Fraunhofer ISST

Fraunhofer Institute for Software and Systems Engineering ISST

DATA CATALOGS

ISST-REPORT
Implementing Capabilities for Data Curation, Data Enablement and Regulatory Compliance - 2022 Edition

Part of the Project

ISST-REPORT

DATA CATALOGS

Implementing Capabilities for Data Curation, Data Enablement and Regulatory Compliance - 2022 Edition

The ever increasing amount of data, the need for developing innovative and data-driven business models and complex regulatory requirements pose challenges to traditional data management practices in enterprises. More than ever enterprise data need to be findable, assessable, interoperable and reusable (FAIR) by a wide range of employees.Data Catalogs have become an essential part of enterprise data management and data governance initiatives that seek to foster the aforementioned data principles. They provide the needed capabilities for data curation, data enablement and ensuring compliant data usage. In the contexts of ongoing data (management) decentralization Data Catalogs serve as a central point of contact for many data-related activities. However, enterprises new to this endeavor are challenged by a large number of different tools, diverging functional scope of Data Catalog solutions, complex integration scenarios and a non-transparent market.

This report seeks to clarify the most urgent questions for data initiative leaders, decision makers and practitioners concerning Data Catalog implementation projects. It introduces the readers to the field, elaborates on central Data Catalog characteristics, presents a market analysis and depicts contemporary market trends.The work is based on a thorough analysis of Data Catalog solution providers and the experience of the researchers at the Fraunhofer Institute for Software and Systems Engineering ISST. The Fraunhofer ISST supports organizations in implementing Data Catalog solutions as neutral intermediary by offering services such as requirements analysis, pre-selection of potential providers, and critical monitoring of the selection process including the integration with data strategy or data governance initiatives.

ISST – Series of Reports:

Within the series of "ISST Reports", the Fraunhofer Institute for Software and Systems Engineering ISST, publishes its white papers. The series examines trends and technologies in computer sciences and takes up innovative subjects from some of the institute's research projects. The publications of this series provide insights into the current state of research concerning "Innovations from Data", the main research topic of Fraunhofer ISST.

AUTHORS

Nils Jahnke
Markus Spiekermann
Behnam Ramouzeh

EDITOR

Prof. Dr.-Ing. Boris Otto

CONTACT

Fraunhofer Institute
for Software and Systems Engineering ISST
Emil-Figge-Straße 91
Germany - 44227 Dortmund
info@isst.fraunhofer.de
+49 231 97677-0

ISST-REPORT

ISSN 0943-1624

Image source

Cover and Page 4,5 : ©piranka - iStock-1196044470

Published: August 2022

Table of Contents

Executive Summary

The importance of data for operational excellence, innovative products and novel business models is widely acknowledged. Further, data can become an economic good itself once they are traded on data markets with third parties. However, many companies struggle to generate value from available data. Reasons are manifold and include organizational data silos, unknown data quality and provenance, limited understanding of data objects, unclear workflows for data access and difficulties in handling personal data. Data Catalogs are information systems that provide a platform for all data-related roles of the enterprise. Based on the ingestion and leverage of metadata different functions are provided in order to connect data supply and demand and enable data users to work with data. These functions can be assigned to areas such as data inventory, data governance, data assessment and data discovery. Therefore, Data Catalogs play an important role in the modern enterprise information landscape.

Innovators and early adopters have started their Data Catalog endeavor but are experiencing challenges in Data Catalog implementation projects. They are faced with a large number of different tools and functions offered on the one hand and on the other hand with a non-transparent market and complex integration options. This report seeks to bring clarity to data management project leads and decision-makers by analyzing the Data Catalog market and answering important questions occurring before and during Data Catalog implementation projects. The report therefore builds upon the previous report by Korte et al. (2019), which was collaboratively created by researchers at Fraunhofer ISST and the Competence Center Corporate Data Quality. Findings of the legacy report such as the vendor assessment are updated based on latest developments of the Data Catalog market. Additionally, new content, e.g. the Data Catalog integration model, seeks to answer further practitioner questions that were left unanswered in the previous version.

Initially, this report scopes the Data Catalog field. Five different types of Data Catalogs were identified. They include enterprise Data Catalogs, platform solutions and cloud platform Data Catalogs, which are able to foster enterprise-wide metadata management. Only these solution types are the focus of the report. The other solution types of tool-specific Data Catalogs and data portals only provide benefits for specific use-cases and are therefore not further considered.

By the analysis of different Data Catalog solutions an updated capability map for Data Catalogs was defined. The capabilities were further elaborated by associating typical functions and classified into three categories ranging from core capabilities to add-on capabilities. Core capabilities are fundamental for the value-add of a Data Catalog and comprise e.g. capabilities in the area of data inventory, the description of roles, data profiling and data tagging. Add-on capabilities can be seen as additional benefits provided by only a small fraction of Data Catalog vendors. They include e.g. data quality assessments, data valuation and data recommendations. The capability map fosters the understanding of what capabilities can be expected from a Data Catalog solution. Further, it can help to structure requirements or user stories during the beginning of Data Catalog implementation projects.

The Data Catalog role model elaborates on current problems of data users in the enterprise and depicts improvements made possible by Data Catalog implementations. The Data Catalog integration model describes generic integration scenarios for Data Catalogs considering cloud and on-premises data sources, data management tools and applications that leverage (meta-) data.

During the following analysis of the Data Catalog market 60 solutions were identified. USA-based companies account for 70 percent of all providers while European companies make up an additional 20 percent. The current market for Data Catalog solutions is highly dynamic as new solutions appear and acquisitions take place very frequently. The market consolidation is therefore still low. According to market theory further acquisitions can be expected as big players will grow in market share and want to extend their core business by further services. Data Catalog innovation will move from product characteristics towards cost optimization, leading to more affordable solutions in the future.

Out of the 60 previously identified providers 15 providers were assessed in greater detail according to the capability model and based on characteristic functions. The assessed solutions include enterprise Data Catalogs (commercial and open-source), data platforms and one cloud Data Catalog solution. While results for individual solutions are provided in a tabular overview, a deeper analysis showed the following findings:

- No vendor was able to demonstrate a satisfying coverage of all capabilities. This indicates trade-offs likely need to be made during the selection process.
- Capabilities in the area of automation and AI, data governance, and data collaboration have been identified as the biggest differentiators between the assessed solutions.
- Open-source solutions currently provide only limited capabilities compared to enterprise Data Catalogs and the best performing platform solutions. Especially they are lacking features in the areas of automation and artificial intelligence and the support of data analytics.
- Although the boundaries between the Data Catalog categories have become “blurred”, it is still possible to differentiate between Data Catalog solutions that focus on data governance versus solutions that focus on analytics productivity.
- Four Data Catalog trends have been identified. They are: adoption of AI-features, support of data mesh and data fabric paradigm, transformation towards active metadata management platforms and increasing possibilities for cloud-deployments.

Based on these findings the authors of this study recommend practitioners to pay special attention to the requirements gathering and vendor preselection phase, as Data Catalog solutions differ highly in supported capabilities, integration patterns and costs. To find a matching solution it is of big importance to align with other data initiatives in the organization. This will also help to get engaged as many stakeholders as possible as the success of every Data Catalog implementation project depends on the willingness of different Data Catalog users to provide content. This will create network effects, and therefore attract further possible content providers or users. Under consideration of the high license and implementation costs of Data Catalogs, smaller organizations or those with a limited budget may consider smaller providers or start with open-source solutions to provide the most required capabilities.

1. Introduction

The amount of data collected is increasing nearly exponentially (Diamantini et al., 2018). It is widely accepted that data are an own asset class with potential value to an organization. However, value from data is only created once data are transformed, enriched, aggregated and brought into context and thereby converted into actionable information (Koutroumpis et al., 2020). To this end, data must be well maintained, of high quality, trusted and available to a wide set of employees, amongst other properties. Yet studies of various industries show that only a fraction of the data collected and stored by enterprises is actually leveraged in value-creating processes (DalleMule & Davenport, 2017). Accordingly, a high proportion of the recorded data remains unused and degenerates into so-called "dark data". This separation of data supply and demand leads to productivity losses in areas such as business intelligence, machine learning and others. In these areas data users especially encounter problems in the data acquisition phase including data discovery and data assessment. Further, regulations such as GDPR/ CCPA or industry-specific standards pose additional challenges regarding the retention, usage and deletion of data.

Metadata management and data governance are essential to tackle these challenges and foster value-generating and compliant data usage (Dinter et al., 2015; Otto, 2011). To manifest and operationalize metadata management and data governance in enterprises, Data Catalogs are an essential tool. In addition to functions in the areas of data inventory and data discovery Data Catalogs support data assessment, data governance and data collaboration. Accordingly, they adopt a holistic metadata perspective. By integrating data supply and data demand Data Catalogs help to eliminate data silos. Further benefits of Data Catalogs include compliance with regulatory standards, enhancing data availability and fostering data accountability. Data Catalogs enable organizations to treat data according to the FAIR-principles and therefore make data findable, accessible, interpretable and reusable (Labadie et al., 2020). Consequently, they support all data-driven activities in companies, such as data science, business intelligence or reporting.

With the rising amount of data and the need to transform into a data-driven enterprise the challenge of implementing data curation capabilities is still current for many organizations. Therefore, more and more organizations have either Data Catalogs already in place or start with a Data Catalog endeavor. Until 2025 the Data Catalog market will grow by about 25% each year (Mordor Intelligence, 2021). Currently, established vendors are challenged by new entrants often backed by venture capital. In total about 60 solutions could be identified. Amongst them are solutions that focus on delivering value for small and medium businesses as well as open-source software enabling Data Catalog initiatives with lower budgets or bigger needs for customization.

With the increasing amount of vendors opportunities and challenges of finding the matching solution according to the organizational needs, and boundary conditions arise alike. Further, due to the constantly changing IT landscape, companies with a Data Catalog already in use need to monitor whether the solution can still fulfill the set goals. This report is supporting these objectives by framing the topic of Data Catalogs, presenting scenarios for integration and giving an overview about Data Catalog functionality, features and use cases. Additionally, Data Catalog

providers are assessed according to a typical set of capabilities. The assessment evaluation highlights capabilities and functions that differentiate between different solutions. Further, the performance of different solution types is elaborated. Lastly, the development of solutions since the year 2019 as well as ongoing and future trends of Data Catalog solutions are illustrated. To conclude, recommendations for practice are given based on the findings of this report.

2. Role of Data Catalogs in the Enterprise

2.1. History and Evolution

The documentation of data to enable users to understand and harness data in enterprise environments has been a relevant topic since the early days of data processing. Data documentation is closely coupled to the concept of metadata (i.e. "data that defines and describes other data" (ISO, 2015)). On a functional level metadata can be divided into business metadata (e.g. field description and business rules), technical metadata (e.g. data type, data structure, data schema) and operational metadata (data processing information) (Diamantini et al., 2018; Sawadogo & Darmont, 2021). Data Catalogs are platforms for locating, evaluating, and making available data across the enterprise using the different types of metadata (Eichler et al., 2021).

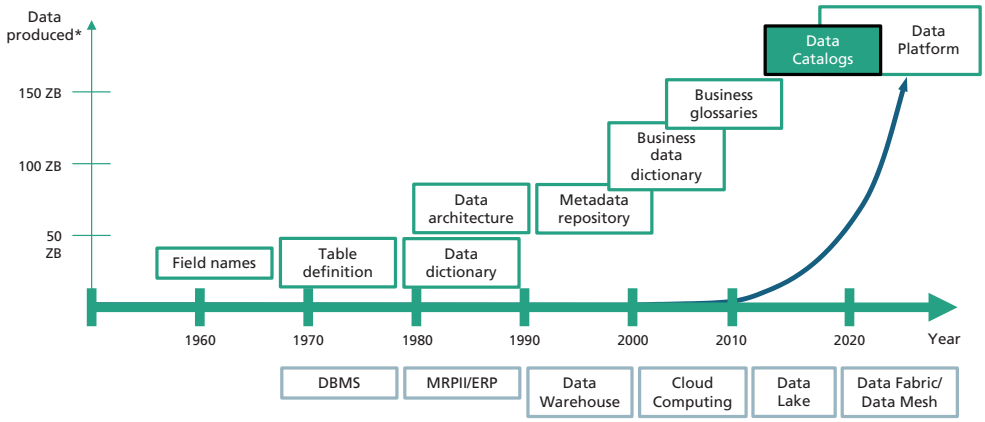


Figure 1: History and evolution of Data Catalogs following Korte et al. (2019)

Data Catalogs today are a result of the evolution of different concepts to describe and provide data in enterprise environments. This evolution is depicted in Figure 1. A first major step towards what is nowadays defined as a Data Catalog was the development of data dictionaries which were used to document technical information of tables in database applications and tied to their respective database.

With the emergence of enterprise information systems, including ERP and early versions of CRM software, more focus was directed towards the integration of business processes resulting in a system landscape of higher complexity. Data architecture became necessary to plan and integrate systems and data according to the business needs. Data warehouses emerged due to the need for data integration and data-based decision-making. As data warehouses integrate data of different formats, structure and semantics and are distributing the consolidated data to users with different expectations, data needed to be documented in data warehouse metadata repositories.

As data storage and information systems became more dispersed due to cloud computing technology, business data dictionaries were established to define technical metadata for the whole enterprise. Subsequently, business glossaries were developed to align business terms and semantics by providing clear definitions within a division or enterprise, allowing users to correctly interpret data for different use cases.

With more companies recognizing the value of data itself and due to the emergence of big data, data lake architectures were implemented to deal with the high amount of data variety and to provide means for data analytics. To exploit the value of data lakes and avoid turning them in so-called data swamps, a holistic metadata management approach is required. Data Catalogs emerged to provide means for metadata storage and curation, enabling data to be found, understood and trusted. Some Data Catalog solutions first focused primarily on data lake use cases, whereas others where directly appealing to the enterprise as a whole.

While many companies leverage a data lake, decentralized and federated data architectures are gaining momentum due to the drawbacks of monolithic data lake approaches. These drawbacks include the high risk of data inconsistencies and the need for data scientist profiles, which may cause bottlenecks in different scenarios (Sawadogo & Darmont, 2021). Examples for emerging architecture paradigms are data fabric and data mesh. Data Catalogs still remain an important part of these architectures, as they provide data curation and discovery capabilities and serve as a central point of contact for several data related activities such as data discovery and data governance. They need to integrate seamlessly into the existing information systems landscape to enable as many users as possible to work with data. Recently, Data Catalogs are increasingly integrated or extended into enterprise data platforms. Data platforms offer end-users further capabilities such as data virtualization and query for data preparation (Gröger, 2021; Wells, 2021) and are based on a modular service offering. In this setup Data Catalog modules focus on their core capabilities (see section 3.1), while other tools and solutions provide extended capabilities for data curation.

These developments emphasize the need for greater interoperability between Data Catalog solutions and further tools (e.g. for data quality, ETL-processes or analytics management) on the one hand, and on the other hand between different Data Catalog solutions in order to federate metadata management in the enterprise.

2.2. Types and Purposes

In the first version of this report in 2019 Data Catalogs were defined as follows:

> A Data Catalog is an integrated platform for data curation, matching data supply and demand. It offers users functions to register data; to retrieve and use data; and to assess and analyze data. A Data Catalog therefore should provide a data inventory (for data supply) and features for data discovery (for data demand) as key components. Additional features should support data governance, data assessment, and data analytics, alongside with appropriate features for catalog administration and data collaboration. (Korte et al., 2019, p. 9)

Due to the growing importance of the mentioned data cataloguing and data curation capabilities in the enterprise, solution providers from different parts of the data value chain incorporated Data Catalogs or data cataloguing functionality in their solution offerings. While these software components fulfill the aforementioned definition and are labelled Data Catalogs by their providers, they often only offer limited capabilities compared to native Data Catalog solutions or are only applicable within a specific context. Nevertheless, within their context, these tools provide useful support.

To provide a better overview about the existing market offerings, a classification of Data Catalog solutions has to be conducted. A distinction is not always easy to make, as the boundaries between the individual categories are "blurry" to some extent. Additionally, it has to be mentioned that today's enterprise Data Catalogs are in some instances as well originating from supporting specific use cases or tools, e.g. data lakes or data virtualization, before they became independent products. Therefore an evolution of a specific tool may lead to a change of Data Catalog classes. In the following, a classification of different types of Data Catalog solutions is presented. In particular, Data Catalogs as data portals, cloud platform solutions and tool-specific Data Catalogs as opposed to enterprise-wide Data Catalogs are being described.

Data Catalogs in the sense of data portals:

One very common misconception not only in popular literature but also in information systems research is the confusion of Data Catalogs and data portals. For instance, Comprehensive Knowledge Archive Network (CKAN), which enables the provisioning of open data, is often labelled as Data Catalog by practitioners as well as scholars (Klímek et al., 2018; The World Bank Group, 2021). However, a more appropriate designation to differentiate this class of Data Catalog is the one of a data portal (CKAN; Máchová & Lnenicka, 2017). The goal of data portals is to provide data to a wider audience and therefore similar to the one of Data Catalogs. However, these data portals are not focused on data curation within a single organization. Instead they are mainly used to provide data created by public institutions to be re-used by enterprises, research organizations or citizens (Máchová & Lnenicka, 2017). In contrast to Data Catalogs The content of data portals is created mainly by the supply side, e.g. by a government agency or citizen initiative, whereas the data demand side takes no active part in data curation activities. So far, data portals usually just provide limited metadata and linkage between data sets. In contrast to Data Catalogs data portals can provide direct access to data, as they are usually available as downloadable files.

Cloud platform Data Catalogs:

To enable data discovery and curation within their own environments, the top three cloud service providers Amazon Web Service (AWS Glue), Microsoft Azure (Microsoft Purview) and Google Cloud (Data Catalog) each implemented their own Data Catalog solution. Main use cases for these Data Catalog solutions are the support of data orchestration and ETL processes.

While these solutions are relatively new and will certainly evolve in the future, they only provide limited data cataloguing capabilities today. One major current drawback of cloud platform Data Catalogs is that they are mainly limited to use cases within the vendors cloud environment. As companies increasingly turn towards complex scenarios including multi-cloud or hybrid cloud architectures, these solutions have their shortcomings. Additionally, they may not be available in all regions of a cloud service provider. However it seems that the vendors have recognized these shortcomings, as they start to offer more possibilities for integration with other tools and enterprise information systems, e.g. Microsoft Purview with its Apache Atlas API.

Tool-specific Data Catalogs:

Another category of Data Catalogs are such that are built-in or available as an add-in for data virtualization or data analytics tools. Especially for business intelligence/ data analytics use cases the understanding of the underlying data is crucial to correctly analyze the created results. An example for such a solution is the business intelligence software Tableau that includes Tableau Catalog as an add-on to improve trustworthiness and findability of data with a potential use for analytics. This implies that this type of Data Catalog solution can only be used within the specific environment of the tool and serve the specific use case. Although tool-specific Data Catalogs may provide means for integration with the wider enterprise environment, they are not suitable for holistic data curation and governance within an organization.

Enterprise Data Catalogs:

In contrast to the Data Catalog types described above, enterprise Data Catalogs support the whole enterprise information systems landscape, either by native or custom integrations. They provide metadata extraction and provisioning for manifold data sources ranging from relational and non-relational databases up to data storage systems from the Hadoop context and event streams. Further, they offer native integration into user systems downstream and upstream the data value chain such as ERP and CRM systems as well as business intelligence and analytics tools and platforms. Enterprise Data Catalogs support most data-related user roles in the enterprise. Amongst others these are data owners, data stewards, data engineers and data analysts. Enterprise Data Catalogs can be further distinguished between stand-alone and platform solutions. Stand-alone solutions cover nearly all of the data cataloguing capabilities in a single tool, offering one holistic solution for data governance and data management in the enterprise. Platform solutions on the other hand integrate a Data Catalog module offering core cataloguing capabilities with other modules or tools, e.g. for data quality or data privacy. Often, these modules can be combined according to the customer's needs. Hence, interoperability of the single modules is a prerequisite for the successful development of a platform solution. Like cloud platform solutions, enterprise Data Catalogs are also available as private or public cloud offerings. Nevertheless they rely on on-premises instances to be able to ingest metadata from locally deployed systems. In addition to commercial enterprise Data Catalogs several open-source solutions are available, which can be deployed without any licensing cost.

The following table summarizes the different categories of Data Catalog solutions and provides an overview about key features and examples. In the following sections of this report enterprise Data Catalogs and partially cloud platform Data Catalogs are in focus.

Type/Purpose	Definition/Explanation	Examples
Enterprise Data Catalog	Data Catalog supporting data management, governance and analytic roles throughout the whole organization while covering key capabilities as well as additional data cataloguing features.	
Stand-alone solution	Stand-alone Data Catalogs offer key and additional data cataloguing components within a single tool. Commercial and open-source offerings are available	Alation Data Catalog (commercial) Apache Atlas (open-source)
Platform solution	Platform solutions combine data cataloguing modules offering key data cataloguing functions with further modules providing additional capabilities.	Collibra Data Intelligence Cloud Informatica Intelligent Data Management Cloud
Cloud Platform Data Catalog	Data Catalog providing Data Catalog key components mostly limited within the cloud service provider environment. Use cases such as orchestration and ETL-processes are the main focus.	AWS Glue Data Catalog Microsoft Purview Google Cloud Data Catalog

Tool-specific Data Catalog (add-ons)	Data Catalog supporting a specific tool, e.g. within the area of business intelligence by providing key components as well as purpose related additional cataloguing features.	Tableau Catalog Databricks Unity Catalog
Data Portal	Data management system to provide open data to a wider audience (citizens, enterprises or research) by leveraging key data cataloguing features and giving direct access to data often in terms of files. Data portals are often deployed by public institutions and aim for societal benefit by data re-use.	CKAN The World Bank Data Catalog (based on CKAN)

3. Functions and Features

In order to understand the functions and features of Data Catalogs in detail, a deeper general understanding of the position of Data Catalogs in the enterprise is required. Figure 2 illustrates the functions and features of Data Catalog systems in the enterprise on a high level. Data Catalogs are situated as intermediaries between data sources (depicted on the left hand side) and data users (depicted on the right hand side). They connect data supply and data demand by offering different capabilities for data curation.

For this purpose Data Catalogs rely on metadata (data about data), while the actual data generally remains in their sources. In some cases however an additional excerpt of sample data may be stored in the Data Catalog environment. Metadata is either ingested by so-called connectors or integrators, imported from spreadsheets or manually entered by roles such as data owners or data stewards.

After the ingestion of a primary set of metadata from its sources, they are enhanced or altered by the capabilities provided by the Data Catalog solution, which will be further explained in section 3.1. While some metadata artifacts may be created automatically by the Data Catalog, other artifacts require manual curation efforts by its users. This holds especially true for the linkage of different metadata artifacts to each other, e.g. business glossary terms to a data set. Additionally, metadata from other data management solutions may be leveraged in a Data Catalog to act as central point of contact for data. This may include data quality information, information from ETL-tools about data transformations or the ingestion of data from other Data Catalog solutions as Data Catalog of Data Catalogs.

On the data usage side metadata maintained in the Data Catalog is leveraged by human users as well as systems and applications. Typical application areas for company-internal users are the search and discovery of data, the assessment of their fitness for purpose or the determination if data is handled appropriately. While external users are not commonly granted access to the Data Catalog itself, they can profit from metadata descriptions provided together with the data set needed. Further, external auditors can be granted access to a Data Catalog in order to accelerate the non-value-adding task of data collection. Connecting applications to the Data Catalog can enhance the value of Data Catalog solutions in two ways. On the one hand metadata of Data Catalogs can be leveraged in business intelligence or advanced analytics applications. On the other hand a support of Data Catalog use cases such as user collaboration or the realization of data governance workflows are possible by integrating with existing instant messaging services or workflow management systems. If data are handled, transformed and processed in applications, more metadata are created. These so-called behavioral metadata give hints on how data are experienced in contrast to how data are designed. Therefore their description starts to become an important part of Data Catalog offerings.

This chapter presents a deep dive into core components of Data Catalog initiatives. Section 3.1 elaborates the capabilities and functions of Data Catalog solutions in greater detail and assesses which capabilities are the core of a Data Catalog and which capabilities are expected to be optional or have an add-on status. In section 3.2 the most important enterprise data roles are presented and their incentives for a Data Catalog implementation are described.

Based on the analysis of architecture models of different vendors, this report also introduces a model for technical integration of Data Catalogs into the enterprise information systems infrastructure. This integration model can serve as a foundation for possible integration scenarios (see section 3.3).

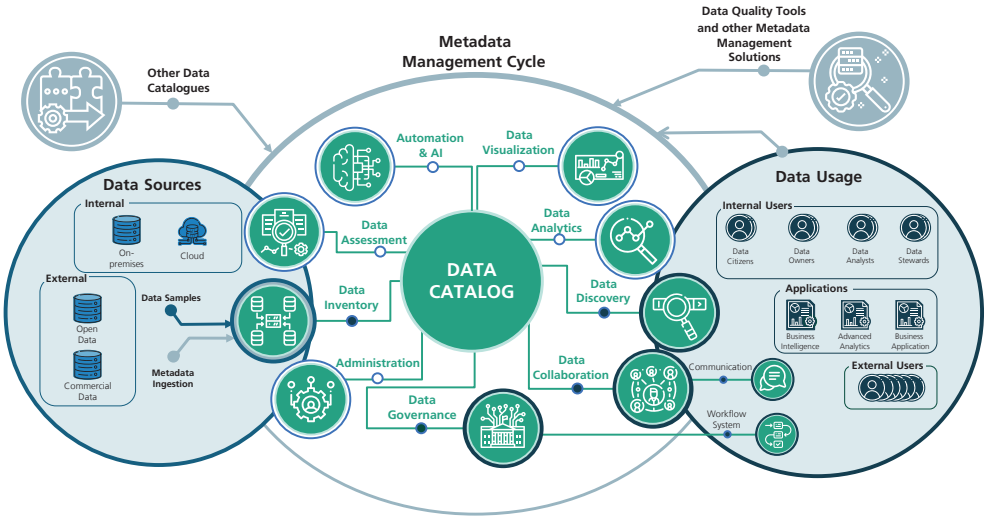


Figure 2: High-level illustration of Data Catalog systems in the enterprise [Icons: Flaticon.com]

3.1. Functional Model

The functional model for Data Catalogs, introduced in the first version of this report (Korte et al., 2019), describes the capabilities offered by Data Catalog solutions for efficient data curation in the enterprise. Based on the findings of the initial report the model has been extended and refined. The data cataloguing capabilities can be classified in superordinate capability groups and are operationalized by distinct functions that enable users to curate enterprise data. The capabilities and superordinate capability groups are further explained according to their allocation in a capability map (Figure 3). To illustrate the capabilities, exemplary functions are elaborated.

The functional model can support persons responsible for Data Catalogs in multiple ways. In general, it offers an overview of the expected functionality provided by Data Catalog solutions and affiliated modules in enterprise data platforms, which support further data curation activities. During the Data Catalog selection process the functional model can serve as basis for requirements gathering, requirements prioritization and vendor assessment as it was already proven by the authors of this study. Additionally, in section 4.3, the functional model is used to examine and compare different Data Catalog solutions to deliver a high-level assessment.

Based on the assessment of solution providers, it was examined to what extend each capability is generally incorporated in Data Catalog solution offerings. Therefore, capabilities are categorized in three categories:

- Core Data Catalog capability: Capability fundamental for the value-add of a Data Catalog for enterprise-wide data curation. Common consensus about the need of this capability exists amongst Data Catalog solution providers, leading to a broad implementation of functions to fulfill this capability in solution offerings.
- Extended Data Catalog capability: Capability that supports the general goal of enterprise-wide data curation but not essential for every Data Catalog use case. While many solutions incorporate functions to support these capabilities, others decide to offer only limited functionality or integrate with other solutions to support such use cases.

- Add-on Data Catalog capability: Capability only featured in small amount of Data Catalog solutions. Features might be hard to implement or are only needed to support use cases outside of the fundamental Data Catalog scope. A small fraction of Data Catalog solutions offer functions to support these capabilities. This can either be due to the origin of a solution provider or the history of a solution, or due to attempts to differentiate a solution from the competition.

Figure 3 shows the functional model for Data Catalogs. The model is subdivided in capability groups represented by green frames and subordinate capabilities depicted as blocks in different colors according to their categorization. The capability groups are separated into such describing core functionalities, which deliver added value for Data Catalog users (Data Inventory, Data Governance, Data Assessment, Data Collaboration, Data Analytics and Data Discovery) and capability groups that support the wide adoption and use of a Data Catalog solution (Automation and Artificial Intelligence (AI), Data Visualization and Administration). In the following, capability groups and capabilities as well as exemplary functions are further illustrated.

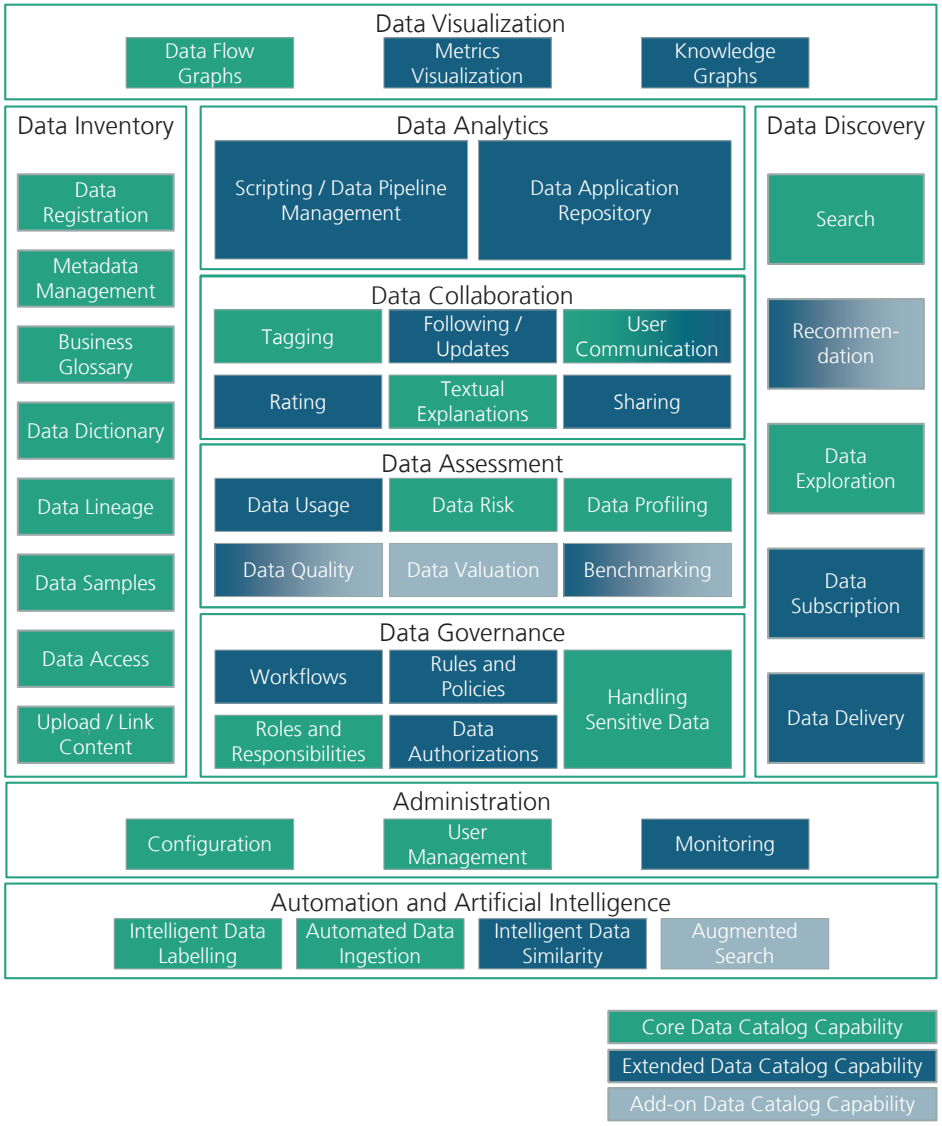


Figure 3. Data Catalogs functional model

Data Inventory:

The capability group data inventory comprises capabilities that are needed to register data, organize data and describe data by means of business, technical and operational metadata as well as further artifacts, e.g. graphics or documents. While metadata is ingested from the source systems, the actual data remains in its source.

Data Registration:
Allows users to register data in the Data Catalog. This can be done manually, via the import of spreadsheets or csv-files, or by the use of data integrators or data connectors that allow for automatic registration of data assets

Exemplary Data Catalog Functions:
- Data connectors: ability to automatically ingest metadata from certain source systems (e.g. relational databases, data warehouses, event streams).
- Data imports: import data asset descriptions that are maintained in spreadsheets.

Metadata Management:
Enables the management of metadata throughout its lifecycle, once it has been registered. Metadata can be automatically updated from files, enterprise systems or other data curation solutions; or manually maintained by eligible users.

Exemplary Data Catalog Functions:
- Manual data descriptions: maintain data descriptions manually (e.g. update business glossary with new definitions).
- Job scheduling: scheduling the automated ingestion of metadata from the source system.
- Versioning: ability to create versions of metadata descriptions, allowing for the assessment of changes.

Business Glossary:
Allows users to describe data from the business point of view. Data are defined and described by business terms in order to understand their meaning and context of use.

Exemplary Data Catalog Functions:
- Enterprise business glossary: define and describe business terms for the whole enterprise.
- Domain business glossary: maintain different versions of business glossaries per domain or business unit.

Data Dictionary:
Documents data from a technical viewpoint on a logical (data type, data format, data schema) and physical (storage system) level.

Exemplary Data Catalog Functions:
- Data schema: obtain the system and data schema in which certain data reside.
- Data type: describe the data type of a certain column (e.g. numeric or string).

Data Lineage:
Allows users to track the history of data from the point of data origin (data provenance) to consumption. It describes where data originated from, how it was transformed and where it was used by whom.

Exemplary Data Catalog Functions:
- Data provenance: ability to trace data back to its original source.
- Data flow and transformation: ability to see how data were transformed and where they were used, tracing data from their origin to data sink (systems level).

Data Samples:
Provides Data Catalog users with a small excerpt of the original data, while respecting data access rights and data protection. Allowing for a clearer picture of a certain data asset.

Exemplary Data Catalog Functions:
- Sample view: ability to preview data under consideration of data security and data protection.

Data Access:
Allows to access the metadata described and maintained in the Data Catalog directly (via downloads) or indirectly (e.g. via API).

Exemplary Data Catalog Functions:
- Data downloads: download metadata directly from the Data Catalog (e.g. as spreadsheet or csv-file).
- Data Catalog APIs: access and obtain metadata through a custom or standardized (e.g. Apache Atlas) API.

Upload/Link Content:
Allows users to provide additional information for a given data asset in the Data Catalog. This can include the attachment of additional documents or media.

Exemplary Data Catalog Functions:
- Embed media: embed graphics or documents in the data description
- Provide links: provide links to further information (e.g. enterprise wiki).

Data Governance:

The capability group data governance comprises capabilities that support or are related to data governance activities in the enterprise such as defining roles and their rights and responsibilities, defining data related policies and monitoring the compliance to internal and external data management regulations.

Workflows:
Allows to define and manage workflows (e.g. propose a business term or assign a user role).

Exemplary Data Catalog Functions:
- Glossary workflow: supports the creation of new glossary terms by Data Catalog users.
- Custom workflows: allows the definition of custom workflows for data governance processes in the Data Catalog (e.g. by a workbench).

Roles and Responsibilities:
Allows assigning different roles and responsibilities to users. This may include the roles of data owner or data steward.

Exemplary Data Catalog Functions:
- Data ownership: assign role of data owner to a user for a specific asset in the Data Catalog, related to certain permissions and responsibilities.
- Data stewardship: assign role of data steward to a user for a specific asset in the Data Catalog, related to certain permissions and responsibilities.

Rules and Policies:
Allows creating, maintaining and publishing rules and policies for handling certain data assets.

Exemplary Data Catalog Functions:
- Data usage policies: ability to describe how data originating from an external source should be handled based on data usage agreements.
- Privacy-conformant data use: ability to describe how to handle a dataset containing sensitive data to avoid identification of natural persons.

Data Authorization:
Allows to control and manage the access to source data depending on the user role. This can be done by automatic means or by a manual approval process.

Exemplary Data Catalog Functions:
- Managing access requests: Data Catalog integrates with access management of data sources and can manage the access to data assets.

Handling sensitive Data:
Allows the identification of sensitive data (e.g. containing personal identifiable information (PII) or confidential information), who has access to it and where this data is used.

Exemplary Data Catalog Functions:
- Mark PII: ability to indicate that a certain data field contains PII.
- Mark classified data: ability to indicate the classification of a data asset.

Data Assessment:

The capability group data assessment comprises capabilities that facilitate the evaluation of data regarding their fitness for use. This includes the basic overview of a data asset by data KPIs, data quality assessments, and evaluation of data risk or the tracking of data usage.

Data Usage:
Allows tracking and assessing of the actual usage of data on a physical level (e.g. data attributes accessed by systems) and on a conceptual level (e.g. tracking data usage in analytics projects).

Exemplary Data Catalog Functions:
- Document data users: maintain a list of data analysts and data scientists that leverage a data set in their analytics projects.

Data Quality:
Delivers predefined data quality metrics and lets users create own data quality metrics to assess the data quality.

Exemplary Data Catalog Functions:
- Default data quality metrics: automatic assessment and presentation of popular data quality metrics (e.g. completeness).
- Custom data quality rules: allows the definition and application of own data quality rules (e.g. for validity of data).
- Quality dashboard: presents an overview of data quality for selected data.

Data Risk:
Allows assessing data-related risks (resulting e.g. from the usage of personal identifiable information or other sensitive business data), e.g. through an impact analysis.

Exemplary Data Catalog Functions:
- Risk dashboard: presents and overview of data risk across the data described in the Data Catalog.
- Valuate PII use: trace the use of PII across the data models of an organization.

Data Valuation:
Allows determining the financial and non-financial value of data (e.g. in terms of its reproduction costs and its value-in-use) using different metrics.

Exemplary Data Catalog Functions:
- Data usage costs: ability to document the price for using certain data assets, once they are not available for free.
- Data value: ability to calculate the value of data maintained in a data catalog based on predefined metrics.

Data Profiling:
Allows automatic generation of data profiles (e.g. number of columns, number of values, value range)..

Exemplary Data Catalog Functions:
- Value distribution: shows the distribution of certain data including minimum and maximum value or most frequent values.
- Dates: shows important dates for data such as the creation date or the date of last change.
- Size of dataset: shows the number of rows of a relational dataset.

Benchmarking:
Allows data comparison and benchmarking based on specific criteria or metrics.

Exemplary Data Catalog Functions:
- Data scores: makes different data sets comparable by giving each column in a data set a custom score based on predefined metrics.

Data Collaboration:

The capability group data collaboration comprises capabilities that enable the collaboration of data-related user groups. This includes features known from other P2P platforms such as tagging, rating or reviewing and communications. These features can source and distribute the knowledge of certain individuals and lead to knowledge spillovers. It further incentivizes good data management practices and enables more people to work with data.

Tagging:
Allows users to mark or label data (e.g. product, plant, finance data). With the help of tags, data can be discovered and filtered.

Exemplary Data Catalog Functions:
- Tagging governance: ability to pre-define tags or tag categories that can be used by Data Catalog users.
- Tagging of business terms: ability to label assets in the Data Catalog with terms from the business glossary.
- User tagging: ability to tag Data Catalog users to make them aware of certain content.

Rating: Allows users to rate data in terms of usefulness from their perspective.	**Exemplary Data Catalog Functions:** ■ Rating: ability to rate data to describe its fitness for use from the user's perspective.
Following/Updates: Allows users to follow data and receive a notification each time changes occur (e.g. when new data were added or their schema was changed).	**Exemplary Data Catalog Functions:** ■ Following data: ability to bookmark or follow a Data Catalog entry. ■ Notifications: automatic notifications for subscribers of data once changes of metadata took place.
Textual Explanations: Allows users to describe data by text artifacts to make it more understandable and actionable for other users.	**Exemplary Data Catalog Functions:** ■ Wiki: ability to collaboratively create and maintain a wiki entry to describe and explain Data Catalog assets. ■ Comments/ Reviews: ability to comment or write a review about data, giving the ability to ask questions, provide feedback or describe own experiences using the data.
User Communication: Allows users to communicate with each other. This can be done by providing contact data of Data Catalog users, by instant messaging, or by setting up a forum for discussion between multiple users directly within the Data Catalog.	**Exemplary Data Catalog Functions:** ■ Contact information: ability to retrieve contact information e.g., of a data owner to ask questions, retrieve further information or to request access. ■ Instant Messaging: provides an integrated chat function to conduct conversations regarding a data asset. ■ Discussion board: provides a discussion board for users to exchange their views.
Sharing: Allows users to share Data Catalog data with others (e.g. a link to a certain Data Catalog view can be shared).	**Exemplary Data Catalog Functions:** ■ Link creation: ability to create a link for a certain Data Catalog entry that can be shared with Data Catalog users (e.g., via email).

Data Analytics:

The capability group data analytics comprises capabilities that primarily support the tasks of data analysts, data scientists and data engineers. They include the management of data lakes, the support of building data models and data pipelines and the management of information from data analytics projects.

Scripting/Data Pipeline Management:
Allows users to write scripts (e.g. for an SQL-query to analyze data) directly in the Data Catalog, or initiate a script to run on a corresponding analytics system (where the data is stored, e.g. a data lake). Furthermore, users can design data pipelines. The results can be displayed either directly in the Data Catalog or via an external visualization tool.

Exemplary Data Catalog Functions:
- Visual query builder: build new logical data models by a graphical user interface and without the need to know SQL operators.
- Query engine integration: integrates with existing query engines to run queries e.g. on top of a data lake.

Data Application Repository:
Allows users to connect the Data Catalog to a data application repository (e.g. machine learning models).

Exemplary Data Catalog Functions:
- Integrate data analytics environments: ability to integrate Data Catalog and data analytics (e.g. business intelligence software or Jupyter notebooks).

Data Discovery:

The capability group data discovery comprises capabilities that enable Data Catalog users to find and obtain the data they need. As the efficient discovery of data represents a core value-add of Data Catalog implementations, these capabilities are crucial for the success of a Data Catalog. Data can be discovered in several ways. They include search, exploration and recommendation.

Search:
Allows data to be found based on a search term. The data are found either based on keywords or a semantic text depending on the underlying data model of the data inventory.

Exemplary Data Catalog Functions:
- Text search: ability to search for data based on a semantic text.
- Filter search results: ability to display/ filter data based on relevant KPIs and metadata (e.g., quantity, quality, ratings).

Recommendation:
Proposes data to the user based on past user behavior, similar user profiles as well as based on data similarity.

Exemplary Data Catalog Functions:
- Similar data recommendations: while viewing a data asset recommendations for a similar data asset are given (e.g. based on column names).
- Behavior-based Recommendations: recommendations of data assets based on previous user behavior.

Exploration:
Enables users to browse and explore data in a specific business area (e.g. within an information system or data storage).

Exemplary Data Catalog Functions:
- Explore data sources: browse the system landscape to find appropriate data for a specific purpose.
- Explore business area: browse data within a business area to find appropriate data for a specific purpose.

Data Subscription:
Offers users an experience similar to an online shop. Users can select data assets and store them in a shopping cart or subscribe to them. During the check-out, data access is requested, which will be granted based on access rights and data license conditions by users outside the Data Catalog.

Exemplary Data Catalog Functions:
- Shopping cart: ability for a user, typically in the role of data analyst or data scientist, to select data assets that they want to access.

Data Delivery:
Allows users to download data directly from the Data Catalog, or access the data via API for a limited period of time (similar to a subscription model).

Exemplary Data Catalog Functions:
- Data downloads: data can be directly downloaded from the Data Catalog e.g. in form of a spreadsheet.
- Data Transfers: data is delivered to business intelligence or analytics software to allow further processing.

Administration:

The capability group administration comprises support capabilities that are needed for the efficient and compliant usage, management and maintenance of Data Catalogs.

Configuration:
Allows configuration of the Data Catalog functionality according to the user's preferences (e.g. customize navigation in the way they better suit the usage purpose).

Exemplary Data Catalog Functions:
- Activating/ Deactivating features: ability to turn some Data Catalog features on or off if they are not needed.
- Customize Views: ability to customize the views of certain data assets to support better user experience.

User Management:
Allows adding new user accounts, editing user profiles and deleting user accounts. Furthermore, access to metadata and functions can be controlled.

Exemplary Data Catalog Functions:
- Integration of existing IAM: ability to support existing security infrastructure in the enterprise (e.g. LDAP/Kerberos).
- Limit Access to Metadata: ability to limit access to metadata objects for specific user groups.

Monitoring:
Allows Administrators to monitor and measure Data Catalog tasks and performance and analyze user behavior.

Exemplary Data Catalog Functions:
- Monitor Data Catalog tasks: ability to get an overview over Data Catalog tasks and their status.
- Analyze User Behavior: analyze the behavior of Data Catalog users (e.g. to assess the most relevant data assets).

Automation and Artificial Intelligence (AI):

The capability group automation and artificial intelligence comprises support capabilities that facilitate data curation by supporting users and taking over manual tasks, enabling Data Catalogs to scale. Each capability group can be supported by AI functions to a certain extend. So-called augmented Data Catalogs comprise automation functions e.g. in the area of data ingestion, data labelling and classification, and search. As Data Catalog providers incorporate more and more automated functions, the inclusion of further AI-capabilities in the near future is highly probable.

Intelligent Data Labelling:
Automatically tags or classifies data in a Data Catalog (e.g. detection of personal identifiable information), learning from the feedback given by Data Catalog users.

Exemplary Data Catalog Functions:
- AI-supported tagging: automated recommendation of tags that are constantly improved by user feedback.
- Recognition of sensitive data: self-improving recognition of PII or other data classifications that can be used for automatic access control or data anonymization.

Automated Data Ingestion:
Automatically ingests data and intelligently determines data lineage, data profiles and data quality metrics.

Exemplary Data Catalog Functions:
- Automatic data ingestion: automatically crawls a data source and populates Data Catalog with metadata.
- Definition of data quality rules: natural language processing can transform verbal descriptions into a technically verifiable rule.

Intelligent Data Similarity:
Automatically identifies similar data e.g. by the use of sematic graphs, supporting more suitable recommendations.

Exemplary Data Catalog Functions:
- Recognize synonyms: recognition of synonyms for data concepts, allowing more elaborated recommendations.
- Linkage of technical and business metadata: connection of data fields to business terms is automated, enabling the discovery of columns with similar features.

Augmented Discovery:
Augments search and discovery by machine learning functionality, e.g. so that search results are ranked according to user behavior.

Exemplary Data Catalog Functions:
- Result rankings: ability to order rankings according to previous user behavior.
- Auto-completion: recommendation of terms to complete a search string.

Visualization:

The capability group visualization comprises support capabilities that facilitate and speed-up the understanding and assessment of data. This includes visualization of data movement, visualization of data metrics, and the visualization of business information and metadata in the Data Catalog.

Data Flow Graphs: Visualizes the movement of data throughout the enterprise including aggregations and transformations.	**Exemplary Data Catalog Functions:** ■ Data lineage graph: visualize the data flow between systems from data source to data sink. ■ Representation of tags in data flow: Integration of tags (e.g. PII) in data flow graph, facilitating data risk assessment.
Metrics Visualization: Visualizes metrics used in data profiling or data quality assessment by means of appropriate charts.	**Exemplary Data Catalog Functions:** ■ Dashboards: ability to present an overview and summary about specific areas of a Data Catalog to reduce complexity. ■ Metrics: ability to visualize metrics in the areas of data quality or KPIs.
Knowledge Graphs: Shows business terms and metadata in a graph visualization in order to explore business concepts and their relation to each other.	**Exemplary Data Catalog Functions:** ■ Concept relationship models: represents the logical relationship between different business concepts in an interactive graph of nodes and edges.

3.2. Role Model

Data Catalogs can be considered as multi-sided platforms, as they are playing an intermediate and matchmaking role to connect several parties that work with data within an enterprise. They thus tie together data supply and data demand side. As the success of multi-sided platforms highly correlates with the number of users that are actively creating content, Data Catalogs need to deliver value-add for many of the data-related roles within an enterprise. More active users joining the Data Catalog platform lead to direct and cross-sided network effects. Direct network effects exist e.g. when a data analyst describes their experience handling a data asset and this experience report is useful to other data analysts of the enterprise. This will lead to more users of the same side joining the platform and hence to an increase of the platform value. Cross-sided network effects are created when participation of one side attracts participants of another side. This is the case when a data steward creates descriptions of data assets that enable data users to work with the data.

Thus, Data Catalogs need to provide useful features and value-add for as many data-related roles in the enterprise as possible to support their respective tasks and tackle current issues in the enterprise data landscape. The following role model describes these roles and their data-related challenges without a Data Catalog. It further explains how a Data Catalog can support the roles to fulfill their responsibilities.

An initial version of a role model was provided in the first version of this report (Korte et al., 2019). This model was enhanced and solidified by insights derived from further Data Catalog projects in industry and interviews with Data Catalog stakeholders in practice. The roles considered in the following have been identified within the practice projects as the main beneficiaries of a Data Catalog implementation. Additionally, the role of a Data Catalog Manager is introduced as an intermediary needed for the seamless functioning of a Data Catalogs as multi-sided platform for data in the enterprise. Following, the roles are presented according to their position in the data value chain.

Role	**Data Owner**
A data owner is assigned to a business unit and accountable for certain kinds of data thereof (e.g. data of a specific product) (Abraham et al., 2019). Responsibilities encompass quality, security and compliance of data (Gröger, 2021).	
Existing Challenges without Data Catalog	As data ownership comes as a task additional to daily business, data ownership tasks are executed in an ad-hoc manner, e.g. by answering questions of data users. Due to shadow IT, data owners do not have a full overview where and in which context data of their domain are used. The lack of a common toolset further leads to problems in standardization (e.g. of data descriptions or policies across domains) and missing automation of repetitive tasks.
Motivation for Data Catalog Usage	With a Data Catalog data owners can register data in their ownership, describe data with the support of automation functions and in a standardized manner. Further, they assess where data of their domain are used. Ad-hoc communication efforts may be reduced. By attaching policies and rules to data assets, awareness for conformant data handling can be created. Additionally, the onboarding of new data roles such as data stewards can be facilitated.

Role	**Data Steward**
Data stewards are in charge to manage data according to the business needs. They are realizing the defined data policies and procedures on a business and technical level (Gröger, 2021). Data stewards have knowledge about the requirements of business and data and translate those requirements into technical specifications (Abraham et al., 2019).	
Existing Challenges without Data Catalog	While the frontend of applications is well known by the business users, the identification of data structures remains a complex task. This holds especially true for ERP and CRM systems. Data curation is a manual process that does not scale throughout the enterprise without appropriate tooling. Because of missing context information, interpretation of data is a time-consuming task, especially if the right contact persons are not identified. This all leads to a time-consuming preparation of data for use in digital processes.

Motivation for Data Catalog Usage	Data Catalogs enable data stewards to quickly identify data ownership. This is especially useful for job entrants without a network. Further, technical and business metadata about a data object can be obtained. In this sense, Data Catalogs act as a central point of knowledge for data stewards. Another helpful feature for data stewards is the ability to break down the data to the actual point of entry and to know the original data source. Through automation features data stewards are supported conducting data curation tasks. Data Catalogs can further foster data quality metrics across the whole organization and support data assessment and implementing the right policies to prepare data for re-use.

Role	**Enterprise Data Steward/ Data Governance Specialist**

Enterprise data stewards take on an overseeing role in data governance activities. Responsibilities include leadership, guidance and coordination of data stewards across domains, development of a data governance vision, ensuring that data governance is conducted according to the vision and business goals, and the creation and management of data governance artefacts (Plotkin, 2014).

Existing Challenges without Data Catalog	The interpretation of data that are not part of the core business is difficult as no definition or context of data is given. This gets even more difficult for data where the contact persons are unknown. Responsibilities for data are maintained in spreadsheets and are not available to all employees. The maturity of data governance may differ across the business units of the enterprise.
Motivation for Data Catalog Usage	Data Catalogs enable transparency about data, including their usage (in processes) and provenance. Enterprise data stewards are able to document roles and responsibilities in a central point of contact. Scaling of data governance measures across the whole enterprise may be improved. With the help of dashboards for data quality and data governance, the effect of data governance initiatives can be assessed. In this sense, Data Catalogs act as a checkpoint for correct data management and provisioning.

Role	**Data Catalog Manager**

A data catalog manager is responsible for the provisioning of a Data Catalog solution. Amongst others their goals are the high availability of the solution, the timely delivery of updates and support of Data Catalog users. All in all the data catalog manager wants to improve the experience of all Data Catalog users and takes over an important intermediary role.

Existing Challenges without Data Catalog	As this role does not exist without a Data Catalog, there are no existing challenges without a Data Catalog solution.
Motivation for Data Catalog Usage	To enhance the satisfaction of business and technical users Data Catalog managers need to provide frequent bug fixes and updates for the selected Data Catalog solution. As experts they support users during their onboarding and their daily business (e.g. supporting the import of data to the Data Catalog or providing data lineage integrations). They are also responsible for personalizing features to enhance the experience of each user. Furthermore they may participate in standardization of data management tasks (e.g. by implementing predefined data management workflows).

Role	Data Protection Officer
Data protection officers are in charge to ensure that data users handle personal data according to the applicable data protection rules (Mateeva, 2019). Data protection officers monitor the data processing activities in the enterprise to ensure data is handled properly. They also engage in activities to proactively prevent the abuse of personal data and promote aspects of data privacy by design (Tsormpatzoudi et al., 2016).	
Existing Challenges without Data Catalog	High quality metadata are an important element to fulfil data protection tasks. However, they are often not available. Data usage within the organization is often not well documented and hence it is not visible where sensitive data are being used. Additionally, data users are not aware that the data they use contains personal identifiable information. This is especially critical when data processing agreements are being revoked and personal data are to be found and deleted. If data are being provided by a third party it is hard to ensure that data usage agreements are being kept and data are only used for the defined purposes.
Motivation for Data Catalog Usage	With the help of Data Catalogs transparency about data usage within the enterprise can be created. This enables Data protection officers to ensure regulatory demands and data usage policies are being kept. Because of rich metadata maintained in the Data Catalog, data can be assessed more easily and other departments, e.g. the legal department, can support the data protection officer in certain situations. As data protection officers are enabled to describe data classifications and guidelines for data handling, Data Catalogs can act as a gatekeeper for data usage within the organization. In this context, Data Catalogs support the question how critical data are and if the context will allow data usage.

Role	Data Architect
Data architects are in charge to define data objects and create, deploy and maintain conceptual and logical data models and map these to physical data models (Abraham et al., 2019; Fadler & Legner, 2021).	
Existing Challenges without Data Catalog	Data are hard to find and identify within the enterprise. Sometimes it is not clear who is the data owner or what is the underlying data schema. Furthermore, the leading system or the data origin are unknown and hence it is unclear how data can be accessed on a technical level. This leads to high communication efforts, especially for people without the required network. In big organizations the use of different terminology across the departments hampers the understanding of certain concepts and the related data.
Motivation for Data Catalog Usage	Data Catalogs help data architects in creating transparency about the enterprise data landscape. They get information about where data originate, where they reside and in which places the data is used. Data Architects are able to determine responsible persons for data objects and address them in case of doubts. They are able to obtain knowledge on how data can be accessed and which interfaces exist. Further, the collaboration with business users will be improved, as terms across departments can be aligned leveraging the business glossary capabilities of a Data Catalog solution.

Role	Data Engineer
Data engineers are in charge of creating and providing the data basis for use in data analysis (Gröger, 2021). This includes tasks like data discovery, data preparation, and implementing and maintaining data pipelines (Cao, 2019).	
Existing Challenges without Data Catalog	Data interfaces and data formats are sometimes changed without further notice. Data fields lack standardization and often have an insufficient data quality or data quantity. The analysis of data suitability needs to be manually conducted before data use. Often professionals rely on their personal network to get information about data context and its location, making it especially hard for job entrants to fulfill their job. Additionally, normalization of data is a tedious task and takes a long time.
Motivation for Data Catalog Usage	Data Engineers obtain a fast overview of data contents and metadata, including data structure. They can easily identify if data are suitable for use and conduct plausibility checks with the support of different Data Catalog capabilities. They can autonomously assess if they are allowed to work with the available data or if it has sensitive content that would permit the further propagation of the considered data asset. Data Lineage capabilities enable data engineers to see if they are dealing with original data or if it has been pre-processed. In case of questions they are able to contact the responsible persons.

Role	Data Scientist
A data scientist is in charge to develop, deploy and maintain advanced analytics models (Fadler & Legner, 2021), therefore being more directed towards long-term and future developments compared to a data analyst.	
Existing Challenges without Data Catalog	Information about data exists as tribal knowledge. Outside of the information silos it is unclear where data originates, where it resides and who owns the data. Additionally, often context about data is missing and data quality is unknown. Access to sensitive data is sometimes prohibited ex-ante to avoid legal conflicts instead of regulating access based on a clear process and depending on data properties. Useful data may get lost during restructuring.
Motivation for Data Catalog Usage	Data Catalogs provide data transparency and enable data scientists to work with data. With Data Catalogs data scientists are able to easier discover and access data assets. They can obtain clear definitions and context of data and see who is responsible for a specific data asset. Data Catalogs further provide features that make data assessable and traceable throughout the enterprise. Data Catalogs support data handling according to privacy legislations, facilitating the process for data scientists to determine if a data asset can be used within a specific contexts. If more specific questions occur responsible roles can be identified and contacted in a timely manner.

Role	Data Analyst
A data analyst is in charge to develop, deploy and maintain reports and ad-hoc analyses (Fadler & Legner, 2021), therefore being more directed towards short term or past developments compared to a data scientist.	
Existing Challenges without Data Catalog	Data are often hard to find for unexperienced users. Further, data quality problems are not immediately recognized. Sometimes it is not clear if data assets contain personal identifiable information and if they can be used for the specific purpose. This leads to high manual efforts evaluating certain kinds of data. Another problem is a lack of information regarding data structure, which may lead to unexpected efforts when data of different formats need to be merged. Missing business or operational metadata makes it hard to grasp the data context (e.g. if data contains a certain bias due the data collection process or data pre-preparation).
Motivation for Data Catalog Usage	With a Data Catalog, data becomes findable. Data Analysts can easily obtain an overview of the properties of a data set, e.g. through KPIs. By these means initial plausibility checks become possible. Data lineage capabilities provide data analysts with information whether data has been pre-processed. Additionally it can be described in which context data was collected. Data Catalogs further maintain indication of whether data analysts are allowed to work with the data in a certain project or if data are too sensitive. Furthermore, Data Catalogs enable organizations to provide clear definitions of responsibilities and therefore accelerate collaboration. In case of doubts contact persons can be identified.

Role	Data Citizen
Data Citizens originate from business areas outside of advanced analytics and are able to combine their domain knowledge with data science skills. Therefore they bridge the gap between the world of business and advanced analytics (Gröger, 2018).	
Existing Challenges without Data Catalog	It is difficult to identify useful data outside of the own business unit and even more unclear how to access it. For existing data, provenance is not always traceable and data quality is unknown. If data comes from outside the business unit or the enterprise, it is sometimes not known how this data can be used.
Motivation for Data Catalog Usage	With the help of Data Catalogs data citizens are able to obtain data more easily and with a higher reliability. Data Catalogs create transparency, especially regarding responsibilities for data assets. Data analytics projects can be accelerated due to data search and discovery as well as data assessment capabilities. Data Citizens can easily obtain an overview if they are allowed to include the data in their use case.

3.3. Integration Model

The integration of the selected Data Catalog solution into the existing enterprise infrastructure is a further critical aspect of Data Catalog implementation projects. In order to describe possible technical integration scenarios this section presents a generic integration model for Data Catalogs based on the analysis of technical architectures and integration options of different providers. Due to its generic approach, the integration architecture of specific Data Catalog solutions may differ slightly from the depicted model. Figure 4 presents the derived Data Catalog integration model.

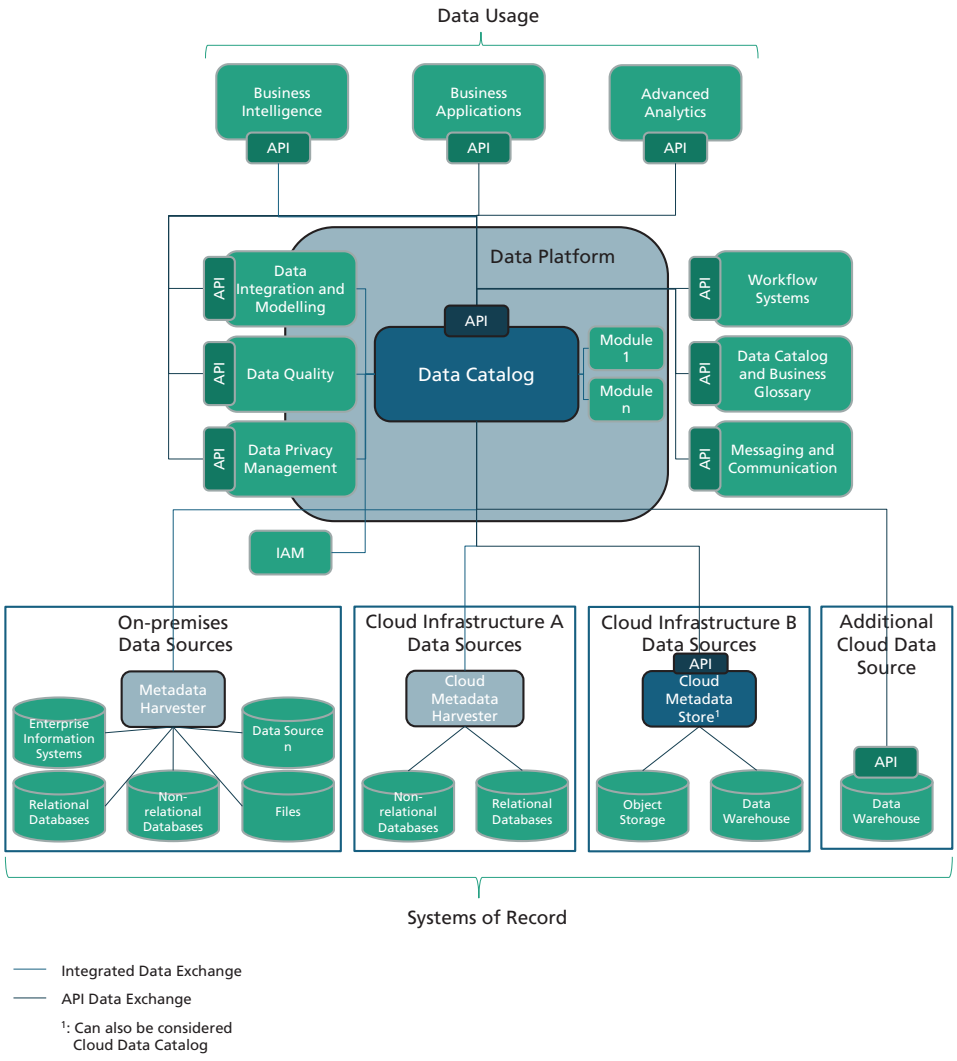


Figure 4: Data Catalog integration model

The integration model can be vertically divided into three parts. The lower part depicts possible data sources occurring in the enterprise. These systems may or may not be authoritative data sources. Their integration pattern depends on the source and deployment type. The middle part shows the enterprise data curation and preparation environment and describes how other data management software can be integrated with Data Catalogs. The upper part showcases the integration of curated metadata for enterprise use cases. The different types of Data Catalogs

and their components are colored in different shades of blue. The first type includes enterprise Data Catalogs as stand-alone solutions or parts of a Data Platform, which can be deployed as managed service or on-premises. The second type includes cloud Platform Data Catalogs, which are shown in the bottom right as part of the cloud infrastructure B. In the following, the three parts of data sourcing, data curation and data usage will be elaborated in greater detail.

As depicted in the bottom part of the integration model Data Catalogs can be automatically populated with metadata of different types of data sources. While, due to limit in space, only a fraction of possible data sources are displayed, a holistic picture of integration patterns between Data Catalogs and enterprise systems is shown. Many businesses, especially those with a background in manufacturing or industrial production, still have a non-negligible amount of on-premises enterprise information systems and databases. They often include essential systems such as ERP or CRM. To gather holistic metadata from on-premises systems so-called metadata harvesters need to be deployed. Metadata harvesters need to be installed in the same environment as the data sources in order to retrieve metadata, run data profiling or create lineage information. After this processing, the results are send to the Data Catalog or data platform solution. For on-premises-deployments of Data Catalogs this circumstance is negligible. However, if Data Catalogs or data platforms are deployed in the cloud, this implies that an additional on-premises component needs to be maintained within the enterprise infrastructure. Some solutions include metadata harvesters that only provide outgoing data, so that no incoming ports need to be opened for data exchange. For the integration of metadata from CRM or ERP systems some providers offer special connectors. This is due to large and opaque data models, which prohibit regular data ingestion, complex discovery of metadata, and the fact that not all data tables are valuable to the business.

When metadata are to be ingested from cloud platforms external to the Data Catalog environment different options for integration are possible. The first option consists in the deployment of a cloud metadata harvester in the cloud provider's environment. Similar to on-premises metadata harvesters these tools profile the data sources in the cloud environment and deliver metadata according to predefined rules to the Data Catalog. The second option is to rely on already existing cloud metadata stores[1] (which can also be called cloud Data Catalogs) to extract metadata via the exposed APIs. As a prerequisite the respective cloud data sources such as relational databases or object storage need to be registered in the metadata store. In this sense the Data Catalog or data platform acts as a catalog of catalogs, being able to integrate metadata already curated in other places. Lastly metadata of several applications deployed in virtual private clouds can be accessed using direct integrations or by leveraging the respective APIs.

In the central part of the graphics the integration patterns between Data Catalogs, which may be part of a data platform, with other applications for enterprise data curation are shown. Data Catalogs and their platform counterparts usually support existing identity and access management solutions in the enterprise such as LDAP and Kerberos. This facilitates single sign-on throughout the enterprise. On the left side further metadata sources that integrate with Data Catalog solutions are depicted. They can include data quality tools, tools for ensuring data privacy, data integration and modelling tools, master data or API management. Once these tools are part of a modular data platform that also includes the Data Catalog, bilateral metadata exchange should be a native feature. In this scenario, the Data Catalog often acts as a first central point of contact for data-related activities. In case the metadata application is not part of a bigger platform solution APIs can be leveraged in a bilateral manner to integrate metadata from each system. The integration of other metadata applications is especially useful in cases where the Data Catalog lacks capabilities in a certain areas, e.g. data quality.

1 One example for such a solution is AWS Glue (https://aws.amazon.com/glue/)

On the right side further applications to be possibly integrated with a Data Catalog are shown. In contrast to the former solutions these applications are not offered regularly as modules of a Data Platform, but provide extended capabilities to a Data Catalog solution. E.g. Data Catalogs integrate with workflow or ticketing systems such as Jira[2] and ServiceNow[3] to manage data issues or data access workflows. Also, existing messaging and communication platforms can be integrated. This fosters a central point of communication for enterprise data and enables to notify users outside of the Data Catalog about issues or developments. Furthermore, metadata existing in other Data Catalogs, e.g. Data Catalogs for a specific business unit, can be integrated, too. This will foster the establishment of a "Data Catalog of Data Catalogs". Usually the mentioned applications can be integrated via their respective API.

In the upper part business applications that leverage metadata for data usage are depicted. Business intelligence software such as Power BI[4] or Tableau[5] are making use of metadata maintained in Data Catalogs for faster data search and interpretation. Vice-versa these systems may be a useful source for metadata themselves. Data Catalogs can maintain a repository of BI projects or dashboards, which makes them available to a broader audience, presents input data of dashboards, and allows lineage tracking in case of data defects. The same benefits apply for integrations of advanced analytics platforms such as Dataiku[6] or other business applications. Usually the exchange of metadata between the systems is based on existing APIs.

2 https://www.atlassian.com/software/jira

3 https://www.servicenow.com/

4 https://powerbi.microsoft.com/de-de/

5 https://www.tableau.com/

6 https://www.dataiku.com/

4. Market Overview

4.1. Market Characteristics

Industry Structure and Dynamics:

Today's Data Catalog market includes about 50 commercial solutions and about ten open-source solutions. According to data derived from crunchbase.com based on the list of commercial Data Catalog providers, the biggest share of solution offerings comes from USA based companies, making up approximately 70 percent of all providers. European companies account for about 20 percent of Data Catalog providers while other companies are headquartered in Canada, Australia and Singapore. A visualization of Data Catalog providers headquarter locations is shown in Figure 5. Currently, about half of Data Catalog providing companies offer a Data Catalog solution as their main product, while the other half also provides additional products, mostly other enterprise software solutions. This means that, even if the Data Catalog offering is not the main product for a company, these companies usually make most of the revenue by providing other information technology and software solutions.

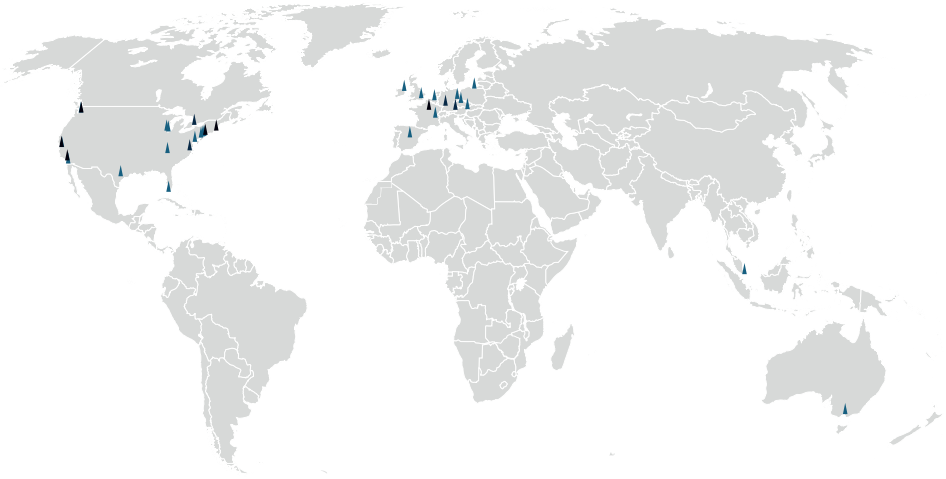


Figure 5. Locations of Data Catalog providers

At the moment, the Data Catalog industry is characterized by two adversarial developments. On the one hand, new product offerings enter the market. These new entrants can be subdivided into three main categories. Solutions in the first category try to enable Data Catalog use for small and medium enterprises. In contrast to the currently leading Data Catalog solutions they promise easier integration and maintenance and lower cost while focusing on the core data cataloguing capabilities. Examples for new market entrants in this category are Zeena and Atlan. Solutions that are directed towards specific use cases represent the second category of market entrants. They include solutions that support primarily the role of data engineers and data analysts. This is achieved by fostering data collaboration and enabling a wider use of analytics e.g. through sharing and explanation of SQL queries. Examples for such Data Catalog solutions are Castor and Secoda, both introduced in 2020. Also Confluent as a solution focused on data streams can be classified into this category. The third category of market entrants is

represented through companies that created business models on top of open-source offerings. These companies originate from teams that created custom Data Catalogs within an enterprise, which were later published as open source. Examples include Metaphor for Linkedin DataHub and Stemma for Lyft Amundsen (now under governance of the LF AI & Data Foundation).

In contrast to this, also market consolidation has been taking place since the release of last report in 2019. In the area of corporate consolidation several acquisitions can be noted. Hitachi Vantara put a high focus on integrating data cataloguing capability into its service offering, acquiring three companies with a Data Catalog offering in terms of Waterline Data, Pentaho and Io-Tahoe. These solutions are now integrated into the Hitachi Vantara DataOps suite. Acquisitions with similar intent were the ones of Unify Software by Dell Boomi (which was spun-off and sold to private equity) and of Infogix by Precisely (Brust, 2021). This emphasizes the importance to incorporate data cataloguing and metadata management capabilities for providers of software solutions in the area of data management and digital transformation. Additionally it is worth mentioning that no relevant solution provider disappeared from the market.

Further, acquisitions are being used to enhance the capabilities of existing Data Catalog offerings. This was the case for Collibra which acquired SQLdep and OwlDQ to enhance its capabilities in the area of data lineage and data quality, respectively. Informatica purchased Compact Solutions to extend its capabilities for more extended metadata ingestion capabilities. These developments can be seen as parts of the ongoing functional consolidation also discussed in section 4.4. It also emphasizes that market leading Data Catalog providers have gained the financial position to take over competitors or companies offering complementary services.

According to different market research organizations the market for Data Catalogs still has high growth potential. While the current and projected market size differs depending on the publication, yearly growth rates between 14.6 (Market Growth Reports, 2021) and 23.1 percent (Mordor Intelligence, 2021) are predicted. The USA are currently the dominating market, accounting for approximately a third of Data Catalog revenue. While the Asian market is currently lagging behind in terms of Data Catalog implementation, it has the highest growth prospects till 2026. The market in Europe and in North America is forecasted to grow approximately at the same rate (Mordor Intelligence, 2021). Due to low market saturation and high growth potential, vast opportunities for established providers as well as newcomers to increase customer base and revenue are envisioned.

Distribution Channels:

Data Catalogs are provided mainly via three distribution channels. First, solutions can be implemented by the consumers directly and with the Data Catalog provider as the only partner. In this case the solution provider also offers additional services such as training, and help during implementation or customization. Another option is the implementation through service partners that have the skills and expertise to deliver the aforementioned services. These partners are often certified or listed by the respective solution provider. A third option is the purchase of a solution over the marketplace of a public cloud provider. These "as a service" offerings use subscription models and are delivered on a suitable infrastructure.

Scientific Evaluation:

To conclude, the Data Catalog industry is summarized according to the so-called consolidation curve (Deans et al., 2002). The consolidation curve represents a model that describes the development and consolidation of an industry from an economy-based perspective. The development of an industry is subdivided into four different stages, each with distinct market properties and behavior and mainly characterized by the market share of the three biggest players. In the opening stage usually one start-up or monopoly emerges in a new industry.

The monopoly is rapidly destroyed as new market participants emerge quickly. In this early stage of an industry first movers are following a strategy of forming a big customer base and creating entry barriers for later market entrants by protecting their ideas or technology. In the second stage companies will build scale regarding markets and industries. Major players are emerging and acquire competitors to form enterprises, leading to strong consolidation. In stage three and with further time, the market is going to consolidate further, as companies try to extend their core business and outgrow the competition. Only five to 12 major players with a high profitability are left in the end of this so-called focus stage. In the last stage of balance and alliance only the biggest players survive. The concentration rate of an industry reaches a plateau or may dip. Once reaching this stage companies try to defend their top position by building alliances and serve new products in other markets.

Based on the industry characteristics described before, the Data Catalog market can be classified into the beginning of the scaling stage, which is characterized by a market share of 15 to 45 percent of the top three players. While mayor players start to emerge, new players still enter the game. To strengthen their position major players are starting to acquire smaller competitors and expand their service offerings. Based on the consolidation curve theory, further acquisitions can be expected until high consolidation sets in. Big players will grow in market share and extend their core business by further services. From a technological perspective and under consideration of the industry life cycle theory (Utterback & Abernathy, 1975) Data Catalogs innovation will move from product characteristics and product quality towards process and cost optimization approaches. This implies more cost effective solutions can be expected in the future.

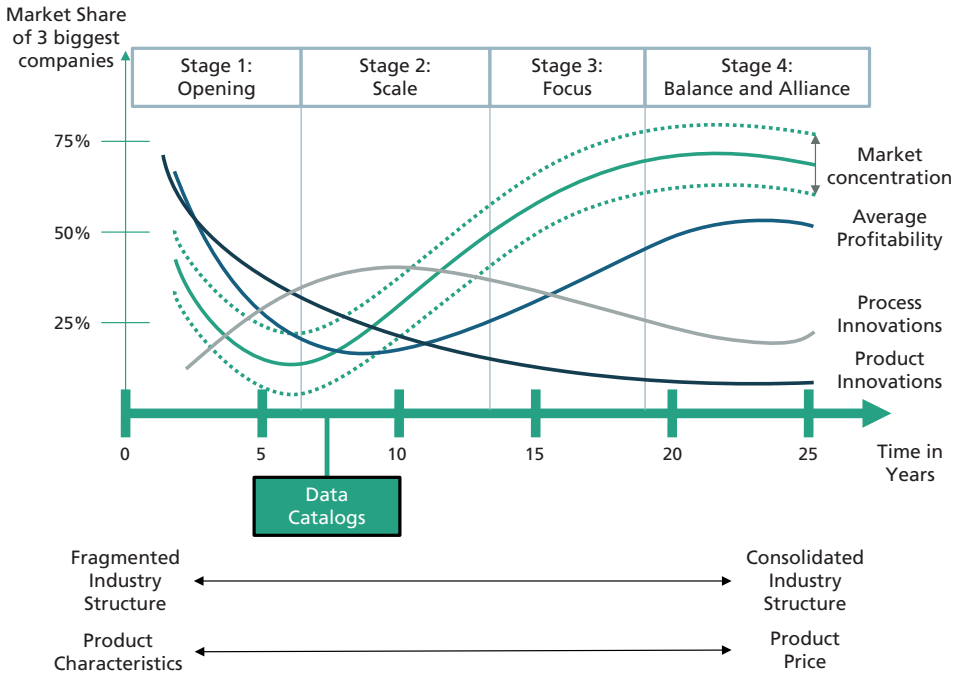


Figure 6. Consolidation curve, industry life cycle and Data Catalog market

4.2. Data Catalog Solutions

The following section presents an overview about currently available Data Catalog solutions. The selection of market offerings was based on the legacy report (Korte et al., 2019). Details about included solutions in comparison to the legacy report can be found in Appendix A.

This overview comprises characteristic vendor data summarized in a tabular overview. It further includes a short description including adoption, reference customers, the vision and recent developments of the solution. In the tabular overview solutions are described by following characteristics:

Commercial solution Property	Open-source solution Property	Explanation
Vendor	Governing organization	Name of vendor/ Name of governing organization for open-source solution
Solution type	Solution type	Type of solution according to characterization in section 2.2
Inclusion in first version	Inclusion in first version	Inclusion of Data Catalog solution in first version of the report
Headquarter	Governing organization location	Location of vendor headquarter or governing organization
Locations	/	Further locations for vendor activities
Vendor foundation date	Foundation date governing organization	Foundation date of vendor or governing organization
Revenue/ Proportion of solution	Stars/ Forks	Approximate revenue of vendor according to crunchbase.com and share of Data Catalog solution in revenue (small, medium, high)/ Stars and forks of open-source solution as of April 2022
Employees	Contributors	Number of employees according to crunchbase.com or the annual report/ Number of contributors and contributing organizations according to Git repository as of April 2022
Website	Website	Link to solution website/ Link to solution website and Git repository

In the following, the solutions are ordered alphabetically according to vendor or governing organization name.

A) Alation Data Catalog

Vendor	Alation Inc.
Solution type	Enterprise Data Catalog
Inclusion in first version	Yes
Headquarter	USA
Locations	Great Britain, India
Vendor foundation date	2012
Revenue/ Proportion of solution	$50M to $100M/ high
Employees	501-1000

Alation Data Catalog represents an extensive enterprise Data Catalog with a broad customer base, especially in the North-American Market. Recently Alation established new offices in Europe to further foster its representation in this growing region. Since the foundation in 2012 Alation was able to raise over $200M in venture capital. Alation is being leveraged by corporations in different industries such as finance, technology or manufacturing with consumers like AON, Cisco and GE Aviation. Originating in the field of data governance, the solution enhanced its capabilities to support data analytics use cases. Most significant recent developments include the enhancement of search capabilities by active metadata, the manifestation as central point of the enterprise data ecosystem, and the transition towards a cloud native architecture.

B) Apache Atlas

Governing organization	Apache Software Foundation
Solution type	Enterprise Data Catalog (Open-source)
Inclusion in first version	No
Governing organization location	USA
Foundation date governing organization	1999
Stars/ Forks	1´200/ 685
Contributors	115 – Cloudera, Hortonworks, IBM i.a.
Website	https://atlas.apache.org/ https://github.com/apache/atlas

Apache Atlas is an open-source Data Catalog with a focus on data governance and metadata management under the governance of the Apache Software Foundation. Apache Atlas was started by Hortonworks (now part of Cloudera) and joined the Apache Software Foundation in 2015. Since 2017 it is a top-level project. Next to companies implementing Apache Atlas as-is,

several Data Catalog providers leverage Apache Atlas to provide their own offering with extended capabilities. Amongst others these vendors include Atlan and Microsoft Purview. Apache Atlas was selected as open-source reference due to its high popularity. Apache Atlas serves as an incubator for Metadata Management initiatives such as Egeria, which aims at building a framework for the exchange of metadata between different tools and platforms in the enterprise data ecosystem[1].

C) Atlan Platform

Vendor	Atlan Pte. Ltd.
Solution type	Enterprise Data Catalog
Inclusion in first version	No
Headquarter	Singapore
Locations	India
Vendor foundation date	2018
Revenue/ Proportion of solution	$50M to $100M/ high
Employees	51-100

Atlan is an emerging vendor in the enterprise Data Catalog landscape. It was founded in 2018 by former data team members. The solution is primarily focused on collaborative data analytics use cases while additionally supporting enterprise data governance. Atlan mainly supports cloud data sources. So far, Atlan raised about $70M in venture capital. Current customers are primarily set in Asia and come from different sectors. Most familiar customers include Unilever and Mahindra Group.[2] While the company is mainly based in Asia, it seeks to enhance its customer base especially in North America.

D) Collibra Data Intelligence Cloud (includes Collibra Data Catalog)

Vendor	Collibra NV
Solution type	Data Platform (including Enterprise Data Catalog)
Inclusion in first version	Yes (Collibra Data Governance Center)
Headquarter	Belgium
Locations	Several European countries, USA, Australia
Vendor foundation date	2008

1 https://egeria.odpi.org/

2 https://techcrunch.com/2019/07/01/atlan-socialcops/

Vendor	Collibra NV
Revenue/ Proportion of solution	$100M to $500M/ high
Employees	501-1000

Founded in 2008 in Brussels, Belgium Collibra is a software company providing data intelligence solutions. Collibra is one of the biggest providers of Data Catalog software on the market. The first version of a Data Catalog was released in 2009 with Collibra Data Governance Center. Collibra Data Intelligence Cloud and its on-premises version Collibra Data Intelligence Platform build upon the existing modules (such as Collibra Data Catalog) to provide a holistic solution for metadata management in the enterprise. Due to historic reasons, the main focus of Collibra lies on data governance. However, it also supports data analytics use cases. Customers of Collibra are mainly situated in Europe and North America and based in many industries such as financial services, healthcare, telecommunications and manufacturing. Most significant developments included the introduction of Collibra Data Intelligence Cloud superseding Collibra Data Governance Center, which was included in last report. While the legacy solution is still supported and receives regular updates, some new features are specific for Collibra Data Intelligence. Further recent developments include the possibility of including data quality capabilities through the former OwlDQ. Additionally, automation features were enhanced and the solution further positioned as central data platform for the enterprise.

E) Lumada Data Catalog

Vendor	Hitachi Vantara Corporation
Solution type	Enterprise Data Catalog
Inclusion in first version	Yes (former Waterline Data Smart Data Catalog)
Headquarter	USA
Locations	Manifold locations worldwide
Vendor foundation date	2017
Revenue/ Proportion of solution	$1B to $10B/ low
Employees	5001-10000

Lumada Data Catalog is an enterprise Data Catalog solution provided by Hitachi Vantara. The solution was incorporated into the service portfolio of Hitachi Vantara by the acquisition of Waterline Data Smart Data Catalog, which was covered in the previous report. Lumada Data Catalog integrates with other Lumada data services but can also be leveraged as a stand-alone solution. The customer base of Lumada Data Catalog is mostly based in North America and is located in sectors such as financial services and healthcare. Being originally a Data Catalog solution for data lakes, Lumada Data Catalog strongly focuses on the capability of data discovery, through which data governance and data analytics can be supported. Recent improvements include the enhancement of automation and artificial intelligence capabilities for duplicate detection and integration of business rules. Further upgrades improved the search and lineage capabilities.

F) IBM Watson® Knowledge Catalog

Vendor	IBM Corp.
Solution type	Enterprise Data Catalog
Inclusion in first version	Yes
Headquarter	USA
Locations	Manifold locations worldwide
Vendor foundation date	1911
Revenue/ Proportion of solution	More than $10B/ low
Employees	More than 10000

IBM Watson Knowledge Catalog is an enterprise Data Catalog part of 'IBM Cloud Pak for Data', a data and artificial intelligence platform based on RedHat OpenShift. IBM Watson Knowledge Catalog can therefore be deployed on-premises or as managed service. Watson Knowledge Catalog is the successor of IBM Information Governance Catalog introduced in 2018. Since then it has been adopted by customers in different industries, most prominently in financial services. Watson Knowledge Catalog supports a holistic view on data, from governance as well as analytics perspective. Developments since the last report include the further integration of machine learning capabilities, e.g. for recommendations or the classification and tagging of data. Additional improvements are the implementation of data quality capabilities, enhanced usability and integration features, and the possibility for integration of sample data.

G) Intelligent Data Management Cloud (includes Enterprise Data Catalog and Axon Data Governance)

Vendor	Informatica Inc.
Solution type	Data Platform (including Enterprise Data Catalog)
Inclusion in first version	Yes (Enterprise Data Catalog and Axon Data Governance)
Headquarter	USA
Locations	Manifold locations worldwide
Vendor foundation date	1993
Revenue/ Proportion of solution	$1B to $10B/ high
Employees	5001-10000

Informatica is one of the biggest providers of metadata management solutions. In 2021 it announced the integration of its wide product spectrum into a modular cloud platform. This so-called Intelligent Data Management Cloud also includes its Data Catalog solutions Informatica Enterprise Data Catalog (focus of technical users) and Axon Data Governance (focus on business

users). These solutions may be combined with other modules of the Informatica ecosystem to provide holistic metadata management capabilities. Informatica has customers all over the world and in diverse industry sectors. Testimonials include companies such as AXA, Avis Budget Group and Rabobank. Informatica Enterprise Data Catalog and Axon Data Governance support mainly data governance activities. Other modules of Informatica Intelligent Data Management Cloud however target the support of data analytics through data integration and data engineering functionality. Recent Data Catalog improvements include further data source connectors, the introduction of artificial intelligence-based data similarity and mapping features, as well as the support for curating Apache Kafka data streams.

H) Amundsen

Governing organization	**LF AI & Data Foundation**
Solution type	Enterprise Data Catalog (Open Source)
Inclusion in first version	No
Governing organization location	USA
Foundation date governing organization	2018
Stars/ Forks	3´200/ 803
Contributors	200 – Lyft, Databricks, ING i.a.
Website	ttps://www.amundsen.io/ https://github.com/amundsen-io/amundsen

Amundsen is an open-source enterprise Data Catalog under the governance of the LF AI & Data Foundation, an umbrella foundation of the Linux Foundation that supports open-source tools for a sustainable AI and data ecosystem. Amundsen was initially started by ridesharing company Lyft to accelerate data discovery in the organization. Code repositories were published in 2019 and in 2020 Amundsen joined the LF AI & Data Foundation. It has so far been adopted by companies in the sectors of finance, information technology and entertainment and became one of the most popular open-source Data Catalogs. Amundsen was selected as community based open-source reference with high maturity. Due to its origin Amundsen is primarily focused on data analytics and productivity use cases.

I) Purview

Vendor	**Microsoft Corp.**
Solution type	Cloud Platform Data Catalog
Inclusion in first version	No
Headquarter	USA

Vendor	Microsoft Corp.
Locations	Manifold locations worldwide
Vendor foundation date	1975
Revenue/ Proportion of solution	More than $10B/ low
Employees	More than 10000

Microsoft Purview, formerly known as Azure Purview, is the successor of Microsoft Azure Data Catalog and was introduced in the end of 2020 in public preview. In fall 2021 its general availability was announced. Microsoft Purview can be categorized as a Cloud Platform Data Catalog aiming primarily at the provisioning of metadata management and data governance capabilities within the Microsoft Azure ecosystem. Being a cloud service itself it comes with benefits such as flexibility, scalability and usage-based billing.

Additional to the support of Microsoft Azure, Microsoft starts to implement capabilities for enterprise-wide metadata management such as connectors to include AWS S3, Google BigQuery and on-premises metadata. Testimonials of Microsoft Purview include companies from sectors such as Financial Services, Manufacturing and Transportation. While its current capabilities are rather limited compared to the established players, Microsoft provides an extensive roadmap and commits to implement further capabilities in the near future. By bringing together the former Azure Purview and the former Microsoft 365 Compliance portfolio under the umbrella of Microsoft Purview a single platform for data management, data governance and data protection is envisioned. Because of this holistic vision Azure Purview was included as cloud platform Data Catalog reference.

J) Oracle Enterprise Metadata Management (OEMM)

Vendor	Oracle Corp.
Solution type	Enterprise Data Catalog
Inclusion in first version	Yes
Headquarter	USA
Locations	Manifold locations worldwide
Vendor foundation date	1977
Revenue/ Proportion of solution	$1B to $10B/ low
Employees	More than 10000

Oracle Enterprise Metadata Management is an enterprise Data Catalog product of Oracle and should not be confused with Oracle Cloud Data Catalog, which is mostly directed towards the use in the Oracle Cloud ecosystem. For Enterprise Metadata Management Oracle profits from its big customer base located worldwide and based in diverse industries. Oracle Enterprise Metadata Management clearly focuses on the governance perspective while it is able to integrate well with other Oracle products supporting data analytics use cases. New features added to the Oracle service offering include collaboration features, semantic mapping and search, metadata

queries and options for incremental ingestion of metadata amongst others. However, the last mayor update for Oracle Enterprise Metadata Management was released in 2020.

K) Precisely Data360® (includes Data360® Data Catalog)

Vendor	Precisely Inc.
Solution type	Data Platform (including Enterprise Data Catalog)
Inclusion in first version	Yes (former Datum LLC Information Value Management)
Headquarter	USA
Locations	Europe, India
Vendor foundation date	1968
Revenue/ Proportion of solution	$100M to $500M/ low
Employees	1001-5000

Precisely Data360® is a data intelligence platform solution by Precisely Inc. Data 360 was investigated in the previous report under the name Datum LLC Information Value Management. In the interim, Datum LLC was acquired by Infogix, which in turn was subsequently acquired by Precisely. As part of its Data 360 platform Precisely provides an enterprise Data Catalog solution in terms of Data360 Data Catalog. Precisely customers are mainly located in North America and based in the financial services industry. Other prominent references are Comcast and General Mills. Due to historic developments the focus of the solution lies on data governance. However, recent enhancements included the initial support of data analytics teams. Further upgrades were related to the improved support of workflows and business processes as well as the integration of data assessment features.

N) Talend Data Fabric (includes Talend Data Catalog)

Vendor	Talend Inc.
Solution type	Data Platform (including Enterprise Data Catalog)
Inclusion in first version	No
Headquarter	USA
Locations	Several European countries, several Asian countries
Vendor foundation date	2004
Revenue/ Proportion of solution	$100M to $500M/ low
Employees	1001-5000

Talend Data Fabric is a modular platform for enterprise data integration offered by the provider of commercial and open-source software Talend Inc. Talend Data Catalog was released in 2018

and can be deployed as part of Talend Data Fabric or as stand-alone enterprise Data Catalog. Customers of Talend Data Fabric come from the financial services and manufacturing sectors and are situated in diverse geographical regions. Talend Data Catalog is clearly directed towards data governance. The integration of other modules of Talend Data Fabric enables the support of data analytics, especially by the facilitation of data engineering. Recent enhancements of the Data Catalog solution included better integration into the enterprise data ecosystem, improvement of search and usability and more options for customization. Furthermore, semantic mapping features were introduced.

O) Zaloni Arena DataOps Platform

Vendor	Zaloni, Inc.
Solution type	Enterprise Data Catalog
Inclusion in first version	Yes (Zaloni Data Management Platform)
Headquarter	USA
Locations	USA, India, United Arab Emirates
Vendor foundation date	2007
Revenue/ Proportion of solution	$1M to $10M/ high
Employees	101-250

Zaloni Arena DataOps Platform (former named Zaloni Data Management Platform) is an enterprise Data Catalog solution provided by Zaloni, Inc. Zaloni has been offering its Data Catalog solution for more than five years. While it was formerly focused on the management of data lakes it now supports data management and governance for the whole enterprise data ecosystem. Additionally to its data governance focus, the transition to Zaloni Arena also included the incorporation of capabilities to support data analytics such as data pipelining and data provisioning. Zaloni has customers notably in the financial and health sector. Recently Zaloni improved its solution by providing data marketplace and data provisioning capabilities, including workflows for data access. Additionally, features for data onboarding and data descriptions were added. Also, Zaloni supports the provisioning of the solution via Kubernetes and teams up with cloud providers for managed service solution offerings.

4.3. Evaluation and Comparison

While Data Catalogs have become an important aspect of many organizations' data management strategy, choosing the right vendor can be a challenging task. An important step in every Data Catalog implementation project is to determine what capabilities are needed and to get an overview of the market options. According to the functional model (section 3.1), Data Catalog capabilities have been classified into the following capability groups: Data Inventory, Data Governance, Data Assessment, Data Collaboration, Data Analytics, Data Discovery, Data Visualization, Automation and Artificial Intelligence as well as Administration.

To provide an overview of the Data Catalog market, the authors assessed the solutions presented in section 4.2 regarding their performance in given capability groups. For this purpose, a representative selection of functions of each capability group was made and the fulfillment of the functions by the individual Data Catalog solutions was assessed. The capability group of Administration was excluded from this analysis as it does not pose an explicit category of functional requirements.

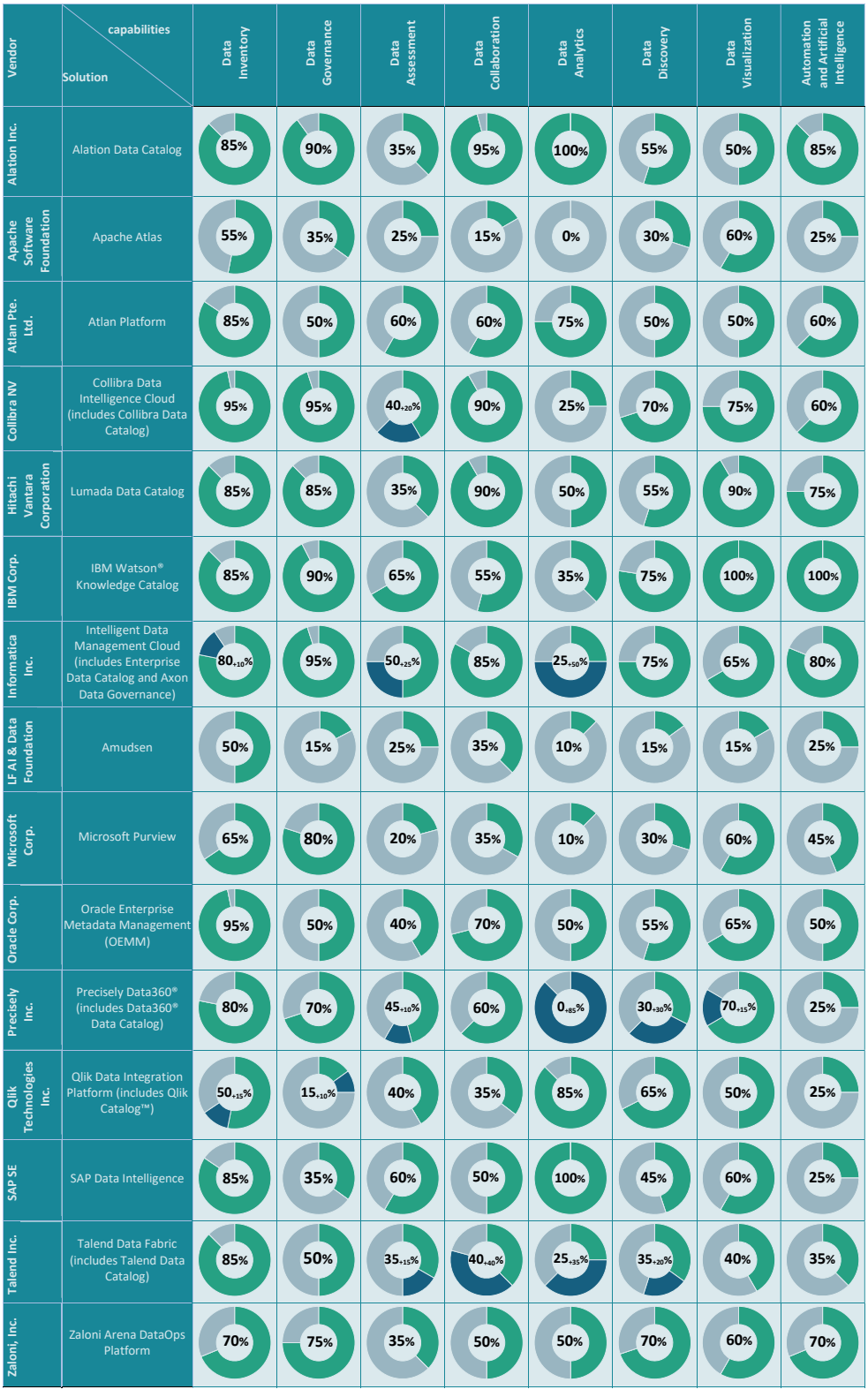


Figure 7: Vendor evaluation overview

Enterprise Data Catalogs, open-source Data Catalogs and cloud platform Data Catalogs were investigated as a whole. For Data Catalogs as part of a wider data platform a two-step approach was conducted. As a first step, the capabilities of the Data Catalog modules was assessed. In a second step additional modules of the data platform with data cataloguing functionality, e.g. for data quality or data integration, were determined and investigated. This approach was taken as it emphasizes the capabilities of the Data Catalog modules, which can also be implemented as stand-alone solutions without neglecting the further possibilities of natively integrating modules of the same provider.

Figure 7 presents the results of this analysis. The performance was rated on a scale of 0 (worst) to 100 (best) rounded to five point increments. The green segments of the ring diagrams visualize the degrees of fulfillment for Data Catalog solutions. Blue segments depict the additional capabilities provided by further modules in case a data platform solutions was assessed. Accordingly, the big numbers quantify the degrees of fulfillment for the Data Catalog while the smaller numbers quantify the additional fulfillment grades provided by further modules of a data platform.

4.4. Evaluation Findings

During the evaluation of the Data Catalog providers, a large number of practice-relevant insights were gained. This section provides a summary of the most important findings. Their description is being supported by Figure 8, which depicts the average fulfillment grades per solution type in each capability group in comparison to the average of all assessed solutions.

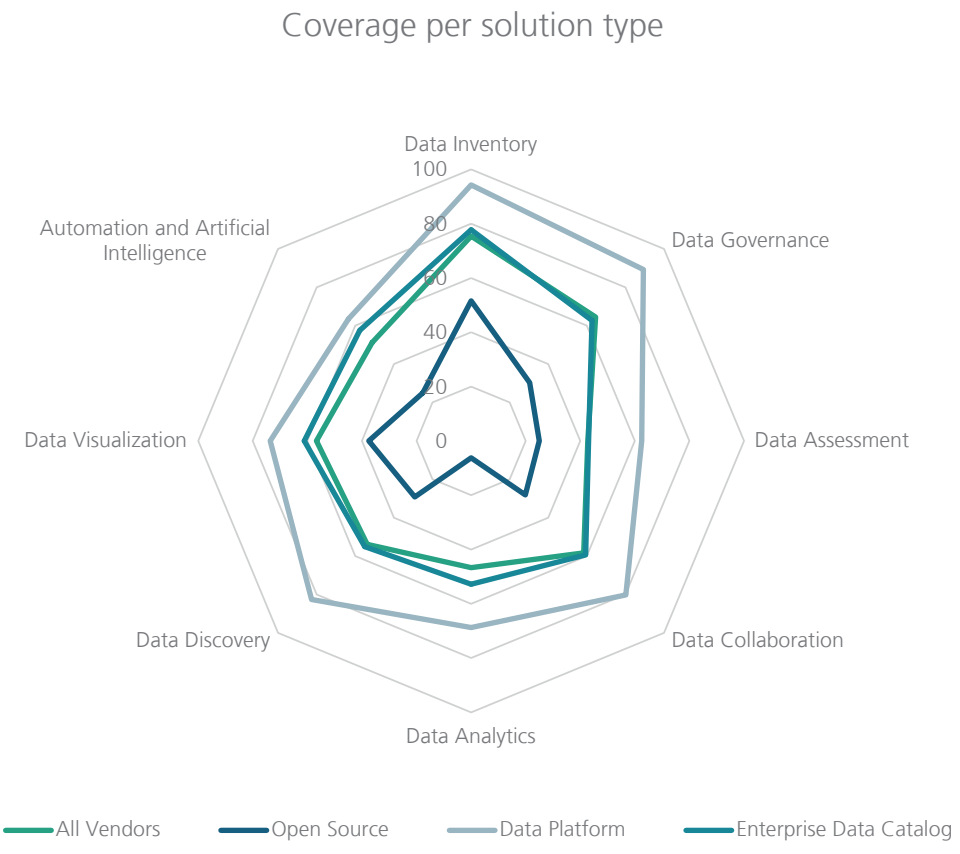


Figure 8: Coverage per solution type

What functions are commodities that every solution provides and what are functions that actually differentiate the solutions?

As illustrated in Figure 8, the highest average maturity of Data Catalog solutions exists in the capability groups of data inventory, data governance and data collaboration. Accordingly, most solutions include functions to register data, document data, define roles and responsibilities, define rules and policies, and for the collaborative metadata management. However, analyzing the standard deviation of fulfillment grades in each capability group shows the highest differences of fulfillment are present in the capability groups of automation and artificial intelligence, the support of data governance and data collaboration. Consequently, these capability groups should be thoroughly assessed in Data Catalog selection processes, once requirements in these areas are given. This finding also emphasizes the ongoing division between solutions that focus on the support of data analytics teams versus solutions that focus on data governance. It has to be mentioned that non-functional criteria may be partially dependent on functional capabilities. As an example, the automation and artificial intelligence features of a solution influence how the usability is perceived. While non-functional criteria can differ highly from solution to solution their evaluation is very much based on personal preference or the enterprise environment and therefore needs to be evaluated by the users, preferably during a proof of concept. Further, financial criteria are a major differentiating factor. But because of the complex and widely varying price factors, costs can only be compared in the context of each organization's own Data Catalog implementation project.

What are the differences between the solution types? How do open-source and data platform solutions perform?

Figure 8 also illustrates the performance of open-source and data platform solutions. Current open-source Data Catalog solutions only cover limited functionalities in comparison to the average score of all the vendors studied. Especially, open-source software achieves only low scores in the capability groups of data analytics and artificial intelligence. Moreover, the lack of documentation and information, as well as the difficulty of customizing and integrating an open-source Data Catalog solution without the support of qualified experts should be taken into consideration. On the other hand, open-source solutions may provide a good starting point for Data Catalog initiatives in small and medium enterprises or larger enterprises with a limited budget. Information maintained in open-source Data Catalogs can later be transferred in bulk via the exposed APIs to a commercial solution.

By implementing data platform solutions enterprises are able to obtain extended data cataloguing capabilities. Especially in the capability groups of data assessment and data analytics further features are provided. These features include holistic data quality analysis, as well as support for data provisioning. Also, some data governance and data privacy features are more likely to be achieved by integrating further data platform modules. However, the implementation of a data platform solution is usually more costly – especially, if many additional modules are implemented. Further, some capabilities provided by these additional modules may already be available in the enterprise due to other software packages. It is therefore necessary to align the Data Catalog initiative with existing or planned data management software initiatives in the context of enterprise architecture management.

How did the solutions develop since last report?

Compared with the results of previous report (Korte et al., 2019), several vendors have been able to improve their solution offering. Examples include IBM Watson Knowledge Catalog for data visualization, Oracle Enterprise Metadata Management for data discovery and analytics,

and Alation Data Catalog for data collaboration and governance. The latter is a good example to emphasize the general development towards including features for data collaboration such as messaging and rating systems as well as for the support of data analytics use cases. However, platform solutions may provide some of these features in form of additional modules (i.e. Informatica, Talend and Precisely). Additionally, vendors try to incorporate further artificial intelligence and automation capabilities in order to reduce the manual effort needed for data curation. However, there is still room for improvement in the areas of artificial intelligence, automation and the support of data analytics that needs to be addressed in order to improve the user experience. While some vendors developed the capability to scan the actual data or excerpts thereof, the majority of solutions is only relying on metadata and therefore cannot derive information based on data contents.

4.5. Trends on the Data Catalog Market

Further inclusion of AI & Machine learning support:

As mentioned above, the integration of AI & ML capabilities to support Data Catalog users conducting labor-intensive or reoccurring tasks such as data labelling or creation of information relationships is continuing to be a major trend amongst Data Catalog vendors. Several analysts and Data Catalog providers promote the term augmented Data Catalog to emphasize the integration of AI capabilities. While vendors were touting these capabilities as selling points for their products, users reported to be unsatisfied due to high remaining manual efforts and partially poor results. Taking into consideration the general developments of increasing private investment in AI as well as performance and cost improvements in the field (Zhang et al., 2022) a continuing adoption and rising maturity of AI methods in the Data Catalog sphere can be expected.

Data Catalog as central Point of Contact in the Data Fabric/ Data Mesh:

As the enterprise information landscape is getting more and more dispersed with multitudes of tools and databases in different cloud environments and on-premises, enterprises are keen to establish a data fabric layer to provide uniform capabilities such as data governance, data analysis or data preparation and query for data users independently of the underlying data source (Woodie, 2021). Recently, the data mesh has been proposed as a new paradigm of data management, where distributed teams manage the data in their domains and publish so-called data products (Deghani, 2019; Machado et al., 2021). While both concepts are envisioned to coexist in larger enterprises (Wells, 2021) there is a need for a central point of contact for data-related activities, e.g. data discovery. Therefore Data Catalogs need to integrate with an increasing variety of data sources such as events-streams and tools.

The Role of Data Catalogs in the Context of behavioral (active) Metadata:

Analysts recently introduced the term active metadata to emphasize the role of behavioral metadata for data governance and data-driven decision making (Simoni & Dayley, 2021). Metadata is produced in the context of many systems. Information about how data is utilized and by whom is included in behavioral metadata, which provides crucial social signals. Previously, this type of metadata has been often overlooked. Recently, Data Catalog providers start to recognize this type of metadata to move their solutions from passive (metadata presentation) to active (e.g. recommendations based on metadata) to foster effective data usage. To this end, metadata needs to a) be collected in the environments where it is created; b) connected with metadata created in other environments; c) be delivered to environments where decisions with and around data are being made. Therefore, Data Catalogs need to support standards for metadata interoperability and may take over the role as metadata orchestrator.

However, this development is still in its infancy and while several providers are advertising their active metadata management capabilities, reality shows that more developments in this area are needed to fulfill the promise.

Data Catalog Moves to the cloud:

Cloud computing[3] is continuing to transform the digital landscape of the nearly all organizations including small and medium enterprises. As investment into cloud solutions is growing year by year, Data Catalog providers are following this paradigm and a continuously increasing amount of solutions is available as cloud service. Exemplary providers for this development are Collibra and Informatica. Data Catalog as a cloud service comes with the advantages of faster setup, reduced maintenance costs, faster updates, increased scalability and the ability to focus on business topics. On the other hand, not all organizations are able to use these solutions due to security concerns. Further, it is hard to forecast the actual cost of these services due to complex pricing mechanisms. Also, for the integration of on-premises systems a locally running component is still needed.

3 For more Information on cloud computing in general, please see the report of Rosian et al. (2021)

5. Recommendations for Practice

Based on the findings of this report and further experiences made in practice projects by the authors, it is possible to formulate recommendations useful to practitioners in the early stages of Data Catalog projects.

Prior to starting a Data Catalog implementation project, the organizational maturity should be assessed. The goal should be to examine whether the prerequisites to generate a return on invest from a Data Catalog implementation are given. E.g. one needs to answer the question if a sufficient maturity of organizational processes and data literacy are already achieved or may be reached within the project lifetime. This is especially important if a Data Catalog is implemented in connection with data architecture initiatives to foster the paradigms of data mesh or data fabric. Another important aspect of the initial stages is to assess whether a cloud-based or cloud-hosted Data Catalog can be selected according to company policies and legal obligations, which exist e.g. for European critical infrastructures.

The market analysis showed that Data Catalog solutions differ highly in supported capabilities, integration patterns and costs. Therefore, the requirements gathering, requirements prioritization, and vendor preselection phase are of great importance to determine the real needs of the organization. A thorough requirements analysis enables practitioners to make informed decisions in the solution selection process under consideration of possible trade-offs and limited budgets. The requirements analysis should especially focus on capabilities that are differentiators between solutions instead of commodities.

Due to the high amount of Data Catalog market offerings, it is necessary to conduct a preselection of vendors before assessing a small amount of solutions in detail. As market transparency is limited, organizations may encounter difficulties to determine the capabilities and functions provided by each solution. Seeking the help of neutral experts is helpful especially during this phase of a Data Catalog implementation project. To gather in-depth knowledge and to make a final decision for a solution conducting a proof of concept is recommended.

During the entire project it is important to involve Data Catalog stakeholders and designated users, and to align with other data initiatives in the organization. This will help to secure support and funding. Further, a Data Catalog relies on the willingness of different Data Catalog users to provide content. As Data Catalogs are a type of platform, the attractiveness of a Data Catalog can be enhanced by extending the number of content creating users. While the integration of a maximum number of stakeholders is an important aspect of Data Catalog projects, it is also necessary to manage stakeholder expectations. Due to statements made by Data Catalog vendors, users expect a solution with a high level of automation and artificial intelligence support. However, the current maturity of these features is low and manual efforts are still needed to populate a Data Catalog with content. This gap may especially disappoint users such as data owners and data stewards that are primarily content creators and will have to deal with an increased initial workload to populate the Data Catalog.

Organizations that struggle with funding or have a low organizational maturity can leverage a lightweight Data Catalog or open-source solution to incorporate basic data cataloguing capabilities. These solutions can then serve as a starting point for implementing a larger product in the future.

6. Conclusion

Due to ever increasing amounts of data generated, regulatory requirements and the pressure to create innovative products and value from data, organizations need to enable their employees to work with data. In this endeavor Data Catalogs play an important role, as they provide a platform for finding, assessing and gaining access to data by the provisioning of holistic capabilities for metadata management. However the complex capabilities, integration options and provider landscape of Data Catalog solutions lead to challenges for data management practitioners. In this report, several concerns of high relevance for practice were answered. In particular, the focus was on answering questions that occur in the different phases of Data Catalog implementation projects.

As a starting point Data Catalogs were classified into the categories of enterprise Data Catalogs (subdivided in commercial and open-source solutions), platform solutions, cloud platform Data Catalogs, tool-specific Data Catalogs and data portals. In the remainder of the report solely the first three categories of solutions were considered, as only these can foster data management and data use across the entire enterprise.

By conducting a vendor analysis, an updated Data Catalog capability map was created. The capability map helps practitioners to understand what can be expected from a Data Catalog solution. It can be applied to structure Data Catalog requirements or user stories. The model consists of capability groups and underlying capabilities. Individual capabilities were illustrated by describing exemplary functions. Additionally, it was analyzed which capabilities represent the core of Data Catalog products and which capabilities are rather provided in the context of additional modules or by the integration of other enterprise IT-systems. Experience from practice projects was used to design an extended role model. The role model describes roles related to Data Catalogs. In particular their data-related problems in the enterprise are presented and solution promises of a Data Catalog implementation are shown. In addition, the integration model for Data Catalogs explains how the various types of Data Catalogs can be integrated into the enterprise IT-system landscape and which interfaces are required.

Based on the created capability map and functions that were deemed representative for each capability an assessment of 15 Data Catalog solutions was conducted. The assessed solutions include enterprise Data Catalogs (commercial and open-source), data platforms and one cloud Data Catalog solution. The analysis showed that none of the providers was able to provide satisfying coverage of all capabilities assessed. Accordingly, it is likely that compromises have to be made when selecting a solution provider and individual capabilities have to be weighed against each other. In particular, there are big differences between vendors in the areas of automation and AI, data governance, and data collaboration. It is therefore still possible to differentiate between data governance focused Data Catalogs and Data Catalogs with focus on analysts' productivity. A comparison between the different solution types shows that platform solutions perform better than stand-alone Data Catalogs while open-source solutions are currently only capable of providing basic capabilities. On the other hand, costs for platform solutions in general are higher than those of stand-alone solutions.

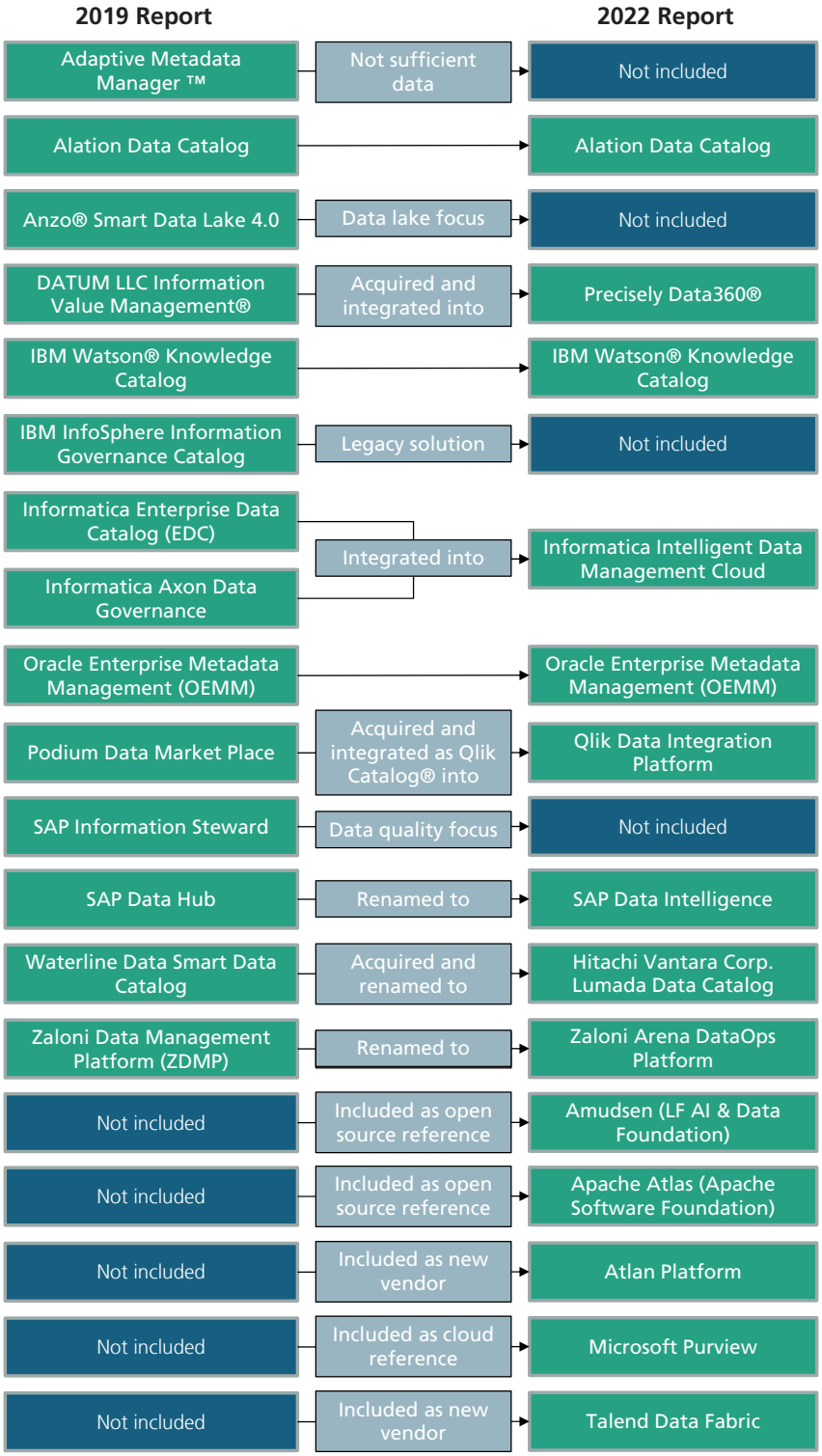
By the analysis of recent developments and provider roadmaps four Data Catalog trends could be identified. These include the increased adoption of AI-features, the support of data mesh and data fabric paradigms, the transformation towards active metadata management platforms, and the increasing amount of vendors providing their tool as cloud solution. In conclusion, the vendor analysis shows that most solutions provide functionality in every capability group. However, the extent of provided functionality differs highly from solution to solution. In general, all solutions show potential for improving the maturity of functions provided.

In the beginning of a provider selection process it is therefore important to carry out a precise requirements elicitation and prioritization. Additionally, prior to provider pitches and a detailed assessment, the market offering needs to be thoroughly analyzed to determine a set of potentially suitable providers that fulfill the requirements. Therefore, Data Catalog implementation initiatives should be aligned with other enterprise data initiatives, e.g. in the area of data governance. In doing so, necessary support of the organization and users can be fostered. This is essential to populate the Data Catalog with content and attract additional users by creating network effects.

While this report is covering many urgent topics for practitioners in Data Catalog implementation projects, further questions remain, e.g.: what are suitable approaches to implement Data Catalog solutions and what are critical success factors of Data Catalog implementations? Additional questions arise during other stages of the system lifecycle, e.g. regarding the progress of a Data Catalog project. Currently, no scientific model to measure the maturity of Data Catalog initiatives and implementations exist. It is planned to continuously extend this report with further insights from practice projects.

Appendix

Solution Inclusion



About Fraunhofer ISST

The Fraunhofer Institute for Software and Systems Engineering ISST identifies and realizes the strategic value of data in cooperation with companies – we offer complete system solutions for your company, from data preparation to the development of new business models.

As part of overall data governance or data strategy initiatives, or in standalone projects the Fraunhofer ISST supports organizations in implementing Data Catalog solutions as neutral intermediary. The Fraunhofer ISST offers services such as requirements analysis, pre-selection of potential providers, critical monitoring of the selection process, and alignment with other enterprise data initiatives.

References

Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. International Journal of Information Management, 49, 424–438. https://doi.org/10.1016/j.ijinfomgt.2019.07.008

Brust, A. (2021). Hitachi Vantara acquires data governance player Io-Tahoe. https://www.zdnet.com/article/hitachi-vantara-acquires-data-governance-player-io-tahoe/

Cao, L. (2019). Data Science: Profession and Education. IEEE Intelligent Systems, 34(5), 35–44. https://doi.org/10.1109/MIS.2019.2936705

CKAN. CKAN - Github Repository. https://github.com/ckan/ckan

DalleMule, L., & Davenport, T. H. (2017). What's Your Data Strategy? Harvard Business Review, 95(3), 112–121.

Deans, G. K., Kroeger, F., & Zeisel, S. (2002). The Consolidation Curve. Harvard Business Review.

Deghani, Z. (2019). How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh. https://martinfowler.com/articles/data-monolith-to-mesh.html

Diamantini, C., Lo Giuidce, P., Musarella, L., Potena, D., Storti, E., & Ursino, D. (2018). A New Metadata Model to Uniformly Handle Heterogeneous Data Lake Sources. In A. Benczúr, B. Thalheim, T. Horváth, S. Chiusano, T. Cerquitelli, C. Sidló, & P. Z. Revesz (Eds.), New Trends in Databases and Information Systems (pp. 165–177). Springer International Publishing.

Dinter, B., Gluchowski, P., & Schieder, C. (2015). A Stakeholder Lens on Metadata Management in Business Intelligence and Big Data - Results of an Empirical Investigation. Twenty-First Americas Conference on Information Systems.

Eichler, R., Giebler, C., Gröger, C., Hoos, E., Schwarz, H., & Mitschang, B. (2021). Enterprise-Wide Metadata Management. Business Information Systems, 269–279. https://doi.org/10.52825/bis.v1i.47

Fadler, M., & Legner, C. (2021). Toward big data and analytics governance: redefining structural governance mechanisms. In T. Bui (Ed.), Proceedings of the Annual Hawaii International Conference on System Sciences, Proceedings of the 54th Hawaii International Conference on System Sciences. Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2021.691

Gröger, C. (2018). Building an Industry 4.0 Analytics Platform. Datenbank-Spektrum, 18(1), 5–14. https://doi.org/10.1007/s13222-018-0273-1

Gröger, C. (2021). There Is No AI Without Data: Industry Experiences on the Data Challenges of AI and Call for a Data Ecosystem for Industrial Enterprises. Communications of the ACM.

ISO (2015). ISO/IEC 11179-1:2015(en): Information technology — Metadata registries (MDR) — Part 1: Framework. https://www.iso.org/obp/ui/#iso:std:iso-iec:11179:-1:ed-3:v1:en

Klímek, J., Škoda, P., & Nečaský, M. (2018). LinkedPipes DCAT-AP Viewer: A Native DCAT-AP Data Catalog. International Semantic Web Conference (P&D/Industry/BlueSky). Advance online publication. https://doi.org/10.1007/978-3-319-47602-5_20

Korte, T., Fadler, M., Spiekermann, M., Legner, C., & Otto, B. (2019). Data Catalogs - Integrated Platforms for Matching Data Supply and Demand: Reference Model and Market Analysis (Version 1.0). Fraunhofer Verlag.

Koutroumpis, P., Leiponen, A., & Thomas, L. D. W. (2020). Markets for data. Industrial and Corporate Change, 29(3), 645–660. https://doi.org/10.1093/icc/dtaa002

Labadie, C., Legner, C., Eurich, M., & Fadler, M. (2020). FAIR Enough? Enhancing the Usage of Enterprise Data with Data Catalogs. In S. Aier & W. Guedria (Eds.), 2020 IEEE 22nd Conference on Business Informatics: CBI 2020 : proceedings : Antwerp, Belgium, 22-24 June 2020 (pp. 201–210). IEEE Computer Society, Conference Publishing Services. https://doi.org/10.1109/CBI49978.2020.00029

Machado, I., Costa, C., & Santos, M. Y. (2021). Data-Driven Information Systems: The Data Mesh Paradigm Shift. 29TH INTERNATIONAL CONFERENCE on INFORMATION SYSTEMS DEVELOPMENT.

Máchová, R., & Lnenicka, M. (2017). Evaluating the Quality of Open Data Portals on the National Level. Journal of Theoretical and Applied Electronic Commerce Research, 12(1), 21–41. https://doi.org/10.4067/S0718-18762017000100003

Market Growth Reports. (2021). Global Data Catalog Market Size, Status and Forecast 2021-2027. https://www.marketgrowthreports.com/global-data-catalog-market-17767832

Mateeva, Z. (2019). Specific Nature of the New Profession "Data Protection Officer" in the Context of Digitalization.

Mordor Intelligence. (2021). Data Catalog Market - Growth, Trends, COVID-19 Impact, and Forecasts (2021 - 2026). https://www.mordorintelligence.com/industry-reports/data-catalog-market

Otto, B. (2011). A morphology of the organisation of data governance. ECIS 2011 Proceedings.

Plotkin, D. (2014). Data stewardship: An actionable guide to effective data management and data governance. Morgan Kaufmann. http://proquest.tech.safaribooksonline.de/9780124103894

Reinsel, D., Gantz, J., & Rydning, J. (2018). The Digitization of the World: From Edge to Core.

Rosian, M., Brinkhege, R., Gür, I., Schleimer, A. M., Scherenberg, F. von, & Spiekermann, M. (2021). CLOUD TRANSFORMATION: TRENDS & IMPLICATIONS. Fraunhofer Insitute for Software and Systems Engineering ISST.

Sawadogo, P., & Darmont, J. (2021). On data lake architectures and metadata management. Journal of Intelligent Information Systems, 56(1), 97–120. https://doi.org/10.1007/s10844-020-00608-7

Simoni, G. de, & Dayley, A. (2021). Market Guide for Active Metadata Management.

Tsormpatzoudi, P., Berendt, B., & Coudert, F. (2016). Privacy by Design: From Research and Policy to Practice – the Challenge of Multi-disciplinarity. In B. Berendt, T. Engel, D. Ikonomou, D. Le Métayer, & S. Schiffner (Eds.), Lecture Notes in Computer Science: Vol. 9484. Privacy Technologies and Policy: Third Annual Privacy Forum, APF 2015, Luxembourg, Luxembourg, October 7-8, 2015, Revised Selected Papers (1st ed., Vol. 9484, pp. 199–212). Springer International Publishing. https://doi.org/10.1007/978-3-319-31456-3_12

Utterback, J. M., & Abernathy, W. J. (1975). A dynamic model of process and product innovation. Omega, 639–656.

Wells, D. (2021). Data Architecture: Complex vs. Complicated. https://www.eckerson.com/articles/data-architecture-complex-vs-complicated

Woodie, A. (2021). Data Mesh Vs. Data Fabric: Understanding the Differences. https://www.datanami.com/2021/10/25/data-mesh-vs-data-fabric-understanding-the-differences/

The World Bank Group. (2021). Data Catalog. https://datacatalog.worldbank.org/

Zhang, D., Maslej, N., Brynjolfsson, E., Etchemendy, J., & Lyons, T. (2022). The AI Index 2022 Annual Report.

Figure Index

Imprint

Editor
Fraunhofer Institute for Software and Systems Engineering ISST
Emil-Figge-Str. 91
Germany - 44227 Dortmund

Authors
Nils Jahnke
Markus Spiekermann
Behnam Ramouzeh

Typesetting and layout
Peter Michatz