

Data Catalog:

Creating a Single
Source of Reference

Introduction

Every business is trying to become data driven, and managing data as a true economic asset is a fundamental part of that transformation. But today's analytic environments are rapidly growing, both in the volume of data that they store and the number of self-service analytic users that they support. When you have hundreds of data sources, potentially millions of different datasets and thousands of people constantly consuming data, the transformation to a data-driven organization quickly becomes overwhelming.

Thirty years ago, the goal was to organize all enterprise data in one location to achieve a "single source of truth." But now there are simply too many data sources and too many self-service tools. Data may be stored in a data warehouse, a data lake, or any number of additional structures; reside in the cloud, on-premises, or both; and drive business insight through a collection of different algorithms, dashboards, spreadsheets, and visualizations. Rather than create a "single-source of truth" today's reality is that the best we can do is create a "single source of reference."

The machine learning data catalog has emerged as a key technology enabler to a "single source of reference" -- one place for anyone within the organization to find curated data, understand how that data has been used and why it was created, and trust that it is right for the analysis at hand, whether they are a data scientist, an analyst, or even a casual business consumer of data.

The Power of a Machine Learning Data Catalog

Organizations leveraging a data catalog see outsized returns from their analytics infrastructure.

According to the Gartner Magic Quadrant for Business Intelligence and Analytics Platforms, 2017, "By 2020, organizations that offer users access to a curated catalog of internal and external data will realize twice the business value from analytics investments than those that do not."

The value comes from a number of areas: increased analyst productivity -- an increase in productivity of 20 to 50 percent on average for an individual analyst; the ability to govern data with greater accuracy and agility; and the gains seen from increased collaboration between different parts of the business that might otherwise not be able to share a common understanding of the data.

In a May 2018 article in the Wall Street Journal, Julie Schiffman, vice president of business analytics at Pfizer, discussed how leveraging a data catalog as part of the company's analytics platform is helping employees from different divisions within the company collaborate. "Data has become an extremely powerful currency for any company, and what we were finding is the data was very fragmented," Schiffman said. With a data catalog as part of the Virtual Analytics Workbench, Pfizer's analytics platform is yielding new insights that were previously difficult to identify, including identifying patients with rare diseases that might previously have gone undiagnosed.

Data catalogs are driving business outcomes across a wide range of industries, even in organizations that have traditionally been analytics leaders. Munich Re is one of the largest reinsurers in the world, staffing thousands of analytic experts called actuaries, and insuring everything from hurricanes and the California wildfires to satellite launches and large building projects. In a September 2018 presentation at Strata New York, Andreas Kohlmaier, head of data engineering at Munich Re, shared that more than 2,000 users leverage a machine learning data catalog to better stay ahead of the rapidly changing risk landscape and to collaborate on new data-driven insurance products. One of the outputs of that work was the founding of a new IoT subsidiary of Munich Re, an IoT-specialized team that leveraged a data catalog to collaborate with other Munich Re experts across the globe to provide end-to-end coverage for anyone who plans to invest, run, and build a wind farm.

“One the advantages of having one global platform is that people can collaborate and share their ideas on the data, which has resulted in the execution of more than 250 business use cases,” Andreas Kohlmaier, head of data engineering at Munich Re.

The Data Catalog Journey

A data catalog represents a critical piece of the analytics strategy for any organization seeking to leverage self-service analytics to become data driven. Once an organization understands that

a data catalog is a critical part of their architecture, the next question is: “Build it or buy it?”

One of the key misunderstandings about a modern, machine learning data catalog is that it is not merely an inventory of data and metadata. The new diversity of sources to store data (SQL, NoSQL, NewSQL, Graph, etc), the ability to process data in more complex ways outside of SQL (Spark, MapReduce, etc), and the dynamic, iterative nature of modern analytics, means that organizations need far more information than a simple inventory or metadata repository can provide. In today’s world, automation is a necessity. Manual documentation efforts to capture details and organizing data appropriately are no longer sufficient. In most organizations, data proliferates much too quickly. Automated data lineage and propagation of appropriate tags on data is required to organize data if it is going to be accessible and properly governed.

The machine learning data catalog also accounts for the broad skill-set diversity of self-service users. Before self-service analytics, business intelligence tools were built for the power-user who was SQL savvy. In today’s world where there are many different types of data user with varying skill sets and wide-ranges of data literacy, organizing data for accurate analysis presents a huge issue to the modern enterprise.

Data catalogs harness machine learning to capture behavioral and usage patterns, crowd source endorsements, and leverage data recommendation engines as techniques for supporting the wide variety of users all through simple, Google-like interfaces.

With these considerations in mind, there is a lot to consider before

building a data catalog. Two of the most data-savvy technology companies on the planet -- eBay and LinkedIn -- undertook the initiative to build a data catalog and their journeys provide important learnings for any organization contemplating a project to build a data catalog.

Build, Build, Build, and then Buy... an eBay Story

eBay connects millions of buyers and sellers around the world with operations in more than 30 countries. At the time that they started their data catalog initiative, eBay was processing 100 PB/day of data, generating 50 TB/day of new data, and running 7M+ queries a day. More than 300 data analysts and 5,000 business users were accessing eBay's analytics platform directly and through more than 10,000 reports in Tableau and 5,000 in MicroStrategy. Data was stored in a Teradata Data Warehouse and an on-premise Hadoop data lake.

The sheer volume of data assets available in eBay's environment made an un-curated data experience at eBay incoherent, disjointed, and untrustworthy. Recognizing that this complexity was detrimental to the company's commitment to enabling every employee to make data-driven decisions, eBay started a data catalog journey that would last for

five years and include numerous iterations that ultimately saw limited success.

The Wiki-approach

First, eBay tried to build an open catalog, an internal wiki-like metadata repository meant to be used to share knowledge on data and manage metadata. The open catalog would act as a crowd-sourced inventory. The project, however, suffered from low adoption. The manual effort required for individuals to populate and keep wiki pages updated, turned out to be too much additional effort for users.

The Social “DataHub”

Next, eBay created a “DataHub”, which included a Facebook-inspired “like” feature for endorsing datasets. By creating a social construct for updating the wiki pages, eBay hoped that users would have more incentive to document their work. “DataHub” had some success but was ultimately too disjointed and too difficult to navigate. Data was getting documented, but the “like” feature fail to capture deeper context and it was still difficult to understand how data assets were being leveraged.

Sharing Stories

The next inspiration, “Storytelling,” enabled users to share stories and best practices. Although “Storytelling” was a novel idea, the application struggled to tie data with context and connect it to the business problem.

The Data Forum

Finally, in the last building effort, eBay built an internal platform called “AnswerHub,” which enabled people to ask questions about data. A dedicated support staff would then attempt to answer those questions. eBay soon found that this method was not scalable. The demands on the team outpaced their ability to answer questions, resulting in a backlog of unanswered questions.

After four attempts to build its own data catalog, eBay was introduced to the Alation Data Catalog. Alex Liang, one of the key architects on the project, said the move replaced the original IKEA-like do-it-yourself model with a governed self-service approach that was more user-friendly—and ultimately, more effective.

eBay’s Data Catalog Journey at a Glance

4 approaches + 5 years

- Environment: Teradata + Hadoop + Many BI Tools
- Data: Petabytes of data, structured & semi-structured logs

Key learnings:

- Analytic metadata is dynamic, therefore machine learning is necessary
 - Wiki-based approaches are too static
 - Analysts do not document their work
- Social collaboration is hard to design
 - What’s the model: Wiki or Facebook or LinkedIn?
 - Context and meaningful engagement are key

LinkedIn Experiments with Open Source

LinkedIn took a different approach than eBay and built its own data catalog with the open source project "WhereHows." WhereHows was developed to address the massive amount of data that LinkedIn was capturing: 50 thousand datasets, 15 PB of storage (across Teradata, Hadoop, and other sources), 14 thousand comments, and 35 million job executions.

WhereHows turned out to be a huge undertaking. In the initial two years of development, the project involved 12 contributors that ranged from software engineers, architects, application developers, and product design. WhereHows did gain adoption among IT users as a good knowledge-based application and a metadata repository but saw little engagement among analysts and business users. As the user-base of self-service analytics grew, LinkedIn saw the need for a catalog designed for engagement and collaboration. LinkedIn purchased Alation to serve the self-service analytics user base that drives many of the analytics decisions in the company.

LinkedIn's Data Catalog Journey at a Glance

Open source project doesn't engage analysts and business users

- 515 PB of storage across Teradata + Hadoop
+ additional sources

Key learnings:

- Effective data catalog requires collaboration and community engagement
 - How to get annotations into a technical platform?
- Huge lift to build and support
 - Team of 12 in LinkedIn data team
+ open-source community
 - Separate team for column-level lineage

What to Consider When Choosing a Data Catalog

Whether deciding to build a data catalog or buy one, a data catalog should be robust enough to support the complexities of your environment, the needs of your users, and the goals of the business. While the needs of each organization will vary, before beginning the data catalog journey consider these key capabilities:

Unified view of your data:

A data catalog provides the most value when it has a robust search capability that covers three types of captured metadata --- technical, operational, and business metadata --- along with user behavior. Make sure that a data catalog offers a combined view of all your data, not just a view of a subset

Data Catalog: Creating a Single Source of Reference

of your data or one type of data. For example, a catalog for just Hadoop or just relational databases will have limited functionality.

Intuitive search and discovery:

Data discovery should be available to everyone in an organization who needs to work with data. To make data discovery more accessible, be sure to include natural language processing (NLP) search, which can help users of different literacy levels find and discover data.

SQL query Tool:

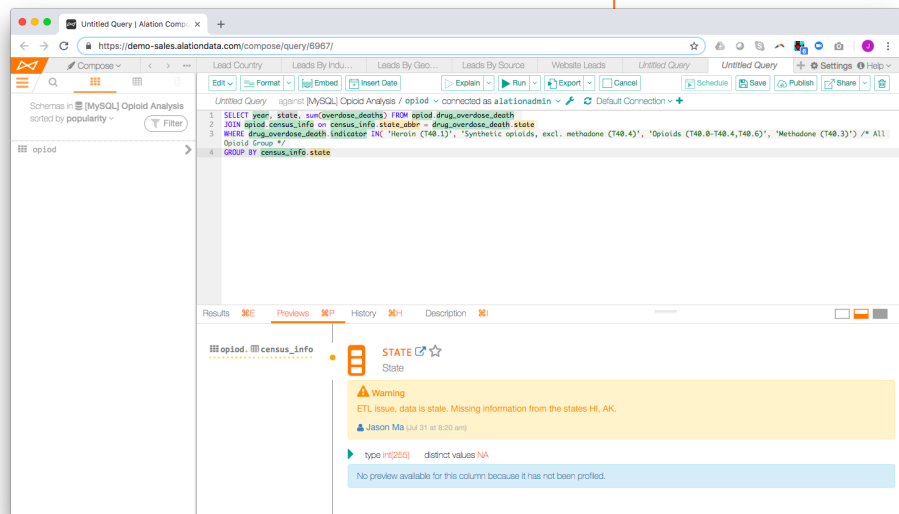
In order to enable self-service analytics for business users, the tools must be intuitive and not require a lot of training or technical knowledge. A data catalog should be designed to

let users immediately begin to write queries in

Standard Query Language (SQL). Adding suggestions to guide users to the best filters and joins, popular columns, and more would be ideal. Data analysts and stewards can write and save queries for use by less technical users, scaling data discovery across an organization.

Machine learning to enhance data context:

Some data catalogs function as simple inventories without the



Alation Compose is an SQL editor with built-in SmartSuggest and TrustCheck capabilities

important behavioral observation capabilities. A data catalog should incorporate machine learning algorithms designed to provide context about the data and how it is used. Machine learning should also alleviate the burden of manually tag and record context on data. Machine learning can enable the data catalog to become smarter over time. As more and more human user behavior is observed and confirmed, the data catalog can provide fine-tuned contextual information, which enhances decision making.

Verification of trusted sources:

Data catalogs should include a data lineage feature to allow users to trace the sources of your data. Look for additional measures of verification, such as data flags or annotations. This capability allows users to endorse an asset of value or provide a warning or deprecation if an asset is outdated or inaccurate. Direct human verification increases trust in data.

Collaborative capabilities to break down silos:

Collaborative capabilities built into the fabric of a data catalog can break down organizational silos. When teams of analysts work in silos, work is re-created rather than re-used, and a great amount of organizational knowledge remains unshared. A data catalog should enable information to be shared across teams with integrated communication tools to enable direct dialogue.

Support for data curation:

Today, data curation is more integral than ever for supporting data governance initiatives but is also becoming increasingly

difficult in this ever changing data landscape, where the volume, velocity, and variety of data is growing rapidly.

A data catalog can give end users the ability to upvote/downvote data assets and a sense of where the data came from, indicating whether or not it can be trusted. Complimenting a crowdsourced model with automated documentation speeds up curation efforts - this includes capturing usage information around top users, popular schemas/tables/columns, and joins/filters.

Native integrations:

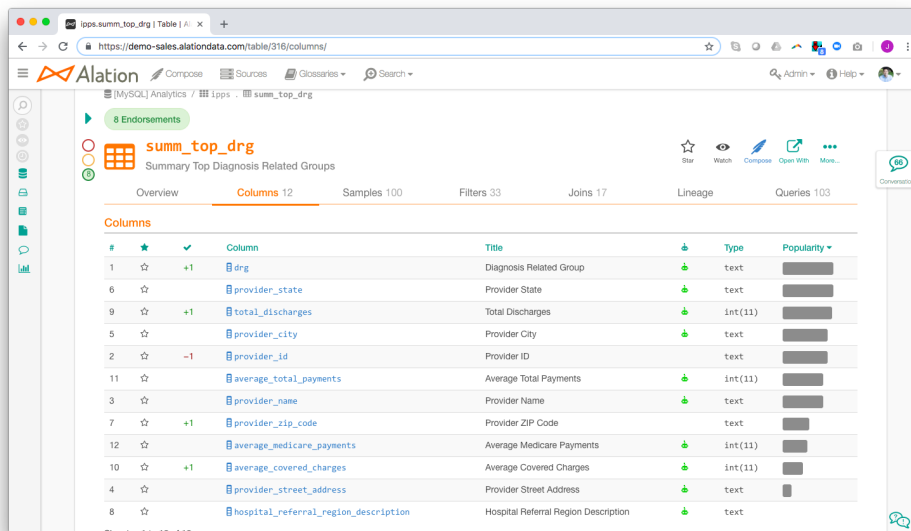
Today's organizations must keep up with what can sometimes feel like an overwhelming number of data sources. Data may be coming from a data warehouse, a data lake, or any number of additional inputs. A data catalog should normalize all of these sources and provide a single source of reference for all of the organization's data, creating an inventory of your data and a single place to access all of your data assets. The ability to connect a data catalog to different sources and ingest data is the bare minimum needed. Also consider connecting to business intelligence (BI) tools to catalog your end to end flow of analysis.

SQL query log analysis:

Part of the context that a data catalog can provide comes from the ability to parse a usage log and track the behavior of people that are accessing the data sets. At a bare minimum, an enterprise should build a data catalog that provides metrics on how many times a data set has been queried and by whom. On top of these basic metrics, behavioral statistics

Data Catalog: Creating a Single Source of Reference

such as which schemas, tables, columns, filters, joins and queries add a wealth of context. Ideally, surfacing machine-learned usage patterns along with the technical metadata will provide end users with a more complete picture.



#	Column	Title	Type	Popularity
1	drg	Diagnosis Related Group	text	██████████
6	provider_state	Provider State	text	██████████
9	total_discharges	Total Discharges	Int(11)	██████████
5	provider_city	Provider City	text	██████████
2	provider_id	Provider ID	text	██████████
11	average_total_payments	Average Total Payments	Int(11)	██████████
3	provider_name	Provider Name	text	██████████
7	provider_zip_code	Provider ZIP Code	text	██████████
12	average_medicare_payments	Average Medicare Payments	Int(11)	██████████
10	average_covered_charges	Average Covered Charges	Int(11)	██████████
4	provider_street_address	Provider Street Address	text	██████████
8	hospital_referral_region_description	Hospital Referral Region Description	text	██████████

A table catalog page showing columns ranked by popularity

Build vs Buy at a Glance

- What are the opportunity costs to leverage internal resources to build a data catalog?
- Can you wait two to three years for development to finish?
- Is the data catalog just for IT or the business? Do you have resources to build the UI/UX?
- Does your development team have the bandwidth to support or maintain the data catalog? Are there resources to train and drive adoption of the data catalog?
- Do we have the machine learning expertise to capture technical, operational, business and social metadata metadata?
- Who will own maintenance and support of the data catalog?