



data.world

Choosing the Right Data Catalog for your Organization

00. Table of contents

For organizations looking to harness the full potential of their data, a data catalog could be the puzzle piece you've been missing.

An enterprise data catalog is a jumping-off point for strong data management and enablement. It creates the map of your data landscape — the mythical “single source of truth” that advanced teams and enterprises seek.

The visibility, should you achieve it, delivers many cross-functional benefits:

1. Enhancing data discovery and accessibility for both business and technical users
2. Ensuring data governance and compliance in increasingly stringent regulatory environments
3. Improving data quality and trustworthiness around analytics and decision-making
4. Facilitating collaboration among data stakeholders
5. Supporting a healthy, long-term data culture
6. Driving AI and machine learning initiatives by providing well-organized, high-quality training data
7. Optimizing data operations and reducing redundancy
8. The enterprise data catalog you choose can significantly impact your organization's ability to derive value from data. Plus, you want to choose the one that people will actually use. That means implementation can't be overly complicated, and the ROI should be seen and felt as quickly as possible.

We wrote this paper for IT leaders, data professionals, and decision-makers tasked with choosing an enterprise data catalog solution. Our objective is to empower you with actionable insights into the key factors that should influence your decision, the essential features to prioritize, and a comparative overview of top vendors in the market.

02. Key factors to consider when choosing a data catalog

The following key factors should be at the forefront of your decision-making process around data catalog platforms.

1. Volume of metadata

Today's organizations are incredibly data-rich. For instance, an organization might have half a million data tables or more spread across various systems. When choosing a data catalog, it's crucial to ensure that the infrastructure can handle this scale efficiently.

Look for catalogs built on robust, scalable architectures that can manage large volumes of metadata without performance degradation. The catalog should offer features like distributed processing, efficient indexing, and optimized search capabilities to maintain responsiveness even as your metadata and data assets grow. Additionally, consider solutions that provide flexible deployment options, such as cloud-native architectures or hybrid models (more on that later), to accommodate your specific infrastructure requirements and future scaling needs.

2. Diversity of data

Modern enterprises deal with an incredibly diverse range of data types and sources. A truly effective data catalog should be able to model and catalog virtually any type of data asset, not just traditional databases, dashboards, etc.

For example, a researcher in molecular biology might need to catalog scientific publications, while a marketing team might need to integrate a dynamic JSON feed. Your chosen catalog should be flexible enough to accommodate all of that data without requiring extensive customization.

Look for solutions that offer:

- The ability to model and catalog structured, semi-structured, and unstructured data
- Support for both static and dynamic data sources
- Capabilities to catalog both metadata and actual data using the same underlying technology
- Flexibility to cater to different user needs, from those who want a fully managed catalog to those who prefer self-service options

The ideal catalog should allow you to self-service at any level you want, providing the knowledge and tools necessary to adapt the catalog to your specific use cases.

3. Integration capabilities

Integration is a critical factor in ensuring your data catalog becomes a central, valuable part of your data ecosystem rather than just another siloed tool.

Consider a scenario where you need to catalog a Snowflake data warehouse. Your catalog should be able to ingest metadata about tables, views, columns, and keys, providing you with a comprehensive understanding of your data landscape. While data catalogs typically don't query the actual data (that's not their primary function), they should offer robust capabilities for metadata integration.

Look for catalogs that provide:

- Wide-ranging native connectors to common data sources and tools
- APIs and extensibility options for custom integrations in an open standards model in order to avoid vendor lock-in
- The ability to capture and represent relationships between data assets across different systems
- Support for real-time or near-real-time metadata synchronization

Remember, the goal is to have a catalog that can provide a holistic view of your data assets, regardless of where they reside.

4. AI and automation capabilities

When evaluating solutions, prioritize those that embed AI, ML, and automation within their core cataloging features. This approach not only manages the growing complexity of data ecosystems but also significantly augments traditional cataloging processes.

Key AI and automation considerations include:

- AI-assisted data discovery
- Generative AI augmentation for tasks like metadata creation and enrichment
- ML-driven data profiling
- Automated anomaly detection
- Predictive tagging
- Proactive compliance reporting

5. Metadata management features

Effective metadata management is at the core of any good data catalog. Your chosen solution should offer comprehensive features for capturing, storing, and managing metadata from various sources.

- Consider whether you might need:
- Automated metadata discovery and ingestion
- Support for both technical and business metadata
- The ability to enrich metadata with custom attributes and tags
- Version control for metadata to track changes over time
- Robust search and discovery features to make metadata easily accessible

6. Data governance and security

In an era of increasing data regulations and privacy concerns, your data catalog must support robust governance and security measures.

- Governance considerations include:
- Fine-grained access controls to ensure users only see data they're authorized to access
- Integration with enterprise authentication systems
- Audit logging to track catalog usage and changes
- Support for data classification and sensitive data identification
- Tools for implementing and monitoring data policies

Your catalog should help you enforce governance policies and maintain compliance with relevant regulations, all while making appropriate data discoverable and accessible to authorized users.

7. Cloud-native and On-prem data

Truly cloud-native SaaS can make a new version every single day, because releases are automatically pushed into the cloud. Modern catalogs often focus on the cloud. At the same time, a data catalog focused only on data from the cloud will miss on-prem data. Ideally, metadata can be collected from both the cloud and on-prem sources.

8. Personas

Are you more focused on business personas? Or data analysts and SQL users? Or technical/engineering resources? Some data catalogs are better for technical users, because they focus on things like metadata management and data lineage. Business personas might be more interested in finding data in the marketplace experience.

Your data catalog should help you today, but it should also ideally position you for future data management challenges and opportunities. The bottom line: the right data catalog will empower your teams to trust and use all of your data assets.

03.

Essential features of a modern data catalog

When evaluating enterprise data catalogs, look for features that work in concert to provide a comprehensive, user-friendly, and intelligent solution. Here are some of the important ones:

Data lineage

Understanding the origin, movement, and transformation of data across your organization is crucial for both governance and analytics purposes. A comprehensive data lineage feature in your catalog can provide this visibility.

Look for lineage capabilities that:

- Automatically capture data movement and transformations
- Provide both technical and business-friendly lineage views
- Allow for manual augmentation of lineage information where needed
- Integrate with ETL tools and data pipelines to capture end-to-end lineage

Automated alerts

Modern data catalogs have the ability to meet users where they are, via alerts. By automatically monitoring data quality issues like missing values or schema changes, alerts help teams address problems before they cascade into downstream processes. They play a crucial role in compliance by flagging when sensitive data is modified or when governance policies might be violated. From a resource management perspective, alerts also warn teams about storage constraints and highlight inefficient usage patterns that could impact performance.

Automated metadata harvesting and ingestion

A modern data catalog should have the ability to automatically discover and ingest metadata from various sources. This includes support for both technical and business metadata, with capabilities for real-time or near real-time metadata synchronization. The system should be able to handle structured, semi-structured, and unstructured data, ensuring a comprehensive view of your data landscape.

Comprehensive inventory

A catalog should provide a complete inventory of diverse data assets including databases, data lakes, cloud storage, SaaS platforms, and AI/ML models. For each asset, it should offer detailed information such as schema, data types, and relationships. The ability to support custom metadata attributes is crucial for capturing organization-specific information and context.

Intelligent search and discovery

Modern data catalogs are usually equipped with AI-powered search capabilities. The system should offer faceted search options for refined results and personalized search results based on user roles and preferences. A recommendation engine for relevant data assets can significantly enhance UX and data discovery.

Data quality and profiling

Automated data profiling helps users understand data characteristics at a glance. The catalog should offer data quality scoring and monitoring, along with anomaly detection and alerting. Integration with data quality tools for detailed assessments can provide deeper insights into data reliability and usability.

Collaborative features

To create a data-driven culture, the catalog should include features that allow users to add comments, ratings, and reviews to data assets. Discussion forums or Q&A sections for each data asset can facilitate knowledge sharing. Integration with communication tools like Slack or Microsoft Teams can further enhance collaboration around data assets.

○ Data governance and compliance

Role-based access controls are essential for maintaining data security. The catalog should support data classification and tagging for sensitive information, along with policy management and enforcement. Audit logging and reporting features are crucial for compliance purposes, as are specific features to support privacy regulations like GDPR and CCPA. Some catalogs are hard core about governance. Some catalogs don't focus on governance as a priority. A middle of the road option might enable many teams to participate in data governance and make suggestions, without being able to enact major change.

○ Business glossary and semantic layer

A centralized repository for business terms and definitions helps ensure consistent understanding across the organization. The ability to link business terms to technical metadata bridges the gap between business and technical users. Support for multiple glossaries or domains, along with versioning and change management for glossary terms, helps manage complex business vocabularies. Ability to turn business glossary into rich semantic layers.

○ AI and ML modeling

As AI and ML become more prevalent, catalogs should include features to inventory these models. This includes capturing model metadata such as algorithms used, training data, and performance metrics. Model lineage and version control are crucial for governance, as is integration with ML platforms and notebooks.

○ Self-service data preparation and analytics

Basic data preparation capabilities such as sampling and formatting can empower users to work with data directly. Integration with popular BI and analytics tools extends the catalog's utility. Features that allow users to create and share datasets, along with tools to promote data literacy, support the goal of data democratization.

API and extensibility

Well-documented APIs are essential for integration with other tools and platforms. Support for custom plugins or extensions allows organizations to tailor the catalog to their specific needs. Webhook support enables event-driven integrations, while SDKs for popular programming languages facilitate custom development. Ability to query the metadata.

Intelligent automation and AI integration

AI-assisted metadata enrichment and tagging can significantly reduce manual effort. Automated data classification and categorization improve organization and searchability. Machine learning can be employed for continuous improvement of search and recommendations, while natural language query capabilities make the catalog more accessible to non-technical users.

Robust visualization and reporting

Dashboards for data asset usage and popularity provide valuable insights into data utilization. Visual representations of data relationships and lineage enhance understanding of the data ecosystem. Customizable reports for various stakeholders ensure that different teams get the information they need, while data asset health and quality visualizations provide at-a-glance understanding of data reliability.

Scalability and performance

The ability to handle large volumes of metadata and data assets is crucial as data ecosystems grow. Support for distributed architectures ensures the catalog can scale with your organization. Efficient indexing and query optimization, along with caching mechanisms, are essential for maintaining performance as the catalog grows.

04. Comparison of top data catalog vendors

Based on research¹ across several platforms, what follows is a comparison of some of the leading data catalog vendors.

data.world

Data.world distinguishes itself with strong semantic capabilities built on knowledge graph architecture. The platform excels in collaborative data exploration and benefits from a large, engaged open data community. Its **AI Context Engine** performs structured data analysis through AI/ML models, which make it a superior tool for analysis and data accuracy. The lack of native data observability may be a drawback for some organizations. Data.world is best suited for companies that value collaborative data exploration and semantic data relationships.

Informatica

Informatica distinguishes itself with its all-encompassing AI-driven data management solution, the Intelligent Data Management Cloud (IDMC). This platform offers robust democratization and self-service features, making it a compelling option for large corporations. Through strategic acquisitions, Informatica continuously enhances its capabilities. The platform stands out for its superior lineage tracking and AI-powered data quality assurance. However, Informatica's strengths come with trade-offs – both in terms of cost and complexity. The steep pricing may impede full product adoption for some organizations, and the implementation process can be intricate, particularly for smaller entities. Informatica is most suitable for large enterprises with intricate data ecosystems and substantial financial resources.

¹ Sources cited: [Collibra](#), [Atlan](#), [G2](#), [TrustRadius](#), [Atlan Comparison](#), [Forrester](#), [data.world](#), [BigID](#)

atlan

Atlan is a dynamic player in the data catalog market, offering a personalized, AI-driven catalog that rapidly brings value to organizations. Its emphasis on data democratization and self-service functionalities distinguishes it from competitors, making it particularly effective for companies striving to foster a data-centric culture. Atlan's automated AI/ML metadata management and AI-assisted discovery features further boost its attractiveness. However, the absence of built-in data quality and observability tools may require integration with supplementary solutions for comprehensive data oversight. Atlan is ideally suited for businesses seeking quick implementation of an AI-enhanced, user-friendly data catalog.

Collibra

Collibra has established itself as a leader with its exceptional data marketplace features and robust capabilities in data governance and AI use case support. The platform excels in offering comprehensive regulatory compliance tools, making it a preferred choice for businesses in heavily regulated sectors. Collibra's user-friendly detailed lineage and versatile marketplace encourage data sharing and accessibility, nurturing a data-centric environment. However, the platform is not without its challenges. Some users encounter integration difficulties with its intelligence platform, and certain organizations have reported user adoption hurdles. There's also potential for enhancement in AI-driven explainability and curation. Collibra is particularly well-suited for enterprises prioritizing data governance, regulatory adherence, and those seeking to implement a sophisticated data product sharing ecosystem.

The Alation logo consists of a stylized orange and yellow 'A' icon followed by the word 'Alation' in a bold, white, sans-serif font.

Alation is a strong data catalog platform focused on collaboration. Alation excels in data stewardship and maintains an extensive catalog of data products, making it ideal for companies aiming to productize their data assets. Its advanced monitoring and dashboard features offer users deep visibility into their data ecosystem. However, Alation has some limitations. The absence of built-in data quality tools might necessitate additional software for complete data oversight. Some users have encountered issues with automatic metadata collection. Furthermore, Alation could enhance its offering with more advanced AI integration and intelligent AI-driven suggestions. The platform is particularly well-suited for enterprises that emphasize teamwork, data product development, and rigorous data governance in their data management approach

The BigID logo features a stylized fingerprint icon on the left, composed of white lines forming a circular pattern, followed by the text 'BigID' in a bold, white, sans-serif font.

BigID distinguishes itself with robust privacy and security capabilities, appealing to organizations with strict data protection needs. Its strengths lie in automated profiling and classification, offering comprehensive categorization features and a wide array of built-in connectors. These attributes make BigID particularly effective for managing sensitive information in intricate environments. However, the platform has some drawbacks. It lacks certain lineage features, which may disappoint organizations seeking end-to-end data traceability. Moreover, BigID's complexity could hinder adoption for some users. The platform also falls short in data productization and marketplace functionality compared to some rivals. BigID is ideally suited for businesses with rigorous data privacy and security demands, particularly those in heavily regulated sectors.

05. Summary table

The following table highlights the diverse strengths and focus areas of the top data catalog vendors. When selecting a vendor, carefully consider specific needs, existing infrastructure, and long-term data management strategy to find the best fit.

Vendor	Key Strengths	Key Weaknesses	Best Suited For
data.world	Semantic capabilities, collaboration, AI-driven capabilities for catalog and governance, governance automations, flexible workflows, a focus on adoption	Lacks some data observability features	Collaborative exploration, semantic focus, agile data governance, large enterprises, complex ecosystems that require flexibility when building AI Apps
Atlan	Dynamic access control, self-service, automation	Lacks native quality features, lacks collaboration features, light on governance features	Rapid deployment, AI-enhanced self-service, data engineers and technical personas
Informatica	Comprehensive platform, strong ecosystem, multi-capability	High cost, complex implementation	Large enterprises, complex ecosystems
Collibra	Partner integrations, governance including AI use cases, compliance	Integration challenges, adoption issues, Legacy technology, light on genAI-assisted productivity and explainability capabilities	Governance-focused, regulatory compliance, Data stewards and technical people building workflows
Alation	Collaboration, search capabilities, integrations	Lacks quality features, metadata curation issues, Confusion around data mapping, light on genAI capabilities	Collaboration, data product creation, SQL analysts
BigID	Privacy and security, Metadata collection	Complex offerings, no end-to-end lineage	Stringent privacy and security needs

06. In summary

Choosing an enterprise data catalog is akin to selecting a skilled guide to an ever-expanding, multidimensional labyrinth of information.

The right guide doesn't just prevent you from getting lost; they unlock hidden chambers of insight.

An ideal data catalog is not a one-size-fits-all solution. It's a chameleon-like tool that adapts to the unique contours of your data. Whether you're dealing with a volume of data that rivals the Library of Congress or a diversity of information types, the best data catalog is the one that becomes an indispensable part of your organization's data diet. Beyond features and functionality, it should nourish a culture of data literacy.

Will it flex and grow as your data ecosystem evolves? Can it keep pace with the relentless march of technological progress? Will it empower your team to ask better questions, not just find faster answers?

In the end, the right data catalog isn't just a tool. It becomes the virtual heartbeat of your organization.



**Ready to start your data
catalog journey?**



data.world