Global Practices
Big Data & Analytics Practice

Overview of Data Catalog Capability

December 2022
Arun Viswanathan

GlobalLogic®
A Hitachi Group Company

Intended Audience

The intended audience of this document is Architects that are looking to understand practical approaches to notable Big Data & Analytics Architecture-level aspects. It is NOT the intent of this document to cover Design and/or Implementation level aspects, however, it IS intended that such detailed levels of associated content could be subsequently developed.

Context

Problem Statement

With the explosion of data and data systems, organizations are struggling to keep up with an increasingly fragmented data environment. Over time, data tends to get siloed across the many different systems that crop up in an organization making it difficult to find the relevant data quickly. And even when it is discovered, it is usually not classified or described properly making it difficult to understand. This can create a risk when data is required to make critical business decisions.

For organizations to solve the challenges around data discovery and generate meaningful insights, it is important to first discover the data across all their data assets including data warehouses, data lakes and the different data stores. The data scientists and business users need the ability to search, discover and access any data asset across the organization. That means a need to see the information about the data (i.e. metadata, refer Taxonomy section below for definition) across data silos, data warehouses, data lakes, and analytical environments.

Additionally, with data privacy laws being enacted it has become more important now to understand data from creation to consumption and the inability to track the data lineage as it moves from the source systems to the consuming systems can cause a significant risk. Adding new data sources and consumption of data by different teams across the enterprise makes the discovery and tracking data lineage even more difficult for data scientists and business users who are unaware of the new changes. Unfortunately, a majority of organizations are not equipped for this reality. In the age of big data, companies can capture large and exponentially increasing volumes & variety of data in data lakes, but they are unable to efficiently search, discover and access all this data. Hence a vast amount of data is unknown to the users in the organization and creating value from those data sets remains unused.

An enterprise data catalog solution can help address this challenge. A Data Catalog is a **company-wide inventory of data assets** that enables discovery, collaboration, governance and establishes trust in the data. It contains information necessary to understand the technical characteristics and the business context of all data assets of a company.

This document intends to provide an introduction to the Data Catalog capability along with the solution approach to implement it across an enterprise. It also briefly describes the different tools available for providing this capability along with the evaluation criteria that can be used in choosing the right tool for an organization.

Taxonomy

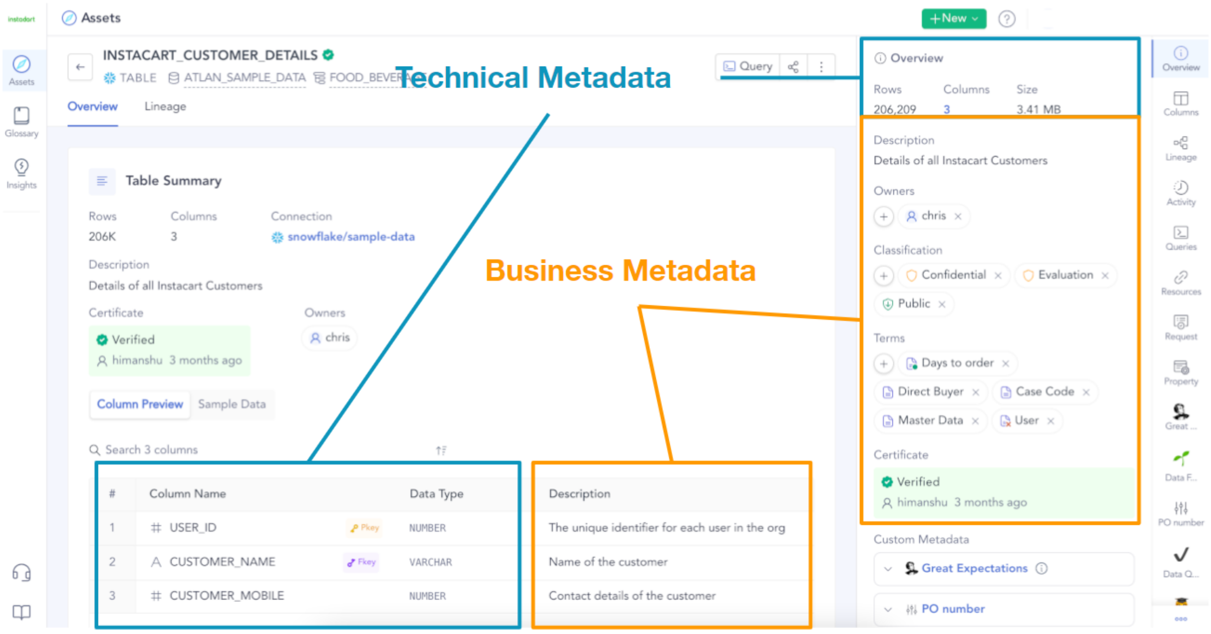
There are a number of similar sounding terminologies that a user would come across when exploring a data catalog. In this section we look at the important terms and how they are related to each other.

Metadata is data that provides more information about the data itself but does not include the actual data. This information could be basic details such as name, type, owner, etc. or more complex information such as statistical data, tags, quality details, etc. There are a number of different types of metadata - technical and business being the most common ones.

Technical metadata is the technical information that is related to format, structure and storage details of the data. Some examples include database name, table name, column name, data type, data lineage, data stats and so on.

Business metadata is the information that provides meaning to the data in everyday language. Examples include keywords, comments, description explaining its significance, quality, tags and so on.

In the following figure, the sample metadata along with the different types of metadata that is collected in a data catalog is shown.



INSTACART_CUSTOMER_DETAILS

TABLE ATLAN_SAMPLE_DATA FOOD_BEVER...

Table Summary

Rows: 206K
Columns: 3
Connection: snowflake/sample-data

Description: Details of all Instacart Customers

Certificate: Verified — himanshu 3 months ago

Owners: chris

Column Preview | Sample Data

Search 3 columns

#	Column Name		Data Type	Description
1	USER_ID	Pkey	NUMBER	The unique identifier for each user in the org
2	CUSTOMER_NAME	Fkey	VARCHAR	Name of the customer
3	CUSTOMER_MOBILE		NUMBER	Contact details of the customer

Overview

Rows: 206,209
Columns: 3
Size: 3.41 MB

Description: Details of all Instacart Customers

Owners: chris

Classification: Confidential, Evaluation, Public

Terms: Days to order, Direct Buyer, Case Code, Master Data, User

Certificate: Verified — himanshu 3 months ago

Custom Metadata

Great Expectations

PO number

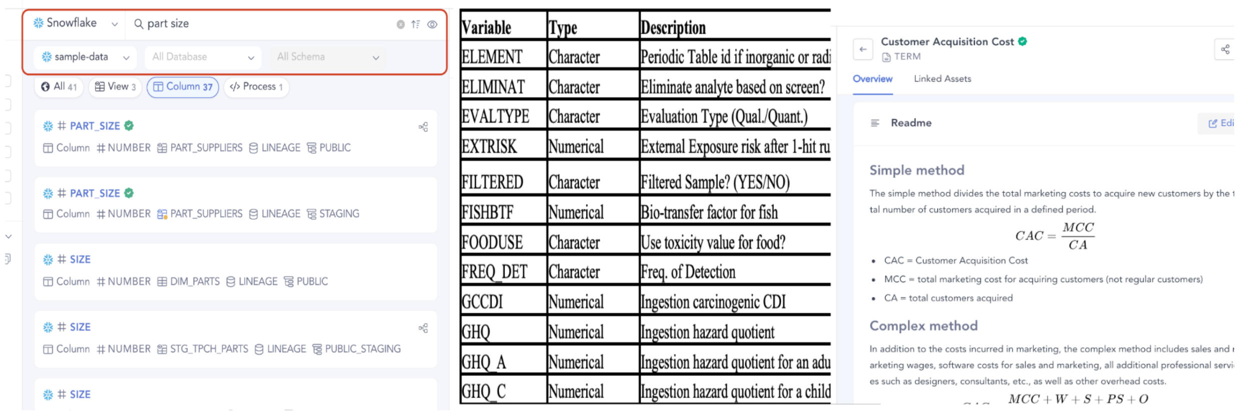
Figure 1. Metadata, Technical Metadata and Business Metadata (Source)

Data Dictionary is a repository that contains the description of the data and its metadata. That is, it contains general information of the dataset (like its statistics, origin, how it relates to upstream and downstream datasets, etc.) and column level details of the dataset (such as data type, constraint, descriptions of individual columns, etc). This is mostly *technical metadata* that can be automatically extracted from the datasets.

Business Glossary is a collection of business terms and concepts that are used across the organization along with their unique definitions and other related useful information. It typically includes the definition along with a number of attributes such as type, status, hierarchy, relationships, data links, etc. This is the *business metadata* that provides a common vocabulary for an organization, helping ensure the right terms are used consistently. While some tools attempt to automatically suggest new and popular business terms, this is mostly created and updated by data stewards manually.

The Data Dictionary and the business glossary together form the core capabilities of a Data Catalog as shown in the following figure. A **Data Catalog** is thus a collection of technical and business metadata across all data assets in the organization that is made easily discoverable and accessible for business users and data scientists. For a more detailed documentation around data dictionary and business glossary refer here.

The following figure shows screenshots of what a data catalog, data dictionary and business glossary looks like.



Snowflake | part size

sample-data | All Database | All Schema

All 41 | View 3 | Column 37 | Process 1

PART_SIZE
Column # NUMBER PART_SUPPLIERS LINEAGE PUBLIC

PART_SIZE
Column # NUMBER PART_SUPPLIERS LINEAGE STAGING

SIZE
Column # NUMBER DIM_PARTS LINEAGE PUBLIC

SIZE
Column # NUMBER STG_TPCH_PARTS LINEAGE PUBLIC_STAGING

SIZE

Variable	Type	Description
ELEMENT	Character	Periodic Table id if inorganic or radi
ELIMINAT	Character	Eliminate analyte based on screen?
EVALTYPE	Character	Evaluation Type (Qual./Quant.)
EXTRISK	Numerical	External Exposure risk after 1-hit ru
FILTERED	Character	Filtered Sample? (YES/NO)
FISHBTF	Numerical	Bio-transfer factor for fish
FOODUSE	Character	Use toxicity value for food?
FREQ_DET	Character	Freq. of Detection
GCCDI	Numerical	Ingestion carcinogenic CDI
GHQ	Numerical	Ingestion hazard quotient
GHQ_A	Numerical	Ingestion hazard quotient for an adu
GHQ_C	Numerical	Ingestion hazard quotient for a child

Customer Acquisition Cost
TERM

Overview | Linked Assets

Readme | Edi

Simple method

The simple method divides the total marketing costs to acquire new customers by the t tal number of customers acquired in a defined period.

$$CAC = \frac{MCC}{CA}$$

- CAC = Customer Acquisition Cost
- MCC = total marketing cost for acquiring customers (not regular customers)
- CA = total customers acquired

Complex method

In addition to the costs incurred in marketing, the complex method includes sales and r arketing wages, software costs for sales and marketing, all additional professional servi es such as designers, consultants, etc., as well as other overhead costs.

$$MCC + W + S + PS + O$$

Data Catalog | Data Dictionary | Business Glossary

Figure 2. Sample screens of Data Catalog, Data Dictionary and Business Glossary (Source)

Data Lineage is another concept that provides details of how the data is transformed as it flows from the source to the target systems. It is visualized as a graph and provides information on the upstream & downstream systems or processes at a given point in the lineage graph.

While these are a few of the core capabilities, a data catalog also provides a range of capabilities as detailed in this later section.

Other related concepts are Master Data Management and Metadata Management.

Master Data Management or MDM is about managing refined lists of records from data acquired and rules to standardize the records in a systematic way to make data trustable. Master data includes unique, business-critical information such as product name, bill of materials, company branches, etc. It helps build a single source of truth for business-critical data across the organization. MDM deals with master data or reference data that plays a role in the core operation of the organization. MDM is different from a Data Catalog which mainly deals with metadata management, that is information related to the data.

Metadata Management is about managing the metadata about the data. It gives meaning and describes the underlying data to make it searchable and usable. Most modern Data Catalog tools typically provide metadata management as a core capability.

Scoping within the Enterprise

When we talk about Data Catalog within Enterprise Systems, we need to consider all data that is ingested/created by Feature Systems as-well-as the centralized Data Platform. In scenarios where an individual data catalog is not maintained, then the data catalog within the centralized data platform needs to consolidate the metadata of all the data across the feature systems. When individual feature systems maintain their own data catalog, this metadata needs to be synced into the data catalog within the centralized data platform. The data catalog is thus the **system of record** for all the enterprise metadata within the enterprise.

The Data Platform and the position of the data catalog(s) in an Enterprise City Map are illustrated in the diagram below.

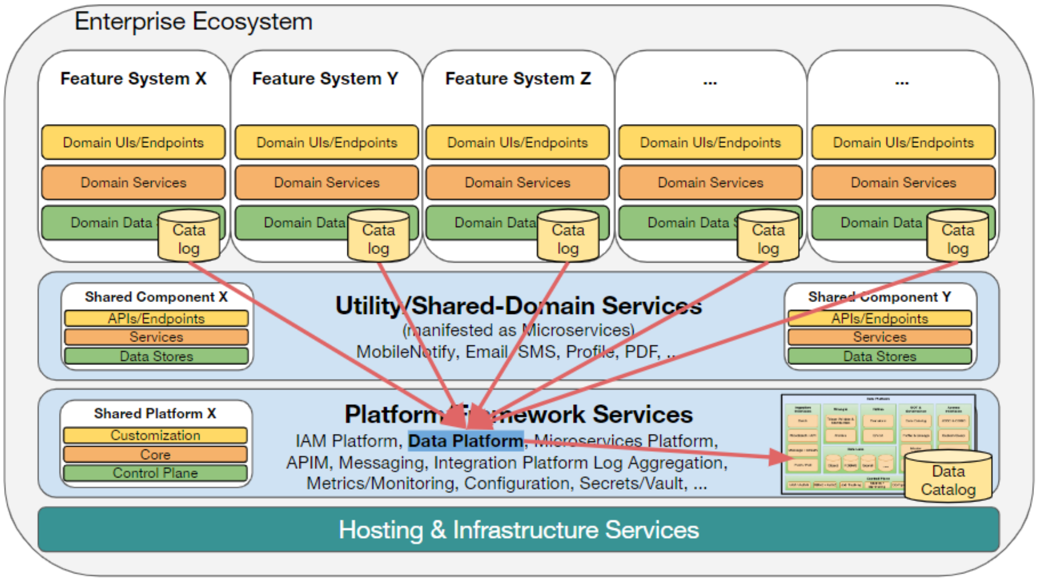

Enterprise Ecosystem
Feature System X
Feature System Y
Feature System Z
...
...
Domain UIs/Endpoints
Domain Services
Domain Data
Cata log
Utility/Shared-Domain Services
(manifested as Microservices)
MobileNotify, Email, SMS, Profile, PDF, ...
Shared Component X
APIs/Endpoints
Services
Data Stores
Shared Component Y
APIs/Endpoints
Services
Data Stores
Platform/Framework Services
IAM Platform, Data Platform, Microservices Platform,
APIM, Messaging, Integration Platform Log Aggregation,
Metrics/Monitoring, Configuration, Secrets/Vault, ...
Shared Platform X
Customization
Core
Control Plane
Data Catalog
Hosting & Infrastructure Services

Figure 3. Enterprise City Map (Source)

Scoping within the Data Platform

Within the centralized data platform, the Data Catalog is a capability that is part of the Data Governance logical pillar which consists of other capabilities such as data quality, data profiling & lineage, master data management and data warehouse / data marts / data models. The Data catalog capability in the Enterprise Data Platform is illustrated in the diagram shown below.

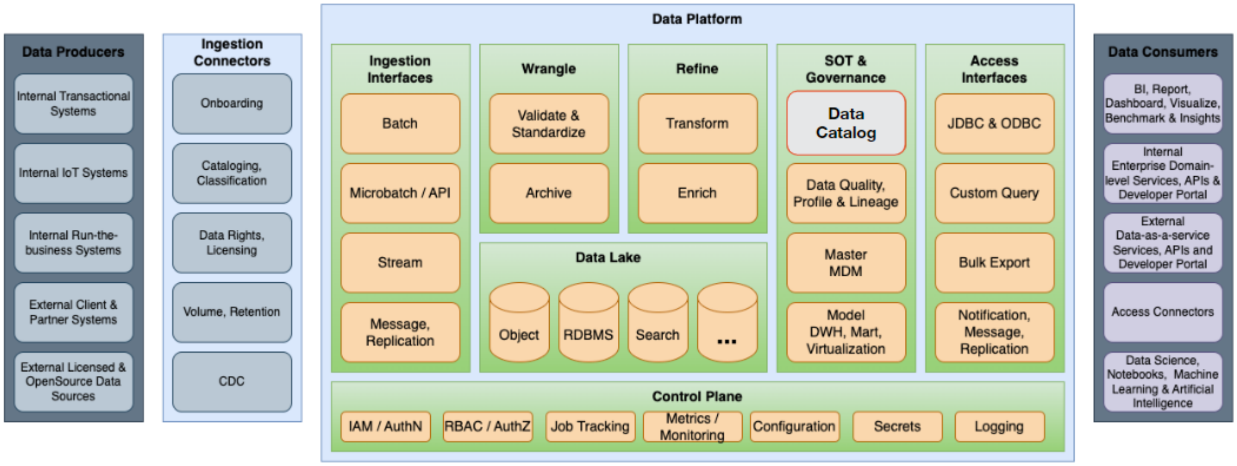

Data Producers
Internal Transactional Systems
Internal IoT Systems
Internal Run-the-business Systems
External Client & Partner Systems
External Licensed & OpenSource Data Sources
Ingestion Connectors
Onboarding
Cataloging, Classification
Data Rights, Licensing
Volume, Retention
CDC
Data Platform
Ingestion Interfaces
Batch
Microbatch / API
Stream
Message, Replication
Wrangle
Validate & Standardize
Archive
Refine
Transform
Enrich
Data Lake
Object
RDBMS
Search
...
SOT & Governance
Data Catalog
Data Quality, Profile & Lineage
Master MDM
Model DWH, Mart, Virtualization
Access Interfaces
JDBC & ODBC
Custom Query
Bulk Export
Notification, Message, Replication
Control Plane
IAM / AuthN
RBAC / AuthZ
Job Tracking
Metrics / Monitoring
Configuration
Secrets
Logging
Data Consumers
BI, Report, Dashboard, Visualize, Benchmark & Insights
Internal Enterprise Domain-level Services, APIs & Developer Portal
External Data-as-a-service Services, APIs and Developer Portal
Access Connectors
Data Science, Notebooks, Machine Learning & Artificial Intelligence

Figure 4. Enterprise Data Platform (Source)

Benefits with using a Data Catalog

The below diagram shows the different problems in an organization (represented in black blurbs) that can be solved by implementing a Data Catalog solution across the organization.

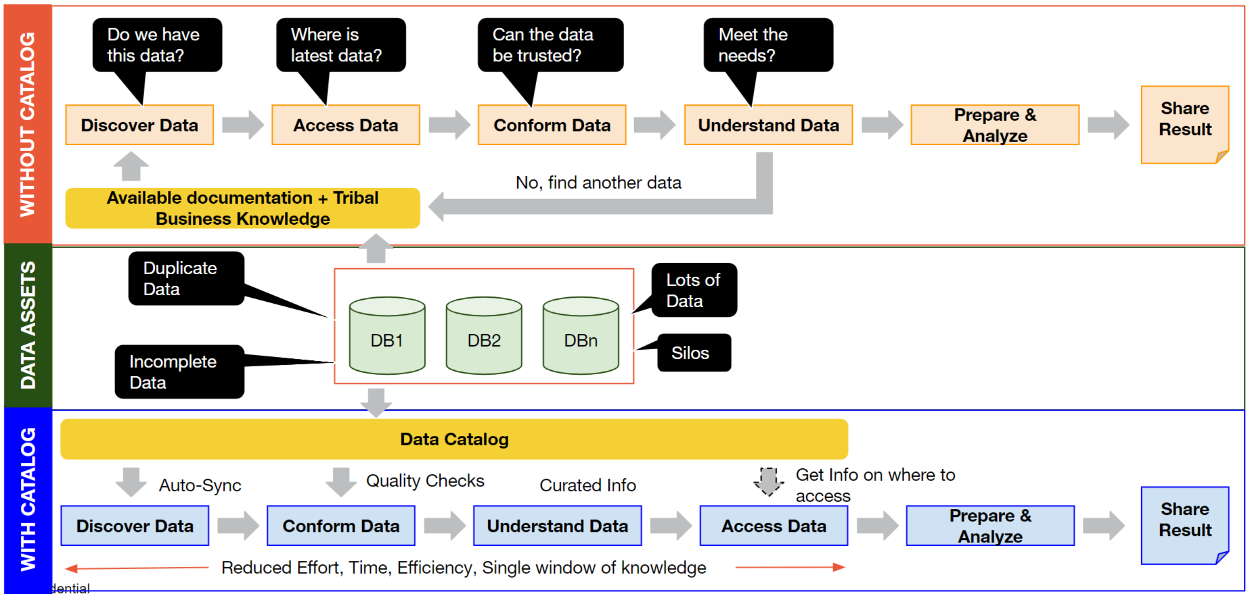


Figure 5. Challenges solved with Data Catalog (Source)

A data catalog can help eliminate many of the pain points which business users and data scientists face when they try to gain business insight from data. The key benefits are as follows:

- **Increase discoverability** of the key datasets across the different siloes. As shown in the above figure, with a data catalog configured to automatically sync data metadata from data sources, searching for the data becomes simpler and covers the entire enterprise landscape.
- Ability to **Find** the right data and **Understand** data better. For example, the datasets in the catalog can be curated and documented properly to enable users to understand them better.
- **Improve productivity and reduce time spent** by teams searching for relevant information or data. As shown in the above figure, through automation and machine learning, data catalogs can reduce the effort & time to discover, conform and understand the data thus improving efficiency.
- **Identify and avoid duplication** of similar datasets by different teams by allowing providing a unified interface containing the complete list of datasets across the enterprise
- **Share consistent knowledge** across the organization which might have been limited to specific users thus reducing the tribal knowledge
- **Improve collaboration** between different business teams and data science teams

- **Facilitate compliance** with growing international privacy and reporting regulations
- Enable **tracking data usage** i.e. how data is being used and by whom
- Help establish proper **data governance** on the data assets. Data governance involves managing data availability, integrity, usability and security based on internal data standards and policies.

However it is important to note that a Data Catalog is not the tool used to access the data itself. That is the function of other capabilities such as JDBC/ODBC connectors, APIs, message queues, etc. A data catalog however provides information on where and how to access the right data for the user, who can then use other means to get access to it.

Solution

In this section we describe the approach to implement a Data Catalog Solution within an enterprise data platform and an overview of different tools available to perform this function.

Core Features of an Enterprise Data Catalog

A number of tools exist that provide a varying set of capabilities with some features overlapping across other tools in the governance block. However it is the opinion of this author that an enterprise data catalog needs to have at least the following core features:

- Ability to **search & discover** Data and visualize details through dashboards & reports
- Provide **out of Box support for connecting** to different types of data sources, integration systems and reporting tools
- Ability to **automatically sync technical metadata** such as structure, schema, ownership, access patterns, etc. across the data assets.
- Capability to **perform all actions** on the catalog **programmatically** through a well supported API
- Support for **adding tags** both automatically & manually to a data entry to provide meaningful context. A tag is a piece of metadata that is used to label an asset, to help categorize the asset. For example, tags can provide information on personally identifiable information (PII), the data retention policy for the asset or a data quality score. This can make it easier to search the required data asset
- **Collaborative** catalog features like adding ratings, providing reviews on data quality and conversations between users regarding the datasets
- Ability to automatically derive **data lineage** from different types of code—ETL jobs, SQL scripts and stored procedures to help users understand the origin and destination of any data asset in a data catalog and how data was transformed or enriched on the way.

- Automatically **profile the data** to give users info about the data and using it to intelligently detect similarities in data, and relationship between data points in different data sources
- **Detect anomalies** or sudden changes in data and notify users about such events allowing errors to be corrected continuously.
- Assist in **Active governance** of metadata access by different users as per their required permissions and need

Solution Approach

As we have seen in the previous sections, an enterprise data catalog tool can solve a lot of the problems faced by organizations with discovering, understanding and using data spread across the organization. But it is important to understand that the tool by itself will not solve the problem automatically. It needs to interact with relevant stakeholders who can review, update and enrich the data collected by it to make it consumable. While it is not the intention of this document to discuss the implementation aspects, a high level solution approach to implementing a data catalog within an enterprise is discussed here. This section provides a generic approach that can be implemented using any of the Data Catalog tools available. Some of the Data Catalog tool options will also be briefly covered in the following sections.

The following diagram shows the process of integrating a data catalog with enterprise data assets and the data governance council to create the required output repositories.

The data governance council is a governing body consisting of people who strategize data governance of which Data Catalog is a function, define data standards, data policies and oversee its implementation across the enterprise. For more details you can refer to this Collibra blog. The members of the council are described further below in this section.

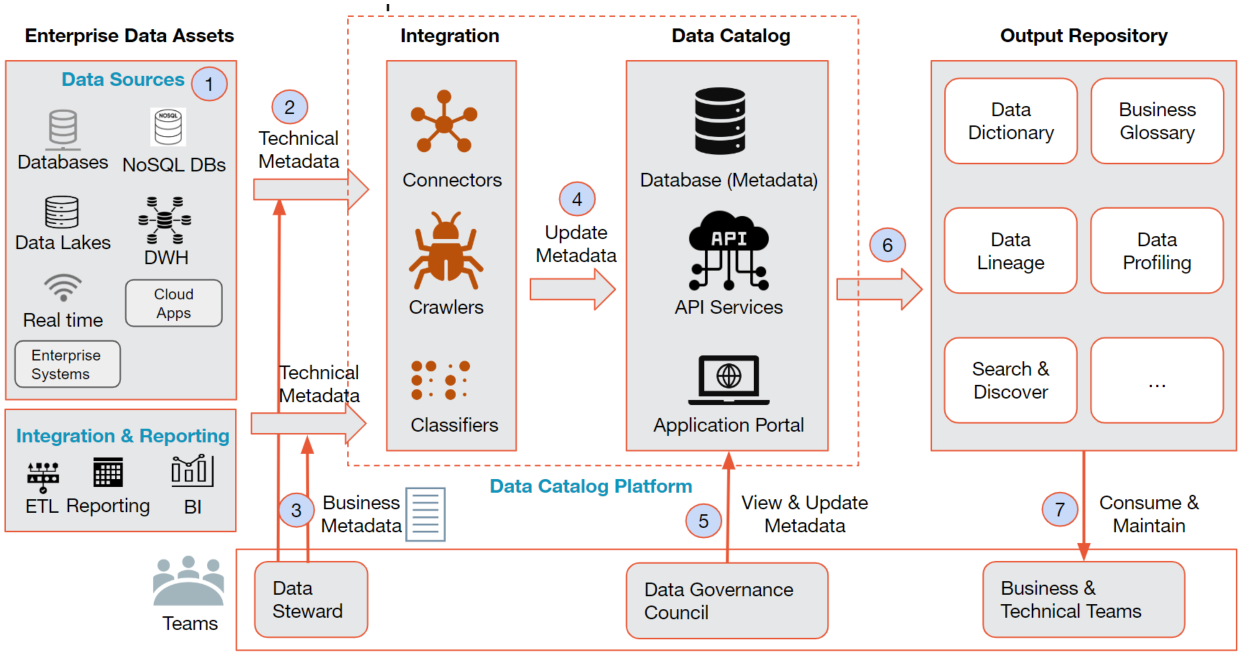


Figure 6. Illustrative Data Flow using Data Catalog (Source)

The steps to be followed while implementing the data catalog are as follows:

1. The first step would be to register the different data sources, integration and reporting tools in the Data Catalog tool using Admin console or a database registry manually
2. The tool can then use connectors, crawlers and classifiers to automatically Ingest the technical metadata from the tables of the registered databases.
3. The Data Steward and/or business user can then add the business metadata to the technical metadata once it has been ingested.
4. Data Catalog tool is used to perform operations like identifying schema changes (like type change or format change over time), generating data lineage, and providing data sampling for the data sources registered. Some tools also provide advanced capabilities to identify quality issues and provide recommendations on improving data quality.
5. A Data Governance council can then manually review the metadata collected in the data catalog, update them and add tags as applicable. The data governance team includes data steward, data engineers and business analysts. These roles and responsibilities are described in the next section.
6. Based on a combination of automated metadata fetching and manual updates the data governance output such as data dictionary, business glossary, data taxonomy, etc. can be generated.
7. The business and technical teams can then consume the data catalog content and maintain them as there are changes in the applications.

It is highly recommended to create a Data Governance Council composed of all relevant stakeholders such as data managers, data producers, data stewards. data architects, process owners, data owners and data consumers. It is the joint responsibility of the data governance council to maintain and manage the data catalog to ensure that it is up-to-date and trustworthy.

The following figure shows the composition of the Data Governance Council.

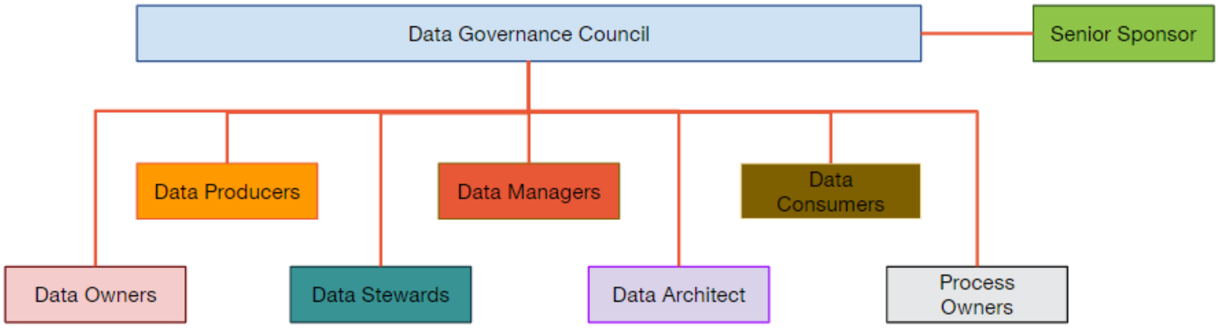


Figure 7. Data Governance Council Role & Responsibilities (Source)

Below are the roles and responsibilities of the members of the Data Governance Council.

Role	Responsibilities
Data Governance Council	Governing body responsible for the strategic guidance of the data governance program, prioritization for the projects and initiatives, and approval of organization-wide data policies and standards
Senior Sponsor	Senior Leadership team member sponsoring the data governance initiative
Data Managers	Establish and govern an enterprise data governance implementation roadmap including strategic priorities for development of information-based capabilities
Data Producers	Group of people who create the data or collect it from different sources that will be consumed by other people or systems.
Data Stewards	Responsible for utilizing data governance processes to ensure the quality of data elements, including content and metadata. ● Fully understand Data Quality, Security & Compliance aspects. ● Teach and Support Product, Development and Quality Teams. ● Create & Own Data Governance Epic/Feature Backlog Items. ● Identify, Clarify, Escalate and Resolve data work-items.
Data Architects	Architects and implements the catalog solution and puts standards &

	quality in place
Process Owners	Group of people that develops an organization's policies and practices to treat data as a strategic asset
Data Owners	An individual who is accountable for a data asset
Data Consumers	Team members with direct responsibility for using data as part of their daily tasks. They may directly access and investigate integrated datasets at the unit record level for statistical and research purposes.

An Overview of the Data Catalog Tools

When we consider the implementation approach, a number of mature enterprise data catalog tools are available. In this section, we will introduce the following select tools spread across the pricing and deployment options.

- **Commercial tools:** Informatica, Alation, Collibra
- **Open Source tools:** Apache Atlas, Amundsen, DataHub
- **Cloud Native tools:** AWS Glue Data Catalog, Microsoft Purview, Google Cloud Data Catalog

Informatica is a market leader in this space. It provides Enterprise Data Cataloging, Data Lineage and AI engine with the following features:

- Automatically catalog and classify all types of data across the enterprise using an AI-powered catalog
- Provide a metadata system of record for the enterprise with a catalog of catalogs
- Automatically extract the granular metadata from a wide array of data sources, including complex enterprise systems
- Find data assets through powerful Google-like semantic search
- Discover and understand the data assets with a holistic view including lineage, relationship views, and data profiling stats and quality scorecards
- Identify domains and entities with intelligent curation
- Enrich data assets through governed and crowdsourced annotations, ratings, and reviews
- Automatically associate business glossary terms to technical data assets
- Provide Open APIs to integrate into existing data landscape
- Provides ability to measure and optimize the the data assets with Analytics
- Informatica products tend to have higher-end pricing.

Below are some screenshots of the different features of the tool.

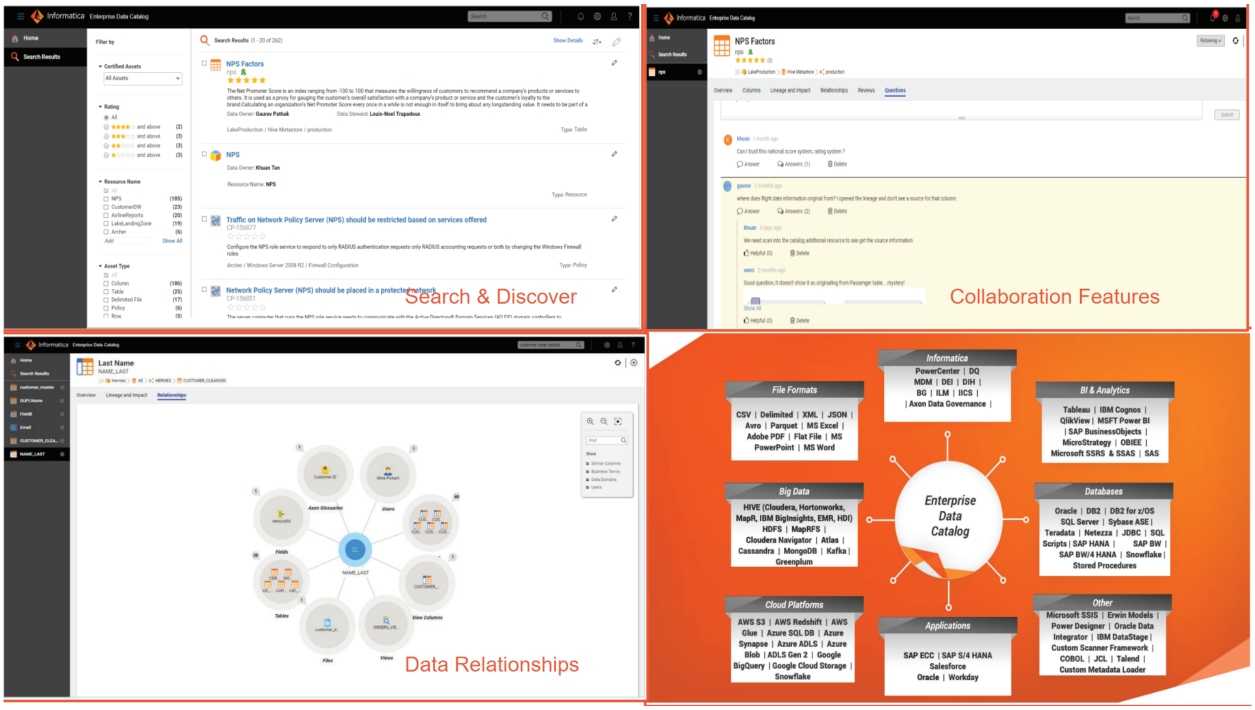

Search & Discover
Collaboration Features
Data Relationships
Informatica
PowerCenter | DQ
MDM | DEI | DIH |
BG | ILM | IICS |
| Axon Data Governance |
File Formats
CSV | Delimited | XML | JSON |
Avro | Parquet | MS Excel |
Adobe PDF | Flat File | MS
PowerPoint | MS Word
BI & Analytics
Tableau | IBM Cognos |
QlikView | MSFT Power BI
| SAP BusinessObjects |
MicroStrategy | OBIEE |
Microsoft SSRS & SSAS | SAS
Big Data
HIVE (Cloudera, Hortonworks,
MapR, IBM BigInsights, EMR, HDI)
HDFS | MapRFS |
Cloudera Navigator | Atlas |
Cassandra | MongoDB | Kafka |
Greenplum
Enterprise
Data
Catalog
Databases
Oracle | DB2 | DB2 for z/OS
SQL Server | Sybase ASE |
Teradata | Netezza | JDBC | SQL
Scripts | SAP HANA | SAP BW |
SAP BW/4 HANA | Snowflake |
Stored Procedures
Cloud Platforms
AWS S3 | AWS Redshift | AWS
Glue | Azure SQL DB | Azure
Synapse | Azure ADLS | Azure
Blob | ADLS Gen 2 | Google
BigQuery | Google Cloud Storage |
Snowflake
Applications
SAP ECC | SAP S/4 HANA
Salesforce
Oracle | Workday
Other
Microsoft SSIS | Erwin Models |
Power Designer | Oracle Data
Integrator | IBM DataStage |
Custom Scanner Framework |
COBOL | JCL | Talend |
Custom Metadata Loader

Figure 8. Screenshots of Informatica Data Catalog (Source)

Alation is another popular and mature data catalog tool that helps organizations identify, understand, and manage their data assets. It provides the following features:

- Ability to Search & Discover relevant information of the data assets
- Support for native built-in and non-built-in JDBC drivers for all the popular databases
- Supports table, column, directory, file level lineage
- Machine Learning based pattern recognition to show insights on data
- Enables data quality & data governance
- Provides collaboration abilities like annotation and discussions
- Some API development is needed for utilizing some features such as Lineage
- While Alation has a number of features and is a mature product, it tends to have higher-end pricing.

Below are some screenshots of the different features of the tool.

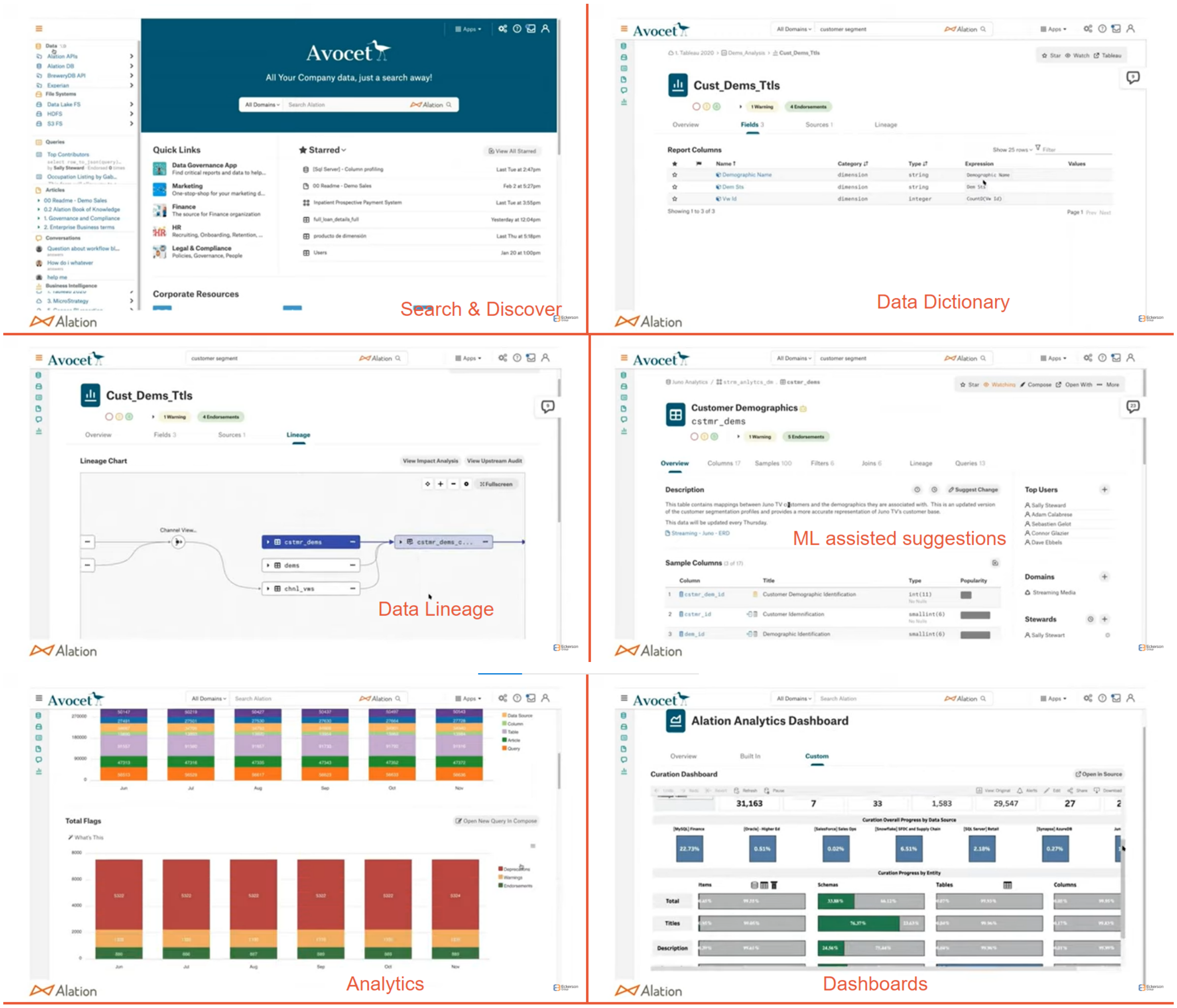


Figure 9. Screenshots of Alation Data Catalog (Source)

Collibra provides solutions for data management through products like data catalog, data governance and data lineage supported with APIs enabling collaboration. Its features include:

- Out of box linking to different data sources, business apps, etc.
- Ability to automatically classify data
- Ability to manually ingest schema metadata
- Support for adding tags to data assets
- Automatically map end-to-end lineage
- Provides common data profiles like type, count, distinctness, stats, etc
- Provides data sampling capability as part of profiling
- Can integrate with data quality user-defined rules, metrics and dimensions
- Support for Data Security
- Ability to perform actions through API that can be done from the user interface
- Provides UI to search and view catalog reports
- Provides collaboration features
- Provides a quality data catalog with embedded governance and privacy option

- The Collibra cloud platform tends to have medium-end pricing

Below are some screenshots of the different features of the tool.

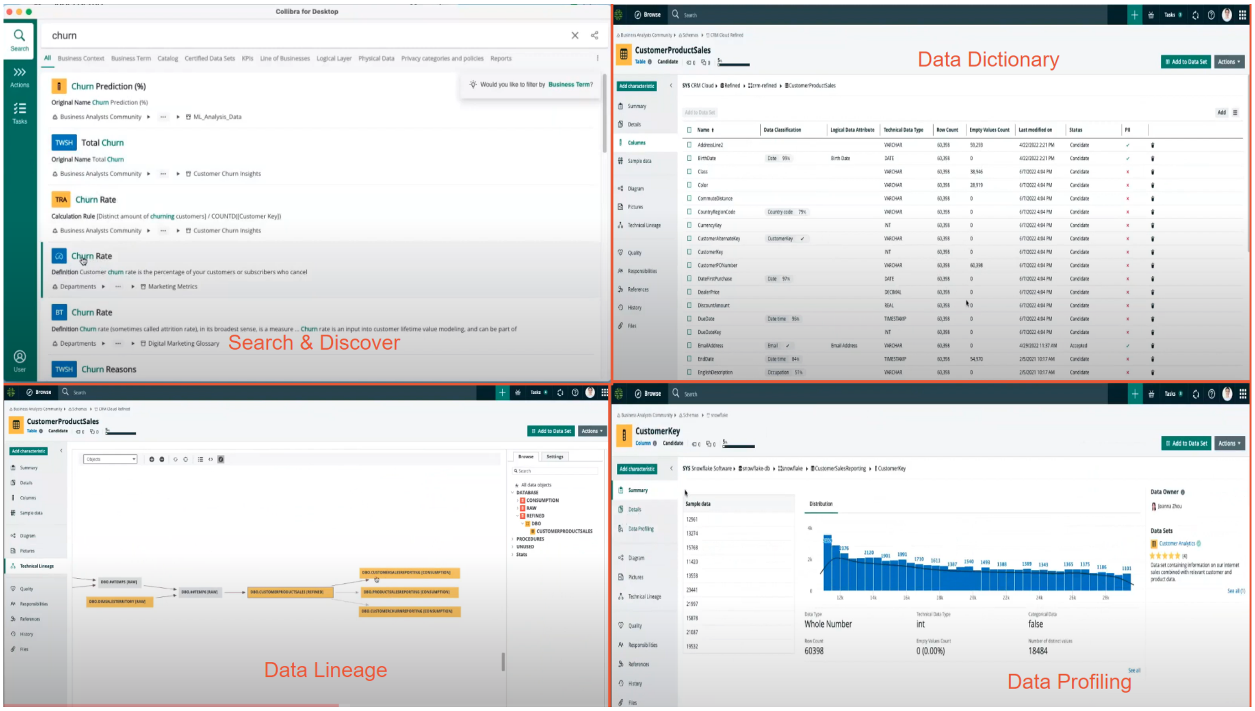
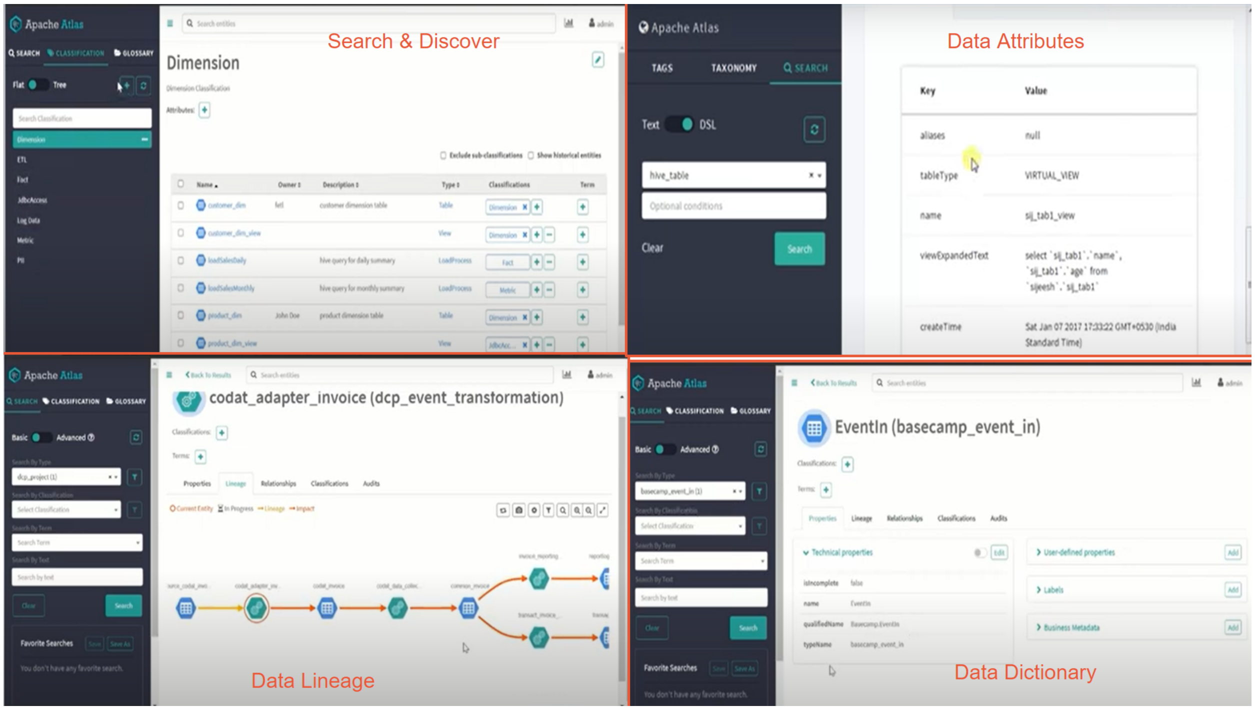


Figure 10. Screenshots of Collibra Data Catalog (Source)

Apache Atlas is an open-source metadata management and governance tool that enables organizations to build a catalog of data assets, classify and govern them. Some of its features are:

- Support for metadata capture from mostlyHadoop and some non-Hadoop data sources
- Ability to plugin new types of data sources to be supported
- Support for cloud based data sources are not available out of box
- Provides REST APIs to integrate with the tool from existing tools in the data landscape
- Ability to dynamically classify attributes like PII, data quality, sensitive, etc.
- Provides Intuitive UI to view data lineage across different steps of the pipeline
- User can search using free-text or SQL like query language
- Integrates with Apache Ranger to enable authorization and data masking
- Being open source, deployment and maintenance needs to be managed by the implementation team with community support

Below are some screenshots of the different features of the tool.


Search & Discover
Dimension
Data Attributes
codat_adapter_invoice (dcp_event_transformation)
Data Lineage
EventIn (basecamp_event_in)
Data Dictionary

Key	Value
aliases	null
tableType	VIRTUAL_VIEW
name	sij_tab1_view
viewExpandedText	select `sij_tab1`.`name`, `sij_tab1`.`age` from `sijeesh`.`sij_tab1`
createTime	Sat Jan 07 2017 17:33:22 GMT+0530 (India Standard Time)

Figure 11. Screenshots of Apache Atlas Data Catalog (Source)

DataHub is an open-source metadata platform for the modern data stack. It provides the following features:

- Provides pre-built integrations with common data sources, orchestrations and ML engines
- Also supports manually ingesting the schema metadata if required
- Supports adding tags, glossary terms and domains
- Provides lineage across platforms, datasets, ETL/ELT pipelines, charts, dashboards
- Provides profiling capability with stats on duplicates, outliers, data sampling, etc.
- Can automatically classify PII data
- Provides Dataset Usage & Query History
- All actions can be done via API
- Provides intuitive user interface to Search, Browse and View/Edit Metadata
- No collaboration features are supported
- It has been architected to manage continuously changing metadata
- It is an Open source project and support is dependent on community
- However, DataHub by AcrylData provides a managed service that has a lower-end pricing.

Below are some screenshots of the different capabilities of the tool.

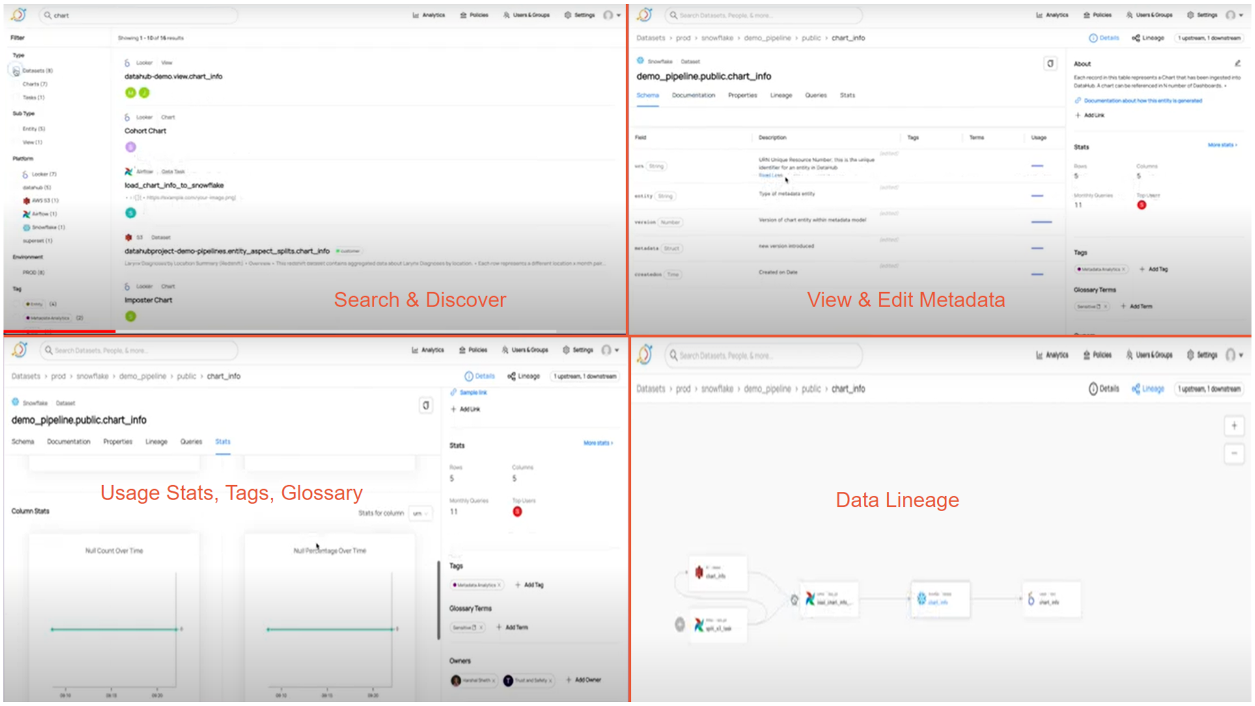
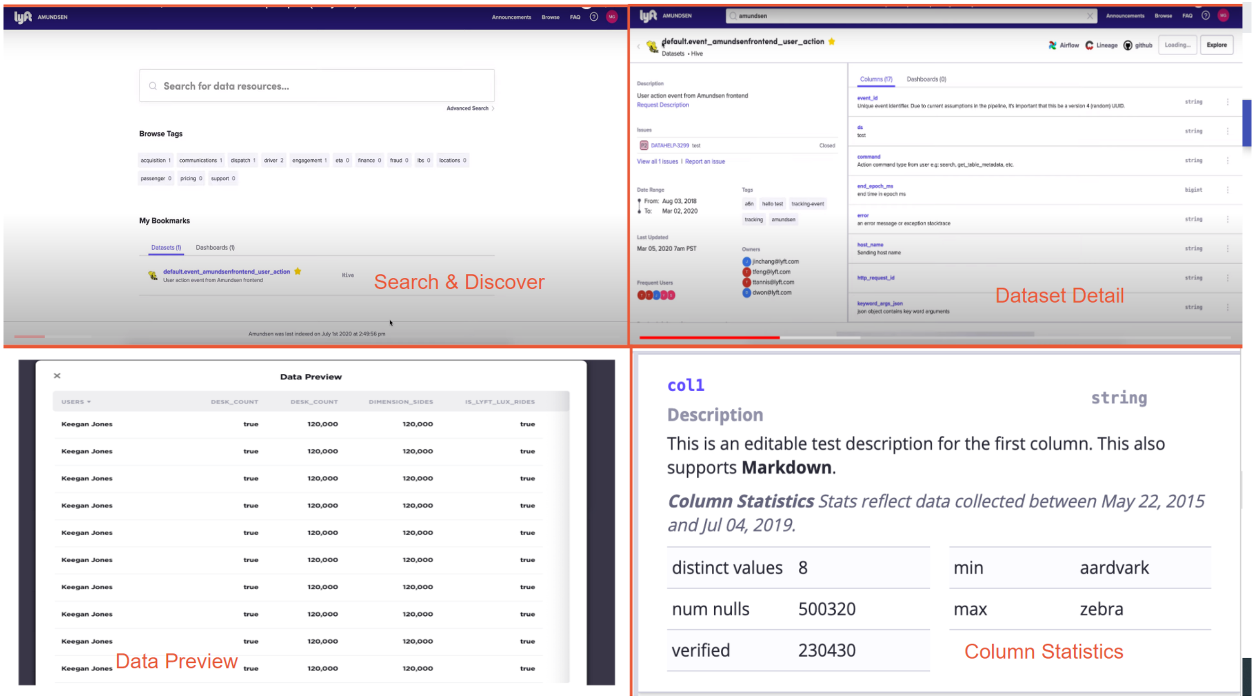


Figure 12. Screenshots of DataHub Data Catalog (Source)

Amundsen is an open source data discovery and metadata engine for improving the productivity of data analysts, data scientists and engineers when interacting with data. Following are some of the core features of the tool:

- Has integration with Hive, Presto, Postgres, Redshift, etc.
- It can index data resources (tables, dashboards, streams, etc.) that powers a page-rank style search based on usage patterns (e.g. highly queried tables show up earlier than less queried tables).
- Can integrate with Apache Atlas in the backend.
- It does not provide fine-grained access control.
- It is Open source, so dependent on community support. Hence deployment and maintenance needs to be managed by the implementation teams.
- Stemma.ai provides a managed service across all clouds that is priced on the lower-end.

Below are some screenshots of the different features of the tool.



Column	Description	Type
event_id	Unique event identifier. Due to current assumptions in the pipeline, it's important that this be a version 4 (random) UUID.	string
ds	test	string
command	Action command type from user e.g: search, get_table_metadata, etc.	string
end_epoch_ms	end time in epoch ms	bigint
error	an error message or exception stacktrace	string
host_name	Sending host name	string
http_request_id		string
keyword_args_json	json object contains key word arguments	string

Data Preview

USERS	DESK_COUNT	DESK_COUNT	DIMENSION_SIDES	IS_LYFT_LUX_RIDES
Keegan Jones	true	120,000	120,000	true
Keegan Jones	true	120,000	120,000	true
Keegan Jones	true	120,000	120,000	true
Keegan Jones	true	120,000	120,000	true
Keegan Jones	true	120,000	120,000	true
Keegan Jones	true	120,000	120,000	true
Keegan Jones	true	120,000	120,000	true
Keegan Jones	true	120,000	120,000	true
Keegan Jones	true	120,000	120,000	true
Keegan Jones	true	120,000	120,000	true

col1 string

Description

This is an editable test description for the first column. This also supports **Markdown**.

Column Statistics *Stats reflect data collected between May 22, 2015 and Jul 04, 2019.*

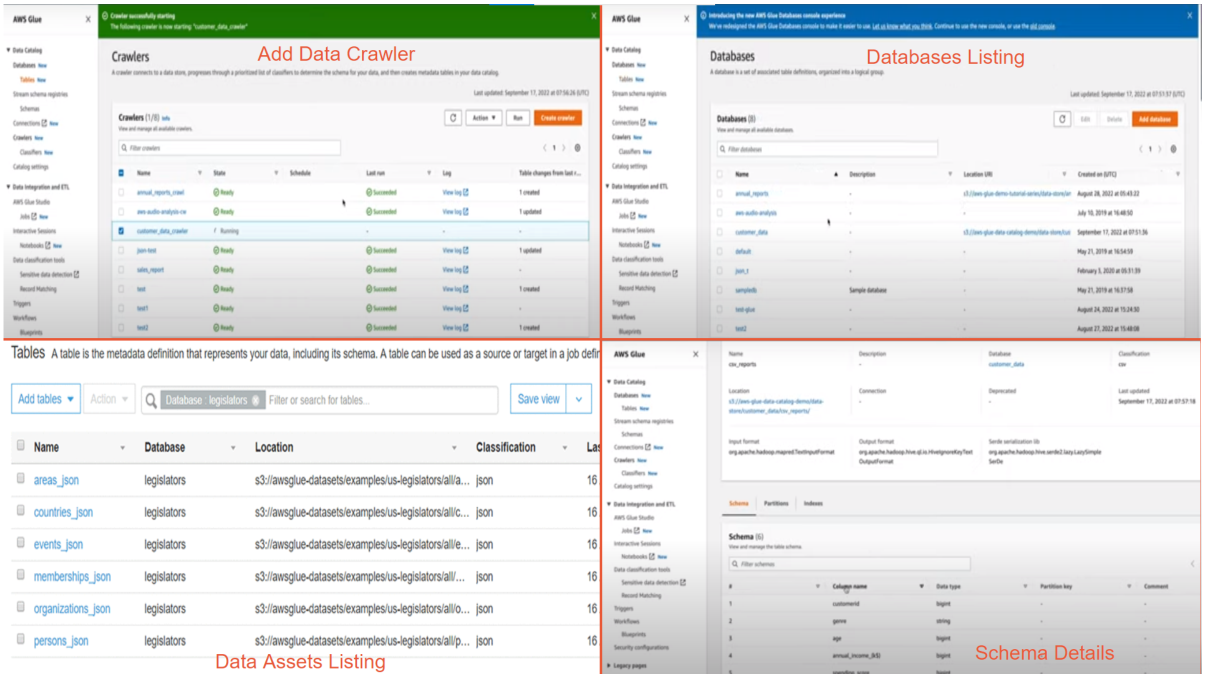
distinct values	8	min	aardvark
num nulls	500320	max	zebra
verified	230430		

Figure 13. Screenshots of Amundsen Data Catalog (Source)

AWS Glue Data Catalog is an AWS cloud native service that is used to store references to data that are used as sources and targets of the extract, transform, and load (ETL) jobs in AWS Glue. It provides the following features:

- Tracks runtime metrics, stores the indexes, locations of data, schemas, tracks ETL jobs on AWS Glue.
- Integrates with AWS services through built-in crawlers for S3, DynamoDB, Redshift, RDS, etc.
- However it only has Integrations limited to AWS services or publicly accessible databases
- It comes with a Schema Registry to manage and enforce schema on data streaming applications through integrations with Apache Kafka, MSK, Kinesis Data Streams, Kinesis Data Analytics and Lambda functions
- The metadata from Glue Data Catalog can be used by Amazon Athena to run SQL queries across data sources
- It is not a full-fledged governance tool. Needs other tools for access control, privacy, governance, visualization, etc.
- Pricing is **pay-per-use** based on crawlers, data catalog storage and consumption.

Below are some screenshots of the different features of the tool.



Crawler successfully starting
The following crawler is now starting: "customer_data_crawler"

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Last updated: September 17, 2022 at 07:56:26 (UTC)

Crawlers (1/8)
View and manage all available crawlers.

Name	State	Schedule	Last run	Log	Table changes from last r...
annual_reports_crawl	Ready		Succeeded	View log	1 created
aws-audio-analysis-cw	Ready		Succeeded	View log	1 updated
customer_data_crawler	Running		-	-	-
json-test	Ready		Succeeded	View log	1 updated
sales_report	Ready		Succeeded	View log	-
test	Ready		Succeeded	View log	1 created
test1	Ready		Succeeded	View log	-
test2	Ready		Succeeded	View log	1 created

Introducing the new AWS Glue Databases console experience
We've redesigned the AWS Glue Databases console to make it easier to use. Let us know what you think. Continue to use the new console, or use the old console.

Databases

A database is a set of associated table definitions, organized into a logical group.

Last updated: September 17, 2022 at 07:51:57 (UTC)

Databases (8)
View and manage all available databases.

Name	Description	Location URI	Created on (UTC)
annual_reports	-	s3://aws-glue-demo-tutorial-series/data-store/an	August 28, 2022 at 05:43:22
aws-audio-analysis	-	-	July 10, 2019 at 16:48:50
customer_data	-	s3://aws-glue-data-catalog-demo/data-store/cus	September 17, 2022 at 07:51:36
default	-	-	May 21, 2019 at 16:54:59
json_t	-	-	February 3, 2020 at 05:31:39
sampledb	Sample database	-	May 21, 2019 at 16:37:58
test-glue	-	-	August 24, 2022 at 15:24:30
test2	-	-	August 27, 2022 at 15:48:08

Tables A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job defir

Database : legislators Filter or search for tables...

Name	Database	Location	Classification	Las
areas_json	legislators	s3://awsglue-datasets/examples/us-legislators/all/a...	json	16
countries_json	legislators	s3://awsglue-datasets/examples/us-legislators/all/c...	json	16
events_json	legislators	s3://awsglue-datasets/examples/us-legislators/all/e...	json	16
memberships_json	legislators	s3://awsglue-datasets/examples/us-legislators/all/...	json	16
organizations_json	legislators	s3://awsglue-datasets/examples/us-legislators/all/o...	json	16
persons_json	legislators	s3://awsglue-datasets/examples/us-legislators/all/p...	json	16

Name: csv_reports
Description: -
Database: customer_data
Classification: csv
Location: s3://aws-glue-data-catalog-demo/data-store/customer_data/csv_reports/
Connection: -
Deprecated: -
Last updated: September 17, 2022 at 07:57:18
Input format: org.apache.hadoop.mapred.TextInputFormat
Output format: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Serde serialization lib: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe

Schema (6)
View and manage the table schema.

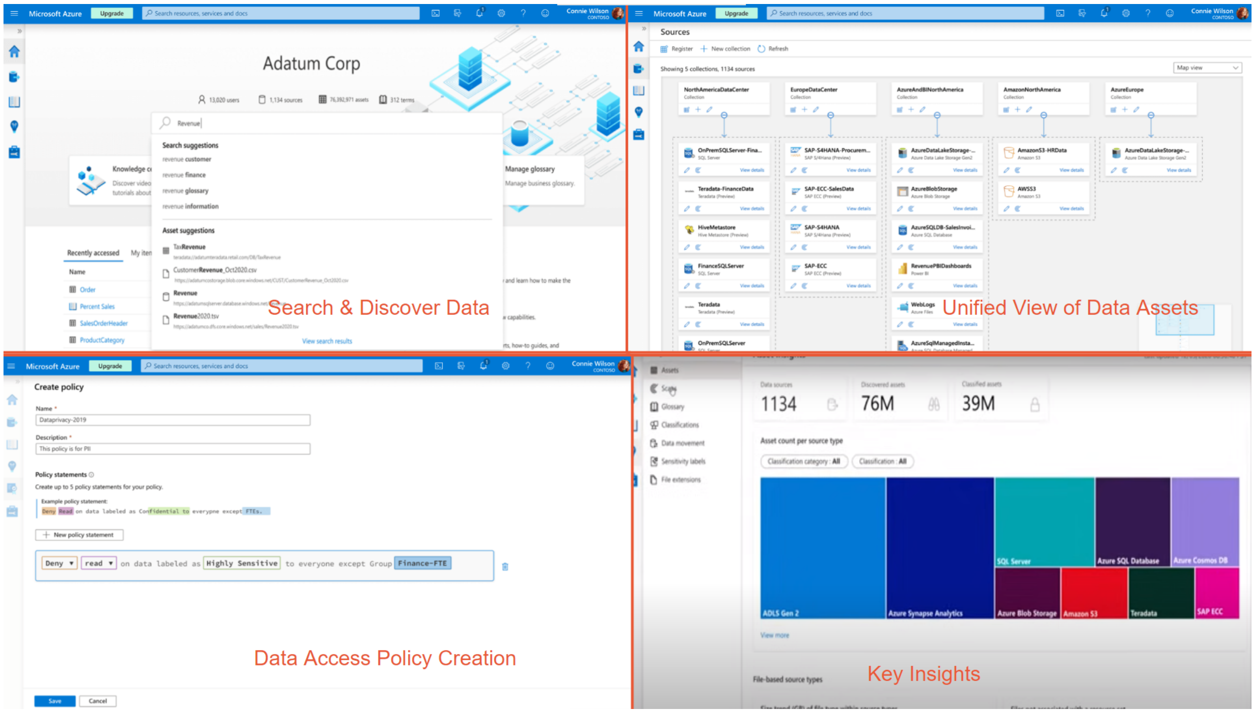
#	Column name	Data type	Partition key	Comment
1	customerid	bigint	-	-
2	genre	string	-	-
3	age	bigint	-	-
4	annual_income_(k$)	bigint	-	-

Figure 14. Screenshots of AWS Glue Data Catalog (Source)

Microsoft Purview is an Azure cloud native service that provides a unified data governance solution including automated data discovery, lineage identification and data classification, unified map of data, business glossary. It provides the following features:

- It integrates seamlessly with all Azure services to create a unified map of data across data landscape
- Provides a number of built-in and custom classifiers to identify and label sensitive data
- Provides a rich user interface for easily searching and discovering data
- Supports interactive data lineage visualization
- Provides insights on key health metrics
- Enable data engineers and owners to provision access to data assets
- Will need to integrate with other Azure services for data masking, data profiling, etc
- Pricing is **pay-per-use** based on data map population, enrichment and consumption.

Below are some screenshots of the different features of the tool.



Search & Discover Data

Unified View of Data Assets

Data Access Policy Creation

Key Insights

Figure 15. Screenshots of Microsoft Purview Data Catalog (Source)

Google Cloud Data Catalog is a fully managed Google Cloud native service that provides scalable metadata management as part of the Google Cloud Dataplex service. It has the following features:

- Provides a predicate-based search experience for discovering technical and business metadata
- Support auto-tagging of sensitive data using the DLP API integration
- Also allows adding user-generated tags for metadata to enrich its understandability
- Enables automatic catalog of assets across BigQuery datasets, Pub/Sub topics, Dataplex lakes, Analytics Hub linked datasets and Dataproc datasets
- It can also catalog select non-GCP data assets like Hive, RDBMS, Teradata, Redshift, Looker, Tableau, etc.
- Provides access control for viewing the metadata content through permissions
- Pricing is **pay-per-use** based on consumption

Below are some screenshots of the different features of the tool.

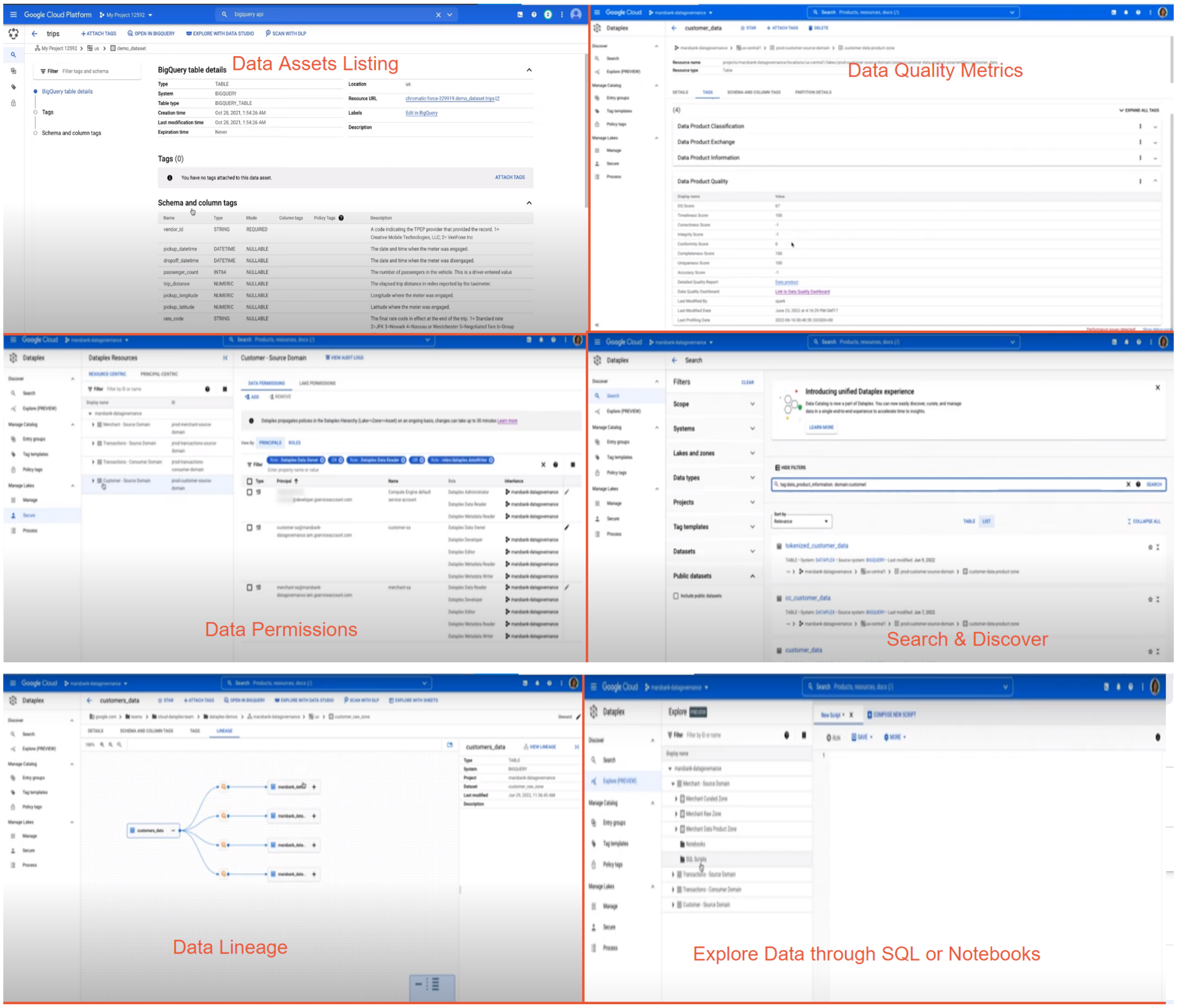


Figure 16. Screenshots of Google Cloud Data Catalog (*Source*)

Criteria for Evaluating Data Catalog Tools for an Enterprise

As seen in the previous section, there exist a large number of tools that provide the data catalog capability. This can create confusion in the minds of users who are looking to implement Data Catalog in their organization. The suitability of tools can vary on a case-to-case basis and all criteria need to be considered before making an educated choice. In this section we provide the important criteria to be considered while evaluating a Data Catalog tool for an Organization. A detailed criteria is documented and available for reference on this page.

Following are some important criterias that need to be considered for evaluating an enterprise data catalog.

- Support for Data Sources
- Ability to easily Search & Discover data assets
- Support for Data Lineage & Quality
- Support for Collaboration
- Programmatic Access to Features
- Automated insights and recommendations for optimizations
- Deployment Options
- Pricing and Support

Under each of the core criteria, following are the detailed features that the data catalog tool should be able to support.

Support for Data Sources:

- Support for common data source types across cloud, on-premise through out-of-box connectors and crawlers.
- Support for a marketplace, that allows integration with 3rd party connectors for data sources not supported out of box
- Ability to add or configure a custom data source that is not already supported.
- Ability to auto discover a schema from an added data source and support for manual/automated re-discovery and schema changes.
- Ability to manually import a schema where connectivity to a data source is not possible.
- Support for “alternate” Data Source mechanisms such as Messaging (e.g. Kafka Topics) and/or APIs (e.g. REST APIs).

Ability to easily Search & Discover data assets:

- Ability to search for data assets using free-text or domain specific language through an intuitive user interface.
- Support for auto tagging of data assets in terms of privacy or sensitivity
- Ability to manually tag schema using built-in tagging standard or custom naming.
- Support for searching based on schemas and tags.
- Support for adding different types of Annotation such as friendly name, description, documentation, ownership assignment

Support for Data Lineage & Quality:

- Support for end-to-end data lineage using out of the box capabilities and identify what is not supported e.g. movement within data sources but not across data sources.
- Support for data sampling capabilities e.g. view a sample set of data from a supported data source.
- Support for data syntax check for structured data schemes such as Avro, Parquet, JSON, XSD, etc, stored in a cloud object storage or document database.
- Support for data profiling and quality capabilities e.g. format, null, range, etc. issue detection.
- Ability to add custom Business Rules on data for validation
- Ability to capture data usage metrics such as schemas/tables which are highly or rarely used

Support for Collaboration:

- Support for collaborative documentation capabilities and how these would be integrated with other documentation platforms such as confluence.
- Support for any collaborative ticketing/workflow capabilities and the philosophy of how these would be used alongside / integrated-with other ticketing/workflow platforms such as Jira.

Programmatic Access to Features:

- Ability to perform through API everything that can be done in the UI.
- Good documentation around the available APIs by category.
- Support for API testing through tools such as Postman.

Automated Insights and Recommendations:

- Support for insights and recommendations wrt data optimization and standardization automatically through AI & ML algorithms

.

Deployment Options:

- Deployment options available for the solution from a On-premise / Cloud PaaS / Marketplace perspective and what Cloud providers it is available on.

Pricing and Support:

- Evaluate the cost areas of the solution i.e. connectors, creator/editor users, viewer users, etc.
- Identify and detail licensing aspects such as floating/named users.
- Availability of rich user guides and demo videos detailing the features and usage of the tool
- Availability of post-sale activities e.g. setup, training and support channels.
- Identify non supported features or features that require paid updates.

In addition to the evaluation of the different data catalog tools across the criteria shared above, the implementation team should also take into consideration customer preferences based on existing relationships with the product vendors to arrive at a decision on choosing the most applicable tool.

Conclusion

Data Catalog is an important capability that helps data scientists and business users quickly discover and access data assets spread across the organization. There are a number of tools available including new cloud-only solutions that are easy to integrate and provide a whole set of features. The implementation team needs to carefully evaluate the tools based on the criteria that is most important for the organization and decide the right tool.

In the experience of this author, while a number of organizations see the benefits of implementing a data catalog it is still something that organizations see as a “Good to Have” capability instead of a “Must Have”. Hence they tend to postpone its implementation. High initial cost and subsequent maintenance cost of some of the established vendors is one reason for it. But there are a number of open source tools and cloud native tools that provide a number of capabilities at a much lower cost. With the explosion of data and creation of data silos in most organizations, data catalog implementation should be considered a “Must Have” capability that will greatly benefit organizations in ways documented in the earlier sections.

Below are some concluding thoughts from an implementation perspective:

- The data catalog should be able to have seamless and out-of-box integration from different producing applications to consuming applications.
- A business user or data scientist should be able to discover and access the data from any collaboration or analytical environment.
- ML capabilities should be leveraged as much as possible to automatically curate the metadata and reduce manual effort.
- Start cataloging with highly used/accessed data assets and cover all the data in a systematic way.
- Creation of business metadata will be mostly manual by knowledgeable business analysts and business users.
- Finding business resources who are willing to spend time, effort and have motivation to curate and define business metadata can be a challenge.
- Data Stewardship should be enabled to ensure that high-quality metadata is being created and maintained.
- Set Up an enterprise data governance council team that is responsible for governing, maintaining and curation of cross-functional metadata definitions & usage policies.

References

- Amundsen Data Catalog: Lyft's Open Source Discovery Tool
- Open Source Data Catalog Software: 5 Popular Tools in 2023
- AWS Glue Data Catalog: Architecture, Components, Crawlers
- DataHub: Popular metadata architectures explained | LinkedIn Engineering
- What is Data Catalog? Definition, Benefits and Use Cases
- What Features Do You Need in A Successful Data Catalog? | Alation
- Data Catalog Implementation: Strategy & Advice | Alation
- Data Catalogs — Unlocking Value in your Data Lakes | by Sajjad Syed | Medium