AWS Whitepaper

Enterprise Data Governance Catalog

Enterprise Data Governance Catalog: AWS Whitepaper

Table of Contents

Enterprise Data Governance Catalog

Publication date: **December 3, 2021** (*Document history*)

This whitepaper outlines the benefits and strategies for implementing an enterprise-wide unified Data Governance Platform to enable business users and stakeholder with the ability to find, manage, understand, access, and trust their data to make better data-driven business decisions.

This whitepaper is for technical and business leaders who are responsible for managing data and analytics platform.

Introduction

Business and data users want the capability to analyze the data scattered across various data assets within their organizations. Data assets are stored across various databases, file systems, servers located on-premises and in the cloud (including data warehouses), data lakes, and big data.

However, many data assets are hidden deep inside data silos without much clarity into the datasets, the classifications associated within the datasets, and their business relationships. Vast amounts of data are created, captured, and consumed by organizations, which further increase the complexities of finding and understanding data assets. Identifying relevant datasets, profiling, and combines the related data to get meaningful technical and business insights is tedious.

Organizations face numerous challenges to analyze data spread across various data assets within their organization to get business insights and drive business decisions related to growth, adoption, and investments. This is a challenge due to the lack of a data-first paradigm, where data is the driver to make key business growth decisions within the organization. There is a lack of understanding the business value of data as a product, and technical design gaps are introduced while managing data.

Data Catalogs have evolved from a promising to essential framework which supports organizations data and analytics. In 2017, Gartner declared Data Catalogs as “the new black in data management and analytics”, and now they are recognized as a central technology for data management. According to International Data Corporation (IDC), four out of five (80%) of the organizations take advantage of data across multiple organizational processes. However, despite increases in innovation, some studies show that workers waste 44% of their time each week struggling with data due to a lack of collaboration, knowledge gaps, and organizational resistance to change.

This whitepaper outlines key considerations to build a Data Catalog, and provides an approach to implement data governance through a Data Catalog using Amazon Web Services (AWS) Cloud technologies. It showcases how a robust Data Catalog empowers data users to explore hidden data insights effectively, while driving their organizations' growth by making data-driven business decisions.

This whitepaper also provides a high-level approach to managing metadata (the data providing information about one or more aspects of the data). Metadata can be used to classify, organize, and access data assets to provide deep technical and business insights. Business insights are essential for organizations to make better business decisions, achieve operational efficiency, and improve data understanding and data quality.

Data governance

Data governance is the process of managing the availability, usability, integrity, and security of the data in enterprise systems, based on the internal data standards and policies that control data usage. Effective data governance ensures that data is consistent and trustworthy. The data governance process enables organizations to ensure that high-quality data exists during the lifecycle of the data. Data governance implements data access rules and policies to improve data security.

Data governance also manages the data lifecycle, which is a sequence of stages that a particular unit of data goes through from its initial generation or capture to its eventual archival or deletion. Traditionally, organizations focused on management of the data scattered around various data assets to meet tactical goals, with less emphasis on strategic business needs. Now organizations are starting to recognize the benefits of well-organized and well-classified data, to get a profound visibility on data as a strategic asset.

Data privacy compliances like the California Consumer Privacy Act (CCPA) and the General Data Protection Regulation (GDPR) require that the organizations put in place appropriate measures to protect and manage customer data, because customers have the right to know if their personal data is being stored, sold, or disclosed.

The GDPR's primary aim is to give individuals control over their personal data and to simplify the regulatory environment for international business by unifying the regulation within the EU. The CCPA states that a consumer should know if their data is being stored by an organization, and consumers can request to delete their personal information. Without a well-defined data governance framework in place, it is challenging for organizations to know all the places it stores customer information to comply with CCPA or other consumer privacy acts. Violations may lead to legal implications and fines or penalties.

A well-defined Data Catalog makes it easier to identify customer data distributed across various data assets, as a Data Catalog tags data and builds relationships between data attributes, enabling the organization to adhere to existing and future data regulatory compliances.

An organization’s multiple business units collect customer communication channel preferences by email, phone call, or text message. The Data Catalog collects metadata associated with all the data stores which handle customer communication preferences. Using the Data Catalog, the organization’s business units can combine all the different rules for customer communication and interact effectively with customers, elevating satisfaction and trust.

Data governance catalog

Realizing the benefits of managing data assets to get deep business and technical insights, organizations are looking for a framework to implement data governance.

A data governance catalog revolves around metadata. Primarily metadata is categorized as *technical* and *business*.

- **Technical metadata** is information such as author, date created, date last modified, source, and size of a dataset.
- **Business metadata** further enriches technical metadata by adding additional details for data classification, structure, data taxonomy, retention period, and other details for a dataset.

Data and business have become inseparable. Absence of data-driven decision-making limits an organization's ability to use their data to its best potential. This results in business decisions being made around growth, investments, and other verticals based on assumptions and personal preferences rather than real data statistics. Business decisions such as how an organization ships goods, interacts with customers, or sends product offerings backed by data empower organizations to run business efficiently.

A data-first approach is essential for a data governance catalog's success. Data stewardship is an important step in aligning data and business processes together. *Data stewards* are product managers, data subject matter experts (SMEs), and business owners, along with data architects and analysts. They are responsible for interpreting collected metadata to derive deep business insights, and promoting a culture of data-driven business decision across the organization.

A Data Catalog aligns people, processes, and technology, helping data users understand and transform data into a business asset. It delivers good visibility into datasets, allowing organizations to comply with global data privacy laws.

A Data Catalog allows an organization to identify data owners and improve data quality, regulatory compliance, and data usage. It enables organizations to orchestrate workflows to incorporate changes to the metadata.

Data ownership ensures that someone in the organization is responsible for the data origin, definition, business attributes, relationships, and dependencies. There are various owners for different business units, such as marketing, supply chain, and finance. Business unit collaboration improves business actions such as launching products on time and on budget, interactions with

consumers, and making new sales and distribution channels easy to build. For example, with improved collaboration, a marketing business unit can consolidate the customer data in the organization and create new marketing campaigns effectively, targeting the right audience.

The following diagram depicts how a data governance catalog can create a relationship between various business and non-business assets across an organization to drive business growth.

Business relationships between organization assets

The Data Catalog creates mature data governance processes and adds value to the organization across several dimensions, including data-driven business decisions. The Data Catalog provides measures and metrics around datasets to guide strategic business decisions that align with the organization's objectives and initiatives.

In organizations where a Data Catalog is not implemented, data is often left fragmented and siloed across numerous sources (such as legacy systems, data warehouses, flat files stored on individual desktops, and modern, cloud-based repositories). Business stakeholders, data analysts, and other users spend too much time trying to discover data due to a lack of easy access and fragmented data environments.

A Data Catalog helps implement data classification, encryption, data masking, and access protection to manage an organization's data securely. Data classification enables data stewards to set up guidelines around the storage and access of classified data, such as customers' Social Security Numbers (SSNs). SSNs are classified as sensitive data, and handled accordingly by the data processing and consumption pipelines.

The preceding diagram shows business relationships between organization assets, enabling business users to understand their data's origins and where it travels over time, without having to understand the underlying technical complexity. It identifies business, data, security and technology assets relationships and is implemented and managed within the Data Catalog.

Data Catalog benefits for business stakeholders

According to International Data Corporation (IDC), organizations are suffering from inefficiencies and ineffectiveness as they turn to data as the lifeblood of their digital transformation, and the workforce is struggling.

Business metadata, such as metadata on ontology which is a way of showing the properties of a subject area and how they are related, by defining a set of concepts and categories that represent the subject for rapidly growing and diverse data. A combination of business and technical metadata provides a unified view of data assets, and reduces the effort of searching for the correct data. A Data Catalog enables business units to discover and access data securely, using data classification and policy management. It helps organizations identify new opportunities for problem-solving, innovation, and revenue growth, using business lineage and reports which are generated on the data asset relationships by data stewards. Data Catalogs help customers make connections that weren't possible before, and drive their business growth while making informed business decisions.

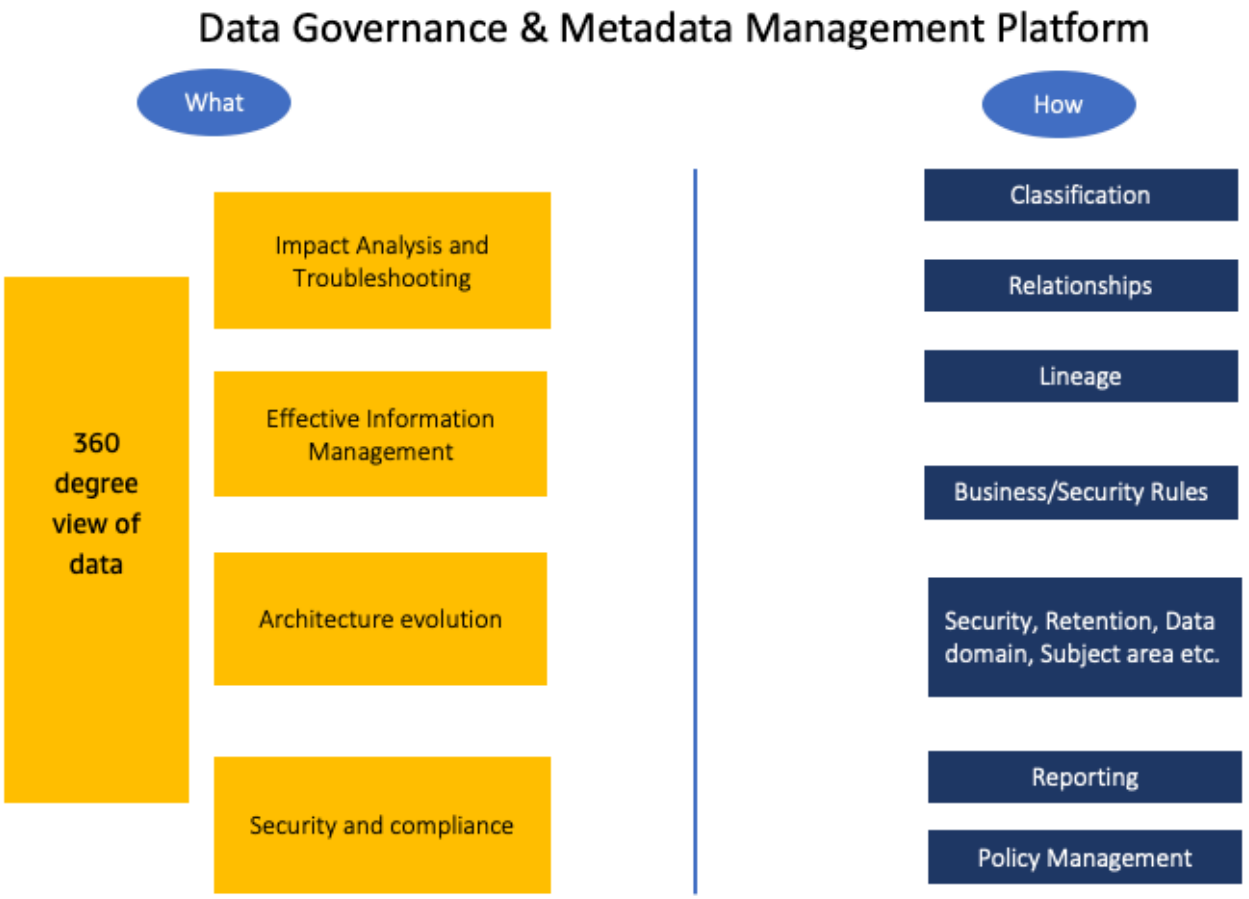
With robust metadata as the foundation of the Data Catalog, the following diagram illustrates the core features and functions supported:

- **Data classification** labels related datasets to business domains, subject areas, and data facets, helping data users understand their datasets better.
- **Data policy management** outlines how data processing and management is carried out to ensure an organization's data is accurate, accessible, consistent, and protected. The policy establishes who is responsible for information under various circumstances, and specifies what procedures should be used to manage it. It incorporates risk management to identify, assess, and control threats to an organization's capital and earnings. It also introduces data ethics principles

to reduce potential business problems from the use of data. Data policy management includes data access management and data retention.

- **Business lineage** highlights the transformation and aggregation of data needed by a business user. Without this capability, business units cannot identify impacted systems and business processes changes.
- **Reporting** allows visualization and dashboarding capabilities around various data assets. This reporting capability provides metrics around data volume growth, data classifications, relationships, and more.

These capabilities make it easier for stakeholders to get an absolute view of their business data, so they can take appropriate business actions.



Data Catalog core features and supported functions

Data Catalog benefits for technical stakeholders

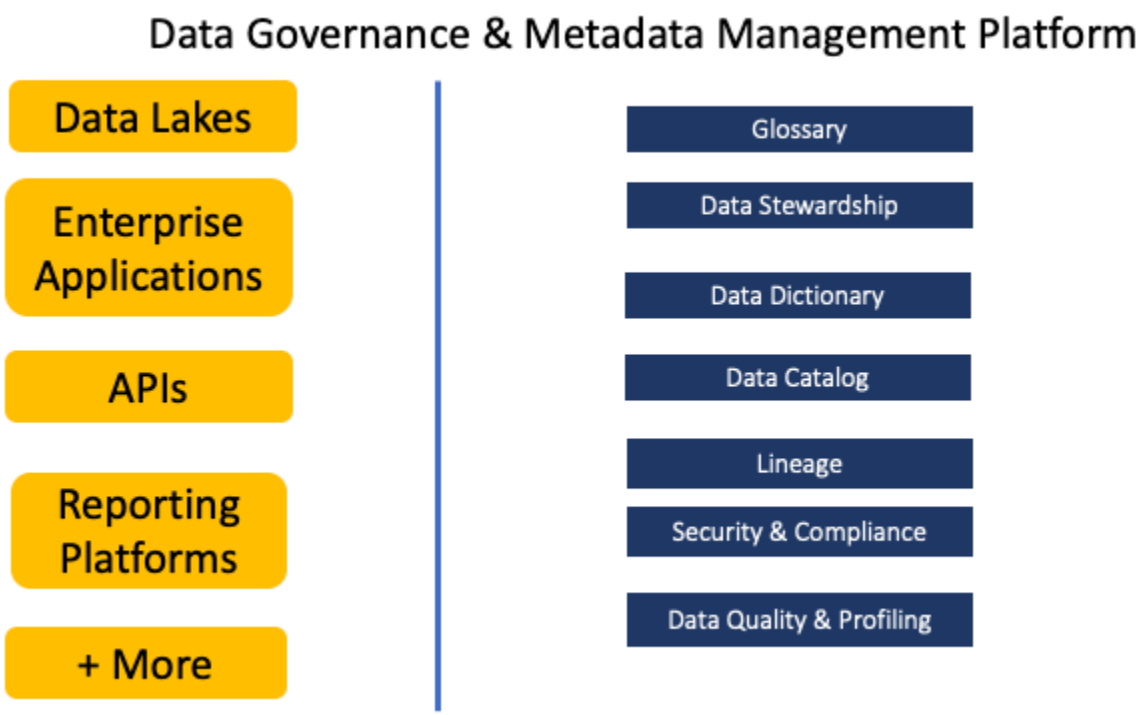
According to Eckerson Group, one of the leading data analytics consulting and research group, data cataloging accelerates analysis by minimizing the time and effort that analysts spend finding and preparing data. Anecdotally, 80% of self-service analysis without a Data Catalog is spent getting data ready for analysis. Using the Data Catalog cuts that percentage from 80 to 20.

A well-managed Data Catalog helps technical teams get more business context on technical and operational aspects of data using a business glossary. It enables technical teams to perform data discovery and analysis efficiently, while building data pipelines for the swift launch of data-driven applications to drive business growth. It provides lineage around the data journey, from origin to consumption.

A Data Catalog provides native and comprehensive data governance capabilities that ensure trust in the data, and proper and compliant use of data across the enterprise. Without the preceding capabilities, it is hard for technical stakeholders to access updated versions of data, and manage the data. This can impact the business user's ability to do data impact analysis.

Using technical lineage, a Data Catalog enables a business units' technical support team to analyze downstream system impacts swiftly. System impacts are caused by a change in a source system feed (a change in attribute type or length). Technical lineage depicts the state of data from origin to consumption. It is a time-consuming exercise to manually identify the business impact due to change in the application or data assets.

A data dictionary and business glossary establish a standard business definition and arrange it consistently across the datasets. With robust metadata as the core of the Data Catalog, many other features and functions are available for an organization's technical stakeholders, as seen in the following diagram.



Data Catalog features and functions

- The **business glossary** establishes standard business definitions to enable a common understanding of data across the organization. A business glossary ensures organizations speak the same language by clearing up ambiguity in business terminology.
- The **data dictionary** is a collection of the names, definitions, and attributes for data elements. The data dictionary defines conventions for the project and consistency throughout the dataset. Without a data dictionary, there's a higher risk of losing crucial information in translation and transition of data. Using a data dictionary helps data users analyze the datasets with ease later on.
- **Technical lineage** shows how data transforms and flows as it moves from system to system, providing additional understanding and trust in data. It empowers users to understand how data was acquired, and how it may have been transformed to establish dependence in the reporting results or insights generated.
- **Granular security controls** are role-based, asset-level permissions and access controls for secure enterprise-wide deployment. With a Data Catalog in place, it's easier to enable the preceding security controls. Any organization that efficiently secures its data builds confidence among internal and external customers, driving business growth.

Key considerations while building a Data Catalog

This whitepaper discussed the challenges which organizations are facing, what data governance brings, and how the data governance framework (Data Catalog) can help. The next step is implementing the Data Catalog. Following are key considerations which an organization should weigh before starting their journey to build a Data Catalog.

Choose a team with domain knowledge and Data Catalog skillsets

A team building a Data Catalog must be a balanced mix of technical and business experts. The team should have a background with data integration, which is the practice of consolidating data from disparate sources into a single dataset, with the goal of providing users with consistent access and delivery of data across all subjects and structure types. This is an essential skill, because a team with domain knowledge and Data Catalog skills can reduce the overall completion time to onboard and maintain a Data Catalog.

Tools, technology, and an approach to build the Data Catalog

There are various tools and approaches to build the Data Catalog. You can start with either a custom build approach, where you choose your own tool sets, or you can use third-party tools. Both have their advantages and drawbacks.

With a custom build Data Catalog, the overall tool and licensing cost is lower, but additional labor is required to acquire, ingest, and present the Data Catalog to users.

Having third-party tools can shorten the metadata acquisition, processing, and presentation time, because it provides out-of-the-box capabilities to achieve these tasks. However, the overall tool and licensing cost is higher.

Implementation reference architecture diagrams

The following are two reference architecture diagrams to help with the design and build phases of implementing a Data Catalog.

- **Reference architecture diagram 1** illustrates how an organization can build a Data Catalog without the use of third-party data cataloging tools, instead collecting technical metadata and enriching it using business metadata with help from a team of data stewards.
- **Reference architecture diagram 2** illustrates how an organization can use third-party tools like Collibra to collect technical metadata, and enrich technical metadata using business metadata with help from a team of data stewards.

A Data Catalog ensures that data is well managed, because data are the building blocks to establish a strong data culture. Data comes in many shapes, sizes, and formats, and each must be captured and depicted in its native format. A Data Catalog captures and depicts data by classifying the data. Implementation of a Data Catalog enables an organization to know datasets' origins, and how data transforms as it flows across different applications, reducing data analysis time.

Implementation reference architecture diagram 1

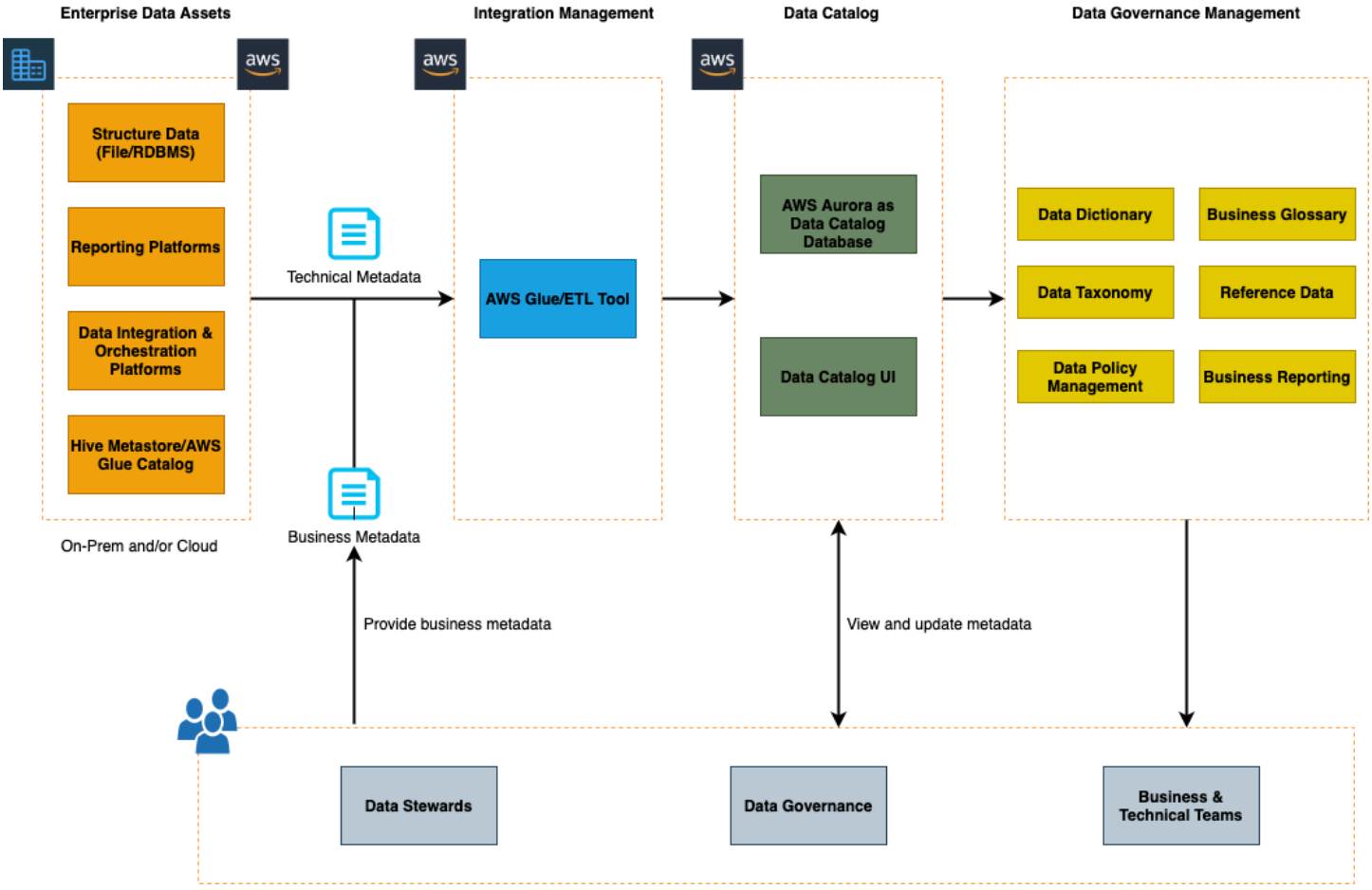
Reference architecture diagram 1 describes the high-level implementation of a Data Catalog using a custom build approach. This approach uses a relational database (Amazon Aurora PostgreSQL/MySQL), and AWS Glue or other available data integration tools.

Technical metadata, which is comprised of tables, attributes, definitions, and so on, are captured by the source and pulled on a scheduled basis from various heterogenous source data dictionaries.

Business metadata, consisting of business context, is prepared and gathered by data stewards who are data architects, product managers, and data analysts. The technical and business metadata is combined together, and it provides a single version of truth for the collected metadata.

Combined business and technical metadata provide details around the data taxonomy, data classification, reference data, business glossary, and security management.

Data taxonomy is the classification of data based on data domains, subject areas, and data facets that introduce common terminologies and semantics across multiple systems.



Reference architecture diagram 1: Enterprise metadata and data governance management catalog

Technical metadata is collected from various enterprise sources by a Database/Application Programming Interface (API). The API/JDBC connection periodically pulls data dictionary details from various relational/application data stores. This technical metadata is stored in a relational database modeled to meet the organizations data governance requirements.

The metadata is enriched by additional business metadata related to the objects and attributes such as description, lineage info, security classifications, ownership, and so on. Data lineage is the process of understanding, recording, and visualizing data as it flows from data sources to consumption. It includes all transformations the data underwent along the way, and how the data was transformed and consumed.

The data stewardship team is an essential part of the data governance process. Stewards update the data taxonomy. Data stewardship is the collection of practices that ensure an organization's data is accessible, usable, safe, and trusted. It includes overseeing aspects of the data lifecycle: creating, preparing, using, storing, archiving, and deleting data. They help promote data quality

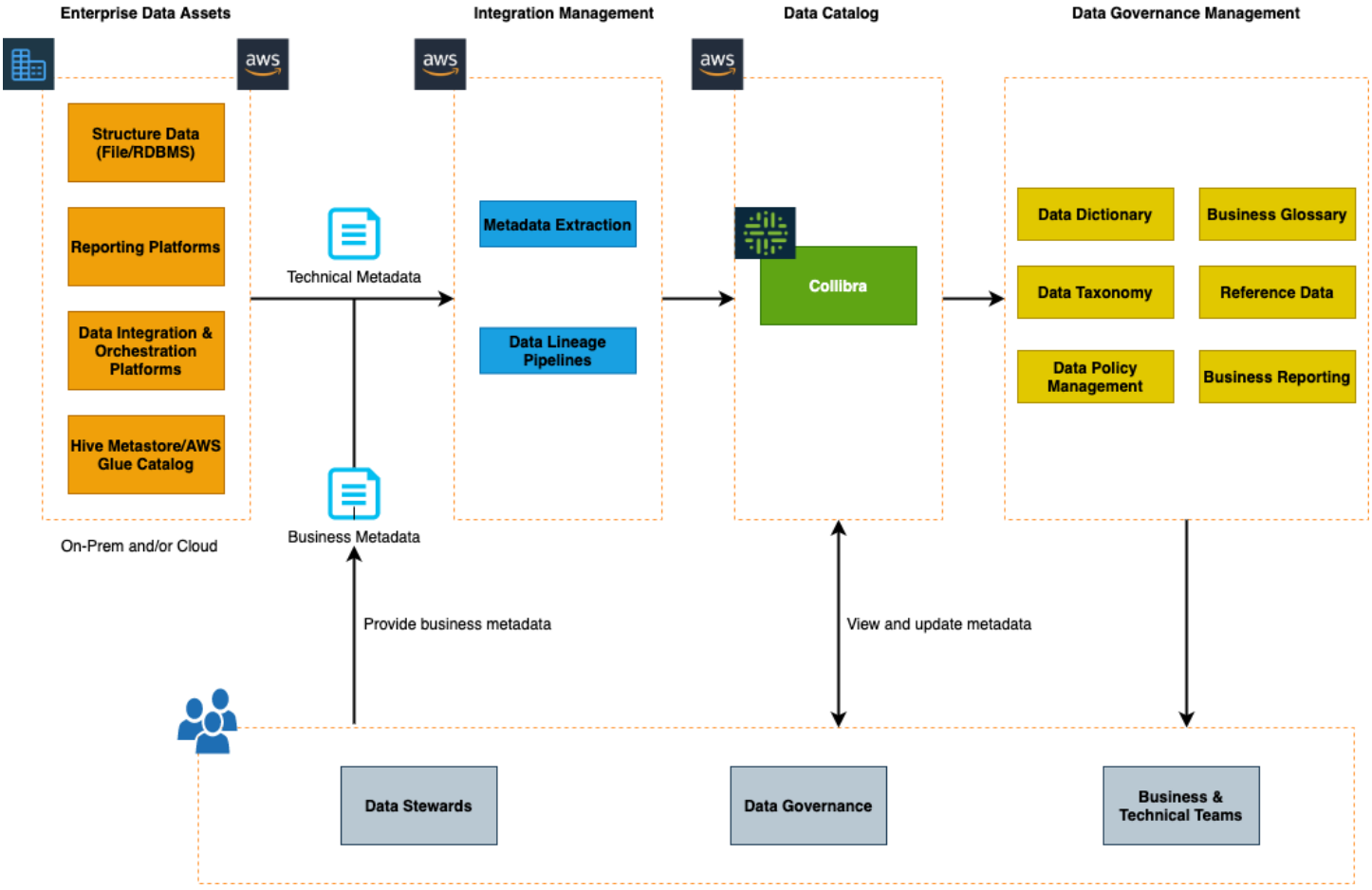
and integrity that is in line with the data governance principles of an organization. They manage and maintain the Data Catalog using the graphical interface build on top of the relational database that stores the metadata.

Implementation reference architecture diagram 2

Reference architecture diagram 2 describes the high-level implementation of Data Catalog using third-party tools like Collibra, relational databases (Aurora PostgreSQL/MySQL), and AWS Glue or other available data integration pipelines. Collibra software is an enterprise-oriented data governance platform for Data Catalog and stewardship. It empowers businesses to find meaning in their data and improve business decisions. Collibra's partnership with AWS makes it possible to unlock the value of data, irrespective of where and how it is stored.

Technical metadata, which is comprised of tables, attributes, definitions, and so on, is captured by the source and pulled on a scheduled basis from various heterogenous sources using Collibra. Collibra centralizes, governs, and certifies reports and metrics on collected metadata. Technical metadata is enriched with business metadata related to the objects and attributes.

Third-party tools provide an out-of-the-box graphical user interface for data stewards and users to view and update business metadata. Most third-party tools also provide machine learning capabilities to find similar matching patterns of data and help them inherit definitions and classifications. As depicted in the reference architecture diagram 1, technical and business metadata is combined to provide users with a meaningful context for various data assets within the organization.



Reference architecture diagram 2: Enterprise metadata and data governance management catalog

Technical metadata is collected from various enterprise sources by a Database/Application Programming Interface (API). The API/JDBC connection periodically pulls data dictionary details from various relational/application data stores.

Technical metadata is enriched by extended metadata related to the objects and attributes. The third-party data cataloging tool's graphical user interface is used to view and update the collected metadata. Third-party tools like Collibra also provide out-of-the-box reports around the collected and enriched metadata.

Conclusion

A Data Catalog benefits an organization by bringing in visibility around the siloed datasets hidden deep within various data stores such as data lakes, data warehouses, and data marts. It helps classify the data assets and make them searchable, evaluable, and useful, to help enterprises make informed business decisions. This whitepaper highlights how a Data Catalog framework combines data and technology together and provides meaningful data insights to stakeholders, enabling them to manage data assets and the security around them effectively.

There are different data protection acts adopted by various countries to put legislations in place regarding data security and privacy protection. This whitepaper describes how a Data Catalog enables organization to adhere to data protection regulations such as GDPR, and CCPA. The reference architecture diagrams showcase multiple approaches to implement a Data Catalog using custom in-house toolsets, utilizing third-party data governance and Data Catalog toolsets.

Contributors

Contributors to this document include:

- Nikhil Jha, Sr. Data Architect, DW & MPP
- Barbra Hale, Sr. Delivery Practice Manager
- Rahul Jani, Data Lake, Data Architect
- Mahesh Goyal, Sr. Data Architect, DW & MPP

Further reading

For additional information, refer to:

- AWS Architecture Center
- Trust your data: why you need a governed data catalog (blog)
- Data Governance vs Data Catalog: What's the Difference? (article)
- Apache Atlas
- Building a Data Catalog For Small and Medium-Sized Businesses (article)
- Data traceability vs data lineage: Understanding the differences (blog)
- Data governance policy (article)
- erwin Expert blogs on Data Catalogs
- What is a Data Catalog and Why You (Definitely) Need it? (article)
- What Is a Data Catalog? (blog)
- Data definitions
- Workers waste half their time as they struggle with data (article)
- Augmented Data Catalogs: A Must-Have for Data & Analytics Leaders (article)
- The Business Case for a Data Catalog (article)

Document history

To be notified about updates to this whitepaper, subscribe to the RSS feed.

Change	Description	Date
Initial publication	Whitepaper first published.	December 3, 2021

Note

To subscribe to RSS updates, you must have an RSS plug-in enabled for the browser that you are using.

Notices

Customers are responsible for making their own independent assessment of the information in this document. This document: (a) is for informational purposes only, (b) represents current AWS product offerings and practices, which are subject to change without notice, and (c) does not create any commitments or assurances from AWS and its affiliates, suppliers or licensors. AWS products or services are provided “as is” without warranties, representations, or conditions of any kind, whether express or implied. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.

AWS Glossary

For the latest AWS terminology, see the AWS glossary in the *AWS Glossary Reference*.