Collibra

Whitepaper

Enterprise data catalogs: seeing the bigger picture

Executive summary

Organizations that effectively utilize their data can gain a better understanding of their customers, improve products and services, and ultimately ensure operational efficiencies and reduce risk. According to the results of a recent global survey of 700 C-suite executives, a mere "25% of organizations are realizing the potential of their data and analytics projects today.[1]" Unfortunately, this revelation is not surprising, as only 29% of survey respondents reported having a high level of trust in their organization's data. Organizations need processes in place to assure data quality and protect data from misuse. They need a deeper understanding of how data flows through an organization. Data catalogs serve that purpose. They offer the means to manage metadata and curate the necessary information to make data assets easier to discover, manage and consume. In doing so, data catalogs have become an essential component of enterprise data architectures.

However, making data easy to discover and consume is just the first step. To get the most out of data, it needs to be trusted, and the right data needs to be shared with the right users, which requires proper governance to assure data quality and define usage policies that can prevent data from being misused or misappropriated.

Governance may seem like a chore. Some may view it as an impediment to an agile data-driven culture. But the opposite is true. Without a foundation of data governance, data consumers will lack the contextual insights necessary to fully understand data. They will run into issues of data quality and consistency, failing to make apples-to-apples comparisons and resulting in flawed analyses. They may also incur significant liabilities from improper use of data. In short, poorly governed data will hamper innovation and constrict business growth.

Implementing a data catalog with embedded data governance helps to drive agile data operations. It enables analytics and AI teams to be truly self-service, safe in the knowledge that analysts are accessing trusted data under permitted circumstances.

1 Source: Accenture, "A new dawn for dormant data"

This paper will explain:

- Different tiers of metadata (and the importance of building a foundation of trust)
- Differences between data catalog implementations
- Challenges in implementing data catalogs — specifically around fostering data quality, consistency and compliance
- How data governance addresses those challenges to ensure trusted data is readily available across the organization
- How users at all levels can quickly discover, understand and trust data to drive impactful business decisions

Building a foundation of trust

The metadata contained within a data catalog can describe technical aspects of data. What format is the data stored in? How large is the file? When was it created or last modified? These are all rudimentary questions that a data catalog can be used to answer. Beyond these basic characteristics, data catalogs need to offer more descriptive and business-oriented information to be of true value to data consumers. That means answering more detailed questions that help data consumers ascertain whether the data can be trusted and useful for their specific analysis. Which categories of data are included - for example, does the data set contain any personally identifiable information (PII), sensitive data or third-party information? What tables are contained within a schema, and what columns are within a table? Which source system (or systems) does the data originate from? How complete and accurate is the data? Has the data undergone any transformations?

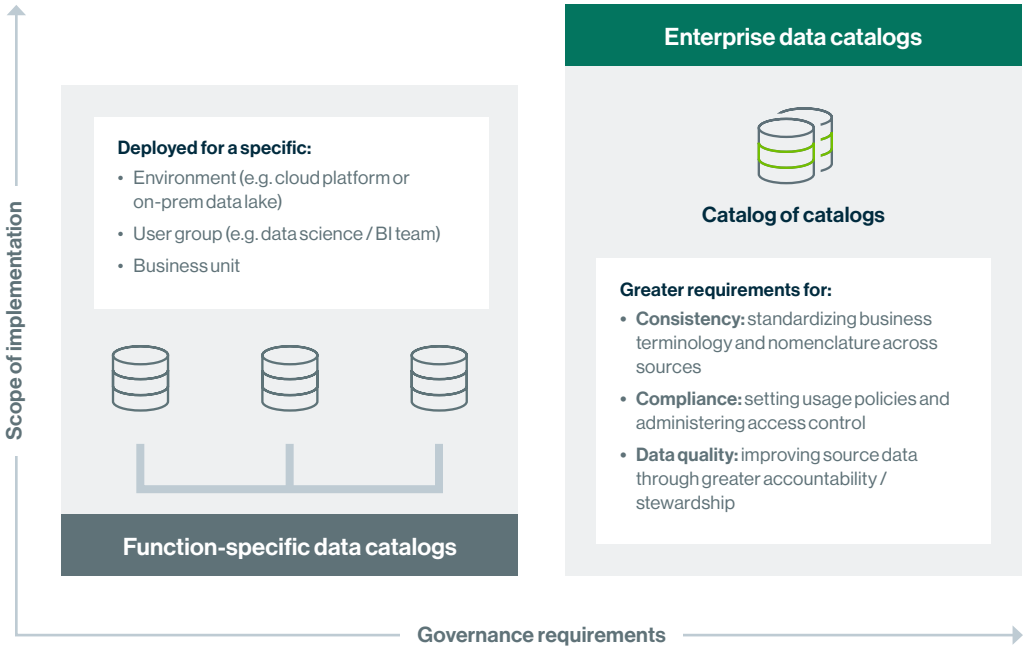
Answering questions of this type requires a combination of automation and human judgment. While automation can help capture basic technical metadata, the most important insights typically require input from experts who can provide the necessary business context to aid understanding. In fact, capturing this tribal knowledge can be an even more effective way to drive productivity than simply automating manual tasks.

By implementing data catalogs with embedded governance capabilities, organizations will truly be able to unlock the value of their data. Data owners and stewards will maintain accountability and ensure the right processes are in place to monitor, manage and promote data quality. Data users will gain a deeper understanding of data with access to richer contextual information provided by experts. They will also be able to trust in the data, knowing that it has been governed and certified. The tools used to derive insights from data (including analytics, models, worksheets, notebooks, dashboards and cubes) can also be certified, helping to foster trust not only in the underlying data, but also in analytical methods. In doing so, organizations will be able to build agile, data-driven operations that positively impact their business and be confident that decisions are backed by trusted data.

Differences in data catalog implementations: from tactical to strategic

Data catalogs can serve a variety of use cases. Some of those are relatively tactical, either helping to ease data discovery for a targeted set of users within a specific environment or for a distinct business unit. Alternatively, data catalogs can be rolled out on a more strategic basis across the enterprise, helping to curate data assets that cross business, regional and operating silos.

Data governance serves as an important foundation to all data catalogs. However, the broader the scope of a catalog implementation, the more onus is placed on governance requirements. Enterprise-wide data catalogs place greater focus on harmonizing data across different environments, for example, by standardizing terminology. Addressing a broader variety of users and business units also emphasizes compliance requirements, as the need to set and administer policies over data usage is magnified.



Tactical, vendor-specific or function-specific catalogs

Some data catalog implementations are relatively narrow in scope. For example, many of the leading cloud service providers offer their own data catalogs, which are well-suited to aiding discovery of data on those platforms, but lack capabilities to support enterprise implementations (particularly for hybrid, multi-cloud architectures). Equally, some specialist data catalog vendors cater squarely to the need of data scientists, helping them to curate technical information on a variety of data assets, but with less consideration to issues such as business context, data quality, consistency and compliance.

Strategic, enterprise data catalogs

Enterprise data catalog implementations serve many of the same purposes as those with narrower remits — helping to make data more easily discoverable, automatically harvesting and curating metadata, and aiding collaboration across consumers. However, enterprise implementations will typically see much more pronounced requirements for data governance. Cataloging a greater variety of data sets and catering to different types of users makes it harder to set and administer usage policies. The need to nurture data consistency is also amplified when there are a wider variety of sources to manage. Finally, assuring data quality and providing contextual insights become more complex within the context of a larger organization.

Given these added complexities, it is important that enterprise data catalogs can offer built-in data governance capabilities, helping to bring in expertise from a range of different roles to drive more detailed understanding, promote data quality and consistency, and ensure compliance.

Key takeaway

While some organizations may begin implementing a data catalog with a narrow remit, most will ultimately discover requirements to expand that approach. It is important to note that enterprise capabilities are not only needed by large corporations. Organizations of all sizes can benefit from proper data governance. Managing the business imperative to extract more value from data in tandem with evolving regulatory responsibilities is a challenge faced by all kinds of companies.

Challenges with data catalog implementations that lack governance capabilities

An increasing amount of descriptive and technical information relating to data sets can be harvested automatically. There are many data catalogs — particularly those deployed to serve a relatively narrow function — that collate metadata in such a way, primarily to aid data discovery. Yet looking to manage data as a strategic asset requires expert insights and human judgments. Data catalogs without built-in, proper governance will invariably run into challenges relating to data quality and consistency. Perhaps even more significantly, they can leave organizations exposed to significant liabilities by failing to comply with relevant rules and regulations.

Data quality

While any business decision can be data driven, the right decisions will be based on the right data. That is why data quality remains such a pressing priority and an integral component of digital transformation strategies. Poor data quality results in a lack of trust, not only in the underlying data, but also in the conclusions garnered from analyzing that data. And a lack of trust will blur an organization's decision making processes.

There are several ways that an enterprise data catalog can promote transparency into data quality metrics, including data profiling, detailed technical lineage, collating user feedback and certification.

However, to truly improve the quality of source data requires expert knowledge and clear accountability. This is what governance brings to the table. By assigning owners and stewards, organizations can begin putting in place the right processes to identify, track and remediate issues relating to data quality. Ultimately, governance ensures there is a foundation of trust, which is what organizations need to drive agile data-driven operations.

> **Every year, poor data quality costs organizations an average $12.9 million. Apart from the immediate impact on revenue, over the long term, poor quality data increases the complexity of data ecosystems and leads to poor decision making.**
>
> Gartner, How to Improve Your Data Quality, July 2021

Consistency

Even small and medium sized organizations run into issues relating to data consistency. It is all too easy for data to proliferate without upfront consideration to standards. When new systems are developed or procured there is often little thought taken into standardizing nomenclature or ensuring KPIs and other analytics are calculated consistently. Achieving data consistency, therefore, needs to be achieved once the horse has bolted, which requires significant amounts of effort and coordination. Given the amount of technical debt faced by most organizations, the task of achieving data consistency can be even more daunting for large enterprises. Without a common language and standardized definitions, it is impossible to ensure coherent analysis of data gathered from multiple source systems or business units. That means organizations risk jeopardizing their analyses by failing to make apples-to-apples comparisons and ultimately yielding invalid conclusions.

Compliance

Compliance with rules and regulations that govern the use of data is a challenge that is not only growing in importance, but also complexity. Privacy and emerging AI regulations around the globe continue to evolve significantly. To further complicate matters, these regional, national and supranational rules (such as GDPR, CCPA and the EU AI Act), have to be managed in conjunction with industry-specific regulations (such as HIPAA and BCBS 239), along with rules from tax authorities that require data retention for audit purposes and a variety of internal policies (for example, prohibiting salary information being shared outside of HR).

Managing the implications of such a complex set of multi-jurisdictional rules requires a granular, data-centric understanding of compliance. Organizations need to know exactly which policies apply to which data sets, so that they can administer access controls and only allow permitted use cases.

Key takeaway

The combination of these three challenges — data quality, consistency and compliance — is what ultimately prevents an organization from being agile and data driven in their decision making. Poor data quality and inconsistency in data definitions leads to a lack of trust in underlying data and analysis. The complexity of compliance requirements can also act as an impediment to agile operations. The only way to enable true self-service business intelligence is for organizations to codify their data usage policies into granular controls that allow access to data only under permitted circumstances.

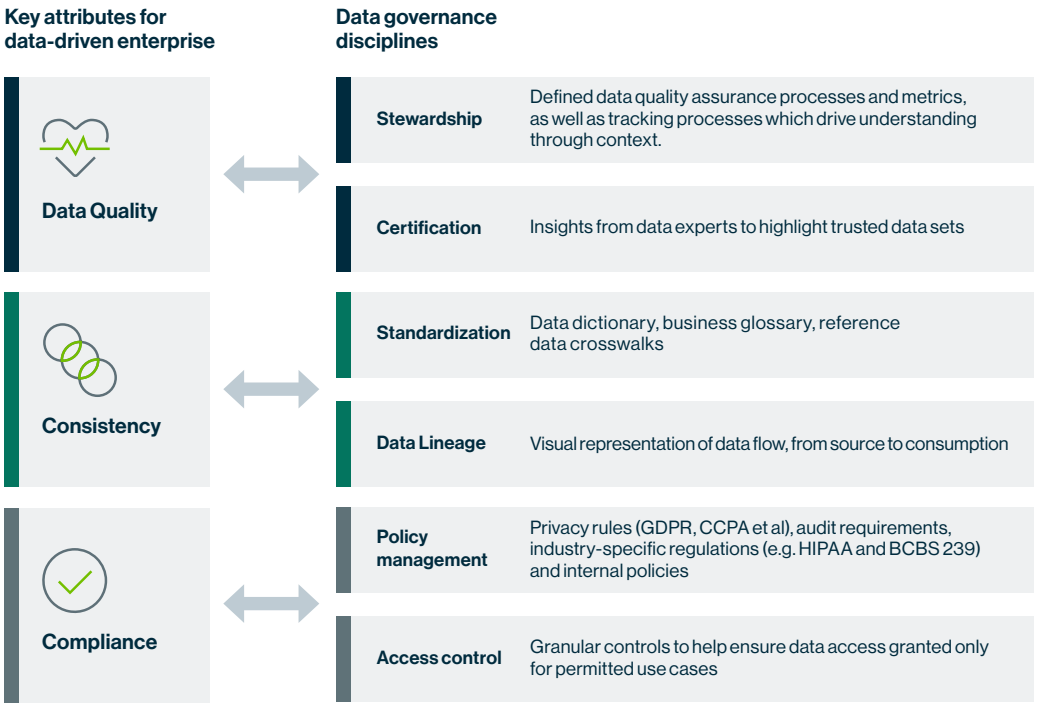
How governance addresses these challenges

44% of CDOs are focused on "establishing clear data governance responsibilities" across the enterprise to bring more value to their organizations.

AWS, Prioritizing business value creation, November 2023

Data governance spans a broad set of disciplines and offers a wide range of benefits. As companies look to take a more data-driven approach in their customer interactions, product development, risk management and operational processes, they need to ensure the data underlying their decisions is well-governed. That means clearing up ambiguities and inconsistencies in definitions, nurturing data quality, and ensuring compliance with evolving regulatory obligations and internal policies.

This section details aspects of data governance that are crucial to successful data catalog implementations of all kinds, but particularly for strategic, enterprise initiatives.



Data governance disciplines

Stewardship and accountability

Data stewardship can be key to driving trust and aiding understanding of underlying source data. Successful data catalog implementations help data stewards share insights across the organization, providing data users with the necessary context to aid understanding. Moreover, assigning individuals responsibility for underlying source data can help improve data quality by helping to ensure appropriate policies, classifications, business rules and quality control processes are in place.

Certification

Certifying data assets including data, KPIs and reports is a crucial process in assuring trust. Certification helps capture the intelligence of data experts and makes it much easier for data consumers to locate trusted sources. Users should not only be able to see if a data set has been certified, but also who certified it, whether they have provided any descriptive notes, which data elements are contained, where they are sourced from and relevant data quality metrics.

Standardization of definitions and metrics

Organizations invariably find that different systems will name the same concept in different ways. Business terminology can be made even more confusing in an environment of constant change. Any ambiguity in definitions causes confusion in analytical processes and can lead to inaccurate conclusions. Tools such as a business glossary, data dictionary and reference data crosswalks help clear up those discrepancies.

Data lineage

Tracking the way data flows through an organization is an increasingly challenging task, not only because of ever-increasing data volumes, but also the complexity of analytical processes and system architectures. Data lineage makes data meaningful and helps to assure trust, mitigate risks and demonstrate compliance. In fact, end-to-end lineage is a necessary and crucial foundation for all data-driven initiatives. Accurate data lineage helps data consumers ensure reports are pointed at the correct sources, and easily trace any potential errors identified through their analysis. From the perspective of compliance, data lineage can also be a useful tool to track data processing activities and monitor where sensitive information resides and how it is processed.

Policy management and access control

Modern organizations are subject to evermore stringent regulations regarding their use of data. The task of devising compliant usage policies is made even more complex given that some rules may conflict with others. For example, privacy regulations that afford individuals the right to have their information deleted can conflict with an organization's duty to keep records for audit purposes. Balancing these requirements requires expert insights, not only understanding which policies apply to which data sets, but also which takes precedence. In turn, ensuring those policies are adhered to requires granular access controls. By adopting a data-centric approach to compliance, organizations will be able to automatically permission access based on all relevant criteria, including the role of the user, their location, as well as their intended use for the data.

Collaborative workflows

Many of the governance disciplines that are mentioned above require collaboration and coordination between a variety of stakeholders, including data owners, stewards and data consumers, to senior management and compliance experts. Enterprises need to automate collaboration across governance and cross-functional teams to establish a common understanding around data. This enables access and visibility into consistent, precise and trusted definitions across various teams in one place.

Key takeaway

Most companies have to deal with technical debt — including legacy information systems and siloed data architectures — that makes the challenge of maintaining data quality, consistency and compliance all the more complex. Tackling these challenges on a piecemeal basis is simply not feasible. It means organizations need to create an abstraction layer that can harmonize their disparate physical data infrastructure. Data catalogs play a key role in that abstraction, particularly if they offer embedded governance capabilities. They enable organizations to establish centralized definitions, mappings, policies and controls that can be applied consistently across disparate physical data stores. They also help to clear up confusion arising from complex data infrastructures by promoting certification and enabling effective and compliant information sharing.

How Collibra can help

Collibra Data Catalog with embedded data governance and privacy capabilities, helps enterprises to unlock the value of data and scale data initiatives across the enterprise. With Collibra Data Catalog, organizations break down legacy data silos to surface relevant data and enable data citizens to understand and trust the context and relevance of the data, supporting the drive toward self-service analytics. It helps organizations:

Build a foundation of trust

Embedded governance capabilities help ensure data consumers have access to reliable, certified data. Increase productivity and make data more accessible by enabling users to find, understand and access trusted data easily, while maintaining compliance with policies.

Foster agile data operations

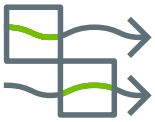
Business users gain rapid access to trusted data for meaningful analysis, while the enterprise can also overlay appropriate policies to mitigate potential liabilities from data misuse.

Drive analytical consistency

Standardized business terminology and data definitions help ensure decisions are backed by consistent data that has been interpreted in the right context.

Adopt a strategic approach

An open architecture enables Collibra Data Catalog to sit on top of tactical solutions designed for narrow use cases, allowing firms to manage data across disparate enterprise architectures and business silos.



Track data as it flows through the enterprise

Insights into the way data flows from source to consumer, including views of both technical and business lineage, help drive data quality, mitigate risk and demonstrate compliance.

Enable enterprise-grade security

Embedded governance within a data catalog allows organizations to define usage policies. This capability empowers organizations to grant appropriate access to data across the enterprise through access controls.

Collaborate across your business

Facilitate a collaborative approach to data governance and self-service analytics with automated workflows. Collaborate to define data, build business context and maintain information as your business changes.

Enable self-service access to data

Share data and enable data consumers to shop for and collaborate on data, reports, models and other assets easily across the organization. Through a business-friendly, self-service data marketplace, allow users to search for, locate and request access to curated, relevant data.

Conclusion

At Collibra, we see data catalogs as a crucial part of an organization's journey to achieving data intelligence and an important factor in driving revenue, improving operational efficiency, and generating innovation and growth. Collibra resolves the challenges that cause decision makers to not trust data and analytics by ensuring the common understanding of terms and metrics, documenting data owners and stewards, establishing data lineage, and enabling the certification of key reports and KPIs/metrics. With trusted data in one place and a common understanding around data, Collibra empowers data citizens to deliver outcomes that drive quality business results.

