Global Practices Big Data & Analytics Practice

Overview of Data Catalog Capability

December 2022 Arun Viswanathan



Intended Audience	2
Context	3
Problem Statement	3
Taxonomy	4
Scoping within the Enterprise	6
Scoping within the Data Platform	7
Benefits with using a Data Catalog	8
Solution	9
Core Features of an Enterprise Data Catalog	9
Solution Approach	10
An Overview of the Data Catalog Tools	13
Criteria for Evaluating Data Catalog Tools for an Enterprise	22
Conclusion	24
References	25

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 <u>Taxonomy</u> 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.



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 <u>here</u>.

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

Snowflake v Q part size	© ↑? ©	Variable	Туре	Description	
🏶 sample-data 🗸 🛛 All Database 🗸 All Schema 🗸		ELEMENT	Character	Periodic Table id if inorganic or radi	Customer Acquisition Cost 🔮
③ All 41		ELIMINAT	Character	Eliminate analyte based on screen?	Overview Linked Assets
⇔ # PART SIZE 💁	eg.	EVALTYPE	Character	Evaluation Type (Qual./Quant.)	≣ Readme 🕑 Ec
Column # NUMBER @ PART_SUPPLIERS @ LINEAGE @ PUBLIC	-0	EXTRISK	Numerical	External Exposure risk after 1-hit ru	
		FILTERED	Character	Filtered Sample? (YES/NO)	Simple method
Column # NUMBER # PART_SUPPLIERS @ LINEAGE @ STAGING		FISHBTF	Numerical	Bio-transfer factor for fish	Ine simple method divides the total marketing costs to acquire new customers by the tal number of customers acquired in a defined period.
		FOODUSE	Character	Use toxicity value for food?	$CAC = rac{MCC}{CA}$
		FREQ_DET	Character	Freq. of Detection	CAC = Customer Acquisition Cost MCC = total marketing cost for acquiring customers (not regular customers)
E CORINI # HOWPER E DIMENNE & ENERGE & LODIO		GCCDI	Numerical	Ingestion carcinogenic CDI	CA = total customers acquired
泰 # SIZE	٥G	GHQ	Numerical	Ingestion hazard quotient	Complex method
Column # NUMBER STG_TPCH_PARTS LINEAGE PUBLIC_STAGING	3	GHQ A	Numerical	Ingestion hazard quotient for an adu	In addition to the costs incurred in marketing, the complex method includes sales and arketing wages, software costs for sales and marketing, all additional professional serv
恭 井 SIZE		GHQ C	Numerical	Ingestion hazard quotient for a child	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 <u>this</u> 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.



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.



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 <u>Collibra</u> 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 lngest 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

<u>Informatica</u> 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 Al-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.



Figure 8. Screenshots of Informatica Data Catalog (Source)

<u>Alation</u> 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.



Figure 9. Screenshots of Alation Data Catalog (Source)

<u>Collibra</u> 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

	Cellibra for Desktop		🍰 🕜 Broose	Q Seath								+	ë ™ 1 0 @ #		
Q	chum	à Baires Andres Cornaria	+ &Schemen + () CBN Goud Refined												
Search	All Business Context. Business Term: Catalog: Certified Data Sets. 1975. Line of Businesses. Logical Layer: Physical Data. Privacy categories and pr	licles Reports I	Custome Productales												
Actions	Chum Prediction (%)	Ý- Would you like to filter by Business Term?	Add characteristic												
3⊟ Testa	Gargenal Name Chum Presicion (n) Business Analysis Community ···· Ti ML_Analysis_Data		Summary	Add to Outs Set									A4 1		
			g beas	Name #	Bata Classification	Logical Duta Attribute	Technical Duta Type	Rew Count	Empty Values Count	Last modified an	Status	PI			
	Total Chum		Columns	AddressLine2			WOM.	60,398	55,293	4/23/2022 2:51 PM	Candidate	1	1		
	Original Name Total Chum		뀸 Sample data		0.01 994	Birth Date	DIES	62,398	0	4/23/0622 2:51 PM	Candidate		:		
	a Business Analysis community +			City City			WOM	\$0.335	21.515	6/7/2022 4:54 PM	Candidate	- 1			
	TBA Churp Rate		<c depan<="" td=""><td>Commuta Ostance</td><td></td><td></td><td>WACKIR</td><td>60.398</td><td></td><td>6/7/2022 4:04 PM</td><td>Candidate</td><td>*</td><td>1</td></c>	Commuta Ostance			WACKIR	60.398		6/7/2022 4:04 PM	Candidate	*	1		
			B Pores	CountryRegionCode	Cauntry code 75%		TACKA	60,398	0	6/7/2022 4:04 PM	Candidate		1		
	Calculation Rule (Distinct amount of churning customers) / COUNTDI(Customer Keyl)		& Technical Larange	Currency/Ay			NT	62,398	0	6/7/2022 4:64 PM	Candidate		1		
	a Business Analysts community + 😐 + 🖬 Customer Churn Insights			Customerkitemaneliay	Curtomeniey 2		WACHINE	\$2,398		6/7/2022 4:564 PM	Candidate				
	Churn Bate		@ Quality	Custome Rey			NT	\$2,398	8	6/7/2522 4:54 PM	Candidate		1		
			a summittee	Custome #ONumber			WACHER	\$2,398	62,218	6/7/2022 4564 PM	Candidate		1		
	Definition Customer chum rate is the percentage of your customers or subscribers who cancel		A separation of	Deterioshathate	Gag 576		DAGE	60,398	0	6/7/2022 4:04 PM	Candidate	к.	1		
			30 References	DealerPrice			DECIMAL	\$2,398	0	6/7/2022 4:64 PM	Candidate	к	1		
	Churre Date		О нахиу	Discount/wount			REN.	62,398	ho	6/7/2022 4564 PM	Candidate		1		
	Churn Race		e 14	Durbine	Date time 195%		TINESDAMP	60,398	0	673,2522,4364 PM	Candidate		1		
	Definition Churn rate (sometimes called attrition rate), in its broadest sense, is a measure Churn rate is an input into customer lifetime value modeling, i	nd can be part of	y	DurDutationy			NT	60,398 0	0	6/7/2022 4566 PM	Candidate	×	1		
	Departments Digital Marketing Glossary Search & DISCOVER			Enal4dtres	Bral 🗸	Email Address	WACKAR	60,398	0	4/29/2022 11 37 AM	Accepted	1	1		
8				C EndDate	Dan time dan		TRUESDARP	\$2,358	54,970	35/2021 10:17 AM	Candidate		1		
User	Chum Reasons			English@escription	Occupation 5%		TACAN	62,398	0	35/2021 10:17 AM	Candidate	×	1		
🛞 🔘 Bre	Ra Q Servit	+ + 1ena + 🗘 🕲 🗐 🗄	🔅 🕜 Browse									+	💥 Taola 🔹 🛟 🕐 🗒 🏭		
à Buines Arajos	Immunity + dischemic + ISOH (Boot Refree)		A Business Analysis Community	+ AScheman + Charpefain											
Custo	merProductSales	The Addition Designation of Additional Addita Additiona Additional Additional Additional Additional	Curtomark												
	Cases 01.01 2	U Add to Luna Set	Customerk	ey Mar ar Da h									It add to Data Ser. Articles a		
Add thankholder		frame letting		01.01	-								IT FOR IN OWN JAK		
d Summary		D Gard	Add characteristic	SHS Snowflake Software + @sno	afaie-db + 10xmvfake + 0	OutomerSalesReporting + 1	Oustomer Key								
S Desirs		+ AT GALLEDUCTS													
8 CO.8755		DATEBASE	Summery	•									Data Owner @		
E Sener des		RAW	S Deals	Sample data	D	tribulion .							🔒 Joana Zhou		
		- E 090	h pertorne	12961											
+d Dagan		CUSTOMERPRODUCTSALES	PC Converse	13274	44								Data Sets		
B nave		> UNUSED > Stats		15768		2178 2129	1811 1911						Castomer Analytics ()		
A Technical Li	Ngt		*# Dagian	11420	24		1710	1611 1107	1540 1400 1340	1389 1343	165 117 11	10 100	Data set containing information on our internet.		
	tee terrowishing the		B Pictures	13558									sales combined with relevant customer and product data.		
Corrig	BOD ANTERNA DAVAL BOD CATEBORDOUTSALES REVAILS - DOD PROCEETING STORE ANTONI		A Technical Linear	23441	0	10			14				See all (1)		
A hepenible			1 concernenti	21997		118 [16]		-14	vot 10	14					
3 triences	Provide a subject of the provide a subject of		C ante	15878	Data Wh	ole Number	int	of Data Type		false					
O Hitty			V (and)	21087											
8 1m			A* Responsibilities	19532	few 60	198	0 (D.	00%)		18484	r en vel				
	Deta Lineare		3 References				. (
	Data Lineage		0.000							Data I	Profil	ind			
			C HISNY							Jala		ing			
			& Res												

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

Figure 10. Screenshots of Collibra Data Catalog (Source)

<u>Apache Atlas</u> 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

🚯 Apache Atlas	Search & D	iscover	Apache Atlas	Data At	Data Attributes				
Q SEARCH	Dimension		TAGS TAXONOMY Q SEARCH						
Flat Tree	Démension Classification			Кеу	Value				
Search Classification Demonster	Allowing 💌	Exclude sub-classifications Show bistorical antities	Text ODSL	aliases	null				
En. Fied	Name Owner 1 Description 1	Type D Classifications Term	hive_table × +	tableType	VIRTUAL_VIEW				
Mochanes	C customer_dim fet customer dimension table	Table Dimension X + +	Optional conditions	0.5756	sii tah1 view				
Log Duta Metric	C 🕘 customer_dim_view	Vev Dimension X + - +				i i			
ы	loadSalesDaily Nive query for daily summary	LandProcess Fact + - +	Ciear Search	viewExpandedText	select 'sij_tab1', 'name', 'sii tab1', 'age' from				
	Control Contro Control Control Control Control Control Control Control Control Co	Loudhooms Metric 🗣 🖛 🔶			'sijeesh'.'sij_tab1'				
	 product_den John Doe product dimension table 	Table Dimension X +		createTime	Sat. Jan 07 2017 17:33:22 GMT+0530 (India				
	C 😡 product_dim_view	Vew AbsAsc X + - +			Standard Time)	L.			
🕞 Apache Atlas	C Seach entities	🔝 🛔 🛔 admin	🕞 Apache Atlas 🔭 🗉 Keak lo Reula 🔍 S	each enlides	6	Al Antrin			
	Codat_adapter_invoice (dcp_eve	nt_transformation)	Q SEARCH & CLASSIFICATION & GLOSSARY						
Basic 🔵 Advanced 🕲 😰	Classifications:		Eventin	(basecamp_event	_in)				
Search by Type	Terma: 🔶		Cassification: +						
Search by Camilication	Properties Lineage Relationships Classifications Audits		basecang_event_in (i) 🗙 🔹 🝸 Terms: 🔶						
Select Classification • T	O Current Entity ≚ in Progress →Lineage → Impact		Search By Classification Select Classification V V	ielationships Classifications Audits					
Search Term •		mout, sporting . separate	Search by Technical properties	() Edt	> User-defined properties	Add			
Search by lest	arra, pola, mo. colal, adapter jeu. colal, motor colal, ada, polac		Search by Tot Intromplete false		N Labola				
dur Sech		Junat more tana	Search by toot name Evention		7 6000				
Favorite Searches Sever As			Corr Search qualified/iame Basecamp_liver	ith et_in	Business Metadata	Add			
You don't have any favorite search.	Data Lineage		Favorite Searches Som Same As	Data D	Dictionary				

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

<u>DataHub</u> 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.

Q chart		j⊴ Analytica 🚖 Policies 🔅 Users&Groups 🔅 Settings () +	Ø Q Search Dataset		Let Analys	cs ≙ Policies (); Users&Groups (); Settings ()) +
Filter	Driving 1 - 10 of 54 results		Datasets > prod > snowfl	ike > demo_pipeline > public > chart_info		Octails of Lineage Tupsteam, I downstream
Type Dutasets (8) Charts (7) Teeks (1) Sub Type	S Lotter View defailshub-demousleeuchant_iteto Φ δ Lotter Chart		forwhite Dataset demo_pipeline.public Schema Documentation	Lohart_info Properties Lineage Queries Stats	٥	Abox C Garrisophin the falls represent a Chart that has been registed and Database A shart can be efferenced in N motion of Databases + Observation abox how the entity is generated + Antium
Entite (3) View (13 Pattlem Š (cotext (7) estanus (3) # Antion (1) Antion (1)	Orbot Court Court Court Version - Sea Sea Sea Court Links, p. proventian - (-) - High searching court of a mage (H) - O		Ped ves.(tong) estity_(tong) version_Number;	Description URL bitlings Resource Number, this is the unity as describe the an entry is Described Toget of restances with a setting Version of other entry and/or nestance method	Tops Terms Usage	Stats More Radio + Revers Columns 5
Socialize (1) Express(1) Express(1) Express(1) PROD (8) Tag extens (k) extens (k) (2)	at an appropriet device specifies and the s	n & Discover	netalete (Bout) createlles (Bout)	ee write intoducer entropy ent	– – lit Metadata	Tags •vendes indexs *) + Add Tag Gossery Terms (means *) + Add Tag
Q Search Datasets Datasets > prod > snowflak Socwtax Dataset demo_pipeline.public.	hreps (hreps,)))))))))))))))))))	Lint Analysis № Pulsosis № Pulsosis № Demote Directory ① Demote Directory ② Demote Directory ③ Demote Directory ③ Demote Directory ③ Demote Directory ④ Demote Directory	Q Search Datasets Datasets > prod > snowfs	Despie & more	la Anayo	cs 🖄 Petrose (k) Uners Giorupe (f) sectorps (f) v (f) Details off Lineage (Tupetners, Tournetween
Schema Documentation	hteerine Leege Geree Ban Usage Stats, Tags, Glossary Data tree	State More datas Row Gitures S S More shares Taxime 11 •		Data Li	neage	
		Top: ••••••••••••••••••••••••••••••••••••	• •	• *===	0117 6 II	

Figure 12. Screenshots of DataHub Data Catalog (Source)

<u>Amundsen</u> 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.

IyA MUNCSEN			·	Announcements Browse FAQ (9 🙆 📕	IYR AMUNOSEN	🔍 arrundsen				Announcements	Brewsø FAQ 🕤 🔤
					<	Catasets - Hive	ndsenfrontend_user	_action 🔶			≷ Airflow 🕻 Lineege 🛞 github	Loading
	Q Search for data resour	rces				Jescription	netrod		Columns (17) Deshboards	0)		
			Advas	need Search >		Request Description	orixeria		event, ld Unique event identifier. Due to current assumptions in the pipeline, it's important that this be a version 4 (andorr) UUD.			string ;
	Browse Tags				1			Orest	da tent			string :
	acquisition 1 communications 1 dispatch 1	triver 2 engagement 1 eta 0	finance 0 fitsud 0 lits 0 locations	0	v	Aew all 1 issues 1 Report an issue			command Action command type from user e.g.	eerch, get_table_wetedate, etc.		string
	passenger 0 pricing 0 support 0				D	Date Range	Tegs		end_epoch_ms end time is epoch ms			bigint :
	My Bookmarks				I	From: Aug 03, 2018 To: Mar 02, 2020	atin hellosest ba backing amundsen	cking-event	error an error message or exception stack	ince		string
	Datasets (1) Deshboards (1)					.est Updated Mar 05, 2020 7am PST	Owners		host_name			string :
	👷 default.event_amundsonfrontend_us	er_action 🚖 Kim		Di			jinchang@lyft.com		Sording hod name			
	 User action event from Amundaen frontend 		Search &	Discover		Prequent Users	tznis8lyft.com dwon@lyft.com		have and here	Da	itaset Detail	
					-				joan object contains key word engurn	ints		string :
	Amund	aen was lest indexed on July 1st 2020	et 2-49:56 pm									
×		Data Preview				col 1						
USERS +	DESK_COUNT	DESK_COUNT	DIMENSION_SIDES	IS_LYFT_LUX_RIDES		Deserie	tion				string	
Keegan Jones	true	120,000	120,000	true		Descrip	tion					
Keegan Jones	true	120,000	120,000	true		This is a	n editabl	e test (description	or the firs	t column. This als	0
Keegan Jones	true	120,000	120,000	true		support		OWII.				
Keegan Jones	true	120,000	120,000	true		Column	Statisti	cs Stat	<i>'s reflect dat</i>	a collected	d between May 22	2, 2015
Keegan Jones	true	120,000	120,000	true		and Jul (04, 2019.					
Keegan Jones	true	120,000	120,000	true		distinct	values	8		min	aardvark	
Keegan Jones	true	120,000	120,000	true		aistinet	, and es	•			aaravant	
Keegan Jones	true	120,000	120,000	true		num nu	ills	50032	D	max	zebra	
Keegan Jones	an Jones true 120,000 120,000 true					verified		23043	n	Colum	n Statistics	
Keegan Jones Da	ata Preview 👞	120,000	120,000	true		vernieu		25045	0	Colum		

Figure 13. Screenshots of Amundsen Data Catalog (Source)

<u>AWS Glue Data Catalog</u> 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.

AWS Glue X	Crawler succ The following	esofully starting crawler is now starting: "custo	ne da coie					× AWS Glue ×	© introdu We've o	ucing the new XIN'S Glue Databases consule experien redesigned the XIN'S Glue Databases console to make I	en It easier to use. <u>Let ut know what you i</u> t	ing. Continue to use the	new consols, or use the gid corpole		×
Outs Catalog Databases New Tables New	Crawle A crawler cor	Y TS svects to a data store, progress	es through a priorit	Add Data	a Crawler	adata tables in your data	assing.	♥ Duta Catalog Databases New Tables New	Dat	tabases base is a set of associated table definitions, organized	Databa	ases L	isting		
Stream schema registries						Let	t updated. September 17, 2022 at 07.56.26 (UTC)	Stream schema registries					6	ot updated. Septem	uer 17, 2022 et 07:53:37 (UTC)
Schemas Connections 🔁 New	Crawler View and it	rs (1/8) tale serage all available conviers.				C Ad	ion V Run Crash crawler	Schemas Connections 🕑 🏎	Da	rtabases (8) a und manage al availatie decilianes.			٥	tilt bie	Add database
Classifiers New	Q, Filter	r crawlers					(1) 0	Classifiers here	0	Filter detaileses					(1)0
Catalog settings		une v	State	Y Schedule	Last run	T Les	Table changes from last r	Catalog settings		No.	Description	T Local	in UR	Constant are \$17723	
• Data Integration and ETL		here shows have	() teaty		(A factored at	View last 68	1 created	• Data Integration and ETL		and much	and down	- 100M	nen el se deren la testal series llates stran lar	Amount 24, 2022 at	
AWS Glue Studio		en auto andeix ce	() Inste		• O termeted	View last 68	Lundated	AWS Glue Studio		an ada antaŭ				1.4. 15 1018 of 16	48.50
Jobs 🔄 New		utures data conder	(Russian					Jobs C New		colore data		0.00	an also dan circles dana Man dan bur	feetender 12 303	
Notebooks 🖉 🚥			Obut		Atomstel	March 10	Luckted	Notebooks 🖉 New		and a second sec		12//4	no for one cand services motion	Sepamole 17, 200	
Data classification tools			Oreany		() sources	rev of D	1 spore	Data classification tools		Brook				May 21, 2019 at 14	5459
Sensitive data detection 🗄		een_report	(C Ready		(c) Succeeded	view top US		Sensitive data detection 🖉		Jan J				February 3, 2020 a	053139
Record Matching	0 1	et	() Ready		(a) Succeeded	View log 🖉	1 created	Record Mutching Triaters		sampledb	Sample database			May 21, 2019 at 16	3758
Workflows	0 1	est1	() Ready		(c) Succeeded	View log 🕑		Workfews		test-glue				August 24, 2022 at	15.24.50
Bueprints	0 1	est2	@Ready		@ Succeeded	View log 🕑	1 created	Buspints		text2				August 27, 2022 at	15-48-08
Add tables Add tables	ne metadat	a definition that i	represents	your data, including i	its schema. A table can	be used as a	Save view v	** Data Catalog Data Catalog Databases New Tables New Stream scheme registriles	¢	Name car, reports Location s3//ms-glue-data-catalog-demo/data- stam/uutamer_data/csr.reports/	Description - Connection		Database customer_data Degrecated		Casofication CNV Lett updated September 17, 2022 at 07:57:18
Name	Ŧ	Database	٣	Location		~ C	classification - I	Schemas Connections 🖾 New Crawlers New		Input format org.apache.hadoop.mapred.TextInputFormat	Output format org.apache.hadoop.hive.qLi OutputFormat).HivelgnoreKeyText	Serde serialization lib org.apache.hadoop.hive.serde2.lazy SerDe	LazySimple	
areas_json		legislators		s3://awsglue-datase	ets/examples/us-legisla	tors/all/a js	son	Classifiers New Catalog settings							
countries_json		legislators		s3://awsglue-datase	ets/examples/us-legisla	ors/all/c js	son	6 * Data Integration and ETL ARIS Glue Studio		Scheme Partitions Indexes					
events_json		legislators		s3://awsglue-datase	ets/examples/us-legisla	tors/all/e js	son	105 (2 New Interactive Sessions Notebooks (2 New		Schema (6) View and manage the table schema.					
memberships_j	son	legislators		s3://awsglue-datase	ets/examples/us-legisla	tors/all/ js	son	6 Data classification tools Sensitive data detection 🗈		Q, Filter schemes	lago name 🛛 🕷	Data type	v Partition key		v Comment
organizations_j	son	legislators		s3://awsglue-datase	ets/examples/us-legisla	tors/all/o js	son	6 Triggers Workflows		1 cm 2 pm	tomerid me	bigint string			
persons_json		legislators	Det	s3://awsglue-datase	ets/examples/us-legisla	tors/all/p js	son	6 Biusprints Security configurations		3 ap	and income, MSI	bigint bigint	Schema De	etails	
			Data	a Assets I	lisung			Legacy pages		1	infor som	history			

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

<u>Microsoft Purview</u> 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.

=	Microsoft Azure Upgrade 🕫	Search resources, services and docs		<i>C</i> .	9 ? 💮 Connie Wilson 🔮	≡ Mice	rosoft Azure Upgrade	${\cal P}$ Search resources, services and docs		5 R	🔓 🐵 🖓 😳 Connie Wilson 🍪
Э					1.	⇒ So	ources				
A					100	†	Register + New collection () Refresh			
			Adatum Corp		~	😰 sh	owing 5 collections, 1134 sources				Map view 🗸 🗸
		0 12000	C 1134 courses		10 - 5.0-	80	NorthAmericaDataCenter Collection	EuropeDataCenter Collection	AzureAndBiNorthAmerica Collection	AmazonNorthAmerica Collection	AzureEurope Collection
		0 - 1			2 - A A S		# + / o	# + / e	# + / o	# + / o	w + / o
•) Revenue				8	,		,,	,,	
		Search suggestions					CoPremSQLServer-Fina.	SAP-SAHANA-Procurem	AzureDataLakeStorage	AmazonS3-HRData	AzureDataLakeStorage Azure Data Lake Storage Gen2
	Knowledge ci	revenue finance			Manage glossary		I € Version	n / E Vendersin	✓ € Ver brais	✓ € Ver deals	2 C Ventests
	tutorials about	revenue glossary			manage ousiness glossary.			SAP-ECC-SalesData SAP-ECC (Preview)	AzureBlobStorage Azure Blob Storage	AWSS3 Amazon SJ	
		revenue information			~		↓ € Vew Sets	in 🦯 🐔 Veri detain	I C Veu details	/ C Ver deals	
		Asset suggestions					Wive Metastore (Preview)	SAP-S4HANA SAP 5/4/tana (Preview)	Azure SQL DB-SalesInvol Azure SQL Database		
	Recently accessed My iter	teradata.//adatumteradata.retail.com/01	8) Tanfeverue				Ver det		Ver Stals		
	Name	CustomerRevenue_Oct2020.cs https://adatumcastorage.blob.come.win	9V vdows.net/CUST/CustomerRevenue_Dct2020.csv		r and learn how to make the		SQL Server	Sub ECC (Preview)	Power Bl		
	Order Percent Sales	Revenue https://adatumoglasmer.database.winds	Sparch & Discou	vor Data			ven Teradata		weblogs Inifi	ad View of	Data Assots
	SalesOrderHeader	Revenue2020.tsv Intps://adatumca.db.com.windows.net/		ver Dala	w capabilities.		Teradata (Preview)		Accertion UTITI		Dala Assels
	ProductCategory		View search results		rts, how-to guides, and		CnPremSQLServer		AzureSqlManagedinsta		
= 1	licrosoft Azure Upgrade 😥	Search resources, services and docs			O Connie Wilson Connie Wilson		Isets				ente alla ser e la subsección de serve e serve e subsección de serve e s
20	Create policy					C ×	1	Data sources Discove	red assets Classified	85465	
n	Name *					00	lossary	1134 🕒 76	M 🙆 39N	A A	
8	Dataprivacy-2019					1 920	assifications				
	This policy is for PII					6.0	ata movement	Asset count per source type			
Ŷ	Policy statements ③					B ×	insitivity labels	Classification category : All Classific	ation : All		
	Create up to 5 policy statements for your po	slicy.				1 0 "	e extensions				
8	Deny Read on data labeled as Conf	idential to everypne except FT	Ts.								
	+ New policy statement										
	Denv x read x on data	laboled as Highly Sens	to execute except Group Fin	ADCO-ETE						SOL Server	Azure SQL Database Azure Cosmos DB
	beny - Fead - on data	cabeted as mighty sen	to everyone except on oup		8						
								ADLS Gen 2	Azure Synapse Analytics	Azure Blob Storage Amazon	s S3 Teradata SAP ECC
								View more			
			Data Access Poli	cy Creatio	n			12	and the starts to		
							Fil	e-based source types	ey insights		
	Save Cancel							fire based of the of the base within strategy		December of the state	

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

<u>Google Cloud Data Catalog</u> 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

Google Cloud Platf	orm 🗦 Mr Project 12502 👻		Q bis	istrery api		x v			e narstani	dataprenance +		Q. Branch Products, resources, docs (/)	B + + ()
(¹) (Max								Dataplex		- customer_data	¢ star	+ ATRACTAS BOLTH	<u> </u>
A My Project 12992	> Hus > E demo dataset		n, mini shin s					Dataver	~	nantari-detagonmenter >	Bernetter	III proficutione outra-domain III rotatine data profici atre	
٩			Г)ata Assats Lie	etina			Q. Seath		mource name project	talmanbark dat	apremance/locations/as central/Lakes (professioner energy demain/tenggis-anterner data graduat anna/indiges/systemer data	
G Titer Filter tog	ps and schema	BigQuery table of	details 🖵		Sung		^	< Explore (PREVER)	1	Resource type Table		Data Quality Metrics	
BigQuery table d	letails	Type System	TABLE BIGQUERY	Local	ion us			Manage Catalog	^ i	NUMBER OF STREET	EMA AND COLU	AN TADS PARTITION DETALS	
©		Table type Creation firms	BIGQUERY_T	XELE 1-14-26 EM	Bet onc	e 1244 rikolerie jeanset oppi	6	 Tag templates 		(4)			V EXPAND ALL THES
		Last medification time	Oct 28, 2021,	1:54.26 AM Desc	iption			C Policy tage		Data Product Classifica	tion		1 -
 Schema and cole 	umn tags	Expiration time	Never					Manage Lakes	^	Data Product Exchange			1 -
		Tags (0)						II Matupa		Data Product Informatio	on		1 -
		O You have no	teos attached to:	this data asset.			ATTACH TAOS	I Deces					
		-								Data Product Quality		144	1 ~
		Schema and col	lumn tags				^			00 Scare		ren Ø	
		Name	Type	Mode Column tags Policy Tags 🔮	Description	day first provided the second 1				Convertiness Score		100 -1	
		0,000	0.0000	NOVPEN	Dreative Mobile Technologies, LI	.C, 2+ VetFone Inc				Integrity Score			
		pickap_datetime	DUTETIME	NULABLE	The date and time when the met	er was engaged. er was disponanted				Completeness Score		130	
		passenger_count	INT64	NULLABLE	The number of passengers in the	e vehicle. This is a driver-enters	ed value			Uniqueness Score Accuracy Score		100	
		trip_distance	NUMERIC	NULLABLE	The elapsed trip distance in mile	s reported by the taximeter.				Detailed Quality Report Data Quality Dashboard		Data product Link to Data Davido Davido and	
		pickup_lang#ude	NUMERIC	NULLABLE	Longitude where the meter was - Latitude where the meter was en	mgaged. gaged				Last Modified By		line and the second sec	
		rate_code	STRING.	MULLABLE	The final rate code in effect at th	e end of the trip. 1= Standard e	sate			Last Profiling Date		June 23, 2022 & 4 19 Jan Hunn P 1 2022 De 14 00: 40 98 333301-00	
E Google Courd & man	dark dagownarce 💌		9.5	earch: Products, resources, docs (/)	2-57K I'menan Amatoso or V	reconnecter or responsed dara	a + • _ 6	= Goode Crust		decomposite a		Search Products measures doo 15	
2 Cutacles	Dataplex Resources		K	Customer - Source Domain Wyray Autor Los				12 Datastas		Fauch			U. C.
4. contro	NENDACE-CONTROL PRINC	PAL-CENTRIC						C: camper	•	odarch			
Dacaver A	The fibricity barrane			DATA PERMISSIONS LARE PERMISSIONS				Dataver	^ 1	Filters	CLEAR	Introducion unified Dataslay experience	×
< Second	Draphy same			ADD -T HEMOLE				Q, Search		Scope	~	Deta Catalog is now a part of Detaplex. You can now easily discover, curvin, and manage	
ng toporpresent	maniferit defagovernance			Outuples propagates policies in the Dataples Hierarchy 1	(aler-Zone-Asset) on an ongoing basis,	changes can take up to 30 minutes	Lean now	< Equine PREVER)				data in a single end to end experience to accelerate time to insights.	
9 Intranes		domain.		- Internet and				Manage Catalog	^ 1	Systems	Ŷ	LEARN MORE	
B Tag templates	 Itersections - Source Dama 	n proditransactions or domain	NOT .	Teely PERCENTS FOLIS				e terypops		Lakes and zones	~		
() Pulky tage	 III Transactions - Consumer De 	nais profitsesactions consumer domain		Y Fiber Chiefe Data Data Data Data Data	Indepies Inda Roader () (11 () (11	in nimulatajin datailitte O	× • =	0 Petrates				E Hot wrank	
Manage Lakes	H Cantoner - Source Domain	prod customer source	ir i	I fee Printed †	Netw	Refer	infertance	Mature Later		Data types	×	Q, top data, product, information, domain contorned	X 🛛 SEARCH
E Manage				jdeveloper pervicescount com	Compute Engine default service account	Dataples Administrator Dataples Data Reader	 manbark datagovernance manbark datagovernance 	II. Manage		Projects	~		
A line						Dataplex Metadata Reader	hanbark delegoremente	1 leve				Anthe Test Last	T COLUMN AND
II Promi				19 Exitative salphanbank detagovernance iam, parvicesecount.com	custome-sa	Dataplex Data Owner		II Promi	1	Tag templates	~		
						Dataples Deverspor	 manbark delaphemanos 			Datasets	~	tokenized_customer_data	Φĭ
						Datapies Metadata Reader	P manbank datagovernance					TABLE - System DATAPLEX - Source system BIGOUERY - Last modified Jun 1, 2022	
					and at a	Dataples Metadata Histor	In manifestive datagovernance			Public datasets	^	-) & sector and because) If had a sector sector and a final sector part	
				Batagovernance lant guerrosaccount com		Dataplex Developer	> mansbank-datagovernance		0	hickude public datasets		ec_oustomer_data	\$ X
						Dataples Editor	P manbank-datagovernance					Volut - System Deliverus (secon system BH20009 - Last modified Jun 7, 2022	
			Data	a Permissions		Dataplex Metadata Reader Dataplex Metadata Writer	 masbark datapoenance masbark datapoenance 					-) > Instant dispenses > (() o cents) > (() o cents) = () o cents of the set of	
													* ž
E Google Courd De mark	bark datagovertance 💌		Q, Se	ardh Products, resources, docs (/)	*		= + + i ()	= Goode Cloud St	nandará é	-		9. Sauch Products resources does (/)	
🕄 Dataplex	← customers_data	± STAR + ATTRONT	ASS Q OPEN	на ворлати 🗰 слагля в начала второ 🖉 всем	autors Erstrat autoratio								
Dane A	Representation a strains a	toud datapleatean > B	dataples-denos	Anatari (appender) Su) Contra a	,014		iner /	R Dataplex	Ex	plore recurs		New Sorget + X 🛛 COMPOSENCE SCRIPT	
9, Seath	DETAILS SCHEMA AND COL	UNIN TAGS TAGS	LMEAGE					Darres .		Re Fielyberane		• • • • • • • • • • • • • • • • • • •	
< Equive (PREVIDA)	1855 @ @ @					customers_c	data 🙏 vew LINEAGE 🖂		Int	ig tana			
Manage Catalog						Type System	TABLE BIGQUERY	9, Search		nanbark datagovenance			
B Entry groups				+ the same at +		Project Dataset	manbank-datagovernance outlother yee, joine	r≰ Equive(PREVER)		Hechart - Source Domain			
 Tag templates 						Last modified Description	Jun 29, 2022, 11:36-45 AM	Manage Catalog		Merchant Curated Zone			
 Policy tags 			- 0+	· E mentant, det. +						Herbart Ray Zone			
II Manaar	E colore	.40 -						A naidorte		Herchart Data Product Zon	4		
1 becare				• 🖬 menterit, data. 🔶				B Tagterylates		Rendeoks			
II Pacasa				-				🖞 Policy tage		10. topa			
				· E nankark, data. +				Manute Lakes		Tareactive-Source Service			
									2	E forsactore Consume Don	net.		
								E Manage	,	E requie conception			
								1 leve					
		Data	a Lin	neade				Poons		Evel	ore	Data through SOL or Notobooks	
				3-						Explo	ле	Data through SQL of Notebooks	
					-11								

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

- <u>Amundsen Data Catalog: Lyft's Open Source Discovery Tool</u>
- 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