.4 PORTABILITY AND EXIT PLANNING ARE ESTABLISHED

### DESCRIPTION

There are several drivers for the need to transfer data from a cloud environment. The transfer may be a planned data movement to a provider that offers new functionality. Data transfers are also a necessary part of an exit from an existing cloud service or technology provider. Another reason for a data transfer may be a regulatory expectation or a necessary response to an internal risk assessment. Whatever the reason, data portability planning and testing are required. Consequently, exit planning and data portability are critical capabilities when designing and implementing a sustainable cloud environment.

### OBJECTIVES

- Document and approve the requirements of data portability and establish a viable exit plan.
- Perform a risk assessment to highlight degrees of data and data process criticality as input to scoping data portability and exit planning.
- Create architectural *standards* on how solutions provide data transfer and *processing* from a cloud provider to an alternative provider or on-premises environment.
- Create, test and gain approval of data portability plans.
- Create, test and gain approval of exit plans for each cloud provider.

### ADVICE FOR DATA PRACTITIONERS

For many organizations, the footprint of data and functionality deployed to cloud environments continues to increase. Typically, the importance of data in the cloud becomes more critical to the organization. Eventually, such an organization will look to transfer data to other *cloud service providers*.

**Contractual provisions for data portability**

Data practitioners should ensure contractual terms and conditions with cloud service and technology providers, including specific rights for the organization to obtain a copy of its data on demand and delete copies of data held by the provider. It is the responsibility of the data practitioner to ensure full removal and deletion of data from the source data location after completion and verification of data transfer *processes*.

**Data architecture considerations**

Data practitioners should enforce architectural *standards* that provide data transfer and *processing* to another provider to facilitate data portability and execution of exit plans. At a minimum, this includes the ability to extract all required data from a provider and, if desired, migrate the data elsewhere. The *standards* may also include measures to enable movement of database and application *processing* components to a new provider rather than rebuilding those components.

The architectural *standards* should also ensure that data portability plans are not cost-prohibitive by requiring plans to incorporate detailed assumptions on volumes and associated costs, be regularly reviewed, and be kept up-to-date. The architecture designs should also provide points of interoperability and easily replicable infrastructure with industry-standard APIs.

The standard should also encourage the use of technology capabilities that will enable data portability. It is important to identify any use of proprietary databases or data *processing* tools that would require reimplementation rather than re-host or port in the event of supplier exit. Consider provider-neutral technologies and services for data transfer instead of depending on provider-specific tools.

In addition, the *standards* should state cases where data should be stored in a common open format to improve portability and address how consistent snapshots of required data can be exported in bulk for transportation to the new provider.

Data practitioners should enforce complete and consistent *data catalog* use. Refer to *CDMC 2.1 Data Catalogs are Implemented, Used, and Interoperable.* Having an accurate inventory of all data and *data flows* across the *data ecosystem* will simplify and mitigate the risk of exit planning and execution. Migrating all relevant *metadata* must also be considered in data migration plans, ensuring accurate and descriptive information is maintained.

**Data portability plan considerations**

Effective data portability plans that ensure data can be relocated should include provisions for data privacy and security to be maintained throughout the transfer process. Establish an assessment process to identify critical business functions, reducing risk and minimizing the business impact of data portability.

Portability plans should also include considerations for data usage and consent requirements in all affected jurisdictions to ensure legal and regulatory compliance (such as FCA FG 16/5 - Guidance for firms outsourcing to cloud[4]).

In addition, plans should specify an approach for transfer in and out of various cloud environments. Specify which toolsets will be available to enable a more efficient data migration. List all the extracted data formats, and indicate if *data transformation* will be necessary before importing data into the target environment.

When practicable, the plans should outline any automations that would increase operational performance and remove the potential for human error.

**Exit plan considerations**

Data practitioners should establish plans to transfer data to another *cloud service provider* or back to an on-premises environment. Effective exit plans should include risk assessments to identify relevant risks and prioritize their mitigation. Risks include legal and regulatory risk, concentration risk, the availability of skills and availability of resources.

An exit plan should also include an up-to-date inventory of services and functionality in use across cloud service and technology providers. There should be tools and capabilities such as cloud discovery technologies necessary to execute the exit plan. Also, document reconciliation *processes* to verify the *accuracy* and *completeness* of data moved to the new provider.

Additional considerations include how well the exit plan aligns with business continuity and disaster recovery plans. In addition, document how the alternate cloud service and technology solutions capabilities will support existing business requirements. Finally, it is important to ensure both the portability and exit plans have been fully tested and validated. The validation involves following formal *procedures* for testing, approval, release, and periodic review, documenting and persisting all test results; and including key *stakeholders* in test result approvals.

### ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Cloud service and technology providers involved in data transfers should offer capabilities that support the various *processes*. Most importantly, the provider should offer systems for transparent bulk data transfers while meeting the organization's security and data protection requirements. The provider must also support data identities and *entitlements* to be exported and imported in bulk to support the migration of access controls between providers. The service contract should stipulate provisions that guarantee visibility into *processes* and methods used in data *destruction*, including a confirmation when data is requested for removal.

In addition, the organization should have access to user interfaces, APIs, protocols, and data formats for cloud services, to reduce the complexity of data portability. There should also be the capability to export *derived data*, such as log *records* or configuration information.

[4] https://www.fca.org.uk/publication/finalised-guidance/fg16-5.pdf

The provider must also support open technologies (open *standards* or open-source) for administrative and business interfaces. Common, open interfaces make it easier to support multiple providers simultaneously. One example is the Cloud Data Management Interface (CDMI) standard.

Lastly, providers should use open standard APIs to ensure broadly interoperable data discovery and consumption across multiple environments. Refer to *CDMC 2.1 Data Catalogs are Implemented, Used, and Interoperable*. This standard is required for structured data, but there is an even greater need for unstructured data to provide transparency of *metadata* elements within the *data catalog* to enable planning for data transfer. This transparency will minimize or eliminate the necessity to rebuild *metadata* between *cloud service providers*.

## QUESTIONS

- Have requirements for data portability and exit planning been documented and approved?
- Has a Risk Assessment been performed to highlight degrees of criticality for data and associated *processes* as input to scoping data portability and exit planning?
- Are there architectural *standards* on how solutions should provide data transfer and *processing* from a cloud provider to an alternative provider or on-premises environment?
- Have data portability plans been created, tested and approved?
- Has an exit plan for each provider been created, tested and approved?

## ARTIFACTS

- Data Portability and Exit Planning Requirements Document – documented requirements, including business impact analysis and business *stakeholder* approval
- Architectural Standards – documented and approved architectural *standards* to support data portability
- Data Portability Plan – documented data portability plan, including associated testing results and appropriate approval(s)
- Exit Plan – documented exit plan, including associated testing results and appropriate approval(s)

## SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal data portability and exit plans exist. | No formal data portability and exit plans exist, but the need is recognized, and the development is being discussed. | Formal data portability and exit plans are being developed. | Formal data portability and exit plans are defined and validated by *stakeholders*. | Formal data portability and exit plans are established and adopted by the organization. | Formal data portability and exit plans are established as part of business-as-usual practice with continuous improvement. |

## 6.2 DATA PROVENANCE AND LINEAGE ARE UNDERSTOOD

The *data lineage* in cloud environments must be captured automatically, and changes to lineage must be tracked and managed. Visualization and reporting of lineage must be implemented to meet the needs of both business and technical users.

### 6.2.1 MULTI-ENVIRONMENT LINEAGE DISCOVERY IS AUTOMATED

#### DESCRIPTION

As cloud data storage becomes more important for more organizations, there is increasing demand for automatic, continuous discovery and detection of *data lineage*. Many cloud environments host very large data volumes, and such environments will benefit from efforts to automate *data lineage* discovery. Automatic *data lineage* discovery employs APIs, specialized software and artificial intelligence to locate *data assets*, identify interdependencies and record *data lineage* automatically. When practicable, automatic *data lineage* discovery should be implemented to operate seamlessly across hybrid and multiple cloud environments.

#### OBJECTIVES

- Implement automated functionality that identifies *processes* that move data.
- Record *data lineage metadata* for data movement *processes* that are discovered automatically.
- Ensure lineage auto-discovery identifies *processes* that move data across jurisdictions, availability zones and physical boundaries.
- Ensure lineage-auto discovery is enabled in hybrid and multiple cloud environments and identifies data movement between those environments.
- Define and implement *processes* for the review of auto-discovered lineage information.

#### ADVICE FOR DATA PRACTITIONERS

To achieve automated lineage discovery, data practitioners should exploit cloud services and third-party tool automation capabilities, wherever possible, to identify data process execution within a cloud environment. Data movement *processes* include ETL, ELT, intentional duplication, data delivery and streaming. The identification should be performed periodically.

Data practitioners should also employ Artificial Intelligence (AI) and Machine Learning (ML) to perform automatic discovery and recording. AI/ML should be used to identify anomalous results in which auto-discovered information may conflict with previously documented lineage—and flag them for review. Existing documentation, previously cataloged *metadata*, cloud environment logs and application logs can be used as sources for automation efforts and detection of anomalous results.

Data practitioners should take a key step to establish an automated quality assessment process to reconcile automatically discovered *data lineage* with the current *metadata* information. Another important step is to provide a written and graphic representation of the automated *data lineage* discovery process results. Practitioners should ensure that recorded lineage *metadata* includes all the facets and dimensions necessary to support the reporting and visualization capabilities when implementing automatic lineage discovery. Refer to *CDMC 6.2.3 Data lineage reporting and visualization are implemented*.

#### ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Cloud service and technology providers should provide organizations with tools and capabilities that enable automated multi-environment lineage discovery. One important capability is the creation of *processes* that automatically discover *data lineage* within the cloud environment. In addition, the provider should offer access for auto-discovery *processes* through APIs to infrastructure log information on data placement and application logs on data movement. Logs should not be hidden or abstracted away from auto-discovery *processes*.

APIs should be available to obtain *metadata* regarding data movements through the cloud environment. Such *metadata* covers movement through data tiers, between availability zones and between geographies. Also, organizations should have an end-to-end view of *data lineage*, typically made possible by stitching or aggregating lineage information from multiple cloud services.

### QUESTIONS

- Does automatic lineage discovery identify *processes* that move data across jurisdictions, availability zones and physical boundaries?
- Does automatic lineage discovery identify data movement between hybrid and multiple cloud environments?
- Have *processes* for reviewing the auto-discovered *data lineage* information been defined and implemented?

### ARTIFACTS

- Artifacts Lineage Discovery Log – demonstrating automated lineage discovery events, lineage review process and review outcome
- Data Catalog Report – demonstrating the recording of lineage information as *metadata*
- Lineage Reports – including data movement across jurisdictions, availability zones and physical boundaries, and movements between hybrid and *multi-cloud* environments

### SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal automated multi-environment discovery of *data lineage* exists. | No formal automated multi-environment discovery of *data lineage* exists, but the need is recognized, and the development is being discussed. | Formal automated multi-environment discovery of *data lineage* is being developed. | Formal automated multi-environment discovery of *data lineage* is defined and validated by *stakeholders*. | Formal automated multi-environment discovery of *data lineage* is established and adopted by the organization. | Formal automated multi-environment discovery of *data lineage* is established as part of business-as-usual practice with continuous improvement. |

## 6.2.2 DATA LINEAGE CHANGES ARE TRACKED AND MANAGED

### DESCRIPTION

The movement of data along a supply change from source to consumption will change as changes to applications, *data assets* and environments are implemented. Changes to the lineage of in-scope data must be tracked and managed for issue investigation, compliance with regulatory requirements and auditing.

### OBJECTIVES

- Gain approval and adopt *data lineage* change management *policy*, *standards* and *procedures* that apply consistently across on-premises and cloud environments.
- Ensure that *data lineage* changes are identified and recorded.
- Record *metadata* that enables historic *data lineage* to be accurately reported.

- Enable changes in *data lineage* to be associated with the underlying business and technology change events.

## ADVICE FOR DATA PRACTITIONERS

Data Practitioners should begin by identifying roles and responsibilities for *data lineage* tracking. The next major step is to document the standardized *data lineage* tracking, version tracking and change management for the organization. It is also important to ensure that the tracking and change management *policy*, *standard* and *procedure* document the balance of responsibilities between the organization and the cloud service and technology providers.

**Validate data elements and ensure data lineage accountability**

Practitioners should define the scope of *data lineage* tracking and accountability within the organization. It is also necessary to define a *data lineage* change management *policy* that establishes *data lineage* accountability on all platforms at appropriate levels of granularity—including on-premises and cloud environments.

Next, practitioners should establish *processes* to monitor *data lineage* changes and track and alert on *data lineage* changes—according to organization *policy* and *data sharing agreements*.

**Automation**

Wherever practicable, employ automation to both record and access *data lineage* changes and versions. Automation can strongly support broader accessibility with a secure URL that is easily distributed to various users. It is easier to scale operations by capturing many relationships and users that need concurrent access lineage information. Automation also enhances the ability to track *data lineage* versions, manage workflows, keep an audit trail, permit concurrent edits from multiple users and prevent the distribution of multiple versions. Ensure that *data lineage* change tracking *procedures* cover events in which manually recorded changes will override automatically recorded lineage.

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Cloud service and technology providers should provide the capability to record *data lineage* changes and *data processing pipeline* changes (such as tagging changes for versioning for *data processing pipeline* releases). It is also important to offer the ability to trigger workflows supporting the organization's *data lineage* change *processes*. Where applicable, providers should offer a change management repository for recording *data lineage* changes. In addition, providers should offer the ability to access *data lineage* *metadata* history to support reporting and visualization of historical lineage for audit purposes.

*Data lineage* and lineage change history should be readily available in common standard formats. Refer to *CDMC Information Model*. Providers should present *data lineage* change management features, interfaces and functionality in clear, accessible documentation.

## QUESTIONS

- Have *policy* and *procedures* for *data lineage* change management been defined?
- Has accountability for *data lineage* change management been established across both cloud and on-premises environments?
- Is *data lineage* change *metadata* identified recorded?
- Is *data lineage* history accurately reportable from recorded *metadata*?
- Are *data lineage* changes linked to underlying business and technology change events?

## ARTIFACTS

- Data Management Policy, Standard and Procedures – defining and operationalizing data lineage change management

- Data Lineage Change Log – recording accountability for *data lineage* changes, recording lineage change *metadata* and linking to business and technology changelogs
- Data Lineage History Reports – generated from recorded *metadata*

### SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal *data lineage* tracking and change management exist. | No formal *data lineage* tracking and *data processing pipeline* change management exist, but the need is recognized, and the development is being discussed. | Formal *data lineage* tracking and *data processing pipeline* change management are being developed. | Formal *data lineage* tracking and *data processing pipeline* change management are defined and validated by *stakeholders*. | Formal *data lineage* tracking and change management is established and adopted by the organization. | Formal *data lineage* tracking and *data processing pipeline* change management are established as part of business-as-usual practice with continuous improvement. |

## 6.2.3 DATA LINEAGE REPORTING AND VISUALIZATION ARE IMPLEMENTED

### DESCRIPTION

In *data lineage* reporting and visualization, *data lineage metadata* is presented in forms that can be analyzed and explored to understand data movement from *data producer* to *data consumer*. Understanding the *data flow* is essential for an organization to assess *data provenance*, perform root cause analysis and impact assessments, validate data integrity and verify *data quality*. In a cloud, *hybrid-cloud*, or multiple cloud environments, it is important that users can know the origin, movement and use of the data that resides in the cloud environment. Visualization of captured *data lineage* data is a critical capability for comprehensive *data management* in a cloud environment.

### OBJECTIVES

- Document and gain approval on *data lineage* reporting and visualization requirements, including requirements for granularity and *metadata* augmentation and labeling.
- Implement functionality to generate lineage visualizations automatically from authoritative sources of lineage *metadata*.
- Provide the ability to augment lineage visualizations with additional *metadata*, such as *data quality* metrics and data ownership.
- Ensure that lineage reports and visualizations provide complete point-in-time histories of key activities.
- Ensure that lineage is represented consistently across different reporting and visualization tools and different lineage discovery methods.
- Gain approval and adopt *data lineage* reporting and visualizations access *policy*, *standards* and *procedures* that apply consistently across on-premises and cloud environments.

## ADVICE FOR DATA PRACTITIONERS

*Data lineage metadata* becomes actionable through data reporting and visualization. Consumers of *data lineage* visualizations include *data owners* and data stewards, who use these reports and visualizations to examine and understand lineage flowing across the boundaries of multiple business units and functions. Consequently, they must be trained and educated to read and interpret data lineage reports and apply them to various business use cases. Data practitioners must ensure that the *data lineage* reports and visualizations are clear, accurate, timely and readily understood.

*Data lineage* reporting and visualization must always present the correct levels of lineage granularity. For summaries, *data lineage* reports and visualizations need to provide visibility into the business systems and the data that interact with those systems before reaching their destination. In detail, the reports and visualizations should provide the details of fields, transformations, historical behavior, and attribute properties for the data on its journey through the *enterprise data ecosystem*. The visualization and reporting capabilities must rely on a comprehensive *data set*, which means collaboration is crucial among the various systems administrators, business groups and department silos.

*Data lineage* reporting and visualization should be *always-on*, employing services and functionality that ensure the availability and recoverability of the reporting systems in case of outages and system faults. Visualization of *data lineage* can help business users spot the connections in *data flows* and thereby provide greater transparency and auditability of the data within the ecosystem.

## ADVICE FOR CLOUD SERVICE AND TECHNOLOGY PROVIDERS

Cloud service and technology providers must provide the features and tools necessary to support *data lineage* reporting and visualizations—across cloud, *hybrid-cloud*, and multiple cloud environments. To achieve this, the provider must support open lineage *data models* and API interfaces to enable an organization to connect and automate lineage data coming from standardized or bespoke sources.

Cloud environment tools and capabilities must enable business users to explore *data lineage metadata* directly. It is important to realize that *data lineage* reports and visualizations are necessary to explore technical and business lineage data. For example, technical lineage shows data movement from a file in a cloud environment to a table in the analytical platform. Data stewards maintaining transformation rules of this data movement also need to know what business *domains* are affected by any upstream changes. The ability to toggle between the technical *data flow* lineage and business impact lineage will enable data stewards to change *data transformation* rules confidently and communicate with affected parties.

Finally, *data lineage* reporting and visualization tools should present historical changes to the data movement *processes*. These tools should offer the ability for the user to recreate the lineage flow back to a specific point in time for audit and *evidence* collection. In addition, the tools should provide the ability to compare versions and highlight changes among versions—which helps evaluate impact to the systems through which the *data flows*.

## QUESTIONS

- Have lineage reporting and visualization requirements been documented and approved?
- Do lineage reporting and visualization requirements include requirements for granularity and *metadata* augmentation and labeling?
- Can lineage visualizations be generated automatically from authoritative sources of lineage *metadata*?
- Can lineage visualizations be augmented with additional *metadata* such as *data quality* metrics and data ownership?
- Can lineage reports and visualizations provide complete point-in-time histories of key activities?
- Is lineage represented consistently across different reporting and visualization tools?
- Is lineage represented consistently regardless of the discovery method used to collect the lineage *metadata*?

- Is access to _data lineage_ reporting and visualizations granted according to defined _policy_ and _procedures_?

### ARTIFACTS

- Lineage Reporting and Visualization Requirements – including requirements for granularity and _metadata_ enrichment
- Lineage Reporting and Visualization Catalog – detailing the granularity and _metadata_ enrichment supported by each solution
- Lineage Reports – including copies of visualizations
- Data Management Policy, Standard and Procedure – defining and operationalizing granting access to _data lineage_ reporting and visualizations

### SCORING

| Not Initiated | Conceptual | Developmental | Defined | Achieved | Enhanced |
|---|---|---|---|---|---|
| No formal lineage reporting and visualization exists. | No formal lineage reporting and visualization standard exists, but the need is recognized, and the development is being discussed. | Formal lineage reporting and visualization are being developed. | Formal lineage reporting and visualization are defined and validated by _stakeholders_. | Formal lineage reporting and visualization are established and adopted by the organization. | The formal lineage reporting and visualization is established as part of business-as-usual practice with continuous improvement. |

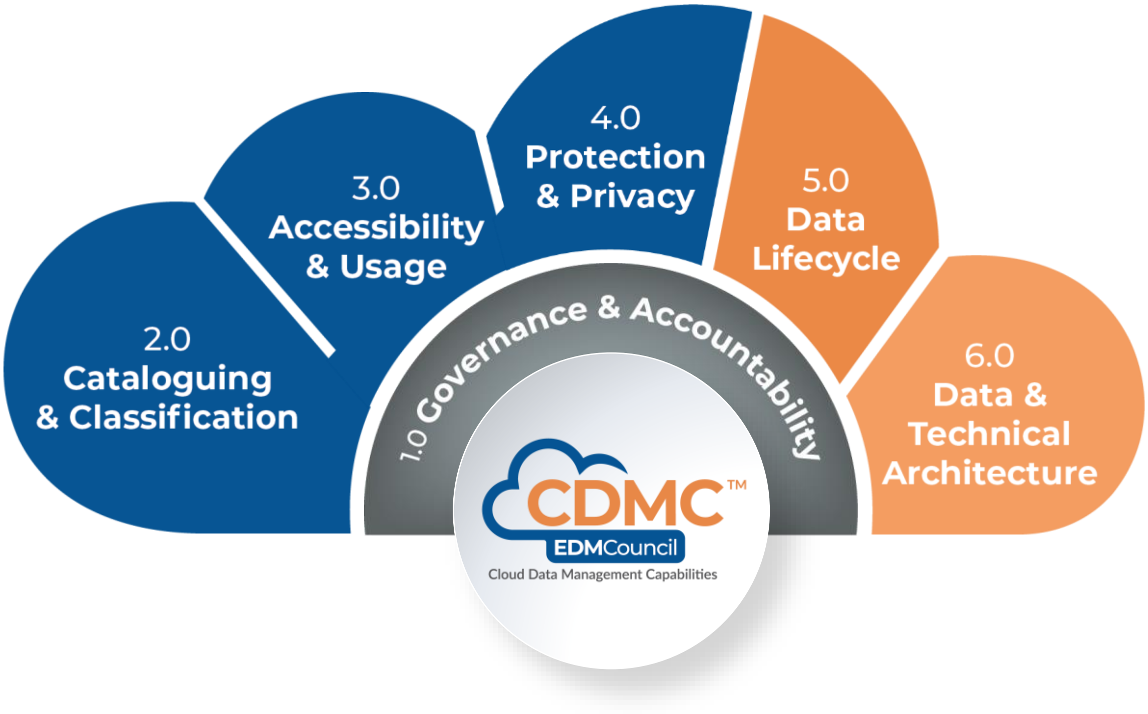
## 6.3 PROTECTION & PRIVACY – KEY CONTROLS

The following Key Controls align with the capabilities in the Data & Technical Architecture component:

- Control 13 – Data Lineage
- Control 14 – Cost Metrics

Each control with associated opportunities for automation is described in *CDMC 7.0 – Key Controls & Automations.*

# 7.0 Key Controls & Automations



# 7.0 CDMC KEY CONTROLS & AUTOMATIONS

## SCOPE OF CONTROLS

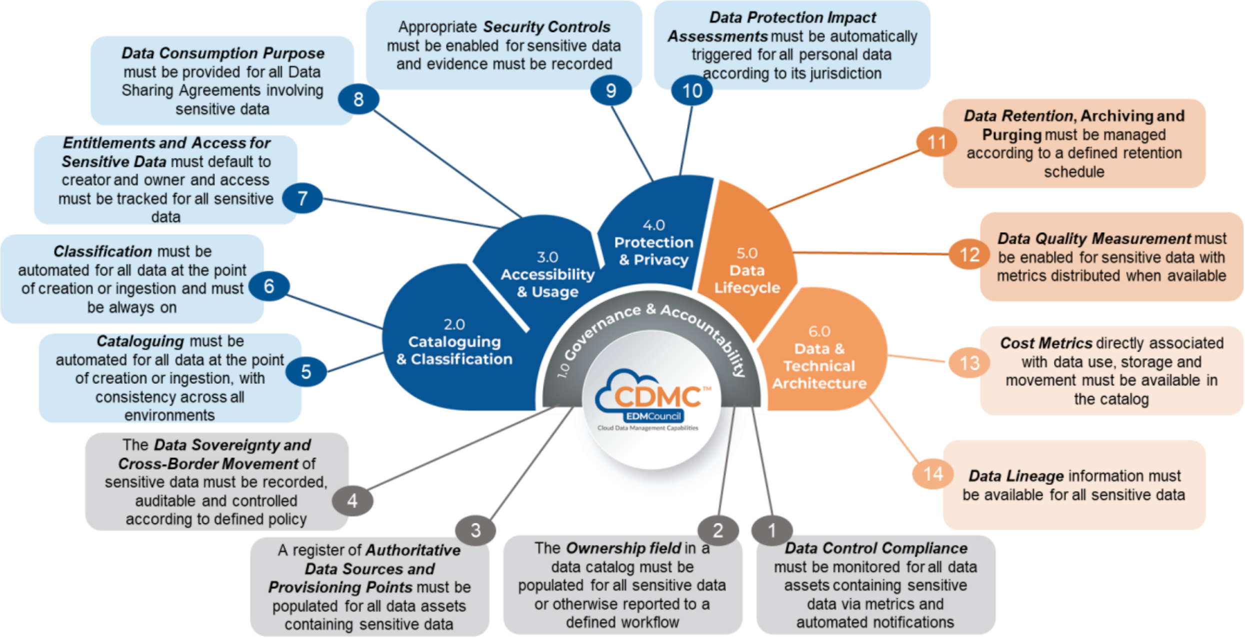
The framework addresses the control of data in cloud, *multi-cloud* and *hybrid-cloud* environments. Controls that address technology risks in other areas such as software development and service management are not within the scope of the document.

Many of the controls refer to applying to ***sensitive data***. Each organization will have a scheme for classifying their sensitive and important data and will determine the specific *classifications* to which the controls must be applied. Examples of *classifications* that may be in scope include:

- Personal Information (PI) / *Sensitive Personal data*
- *Personally Identifiable Information* (*PII*)
- Client Identifiable Information
- Material Non-Public Information (MNPI)
- Specific *information sensitivity* *classifications* (such as 'Highly Restricted' and 'Confidential')
- *Critical Data Elements* used for important business *processes*[5] (including regulatory reporting)
- Licensed data

## KEY CONTROLS SUMMARY

The key controls are summarized in the following diagram:


1.0 Governance & Accountability
2.0 Cataloguing & Classification
3.0 Accessibility & Usage
4.0 Protection & Privacy
5.0 Data Lifecycle
6.0 Data & Technical Architecture
CDMC™
EDMCouncil
Cloud Data Management Capabilities

1 ***Data Control Compliance*** must be monitored for all data assets containing sensitive data via metrics and automated notifications

2 The ***Ownership field*** in a data catalog must be populated for all sensitive data or otherwise reported to a defined workflow

3 A register of ***Authoritative Data Sources and Provisioning Points*** must be populated for all data assets containing sensitive data

4 The ***Data Sovereignty and Cross-Border Movement*** of sensitive data must be recorded, auditable and controlled according to defined policy

5 ***Cataloguing*** must be automated for all data at the point of creation or ingestion, with consistency across all environments

6 ***Classification*** must be automated for all data at the point of creation or ingestion and must be always on

7 ***Entitlements and Access for Sensitive Data*** must default to creator and owner and access must be tracked for all sensitive data

8 ***Data Consumption Purpose*** must be provided for all Data Sharing Agreements involving sensitive data

9 Appropriate ***Security Controls*** must be enabled for sensitive data and evidence must be recorded

10 ***Data Protection Impact Assessments*** must be automatically triggered for all personal data according to its jurisdiction

11 ***Data Retention*, Archiving and Purging** must be managed according to a defined retention schedule

12 ***Data Quality Measurement*** must be enabled for sensitive data with metrics distributed when available

13 ***Cost Metrics*** directly associated with data use, storage and movement must be available in the catalog

14 ***Data Lineage*** information must be available for all sensitive data

Detail for each control is provided in the sections below.

[5] Important business *processes* and the threshold to be considered important are dependent on the maturity of the organization's data management and the extent of its data strategy.

## CONTROL 1: DATA CONTROL COMPLIANCE

| **Component** | **1.0 Governance & Accountability** |
|---|---|
| **Capability** | **1.1 Cloud Data Management Business Cases are Defined and Governed** |
| **Control Description** | **Data Control Compliance** must be monitored for all *data assets* containing sensitive data via metrics and automated notifications. The metrics must be calculated from the extent of implementation of the CDMC Key Controls specified in subsequent sections. |
| **Risks Addressed** | An organization does not set or achieve its value and risk mitigation goals for cloud management. Data is uncontrolled and consequently is at risk of not being fit-for-purpose, late, missing, corrupted, leaked and in contravention of data sharing and retention legislation. |
| **Drivers / Requirements** | Organizations are required to demonstrate adequate control of data being created in or migrated to the cloud. |
| **Legacy / On-Premises Challenges** | Significant tranches of on-premises data do not have *data management* applied to them and consequently do not realize maximum value for the organization or can potentially pose an unquantified risk.<br>When moving data to a new cloud environment, it is critical that organizations actively assess and apply the appropriate levels of *data management* to achieve their stated outcomes, apply controls to achieve this and measure compliance and value realization with those outcomes. |
| **Automation Opportunities** | • Where *evidence* of the existence of controls can be gathered automatically (including those controls referenced in subsequent sections of this document), the **Data Control Compliance metrics** may be calculated automatically.<br>• Where the metrics fall below specified thresholds, alerts should be generated with automated notification to specified *stakeholders*. |
| **Benefits** | Cloud data is demonstrably controlled and supports the Cloud *Data management* business cases and risk mitigation requirements of the organization. |
| **Summary** | Organizations can demonstrate an awareness of the intended outcomes of cloud *data management* and focus on quantifiable value realization and risk mitigation. |

### CONTROL 2: OWNERSHIP FIELD

| | |
|---|---|
| **Component** | **1.0 Governance & Accountability** |
| **Capability** | **1.2 data ownership is Established for both Migrated and Cloud-generated Data** |
| **Control Description** | The **Ownership field** in a *data catalog* must be populated for all sensitive data or otherwise reported to a defined workflow. |
| **Risks Addressed** | Accountability for decisions on and control of sensitive data is not defined. Sensitive data is not effectively owned and consequently is at risk of not being fit for purpose, late, missing, corrupted, leaked and in contravention of data sharing and retention legislation. |
| **Drivers / Requirements** | Organizations have *policies* that require explicit ownership of data that is classified as sensitive. |
| **Legacy / On-Premises Challenges** | Significant amounts of legacy data do not have ownership recorded. |
| **Automation Opportunities** | The **Ownership field** in a *data catalog* must be populated “eventually” for sensitive data that is migrated to or generated within the cloud.<br>• Automatically trigger workflows to enforce population when new *data assets* are created.<br>• Provide the capability to automate workflows to review and update ownership periodically for sensitive data or when an owner leaves the organization or moves within the organization<br>• Automatically trigger escalation workflows to address population gaps.<br>• Implement ownership recommendations driven by the nature of data and ownership of similar data. |
| **Benefits** | Increased compliance with data ownership *policy*. |
| **Summary** | Infrastructure that supports the completion of data ownership information for sensitive data drives *policy* compliance. |

## CONTROL 3: AUTHORITATIVE DATA SOURCES AND PROVISIONING POINTS

| | |
|---|---|
| **Component** | **1.0 Governance & Accountability** |
| **Capability** | **1.3 Data Sourcing and Consumption are Governed and Supported by Automation** |
| **Control Description** | A register of **Authoritative Data Sources and Provisioning Points** must be populated for all *data assets* containing sensitive data or otherwise must be reported to a defined workflow. |
| **Risks Addressed** | Architectural strategy for an organization is not fully defined. Authorized sources have not been defined or suitably controlled.<br>Data is duplicative and/or contradictory, resulting in process breaks, architectural inefficiencies, increased cost of ownership and accentuating existing operational risks on all dependent business *processes*. |
| **Drivers / Requirements** | An important responsibility of a *data owner* is to designate the *authoritative data sources* and *provisioning points* of data for a specific scope of data.<br>*Policy* controls require a *data asset* to be identified as authoritative or not when it is shared. |
| **Legacy / On-Premises Challenges** | Identification and remediation of the use of non-authoritative sources or copies of data require significant manual effort. |
| **Automation Opportunities** | • Automatically enforce the labeling of sources of data as authoritative or non-authoritative.<br>• Control the consumption of sensitive data from sources that are non-authoritative.<br>• Default the labeling of sources to non-authoritative until reviewed and updated by the *data owner*. |
| **Benefits** | Infrastructure that can run automated workflows to identify and retire non-authoritative data provides a cost savings opportunity to eliminate the manual effort involved in this work. |
| **Summary** | *Data assets* automatically tagged as authoritative or non-authoritative will greatly simplify *policy* compliance and eliminate manual costs of controlling data sourcing and consumption. |

## CONTROL 4: DATA SOVEREIGNTY AND CROSS-BORDER MOVEMENT

| | |
|---|---|
| **Component** | **1.0 Governance & Accountability** |
| **Capability** | **1.4 Data Sovereignty and Cross-Border Data Movement are Managed** |
| **Control Description** | The **Data Sovereignty and Cross-Border Movement** of sensitive data must be recorded, auditable and controlled according to defined *policy*. |
| **Risks Addressed** | Data can be stored, accessed and processed across multiple physical locations in cloud environments, increasing the risk of breaches to jurisdictional laws, security and privacy rules, or regulation.<br>Breaches can result in various penalties, including fines, reputational damage, legal action and removal of licenses. |
| **Drivers / Requirements** | The *data owner* should understand the jurisdictional implications of cross border data movement and any region-specific storage and usage rules for a particular *data set*. *Policy*-specified controls must be applied when establishing cross-border *data sharing agreements* to support requests to use data from a particular location. |
| **Legacy / On-Premises Challenges** | Maintaining data about the physical location of data stores and *processes* is a significant undertaking and applying rules consistently across multiple different technologies is prohibitive. |
| **Automation Opportunities** | • Automatically capture and expose the physical location of all storage, usage and *processing* infrastructure applying to a cataloged *data set*<br>• Provide the ability to trigger cross border *data sharing agreement* workflows (for international data transfer and international data requests).<br>• Automatically trigger regional storage, *processing* and usage constraints, with the ability to escalate to a *data owner* where required.<br>• Automatically audit and allow workflow to be triggered when sensitive data is being accessed from a location without a *data sharing agreement*. |
| **Benefits** | Reducing the manual *processing* and audit of *data sharing agreements* will significantly reduce the cost and risk of data *processing* in the cloud. |
| **Summary** | Codifying and automatically applying jurisdictional *data management* rules and cross border sharing agreements will significantly reduce the risk of *processing* data in the cloud. This will increase the adoption of cloud services and reduce complexity in the day-to-day *processing* of data in the cloud. |

## CONTROL 5: CATALOGING

| | |
|---|---|
| **Component** | **2.0 Cataloging & Classification** |
| **Capability** | **2.1 Data Catalogs are Implemented, Used, and Interoperable** |
| **Control Description** | **Cataloging** must be automated for all data at the point of creation or ingestion, with consistency across all environments. |
| **Risks Addressed** | The existence, type and context of data are not identified, resulting in the inability of all other controls to be applied that are dependent on the data scope.<br>Data is uncontrolled and consequently is at risk of not being fit for purpose, late, missing, corrupted, leaked and in contravention of data sharing and retention legislation. |
| **Drivers / Requirements** | Organizations must ensure the necessary controls are in place for large or complex workloads that involve sensitive data such as client identifiers and transactional details.<br>Knowledge of all data that exists is foundational to ensuring that all sensitive data has been identified. |
| **Legacy / On-Premises Challenges** | Organizations cannot scan and catalog the significant variety of *data assets* that exist in legacy on-premises environments. Without comprehensive catalogs of all existing data, organizations cannot be confident that all sensitive data within their *data assets* have been identified. |
| **Automation Opportunities** | • Ensure that catalog entries are generated for all data migrated to or created in the cloud.<br>• Ensure catalog entries are generated for data in development, test and production environments and for both online and archived data.<br>• Generate *evidence* of the comprehensiveness of the *data catalog*.<br>• Implement APIs and support open data *standards* for *metadata* sharing and catalog interoperability. (Refer to the *CDMC Information Model*). |
| **Benefits** | An organization can guarantee that all data has been cataloged and can use this as the foundation on which to automate and enforce controls based on the *metadata* in the catalog. |
| **Summary** | This is the infrastructure describing what data exists, to see how much there is and how many different types there are. It is the foundation of all the other controls. |

## CONTROL 6: CLASSIFICATION

| | |
|---|---|
| **Component** | **2.0 Cataloging & Classification** |
| **Capability** | **2.2 Data Classifications are Defined and Used** |
| **Control Description** | ***Classification*** must be automated for all data at the point of creation or ingestion and must be always on.<br>• *Personally Identifiable Information* auto-discovery<br>• *information sensitivity classification* auto-discovery<br>• Material Non-Public Information (MNPI) auto-discovery<br>• Client identifiable information auto-discovery<br>• Organization-defined *classification* auto-discovery |
| **Risks Addressed** | Sensitive data is not classified, resulting in the inability of all other controls to be applied that are dependent on the *classification*.<br>Data is uncontrolled and consequently is at risk of not being fit for purpose, late, missing, corrupted, leaked and in contravention of data sharing and retention legislation. |
| **Drivers / Requirements** | *Information sensitivity classification* (ISC) is required by most organizations' information security *policies*. An organization is required to know whether data is highly restricted (HR), classified (C), internal use only (IUO), or public (P), and if it is sensitive.<br>Knowing whether data is sensitive is the foundation of most other controls in the framework. This requires certainty that all data has been cataloged and certainty that the sensitivity of the data has been determined. |
| **Legacy / On-Premises Challenges** | The variety of *data assets* in legacy environments impacts the ability to ensure that all data has been identified. Sensitive data may exist in *data assets* that have not been identified.<br>*Classification* of *data assets* is often manual and can be both error-prone and expensive. Even where assets are identified, there may be gaps or errors in the *classification*.<br>The proliferation of copies of data in legacy environments can lead to *classifications* in data sources not being carried through to copies of the data. |
| **Automation Opportunities** | • Apply *classification processing* to all data migrated to or created in the cloud.<br>• Use automated *data classification* to identify the *classification* that applies.<br>• Support organization-specified *classification* schemes.<br>• Default *classifications* to the highest level until explicitly reviewed and changed. |
| **Benefits** | The operations team that is responsible for classifying data is expensive. Auto-*classification* can significantly streamline and reduce the amount of manual effort required to perform this function. |
| **Summary** | Auto-*classification* of data provides confidence that all sensitive data has been identified and can be controlled. |

### CONTROL 7: ENTITLEMENTS AND ACCESS FOR SENSITIVE DATA

| | |
|---|---|
| **Component** | **3.0 Accessibility & Usage** |
| **Capability** | **3.1 Data Entitlements are Managed, Enforced, and Tracked** |
| **Control Description** | 1. **Entitlements and Access for Sensitive Data** must default to creator and owner until explicitly and authoritatively granted.<br>2. Access must be tracked for all sensitive data. |
| **Risks Addressed** | Access to data is not sufficiently controlled to those who should be authorized. This could result in data leakage, reputational damage, regulatory censure, criminal manipulation of business *processes*, or data corruption.<br>Data is uncontrolled and consequently is at risk of not being fit for purpose, late, missing, corrupted, leaked and in contravention of data sharing and retention legislation. |
| **Drivers / Requirements** | Once the auto-classifier has identified sensitive *data assets*, enhanced controls should be placed on those *data assets*, including how *entitlements* are granted.<br>The users that have access to data and how frequently they access it needs to be tracked. |
| **Legacy / On-Premises Challenges** | It is difficult to track which *data consumers* are using which *data assets* unless tracking is turned on and is consistent across all the data in the catalog. |
| **Automation Opportunities** | • Automate the defaulting of *entitlements* to restrict access to the creator and owner until explicitly and authoritatively granted to others<br>• Automatically track which users have access to which data and how frequently they access it and store that information in a *data catalog*.<br>• Provide all *data owners* access to the usage tracking tool<br>• Hold *entitlements* as *metadata* to enable their use by any tool used to access the data. |
| **Benefits** | Tracking of data consumption enables consumption-based allocation of costs. Automation can reduce the cost of performing these allocations manually. |
| **Summary** | **Entitlements and access for sensitive data** at a minimum should be automated to default to being restricted to just the creator and owner of the data until they grant permissions to other people. Once other people have access to that data, monitoring should be in place to track who is using it and how frequently they are accessing it. Costs can then be correctly allocated. |

## CONTROL 8: DATA CONSUMPTION PURPOSE

| | |
|---|---|
| **Component** | **3.0 Accessibility & Usage** |
| **Capability** | **3.2 Ethical Access, Use, & Outcomes of Data are Managed** |
| **Control Description** | **Data Consumption Purpose** must be provided for all *data sharing agreements* involving sensitive data. The purpose must specify the type of data required and include country or legal entity scope for complex international organizations. |
| **Risks Addressed** | Data is shared or used in an uncontrolled manner with the result that the producer is not aware of how it is being used and cannot ensure it is fit for the intended purpose.<br>Data is not shared in compliance with the ethical, legislative, regulatory and *policy* framework where the organization operates. |
| **Drivers / Requirements** | There are emerging ethical-use frameworks and *guidelines* that include specifications for what should happen when the use of data changes. |
| **Legacy / On-Premises Challenges** | It is difficult for human capabilities to recognize when the use of data has changed into a new kind of *processing* that could be protected under some regulatory or legal basis without specific authorization. |
| **Automation Opportunities** | • Record data access tracking information for sensitive data.<br>• Enforce the capture of purpose, for example, integrated with *model* governance.<br>• Provide alerts to the *data owner* or data governance teams when there is an additional use case for existing user access to sensitive data.<br>• Recognize when specific technologies are employed (e.g., Machine Learning) and leverage usage and cost tracking to highlight potential new use cases. |
| **Benefits** | Streamlined ethical data accountability for data that is accessed for new purposes. |
| **Summary** | A *data sharing agreement* between a conumer and the authoritative source expresses the intent to use the data for a specific purpose. Automated tracking and monitoring of **data consumption purpose** can alert *data owners* and data governance teams when there is new or changed use. |

CONTROL 9: SECURITY CONTROLS

| | |
|---|---|
| **Component** | **4.0 Protection & Privacy** |
| **Capability** | **4.1 Data is Secured, and Controls are Evidenced** |
| **Control Description** | 1. Appropriate **Security Controls** must be enabled for sensitive data.<br>2. Security control *evidence* must be recorded in the *data catalog* for all sensitive data. |
| **Risks Addressed** | Data is not contained within the parameters determined by the legislative, regulatory or *policy* framework where the organization operates. Data loss or breaches of privacy requirements resulting in reputational damage, regulatory fines and legal action. |
| **Drivers / Requirements** | The sensitivity level of the data dictates what level of *encryption*, obfuscation and data loss prevention should be enforced. The requirements for **Security Controls** and Data Loss Prevention become increasingly more stringent as the sensitivity level of the data increases. |
| **Legacy / On-Premises Challenges** | It is difficult to ensure that *encryption* is always on for sensitive data. |
| **Automation Opportunities** | • Provide **security controls** capabilities including *encryption*, masking, obfuscation and *tokenization* that are turned on automatically based on the sensitivity of a *data set*.<br>• Automate recording of the application of security controls. |
| **Benefits** | *Evidence* that the appropriate level of *encryption* is on and has been consistently applied is easy to produce.<br>During a security audit, a *data owner* has a list of their data and how much of it is sensitive. Every piece of sensitive data can provide *evidence* that the data is encrypted, and there is a data loss prevention regime in place for all the compute environments it resides.<br>Having security control *evidence* to deliver through the catalog rather than performing a forensic cyber review is a cost savings opportunity. A full-time team of employees typically handles this work. |
| **Summary** | Automation that enforces and records the appropriate *encryption* level based on a data asset's sensitivity level ensures security compliance and reduces manual effort to provide *evidence* of the controls. |

## CONTROL 10: DATA PROTECTION IMPACT ASSESSMENTS

| | |
|---|---|
| **Component** | **4.0 Protection & Privacy** |
| **Capability** | **4.2 A Data Privacy Framework is Defined and Operational** |
| **Control Description** | ***Data Protection Impact Assessments*** (DPIAs) must be automatically triggered for all *personal data* according to its jurisdiction. |
| **Risks Addressed** | Data is not secured to an appropriate level for the nature and content of that *data set*. This results in either data being secured at greater cost and inconvenience than required or data loss or breaches of privacy requirements resulting in reputational damage, regulatory fines and legal action. |
| **Drivers / Requirements** | If a *data set* is classified as containing personal information, an organization needs to be able to demonstrate that it has performed a *data protection impact assessment* on it in certain jurisdictions. |
| **Legacy / On-Premises Challenges** | It is a very expensive workflow to initiate and complete a *data protection impact assessment* for the *data assets* classified as containing personal information.<br>Identifying the DPIAs that need to be performed can be challenging, and completing those DPIAs can be very expensive. |
| **Automation Opportunities** | • Automatically initiate ***Data Protection Impact Assessments*** based on factors such as the geography of the data infrastructure, *classification* of the data or the specified consumption purpose. |
| **Benefits** | *Evidence* that all privacy requirements have been met for sensitive data is easy to produce since DPIAs are automatically initiated.<br>Cost savings opportunities arise from more efficient identification of the need for DPIAs. |
| **Summary** | Automatically enforcing a DPIA on data that is classified as personal ensures *policy* compliance and reduces manual labor costs for that function. |

## CONTROL 11: DATA RETENTION, ARCHIVING AND PURGING

| | |
|---|---|
| **Component** | **5.0 Data Lifecycle** |
| **Capability** | **5.1 The Data Lifecycle is Planned and Managed** |
| **Control Description** | **Data Retention, Archiving, and Purging** must be managed according to a defined retention schedule. |
| **Risks Addressed** | Data is not removed in line with the legislative, regulatory or *policy* requirements of the organization's environment, leading to increased cost of storage, reputational damage, regulatory fines, and legal action. |
| **Drivers / Requirements** | Organizations have a master retention schedule that determines how long data needs to be retained in each jurisdiction it was created  based on its *classification*. |
| **Legacy / On-Premises Challenges** | Organizations will have huge repositories of historical data, often retained to support the requirements of potential future audits. *Data sets* in different jurisdictions will have different retention schedules. It is difficult to comply with these requirements manually since different applicable legal requirements can modify the retention schedule. |
| **Automation Opportunities** | • Automate **data retention, archiving and purging** *processing* based on the data's jurisdiction, purpose and *classification* and according to a defined retention schedule.<br>• Collect and provide *evidence* of the data retention, archiving and purging plan and execution. |
| **Benefits** | Automatically retaining, archiving, or purging data based on its *classification* and association retention schedule will reduce the manual effort required to perform this function and ensure *policy* compliance. |
| **Summary** | Organizations with this automation and control can provide the necessary *evidence* to verify that their data is being retained, archived or *purged* based on the retention schedule of its *classification*. |

CONTROL 12: DATA QUALITY MEASUREMENT

| | |
|---|---|
| **Component** | **5.0 Data Lifecycle** |
| **Capability** | **5.2 Data Quality is Managed** |
| **Control Description** | **Data Quality Measurement** must be enabled for sensitive data with metrics distributed when available. |
| **Risks Addressed** | Data is not consistently fit for the organization's purposes, resulting in the inability to provide expected customer service, process breaks, the inability to demonstrate risk management, inefficiencies, and a lack of trust in the data and decisions based on flawed information. |
| **Drivers / Requirements** | *Data quality* metrics will enable *data owners* and *data consumers* to determine if data is fit-for-purpose. That information needs to be visible to both owners and *data consumers*. |
| **Legacy / On-Premises Challenges** | The limited application of *data quality* management in many legacy environments results in a lack of transparency on the quality of data and an inability for *data consumers* to determine if its fit-for-purpose. *Data owners* may not be aware of *data quality* issues. |
| **Automation Opportunities** | • Automatically deliver *data quality* metrics to *data owners* and *data consumers*.<br>• Make *data quality* metrics available in the *data catalog*.<br>• Automatically alert *data owners* to *data quality* issues. |
| **Benefits** | *Data consumers* can determine if data is fit-for-purpose. *Data owners* are aware of *data quality* issues and can drive their prioritization and remediation. |
| **Summary** | Providing clarity on *data quality* and support to ensure data is fit-for-purpose will help *data owners* address *data quality* issues. |

## CONTROL 13: COST METRICS

| | |
|---|---|
| **Component** | **6.0 Data & Technical Architecture** |
| **Capability** | **6.1 Technical Design Principles are Established and Applied** |
| **Control Description** | **Cost Metrics** directly associated with data use, storage, and movement must be available in the catalog. |
| **Risks Addressed** | Costs are not managed, detrimentally impacting the commercial viability of the organization. |
| **Drivers / Requirements** | As the cloud changes the cost paradigm from Capex to Opex, organizations require additional visibility on where data movement, storage and usage costs are incurred.<br>Poor data architectural choices concerning data placement can incur additional costs through ingress or egress costs. For example, extra compute costs will be incurred when running data warehouse workloads on OLTP infrastructure. |
| **Legacy / On-Premises Challenges** | Limited need to manage data *processing* or storage costs at a *data asset* level.<br>There is no line-item costing on the assets in a *data catalog*, so organizations cannot run a cost-analysis to understand where their *data management* costs are specifically being incurred. |
| **Automation Opportunities** | • Automatically track data assets' movement, storage, and usage costs and make this information available via the *data catalog*.<br>• Support automated *policy*-driven cost management and optimization of data *processing*. |
| **Benefits** | *Data owners* would be able to understand who is using what data, the frequency of that access and the cost incurred to provide that data. |
| **Summary** | The financial operations infrastructure of *cloud service providers* is robust enough to identify accounts and operations that are incurring costs and associating those costs to specific *data assets* as line items in the *data catalog*. |

CONTROL 14: DATA LINEAGE

| | |
|---|---|
| **Component** | **6.0 Data & Technical Architecture** |
| **Capability** | **6.2 Data Provenance and Lineage are Understood** |
| **Control Description** | ***Data lineage*** information must be available for all sensitive data. This must at a minimum include the source from which the data was ingested or in which it was created in a cloud environment. |
| **Risks Addressed** | Data cannot be determined as having originated from an authoritative source resulting in a lack of trust of the data, inability to meet regulatory requirements, and inefficiencies in the organization's system architecture. |
| **Drivers / Requirements** | Organizations need to trust data being used and confirm that it is being sourced in a controlled manner.<br>Regulated organizations produce lineage information as *evidence* that the information on regulatory reports has been taken from an authoritative source for that type of data.<br>Consumers of sensitive data must be able to *evidence* sourcing of data from an authoritative source, for example, by showing lineage from the authoritative source or providing the provenance of the data from a supplier. |
| **Legacy / On-Premises Challenges** | Lineage information is produced manually by tracing the flow of data through systems from source to consumption. The cost of this approach and the consequences of producing incorrect data can be significant. |
| **Automation Opportunities** | • Record ingestion source of all data of specific *classifications* migrated to the cloud.<br>• Record source-to-target lineage of all movement of data of specific *classifications* within the cloud environment.<br>• Record destination lineage of all data of specific *classifications* egressing from the cloud (whether to on-premises or another cloud). |
| **Benefits** | Easy to produce *evidence* of the ***data lineage*** for regulatory reports. Major financial organizations incur significant costs producing this information manually and retrospectively. |
| **Summary** | Automatically tracking lineage information for data that feed regulatory reports would streamline the reports' data and eliminate cost by replacing the manual labor required to produce that information. |