Alation

Using a Machine Learning Data Catalog to Reboot Data Governance

David Loshin
President of Knowledge Integrity, Inc, and consultant in the areas of data quality, master data management, and business intelligence

Rebooting Data Governance

Early attempts at data governance were largely predisposed to "organization": developing organization charts mapping out the roles of the "data governance council," the data owners, and the data stewards, along with ironing out the details of processes for defining and approving data policies. In some cases, these organization activities were accompanied by rote operational tasks such as manually surveying tables and documenting data element metadata. In general, these activities are focused on what could be called the "data production lifecycle," or the processes applied from data acquisition or creation to its delivery to some sort of database, data warehouse, or other type of reporting system.

The challenge is that limiting the data governance activity to manipulating org charts or mindless manual tasks does not contribute to any of the key objectives of operational data governance such as:

- **Expanded data availability and simplified data accessibility**, allowing more data consumers to find and access shared data assets;
- **Standardized data semantics**, providing a common understanding for how data consumers use shared data assets; and
- High confidence in the **trustworthiness of the data**, implying measurably high data quality.

These data governance goals share a common driver: they are motivated by data consumption and information use. Yet the inability to achieve these fundamental goals diminishes the organization's ability to effectively share and leverage the corporate data assets, let alone address compliance and auditability of higher-order information directives stemming from internal business policies, externally-defined laws and regulations, or conformance to industry standards.

In essence, current data governance practices and procedures are insufficient to meet the needs of the enterprise. Data governance must be reframed to address business requirements to support the growing communities of data consumers in meeting these foundational objectives. And while the boundaries of the technology horizons for data governance are still being explored, an expanding pool of technology vendors are publicizing their products as data governance solutions.

In essence, current data governance practices and procedures are insufficient to meet the needs of the enterprise.

In recent years, many organizations have adopted and filled the position of the Chief Data Officer (CDO), yet within that short time, the responsibilities have subtly shifted away from a purely operational set of technical data oversight activities to ensuring data democratization enabling innovative data-centric problem solving. When considering technical solutions, the challenge for the chief data officer, though, is to differentiate between the features promoted by each vendor as being *sufficient* for operationalizing data governance and the capabilities truly *necessary* to satisfy the organization's operational data governance needs.

Recent industry trends show a continuing interest in defining enterprise data policies and practices coupled with processes and procedures for operationalizing oversight over data asset acquisition, transformation, persistence, access, and protection. Data governance increasingly encompasses solutions that blend traditional stewardship priorities (such as understanding data asset usage and harmonizing business glossary terms) with emerging critical demands to assess the organizational data landscape (including identifying, classifying, and documenting the myriads of data assets across the enterprise) to guide the availability, accessibility, and proper use of data. At a high level, data governance implies adherence to data policies, but at a finer level it means ensuring *data utility* – that data assets are consistent, access to shared assets

is transparent, and data consumers are provided with all the information they need to make proper choices for data asset selection and use.

This paper examines the need to revisit your data governance strategy and clearly identify the core requirements for operationalizing data governance. No single tool is likely to satisfy all of these requirements. And although it is to your organization's benefit to consider the technical capabilities that are derived from blending the use of selected tools to achieve data governance goals, this paper will then discuss how data catalog tools blend collaborative metadata, data discovery and assessment, and information policy management in a holistic and systematic manner.

Machine learning capabilities enhance the ways that data catalogs can be used to evolve a comprehensive data governance strategy. By addressing organizational requirements for data governance in a way that scales with the growth of the user community, an intelligent data catalog helps motivate data producers and consumers to work together, reduces the burden on the data stewards in continuously describing data, and streamlines operationalizing compliance with data governance expectations.

Machine learning capabilities enhance the ways that data catalogs can be used to evolve a comprehensive data governance strategy.

Data Governance Requirements

There are foundational requirements associated with each of the aforementioned data governance goals. Expanding data availability and accessibility are factors of raising enterprise data awareness. Most enterprises are home to a variety of data assets, yet many of these assets are effectively hidden when there is no organizational inventory. This implies the need for a data discovery process to crawl across the enterprise, identify, and then catalog corporate data assets. A robust data discovery process will scan the contents of each data asset, determine

whether it is structured, semi-structured, or unstructured, infer the data asset's metadata, and even categorize the data asset in terms of the sensitivity of the embedded content.

Data attributes and meta-tags need to be delineated and documented. Inferred metadata from early scans can be presented to data professionals to classify and label according to known reference domains and metadata and alignment with defined business terms. This contributes to standardizing semantics and providing common understanding of shared data.

Mapping the data production workflows and encapsulating this data lineage provides a valuable service for data quality assessment and remediation. When attempting to identify the root cause of an identified data error, being able to traverse the data lineage backwards from the point of discovery allows a data steward to determine the point where an error is introduced. For example, if when perusing a report, a business analyst might be alerted to values that seem inconsistent with her intuition. In this case, she can trace through the data lineage to review how those values were created, and potentially determine that fixing the source of a data flaw that was introduced at a prior processing stage would fix the report.

Alternatively, data lineage mappings are an excellent resource when analyzing the impacts of changes to business processes or data sources. For example, a system analyst might want to determine which applications are affected when there are changes to a state's tax code.

Assessment of sensitive data is rapidly becoming a requirement, especially as the number of global regulations mandating protection of individuals' personal and private data increases. It is critical to be able to identify which data assets contain information about individuals, and whether that information is categorized as protected according to one or more regulations. Combining information about data asset sensitivity and data lineage allows data stewards to determine

who the users are, what their access rights are in relation to data assets with sensitive data, and to augment data processing workflows with controls to ensure against unauthorized access.

A searchable inventory of corporate data assets collected into a data catalog enhances data accessibility and availability. Enabling a data consumer to search data assets by structural attribute or by semantic topic empowers a broad set of data users to find the data sets that best meet their needs. A data catalog can be used to share different types of metadata, including:

- Physical metadata describing the structure of the source system, such as table and column names.
- Logical metadata describing semantic information such as database descriptions, data quality expectations, and associated data policies. Logical metadata, which is often fed from a lexicon derived from documentation, accounts for specification of business terms and their variants (and corresponding meanings), such as the term "adj" referring to "adjustment," "adjacent," or "adjutant."
- Behavioral metadata describing how data assets are used in a variety of use case scenarios. Behavioral metadata may be the most important, since it gives automatic insight into every object in the system like popularity of schemas, tables and top users.

Machine Learning Data Catalogs to Operationalize Data Governance

Technology can help satisfy data governance requirements by simplifying data discovery, automatically inferring metadata, improving the fidelity of those inferences, and providing visibility to business glossary, data element definitions, data

lineage, and data assets, data obligations (such as privacy protection). These tools help surface the right data assets to individuals and simplify data consumers' ability to find and use corporate data assets.

And while a conventional metadata repository or simplistic data catalog provide some of these capabilities, there is an emerging class of *intelligent data catalogs* that employ machine learning (ML) and artificial intelligence (AI) algorithms to enhance the data catalog's ability to support operational data governance activities. In addition to discovering the physical and logical characteristics of data assets, machine learning can be used to mine auxiliary assets such as transaction and query logs for behavioral insight – different classes of data consumers, which data assets are accessed more frequently, the types of queries users are executing, as well as tracking collaboration associated with documented physical and logical metadata. Machine learning data catalogs are able to use advanced analytics algorithms in a number of ways, such as:

- **Improving automated data discovery and classification:** One can seed the process of automated data discovery, providing an initial taxonomy of data structures, types, and sensitivities. The results of early iterations of the discovery process can be presented to data stewards and domain subject matter experts, who will review the inferences, make corrections when necessary, and provide additional categories and labels. The ML algorithms will learn from these human interactions to refine the discovery and classification processes, improving inference fidelity while reducing the need for human interaction.

- **Making data consumer recommendations:** Different communities of data consumers can use the data catalog to search for data assets that meet their needs. As search results are presented to the different types of data consumers, the ML algorithms leverage active learning by incorporating user selections and actions to iteratively

refine predictive models to improve search results and recommendations. Similar techniques can be used to confirm existing business term definitions, make predictions about data assets that might meet the data consumers' needs to surface the right data assets more quickly to the right set of users for their reporting and analytics needs.

- **Assessing data sensitivity to support compliance:** Different laws have different definitions of what individual data is considered to be "personal" or "private." Through interaction with humans, intelligent data discovery tools can learn to automatically categorize data attributes as personal/private and determine which data assets contain sensitive data that is subject to regulatory compliance.

Intelligent data catalogs blend traditional metadata management capabilities (such as business glossary, structural metadata management, object metadata, and data lineage) with machine learning and artificial intelligence algorithms to learn from human interaction to continually contribute to data governance operationalization.

Intelligent data catalogs blend traditional metadata management capabilities with machine learning and artificial intelligence algorithms to learn from human interaction to continually contribute to data governance operationalization.

Considerations

One might say that the first generation of data governance practices have focused on aligning the organization with fundamental data management principles, such as documenting structural metadata, ensuring data quality, or instituting master data management. And while these operational tactics are necessary for effective digital transformation, they are certainly not sufficient.

Organizations that only embrace techniques and tools to facilitate reactive data policy compliance will find that acute data issues might be addressed, but chronic impediments to

Case Study:
GoDaddy's Data Accessibility is Powered by Alation's Machine Learning Data Catalog

GoDaddy originally started as a company that sold web domains and over time has evolved into a company whose primary focus is helping small business get online. GoDaddy helps its customers by providing a complete end-to-end web presence solution including web domains, domain hosting, help to build out business websites, supporting the creation of eCommerce capabilities, promote personas, and whatever else is necessary to enable entrepreneurs to make their online presence a reality.

Early on in GoDaddy's BI organization tenure analyst were frequently asked to perform simple charting or trending analysis for internal partners. It became apparent that in order to meet the sheer volume of these reporting requests, internal partners needed self-service tools. At that time, there were many different sources of data, but essentially the data was everywhere, put into different platforms such as SQLServer, MySQL, and eventually Teradata and Hadoop Hive, with a lot of data moving among these different platforms. This made it very difficult for individuals to find the data sources they needed to answer their reporting and analytics questions.

In 2013, the group implemented Tableau, in a collaborative effort with the BI team and the Enterprise Data Organization. Together they began to gather, prepare, and surface corporate data for reporting and analysis. The introduction of Tableau was empowering, allowing end users to perform simple reports and analyses that previously required the support of members of the BI/BA organization. Yet as empowering as the tool is, the challenge changes from *how* to do analysis to **finding the right set of data sources that can be analyzed.**

As with any growing organization, the ability to hunt down and find the right data source is difficult. Manual approaches to data

discovery are slow and incomplete, requiring experts to help find, understand, and trust the data, and is additionally complicated by subtle differences in data element definitions that confound self-service data consumers. This challenge is difficult because of two key issues – not only is there a need to identify those data sources, they must be adequately described so that end user can determine which data sources are right for their tasks, understand the levels of governance that have been applied to each data source, learn which data elements are available for use, and importantly, how the data source should be used. Data analytics tools such as Tableau and Google Analytics surface the data for more than 2000 consumers across dozens of projects to use, but the challenge is that with this proliferation of data assets, individuals were no longer able to determine which of the plethora of data sources to look at!

The majority of these 2000 people are generally less technically-skilled resources and have limited knowledge of SQL or how to join or otherwise manipulate these data sources. Yet a desire to help data consumers access the data they needed led to the introduction of curated data packages referred to as a Unified Data Set, or UDS. UDS packages are curated, maintained, monitored, documented, and have full support and oversight. These UDSs are used to package data by domain, such as orders, web traffic, marketing details to simplify the ability to access and answer 80% of their questions, freeing the BI/BA organization to allocate more time to satisfying the more complicated requests or performing deeper analysis by digging into the raw data.

To provide visibility into the breadth of the data sources, GoDaddy implemented the Alation Enterprise Data Catalog. Alation was able to connect to different data sources and pull in the metadata, table structures, database schemas, data types, as well as query logs (that show how people are using the data). In addition, Alation connects to Tableau and pulls in information about workbooks used for creating dashboards and visualizations along with Tableau calculated fields to surface how the field is calculated.

Alation allows users to search for data sources by concept and presents characteristic metadata about the found data sources, especially the UDSs that typically bubble up to the top of the list. Data stewards and analysts can augment the data object description in the catalog to provide details about what information is contained within the data object, how it was constructed, how it can be accessed, and ways that it can be used. This provides insight into the full context around each data source: what is in each data set, who is using it, data glossary items, columns by popularity of use, sample data (limited by the users' access rights), published popular queries that could be reused, all of which provide guidance to the users as to what data sources they should be using.

Data stewards can validate and certify data sources, and these can be managed as endorsed data sources within Alation. That means that when data consumers review a data source or a Tableau report, they can see the endorsement (at the data source level). In addition, data stewards can endorse specific columns. The product also allows the data steward to deprecate columns to indicate that they should no longer be used; reports and published that are dependent on deprecated data elements are automatically flagged and Alation provides recommendations for replacements.

Data stewards standardized 150 top terms that needed further documentation and moved them into Alation's data dictionary to serve as a business glossary. Fields are tagged with characteristic information, such as whether the column include personal information requiring protection, whether the data element is authoritative, and definitions that are applied consistently across collections of uses across the application landscape.

Through machine learning, Alation surveys the breadth of the data landscape, identifies where data elements are likely to be associated with known data classes or data types, and makes recommendations to the data stewards about definitions and glossary terms. These capabilities allow the data stewards to harmonize definitions across different sources and applications

and automatically enforce conformance to agreed-to specifications. At GoDaddy, Alation provides users with endorsed, managed, and maintained data sources that can be used to address their analytic needs. It automatically helps to populate a data element catalog and interacts and learns from the data stewards to assign definitions. And in combination with Tableau, Alation empowers users to leverage self-service data access with confidence.

effective democratization of reporting and data analytics will remain. These organization will stay stuck in their ingrained reactive data stewardship tasks while their nimbler competitors leap ahead in the market.

That being said, "next-generation" data governance initiatives are maturing from exercises in laying out organization charts and defining the roles and responsibilities of data stewards to more comprehensive programs for translating business directives such as improving the customer experience, increasing revenues, reducing costs, or regulatory compliance into operational and enforceable corporate information policies. This transition has inspired a focus on data consumer-facing facets of aspects of data usability, manifested through simplified data accessibility, standardized data semantics, methods supporting data quality, classifying data assets to support regulatory compliance, and importantly, speeding the ability to surface the right data assets for reporting, business intelligence, and analytics. Observing these policies demands more agile accessibility and integrated oversight in enabling and speeding self-service data accessibility.

Delineating operational requirements supporting these data governance facets reveals that conventional metadata management tools must be augmented with technology for assessing and characterizing the data assets across the

enterprise data landscape. This include technology for data discovery, assessment of physical, logical, as well as behavioral metadata – a more comprehensive, intelligent accumulation of information to enhance the data consumers' collective experiences.

Look for tools that facilitate communication and collaboration among the different corporate stakeholders – business consumers, business analysts, and data governance staff – to make sure that all of their needs can be addressed. Avoid tools that skew too far to one of these constituencies, either by imposing restrictions on data use or constraining access to data asset metadata. Instead, embrace those tools that can balance data control, provide accuracy in surfacing the right data assets, and generally raise data awareness.

Machine learning data catalogs internalize these guidelines. They use machine learning and artificial intelligence algorithms to learn from human interaction to improve automated data discovery, map and track data lineage, and even identify sensitive data subject to regulatory compliance. More importantly, a data catalog leverages machine learning to understand usage patterns, correlations between user requests and selected data assets, and user affinity to data sources based on classification and content. The integration of intelligence into the data catalog environment helps to automate significant aspects of operational data governance and helps data users find the right sets of data sources to meet their reporting and analytics needs.

More importantly, a data catalog leverages machine learning to understand usage patterns, correlations between user requests and selected data assets, and user affinity to data sources based on classification and content.

Alation

Alation, the data catalog company, is building a data-fluent world by changing the way people find, understand, trust, use, and reuse data. The first to bring a data catalog to market, Alation combines machine learning and human collaboration to bring confidence to data-driven decisions. More than 100 organizations, including eBay, Exelon, Munich Re and Pfizer, leverage the Alation Data Catalog. Headquartered in Silicon Valley, Alation is funded by Costanoa Ventures, DCVC (Data Collective), Harmony Partners, Icon Ventures, Salesforce Ventures, and Sapphire Ventures. For more information, visit alation.com.

Contact Us

info@alation.com
(650) 410-7164