WHITEPAPER

Peraton

DATA CATALOG: KEY TO A MODERN FRAMEWORK

DO THE
CAN'T BE DONE.

INTRODUCTION

Organizations strive to maximize the value of their data for consumers by making data and data products more discoverable, accessible, and understandable. Data fabric architectures and data mesh approaches are prevailing tactics for modern, distributed data ecosystems. Successfully implementing these types of systems requires a capability that supports rapid data discovery and integrates with enterprise data management processes. Today's data catalogs act as the marketplace for governed data enabling collaboration between data providers and consumers with verifiable and measurable products.

More importantly, the modern data catalog is a vital link between data and metadata and provides the connective tissue between enterprise data management functions and the consumer-focused insight services which are necessary to anchor a modern data framework. A catalog delivers data in the context of the organization's overall information management ecosystem enabling more visibility, reuse, and metrics on leveraging data within the enterprise. Here we describe the key tenets of a modern data catalog and the enhanced outcomes from implementation of modern data catalogs. We'll outline best practices and keys to success in employing a modern catalog to accelerate a government agency's data management objectives.

CATALOG–A NEW GENERATION

Modern Data Catalogs-Characteristics and Features

Today, data catalogs allow users to access high quality data quickly and easily, within a controlled, secure, data governance model. The catalog is a nexus for all types of users—data analysts, stewards, scientists, and business consumers—to access trusted organizational data and metadata. They aid data consumers, allowing them to search for, identify, subscribe to, and ultimately consume enterprise data sets in a streamlined fashion. Data catalogs also instantiate a business glossary so that users can easily understand not only what data assets are available, but how they relate to business processes and how they should be used. Consumers can refer to contextual business terms associated with the dataset, browse a dataset's lineage, and review any usage restrictions for the data. A primary feature is the ability for data owners and consumers to collaborate on data sets, making them discoverable and reusable.

A catalog provides the facility for implementing operational and compliance policies. It tracks sensitive data, reinforcing governance policies and keeping track of sources for privacy and compliance. Modern catalogs are embracing AI/ML and graph technology to provide enhanced knowledge representation of data and data assets across the distributed data ecosystem.

Modern data catalogs come in a variety of configurations. Some have been purpose-built to fulfill needs in leveraging data assets. Others are integrated into cloud provider platforms while others are included in larger data management platforms. All have an extensive set of connectors to link to a variety of data sources and data stores. Each has their own strengths and choosing is dependent on an organization's data strategy and priorities. Several considerations include features compatible with enterprise objectives and ability to easily integrate with the existing or future data environment.

Feature Sets

Data catalogs support the discovery, enrichment, and usage of enterprise data. They include a range of capabilities with diverse configurations and user interfaces. A catalog should serve today's needs and grow in an organization's distributed data framework. Features to consider when selecting a catalog include:

- Semantic layer: A semantic layer functions as a source of enterprise data knowledge combining a business glossary with specific information about data and data fields to provide context to an organization's data
- ML Augmented: The ability to automate metadata tagging and management identifying relationships between enterprise data
- Governance: Catalogs enable reuse and control of an organization's distributed data and many represent the lineage of data assets
- Visibility and collaboration: Catalogs are ideal to enable self-service user needs—data owners, consumers, etc. to access trusted data with quality scores and usage metrics. Users and data stewards more easily collaborate across the enterprise on new data products that support business/mission needs
- Privacy and compliance: Some catalogs provide discernibility to PII and GDPR data to assist in regulatory compliance. If your organization requires adherence to privacy laws and regulations, features that highlight sensitive data assist in enabling compliance

Integration

Ease of integration into enterprise data management processes is an important consideration in selecting a catalog that is compatible with, and optimized for, an organization's data management system and processes.

- Evaluate a catalog in terms of your overall data strategy. Are you fully moving to cloud? Will you have a hybrid environment? What stores and applications do you need to connect to? Do you have existing tools in your platform that offer a full-featured catalog option? Will the data catalog be the primary collaboration point for enterprise data? Consider the answers to these questions to indicate a short list of tools that fit seamlessly into your ecosystem reducing costly integration, support, and maintenance.

Preference for open source or COTS

There are robust choices for enterprise data catalogs that are both commercial and open source. Regardless, of whether you use COTS or open-source software, maximize flexibility by subscribing to open architecture standards.

Assess your organization's requirements for a data catalog and match them to key features and strengths. Evaluate the fit from a governance, business, and technology perspective. It's important to consider catalog selection from a "people" perspective, too. In alignment with your data strategy, make certain the catalog serves the variety of needs of different roles and responsibilities to make data more discoverable, accessible, and trustworthy.

Figure 1. Enterprise Data Catalog Dimensions

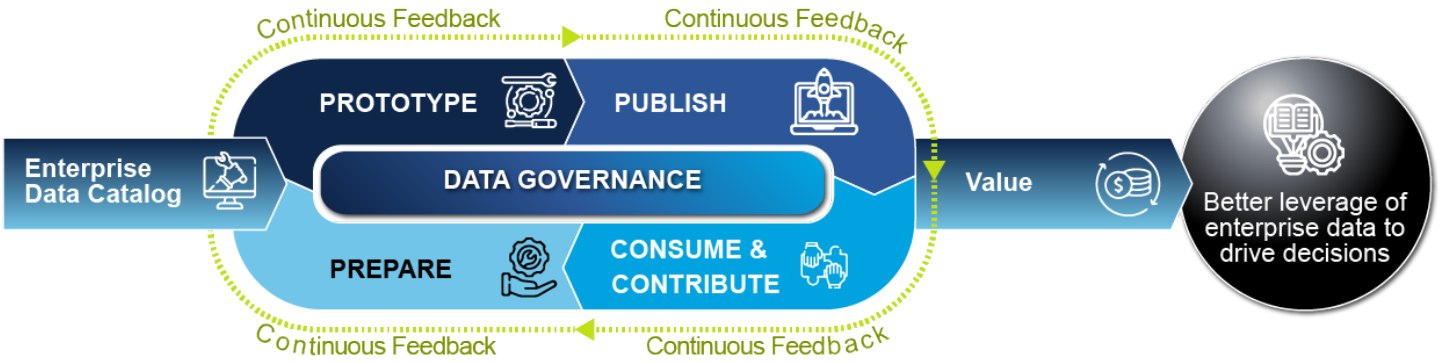

Figure 2. Peraton's Enterprise Catalog Implementation Lifecycle

GETTING STARTED– A LIFECYCLE APPROACH

Proper planning for a catalog implementation is essential for success. An enterprise catalog implementation requires an agile lifecycle approach. Peraton's five-phase lifecycle in Figure 2, maximizes the value to enterprise stakeholders ensuring data is discoverable, accessible, and reusable. Consumers gain increased trust and assurance in the data. Catalog implementation typically eliminates manual tasks, improving time to insight for business/mission knowledge.

Phase 1: Selection

An enterprise data catalog supports an organization's collaborative data culture. The ability to handle different roles, in accessing and contributing data are to be highly considered in the selection process. From a systems perspective, the catalog must be scalable and enable distributed storage and handle decentralized models. Keep in mind the key tenets around features and usage, ease of integration and open architecture. Tools must have an open architecture with APIs and allow applications to update the catalog and enable building upon the catalog to enable automated updates of data assets. The product must have a strong vision and roadmap so that future upgrades can take advantage of the continuing advancements in AI/ML and data technology.

Phase 2: Preparation

A catalog becomes a hub for the data community. Key data stakeholders must be involved in defining the priority data domains and products to be cataloged first, then demonstrated in a subsequent prototype to anchor the catalog to the data types and usage requirements of the organization. Initial iterations focus on identifying high value mission critical data assets. These datasets and assets should be of high quality to foster trust, encourage adoption, and reuse. Initial preparation includes setting the foundations for successive iterations.

Our approach uses collaborative workshops with stakeholders to elicit design/configuration requirements and use cases to support data governance, access management and security, privacy, compliance, and analytics.

The outcome yields a plan for transitioning workflows from source systems to be accessible in the catalog and provides that deep connection to data domains to foster adoption across the organization. The architecture is set during the preparation phase and addresses non-functional requirements to ensure a performant system in the data environment. This informs the implementation roadmap to configure and deploy the catalog, establishing the foundation for data and business knowledge.

Early involvement with stakeholders is key to adoption. In the preparation phase, our approach establishes a training and education program to support the various roles of the implementation, stewards, experts, contributors, administrators, etc. and others. From inception, we set up a measurement program to quantify and communicate improvements in usage and adoption as well as efficiency in leveraging data assets.

Phase 3: Prototype

During prototyping a skilled data engineering team installs the catalog in the data environment enabling appropriate roles and security controls. Additional configuration elements, including links to data assets and applications are needed to be created to address the use case requirements. This includes rules for integration with the identity management application and set up request, review, and approval workflows for additional access. Any customizations needed for mission/business and technical requirements such as specialized data types and workflows are important to model in this phase for alignment to organizational standards and processes.

Next comes the identification of data assets populating the metadata and deriving a business glossary by associating business terms to establish a common semantic foundation. Auto discovery using AI/ML of data assets for the targeted domain area is helpful to rapidly "stock" the catalog. Figure 2 shows a logical data model of metadata content for the catalog. Data stewards identify any manual data asset additions and ensure fields are defined in according to source systems. These could be operational data sources (ODS), enterprise applications, cloud data stores, non-relational data stores, business intelligence (BI) tools, and many others. It's critical to connect the catalog to data from existing metadata repositories to build relationships between these assets.

Our agile lifecycle approach pilots use cases for data domains but maintains a governance focus across the organization. The benefit of this approach prioritizes key data domains while incorporating lessons learned throughout the implementation. In turn, this creates a foundation enabling an increase in velocity to drive success. Quick successes advertised across the organization gain momentum in the adoption the enterprise data catalog.

Governance centers our agile implementation approach. Collaboration across data governance boards, data stewards, and other stakeholders ensures that metadata and other assets are consistently aligned to leverage distributed data as an asset. Rules are determined on what data is made available, who can access it and how it should be used. Information that contains personally identifiable information (PI)I and sensitive data can be identified and carefully managed. The catalog provides the facility for implementing operational and compliance policies.

Testing the prototype against the use cases confirms the connection to data sources, the discovery of domain information, the quality of data, lineage, policies, and the ability of users to easily find and utilize the data. The goal is to drive value and user adoption to accelerate discovery of knowledge and decision making across the organization. Applying DataOps principles at the outset achieves rapid, flexible, and reliable delivery of data, delivering information into the hands of the users in a more consistent and timely manner. This provides a connected data management approach across the lifecycle to support a consistent process which reduces risk, gains productivity, and expedites delivery.

Phase 4: Publish

At the Publish phase, data assets are identified, tagged, enriched, linked, tracked, and made visible and accessible. At this point, data stewards, owners, and stakeholders have already been working in the staged catalog to ensure the business glossary is complete and accurate. Quality data sets and assets are available from the sources and approved data pipelines. Roles and access rules are populated. Documentation is available and experts are identified to help consumers with the assets. Several critical activities at this stage support adoption.

- Outreach: A communication plan and campaign are developed with leadership to drive awareness, training, and education to users and stakeholders to adopt and contribute to the catalog
- Workflow Scenarios: Use case demonstrations show consumers how the catalog advances their work processes, increases efficiency, and improves data quality. Users see how others have exercised data in similar projects. This approach fosters reuse and streamlines processes on new projects.
- Measurement: A set of performance measures are put in place to determine usage, quality, and efficiency and are reported on a visible dashboard. This demonstrates transparency and builds momentum for the catalog.

In the Publish phase, the catalog moves from pilot to production. The implementation is monitored for service performance, identifying who is using catalog, and ensuring that priority assets are identified seamlessly into the catalog. Role-based access is administered by the identity and access management system to enforce policies and safeguard proper handling of PII/sensitive data. Automated metrics are operationalized to support performance measurement and address any issues that may be discovered.

Phase 5: Consume & Contribute

In the Consume and Contribute phase, users/consumers can find assets and view data relationships by searching the catalog using keywords, filters, tags, and business terms. Users can preview asset content and observe data quality through content reviews. They can also select content utilizing filters such as “most recommended,” or “highly rated.”

The catalog becomes a collaboration tool via a crowd sourcing model, where users can team up with data stewards, owners, etc. to add or refine assets. Once the enriched metadata is migrated to the catalog, business users consume data by querying across datasets using embedded query editors that highlight sensitive attributes and promote the usage of certified schema. A catalog can help create a trust-based governance model. By users assigning quality/usefulness scores to their authoritative data assets, they become a good quality indicator of the of those data assets over time.

Monitoring the catalog tracks the activities of data consumers to understand the actual frequency of usage and determine which datasets are most important, which ones are related, and the nature of those relationships. Published metrics on query usage analytics, and certification metrics help business users identify which integrated datasets have been certified, and which ones have not and understand the effectiveness of adoption.

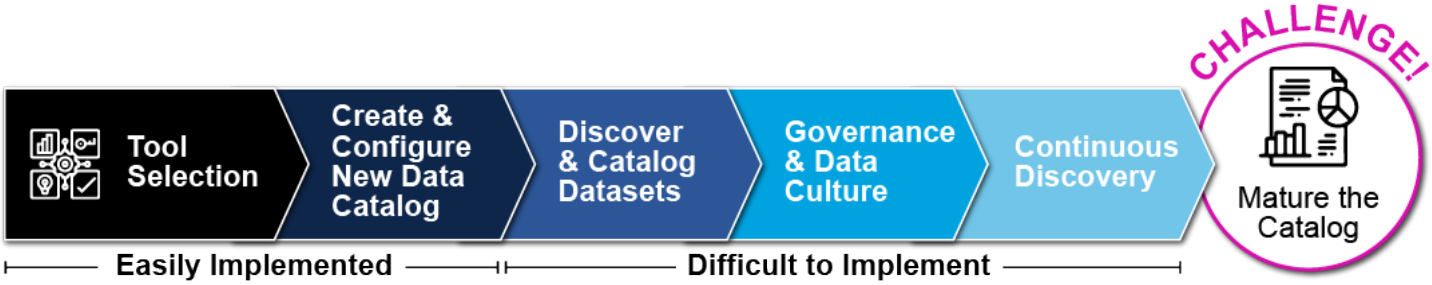

Figure 3. Barriers to Data Catalog Implementation Success

CRITICAL SUCCESS FACTORS AND BEST PRACTICE

While technology implementation is usually the primary focus in establishing a data catalog, human and cultural dimensions are equally important to consider. They are often not given enough attention and they can affect the success and overall return on investment. Figure 3 underscores that working with enterprise data owners, populating data sets, and engaging with users can be the most challenging part of deployment.

Achieving a strong foundation for success depends on working closely with the business domains, identifying the use cases for implementation and "advertising" the purpose and benefits through data governance activities.

As we all know, a "build-it-and-they-will-come" approach fails to engage people to actively use an enterprise catalog. Planning a "campaign" to drive adoption of the catalog is key.

Executing on use cases populates the catalog with timely and business relevant data that is now visible and accessible across the enterprise data fabric. In our experience, there are several other critical success factors and we've employed best practices to address them.

New Data Access Methods

Data consumers may resist a common enterprise approach and continue to rely on personal networks and tribal knowledge for their data sets. The resulting risk arises from is using data that has not been governed or vetted for quality standards. Consumers may not take advantage of new curated data sets that meet their needs, and the organization fails to reduce manual and duplicated data wrangling and integration work. Peraton works collaboratively with the key stakeholders to promote a user enablement and training workshop that includes business specific use cases to ease the onboarding process.

Culture Shift

An enterprise data catalog is most successful in a culture of data sharing, knowledge sharing, and collaboration. Participation is a key element of data culture at all levels. Working closely with the domain subject matter experts and data stewards to promote the catalog for the most used data sources accelerates the shift toward a data-driven culture. We advocate hosting regular workshops for business components to convey the value of the catalog. These forums initiate new use cases and requirements for reusable data products to address the individual data needs of the consumer. These sessions can substantially reduce resistance to data sharing, resistance to knowledge sharing, reluctance to participate in collaborative curation, and reluctance to post ratings and reviews.

Data Literacy

Many users have various responsibilities that depend on data analysis, but they have not been trained in data literacy skills. The skills needed to produce useful information from raw data may not be clear to them. For example, data selection, data understanding, data preparation, data analysis, data visualization, and data storytelling are not intuitive to all users. Executive sponsorship for investments in data literacy activities throughout the organization can answer common data-related questions for users and provide positive outcomes for the agency, including acceptance and usage of the data catalog as the primary marketplace for an organization's trusted data.

IMPACT OF ENTERPRISE CATALOG

Better Access, Reduced Silos

At a large federal civilian agency, an enterprise data catalog implementation decreased information silos and increased data accuracy, integrity, and completeness. The initial prototype leveraged machine learning to uncover relationships within the data and obtained a detailed inventory of all data assets in a domain area to be curated and managed. A strong security and compliance foundation supports privacy while providing authorized users the data they need. In our approach, we worked closely the Office of the Chief Data Officer, project sponsor, data stewards, business and application owners and security SMEs to provide data assets to enable evidence-based decision making. A centralized metadata repository of information was used to discover, tag, and classify data and map it to a business glossary for consumers across all data sources. including Oracle, MySQL, PostgreSQL, Tableau and OBI.

With the data catalog, user auto-discover data assets either on-premises or in the cloud and pull associated metadata for the data assets. There is an end-to-end data lineage view with a graphical representation. This enables data stewards to prototype within the tool and create trusted datasets. By establishing a centralized repository for trusted datasets that is searchable, there is an ability to crowdsource and rate datasets and gather statistics on usage. The data catalog implementation spurred further data modernization efforts to drive data knowledge and sharing seamlessly across data domains.

Peraton

ABOUT PERATON

Peraton is a next-generation national security company that drives missions of consequence spanning the globe and extending to the farthest reaches of the galaxy. As the world's leading mission capability integrator and transformative enterprise IT provider, we deliver trusted, highly differentiated solutions and technologies to protect our nation and allies from threats across the digital and physical domains. Peraton supports every branch of the U.S. Armed Forces, and we serve as a valued partner to essential government agencies that sustain our way of life. Every day, our employees do the can't be done by solving the most daunting challenges facing our customers. Visit Peraton.com to learn how we're safeguarding your peace of mind.

Scan to learn more at
peraton.com/capabilities/cloud/