DataHub

7 Reasons to Rethink your Data Catalog

Modern Metadata Platforms Have Answers

A GUIDE TO MODERNISING YOUR METADATA PLATFORM FOR AI-READY DATA MANAGEMENT.

DataHub

Introduction

Data catalogs have emerged as a critical component of enterprise data management over the past decade. Initially designed to address relatively straightforward needs around data compliance and SQL querying, traditional data catalogs served as static inventories of data assets, providing basic metadata management capabilities. However, as organizations embrace digital transformation, adopt artificial intelligence, and face increasingly complex data governance requirements, the limitations of traditional data catalogs have become apparent.

Today's data landscape is characterized by unprecedented volume, velocity, and variety. Data architectures have evolved from centralized repositories to decentralized ecosystems, creating new challenges around data fragmentation, governance, and accessibility. Modern use cases demand metadata platforms that function as integral components of the production fabric, enabling systems and AI models to orchestrate, access, and record their use of data assets in real-time.

This paper examines seven compelling reasons why organizations should reconsider their current data catalog implementations and adopt modern metadata platforms. We explore how these next generation solutions address critical limitations in scalability, extensibility, unification, accessibility, AI-readiness, security, and future-proofing, enabling enterprises to unlock the full potential of their data assets in an increasingly AI-driven world.

DataHub

The Evolution of Data Catalogs

Before diving into the reasons for reconsideration, it's important to understand how data catalogs have evolved and why traditional approaches are struggling to meet contemporary demands.

Traditional Data Catalogs: Origins and Limitations

Traditional data catalogs emerged to solve specific problems in the early days of big data adoption. They were primarily designed to:

1. Create an inventory of data assets across various storage systems
2. Enable data discovery through basic search and browsing capabilities
3. Provide context through technical metadata (schema, format, location)
4. Support compliance and audit requirements
5. Facilitate SQL query development through data exploration

These catalogs operated largely as passive, offline systems that users would consult periodically during specific tasks rather than as active components integrated into daily data workflows. Their architecture reflected this occasional usage pattern, with designs optimized for:

- Low metadata volume and velocity
- Batch processing of metadata updates
- Simple relationship modeling
- Manual documentation and curation
- Limited integration with operational systems

As data architectures have become increasingly complex and decentralized, and as organizations push to extract more value from their data assets through advanced analytics and AI, these architectural limitations have become significant barriers to effective data management.

The Shift Toward Modern Metadata Platforms

Modern metadata platforms represent an evolution beyond traditional data catalogs, addressing fundamental limitations through advanced capabilities that align with contemporary data management requirements. These platforms are characterized by:

- Real-time metadata ingestion and processing
- Comprehensive relationship modeling and lineage tracking
- Deep integration with production data pipelines and AI systems
- Support for diverse metadata types beyond technical descriptions
- Automation of metadata capture, enrichment, and governance
- Active participation in data workflows rather than passive documentation

With this evolutionary context in mind, let's explore the seven key reasons organizations should reconsider their data catalog strategy and embrace modern metadata platforms.

DataHub

Reason One

Scalability

Meeting the Demands of Modern Data Ecosystems

Every organization's data ecosystem is growing larger, more diverse, and faster than ever before. Traditional catalogs were built for simpler environments, while modern platforms must handle scale across many sources, massive volumes, and real-time updates.

Comprehensive Source Coverage

Modern enterprises manage data across an increasingly diverse landscape of systems, from traditional databases and data warehouses to cloud- native storage, streaming platforms, and specialized AI repositories. A modern metadata platform must be able to connect to and extract metadata from all these sources without significant custom development or maintenance overhead.

Traditional data catalogs were designed for environments with relatively stable data sources, predictable growth, and limited velocity. Today's data ecosystems present radically different challenges that demand new approaches to scalability.

THE SCOPE OF REQUIRED CONNECTIONS INCLUDES:

- Traditional relational databases (Oracle, SQL
- Server, MySQL, PostgreSQL)
- Legacy data warehouses (Teradata, IBM DB2)
- Modern cloud data warehouses (Snowflake,
- Amazon Redshift, Google BigQuery)
- Data lakes and object storage (AWS S3, Azure
- Data Lake, Google Cloud Storage)
- Big data platforms (Hadoop ecosystems, Spark)
- Streaming platforms (Kafka, Kinesis, Pulsar)
- BI and analytics tools (Tableau, Power BI,
- Looker)
- ETL/ELT and data integration tools (Informatica,
- Talend, dbt, Fivetran)
- AI/ML platforms and feature stores (Databricks,
- Sagemaker, Tecton)
- SaaS applications with valuable data assets
- (Salesforce, Workday, ServiceNow)

For each of these sources, the platform must capture not just basic metadata but deep technical details, usage patterns, data quality metrics, and governance information.

DataHub

Extensible Connector Framework

No pre-built connector ecosystem can anticipate every data source an organization might use, especially as technology continues to evolve rapidly. Modern metadata platforms must provide robust frameworks for creating custom connectors that:

- Follow standardized patterns for metadata extraction and normalization
- Support authentication and security requirements
- Support both pull and push-based metadata capture for efficiency
- Include monitoring and error handling capabilities
- Can be maintained and upgraded alongside platform evolution

This extensibility ensures organizations can incorporate metadata from proprietary systems, legacy applications, and emerging technologies without waiting for vendor support.

High-Volume Metadata Ingestion

As data assets grow into the petabyte range and beyond, the associated metadata becomes substantial. Modern platforms must handle metadata at scale, including:

- Billions of data objects and their associated technical metadata
- Complex relationships between data assets, processes, and users
- Historical versions of metadata for temporal analysis
- Enrichment data from automated classification and tagging
- Usage metrics and access patterns
- Quality measurements and validation results

This requires efficient storage mechanisms, optimized indexing, and careful performance engineering to maintain responsive user experiences even as metadata volumes grow exponentially.

Low-Latency Metadata Updates

Perhaps most critically, modern use cases require near real-time metadata updates to support operational decisions and AI workflows. Traditional catalogs with nightly or weekly refresh cycles cannot support scenarios such as:

- Data pipelines checking data quality and lineage before processing
- AI systems validating training data compliance before model building
- Automated governance tools enforcing policies on new data assets
- Active data monitoring for anomaly detection
- Just-in-time access control based on current data classification

Meeting these requirements demands architecture designed for continuous metadata ingestion, efficient processing, and immediate availability of updated information.

DataHub

Reason Two

Extensibility

Adapting to Unique Organizational Needs

Every organization's data landscape, governance requirements, and operational patterns are unique. Modern metadata platforms must provide deep extensibility to accommodate these differences without requiring compromises or workarounds.

Flexible Metadata Models

While standard metadata types (technical, operational, business) provide a foundation, organizations increasingly need to capture specialized metadata reflecting their particular:

- Industry-specific data attributes and classifications
- Custom data quality definitions and measurements
- Proprietary business glossaries and taxonomies
- Organization-specific ownership and stewardship models
- Unique compliance and regulatory requirements
- Special relationship types between data assets

Modern platforms must enable seamless extension of base metadata models without breaking standard functionality or requiring complex customization. This typically involves:

- Schema-on-read approaches that accommodate varying metadata structures
- Support for custom attributes at all levels of the metadata hierarchy
- Flexible relationship modeling beyond predefined patterns
- Ability to define and enforce custom validation rules
- Templates and inheritance mechanisms for efficient extension

DataHub

Robust APIs and SDKs

For true integration into organizational workflows, modern metadata platforms require comprehensive API capabilities that enable:

- Programmatic management of all platform functions
- Metadata capture from custom processes and applications
- Integration with existing data governance workflows
- Extension of platform functionality through custom modules
- Embedding metadata capabilities into other applications
- Real-time notification and event processing

These APIs should follow modern design principles (RESTful, GraphQL) with comprehensive documentation, consistent patterns, robust security, and versioning to support long-term stability of integrations.

Similarly, software development kits (SDKs) in major languages (Python, Java) enable developers to interact with the platform efficiently without managing low-level API details. These SDKs should include:

- Abstraction of complex API workflows into simple function calls
- Handling of authentication, pagination, and error conditions
- Type-safe interfaces reflecting the metadata model
- Examples and tutorials for common integration patterns
- Mechanisms for extending the SDK for organization-specific needs

Every organization's data landscape, governance requirements, and operational patterns are unique. Modern metadata platforms must provide deep extensibility to accommodate these differences without requiring compromises or workarounds.

DataHub

Reason Three

Unification

Breaking Down Metadata Silos

Every data team depends on discovery, quality, governance, and observability to manage their data. Traditional catalogs treat these as separate tools, but modern platforms bring them together in one unified system.

Integrated Discovery, Observability, and Governance

Managing disparate tools for data cataloging heaps an unwarranted load onto overburdened data teams. Some insightful leaders are having their iPhone moment for Data Catalogs, struck by the immense value of a unified metadata platform for discovery, observability, and governance like the obvious appeal of a smartphone unifying a mobile phone, camera, and navigation device into one sleek package.

Modern metadata platforms, like DataHub, are built with a deep understanding of data management workflows and the inter-dependencies of data consumers, engineers, and stewards.

Data should be searchableAND clearly assert its quality and freshness. It should be fresh AND trusted.

This unification enables transformative workflows where:

- Lineage information helps assess quality impacts,
- Governance policies apply automatically based on detected sensitivity,
- Quality improvement investments align with usage patterns, and
- Users receive relevant, governance-compliant asset recommendations

A unified approach to data discovery, observability, and governance is the only comprehensive way to deliver context-aware data management.

DataHub

Comprehensive Metadata and Lineage Graphs

At the core of this unification is a comprehensive graph representation of all metadata relationships, particularly lineage information. Modern platforms must capture:

- Technical lineage showing how data flows between systems
- Transformation lineage documenting how values are calculated
- Process lineage connecting data to the workflows that create and use it
- Impact lineage revealing dependencies between data assets
- Version lineage tracking changes to data and metadata over time

This graph structure enables users to navigate the complex web of relationships in modern data ecosystems, answering critical questions such as:

- Where did this data originate?
- What processes and transformations created this value?
- Which dashboards and applications would be affected by changes to this table?
- Who is using this data and for what purposes?
- How has this dataset evolved over time?

DataHub

Reason Four

Common Language

Supporting Diverse User Personas

Traditional data catalogs often serve a narrow audience of technical users. Modern metadata platforms must support diverse stakeholders with varying technical expertise, responsibilities, and objectives.

Tailored Experiences for Different Personas

For Data Consumers (business analysts, data scientists, application developers):

- Intuitive search and discovery focused on business context
- Self-service access to relevant data with appropriate controls
- Clear documentation of data meaning and appropriate usage
- Simplified lineage views focused on business processes
- Tools to request access, report issues, and provide feedback

For Data Engineers and Platform Teams:

- Detailed technical metadata and system-level lineage
- Performance and usage metrics to inform optimization
- Integration with data pipeline and orchestration tools
- API access for automated metadata capture and validation
- Monitoring and alerting for data quality and pipeline issues

For Data Governance Teams:

- Policy management and compliance reporting
- Automated classification and sensitive data detection
- Access control and entitlement management
- Audit trails and usage monitoring
- Risk assessment and remediation tools

For Data Owners and Stewards:

- Ownership assignment and transfer workflows
- Stewardship dashboards showing asset health
- Access request review and approval processes
- Impact analysis for proposed changes
- Quality and usage metrics for owned assets

Each of these personas requires not just different information but different interaction patterns, terminology, and visualization approaches. Modern platforms provide tailored experiences through:

- Role-based user interfaces highlighting relevant functionality
- Contextual help and guidance appropriate to expertise level
- Customizable dashboards and reports for specific responsibilities
- Terminology management ensuring consistent communication
- Workflow tools supporting cross-role collaboration

DataHub

Reason Five

AI-Ready

Preparing for the Intelligent Enterprise

The rapid adoption of artificial intelligence across enterprises creates new requirements for metadata management while simultaneously offering new opportunities to enhance metadata capabilities through AI itself.

Native AI/ML Asset Support

Modern metadata platforms must catalog and govern the unique assets involved in AI/ML workflows:

- Machine learning models and their versions
- Training, validation, and test datasets
- Feature definitions and feature stores
- Model performance metrics and monitoring data
- Model cards documenting intended use and limitations
- Experiment tracking information
- Deployment configurations and serving infrastructure

These assets have distinct metadata requirements, relationship patterns, and governance considerations compared to traditional data assets. Platforms must provide:

- Specialized entity types with appropriate attributes
- Relationship models capturing AI-specific dependencies
- Lineage tracking for the full ML lifecycle
- Governance controls addressing AI ethics and responsible use
- Integration with popular ML platforms and tools

DataHub

AI-Enhanced Metadata Management

Beyond supporting AI assets, modern platforms leverage AI capabilities to enhance metadata management itself:

- Natural language generation for documentation creation
- Anomaly detection in data quality and usage patterns
- Intelligent search and recommendation systems
- Automated policy enforcement and risk assessment
- Entity resolution and duplicate detection
- Relationship inference from usage patterns
- Automated data classification and tagging

These capabilities dramatically improve the scalability and effectiveness of metadata management, reducing manual effort while improving coverage and accuracy.

Balanced Processing Mechanisms

AI workloads introduce new patterns of metadata interaction requiring flexible processing approaches:

- Pull-based mechanisms for periodic batch processing
- Push-based updates for real-time metadata changes
- Event-based processing for workflow integration

Modern platforms support all three patterns, enabling AI systems to:

- Discover available data through catalog queries
- Validate data quality and lineage before use
- Record new assets created during processing
- Trigger workflows based on metadata changes
- Update documentation and context information

AI-Agent Interaction

As organizations adopt AI agents and assistants, metadata platforms must support these new interaction patterns:

These interfaces enable AI systems to autonomously discover, understand, and properly use organizational data assets while remaining within governance boundaries.

- APIs designed for AI consumption with appropriate context
- Natural language interfaces, using emerging standards such as the Model Context Protocol (MCP), for metadata queries
- Semantic understanding of data meaning and relationships
- Authorization frameworks for AI-initiated actions
- Explanation capabilities for human oversight

DataHub

Reason Six

Security

Protecting Sensitive Metadata

As metadata becomes more comprehensive and integral to operations, its security becomes increasingly critical. Modern platforms must address these concerns through multiple complementary approaches.

Secure Metadata Ingestion

Metadata collection often requires access to sensitive systems and may itself contain sensitive information. Modern platforms provide:

- Remote executor architectures separating connectivity from metadata storage
- Agent-based collection operating within security boundaries
- End-to-end encryption for metadata in transit
- Credential isolation and secure secret management
- Least-privilege principles for connector operations

These mechanisms ensure metadata collection doesn't create new security vulnerabilities or require excessive permissions.

Deployment Flexibility

Organizations with stringent security requirements need deployment options that accommodate their security architecture:

- Private cloud deployment within customer-controlled environments
- On-premises options for high-security scenarios
- Hybrid architectures separating sensitive and non-sensitive metadata
- Integration with existing identity and access management systems
- Support for virtual private cloud networking

This flexibility ensures metadata platforms can meet even the most stringent security requirements while maintaining full functionality.

DataHub

Enterprise-Grade Security Standards

Beyond specific security features, modern platforms demonstrate commitment to security through:

- SOC 2 compliance and regular third-party security audits
- Comprehensive encryption for data at rest and in transit
- Fine-grained access control at the metadata level
- Detailed audit logging of all platform activities
- Vulnerability management and regular security updates
- Support for enterprise security requirements like SSO, MFA, and RBAC

These standards ensure the metadata platform itself doesn't become a security liability as it becomes increasingly central to data operations.

DataHub

Reason Seven

Future-Proof

Adapting to Evolving Requirements

Perhaps most importantly, modern metadata platforms are designed to evolve alongside rapidly changing data management practices, technology landscapes, and regulatory environments.

Continuous Innovation

Modern platforms demonstrate ongoing commitment to innovation through:

- Regular feature updates addressing emerging requirements
- Incorporation of technological advances in data management
- Adoption of new security capabilities as they mature
- Evolution of user experiences based on feedback
- Extension of connector ecosystems to new data sources

This continuous evolution, usually with thriving open-source communities, ensures the platform remains relevant as organizations' data strategies advance rather than becoming technical debt that eventually requires replacement.

AI Transformation of Data Management

As AI capabilities mature, they will fundamentally transform how data is managed. Future-proof platforms are positioned to leverage these advances through:

- AI assistants for metadata curation and governance
- Autonomous metadata collection and enrichment
- Predictive analytics for data quality and usage
- AI-driven optimization of data architectures
- Natural language interfaces for all user personas

Organizations investing in modern metadata platforms position themselves to adopt these capabilities as they mature rather than facing disruptive replacements.

DataHub

Standards Adoption

The metadata management platforms continue to evolve through industry standards and best practices like:

- Contribute to open metadata ecosystems
- Actively participate in standards development
- Quickly adopt relevant standards as they emerge
- Provide migration paths from proprietary approaches
- Support interoperability with complementary tools

This standards orientation ensures organizations avoid vendor lock-in while benefiting from industry-wide advances in metadata management practices.

DataHub

Conclusion

Traditional data catalogs served important purposes in the early phases of enterprise data management, providing basic inventories and discovery capabilities for relatively static data landscapes. However, as organizations embrace digital transformation, adopt artificial intelligence, and face increasingly complex governance requirements, these traditional approaches have reached their limits.

The seven reasons outlined in this paper highlight why organizations should reconsider their current data catalog implementations and adopt modern metadata platforms designed for today's data management challenges:

1. Scalability to handle the volume, variety, and velocity of metadata in modern data ecosystems
2. Extensibility to adapt to each organization's unique data landscape and requirements
3. Unification of discovery, observability, and governance on a shared metadata foundation
4. Common Language supporting diverse user personas with tailored experiences
5. AI-Ready capabilities for both managing AI assets and enhancing metadata through AI
6. Security features protecting sensitive metadata while enabling appropriate access
7. Future-Proof design ensuring continued relevance as data management evolves

Organizations that embrace modern metadata platforms position themselves to achieve greater value from their data assets, reduce governance risks, improve operational efficiency, and prepare for an increasingly AI-driven future. Those that remain with traditional catalogs risk falling behind as data complexity grows and new use cases emerge.

The transition to modern metadata platforms represents not just a technology upgrade but a fundamental shift in how organizations think about metadata - from static documentation to dynamic, operational assets that actively participate in data workflows and enable new capabilities across the enterprise.

DataHub

About Us

DataHub is an AI & Data Context Platform adopted by over 3,000 enterprises including Apple, CVS Health, Netflix, and Visa. Innovated jointly with a thriving open-source community of 13,000+ members, DataHub's metadata graph provides in- depth context of AI and data assets with best-in-class scalability and extensibility. The company's enterprise SaaS offering, DataHub Cloud, delivers a fully-managed solution with AI-powered discovery, observability, and governance capabilities. Organizations rely on DataHub solutions to accelerate time-to-value from their data investments, ensure AI system reliability, and implement unified governance—enabling AI & data to work together and bring order to data chaos.