IDERA® | Data Governance Whitepaper

EMPOWERING DATA GOVERNANCE USING DATA LINEAGE

BY DAVID LOSHIN

INTRODUCTION

Much of the conventional guidance about implementing a data governance program focuses on organizational structure and the protocols for defining and agreeing to corporate data policies. Yet as organizations get their data governance councils set up, it is becoming clear that an effective data governance initiative must go beyond defining roles and holding monthly council meetings. Operational data governance requires methods and tools for enforcing compliance with corporate data policies.

While data modeling, metadata management, and data quality tools have been the typical tools in the data steward's arsenal, these technologies often focus on static data or data at rest. Organizations are increasingly becoming aware of the need for visibility into the way that information flows across the extended information enterprise and the governance characteristics of data in motion.

Data lineage is the key to providing this visibility, and this paper examines ways that data lineage empowers data stewards to operationalize key aspects of data governance. The paper begins with a definition of data lineage, which is followed by a description of how data lineage supports a variety of key operational data stewardship activities. The paper concludes with suggestions of some technical capabilities to look for in a solution providing support for data lineage.

WHAT IS DATA LINEAGE?

Traditionally, metadata has been limited to descriptions of data object characteristics (such as when the data object was created, the number of records it contains, or who the data owner is) and structural information (such as data element names, definitions, data types, and lengths). Traditional metadata describes a data object at rest and can be used to support a data steward's responsibilities in managing a business glossary, business term definitions, and how enterprise data elements map to that business glossary. Yet this type of metadata does not provide insight about data object production: the ways that data values move from the various acquisition points to an assortment of files, database tables, reports, and analyses. And this is a critical gap in empowering data stewards to operationalize data governance.

Data lineage is critical in mapping data architecture in the context of the enterprise data landscape, and effectively rounds out the knowledge about data sources, connectivity, and flow to help data stewards in enforcing compliance with data policies. Data lineage augments the traditional concepts of data object metadata to provide a holistic description that also describes a data object's sources, information pipelines used to produce a data object, as well as how that object is made accessible (such as where the data is stored, how the data object got there, the transformations applied, and the types of controls that are integrated around the data).

Data lineage empowers data stewards in their governance efforts in a number of ways, including:

- Analyzing data dependencies
- Validating semantic consistency
- Impact analysis
- Data quality root cause analysis
- Integrating data controls
- Enforcing regulatory compliance

ANALYZING DATA DEPENDENCIES

In an ungoverned information environment, there are two different types of data dependency risks. One is where siloed business function application development belies the fact that there are undiscovered hidden data dependencies. Reports, dashboards, and analyses may appear to be derived from data sets from isolated systems, but in many cases there is a chain of processing that ultimately originates with data taken from a shared data source. Analyzing and documenting data lineage allows the data steward to trace the data pipelines and expose these dependencies.

Alternatively, there is what could be called "virtual dependence," in which multiple data assets are populated using data from distinct, yet structurally and semantically equivalent sources. For example, one organization might have two systems that capture and manage customer data, and two downstream reports that respectively are sourced from the distinct customer data sets. Data lineage may expose the fact that the two customer data sets are effectively equivalent, allowing the data stewards to integrate the two sources into one sharable customer master data set.

VALIDATING SEMANTIC CONSISTENCY

Alternatively, data lineage coupled with metadata management will highlight where similarly-named data elements actually represent different concepts, recommending against their conflation through an integration process. This will help identify where there are overloaded concepts within the environment and will provide the contexts for data architects to see where data elements are not aligned from a semantic perspective and provide recommendations for eliminating any inconsistencies.

IMPACT ANALYSIS

A common challenge for application owners is understanding what systems are affected when there is a change to a data source or modifications to business process requirements. Often, there is little visibility as to which applications depend on data elements that originate from the changed data source or are associated with a set of system requirements, especially when there are multiple extractions and transformations across an extended data processing pipeline.

Fortunately, data lineage provides a mapping from each data source to all of the downstream data objects as well as the data pipeline processing stages and application touchpoints. Data lineage can indicate what applications and databases are affected when there is a change to a business definition, data element structure, or associated system requirement. When there is a proposed change to a data source, one can use data lineage to trace the information flow forward from the origination point within the organization to any application (including generated reports and analyses). By identifying the applications that depend on the modified data source, the data steward can determine what the actual impact would be and assemble a plan for proactively planning application updates across all impacted systems to accommodate the changes.

DATA QUALITY ROOT CAUSE ANALYSIS

When a data flaw is detected by a data consumer, it can be challenging for the data steward to identify the source of the introduction of that data flaw. Data lineage simplifies the root cause analysis process by providing visibility into the sequence of processing stages through which the data flow. Data lineage is used to produce a backwards trace of the information flow, allowing the data steward to examine the quality of the data at the entry and exit points of each processing stage. By identifying the point at which the values no longer comply with expectations, the data steward can determine the processing stage in which the error was introduced.

When the data steward has identified the processing phase where the data flaw was introduced, the application owners can be engaged to develop small tests, review the application code, and consider the ways that the quality of the data could have been impacted. As soon as a root cause is determined, the application owners can plan and implement code improvements that will prevent that specific error from being introduced again.

Encapsulating the use of data lineage for reverse tracing of data errors creates a repeatable process for data quality root cause assessment and identification. Data lineage enables the enterprise to reduce and possibly eliminate recurring data quality issues in a systematic way, thereby reducing impacts to important decisioning processes.

INTEGRATING DATA CONTROLS

The data steward's role is not limited to only addressing data quality problems when they are found by data consumers. The data steward is also responsible for engaging the data consumers to identify their data quality requirements and to formalize the collected data quality rules. These data quality rules can be used to institute data controls that measure and quantify the levels of data quality at different points of the data production processes.

At any of these control points, if a record is found to violate a data quality rule, the responsible data steward and the application owners can be alerted that there may be a potential issue with the data production process. Automated monitoring and notification not only alerts the data steward about where in the process flow the error originates, it will classify the error type based on the specific rule (or rules) that were violated. That provides valuable intelligence to the data steward and the application owner, allowing them to rapidly investigate the identified process failure point and attempt to resolve the issue before invalid data objects are delivered to downstream consumers.

The typical challenge, though, is that the data stewards may not know where to embed these data controls. Naively, they can attempt to implement all potentially relevant controls for all data values across all the data pipelines, but obviously this is overkill. Instead, one can use data lineage to select critical data elements used by the data consumers, trace the data production flow, and identify specific data processing tasks that are prime targets for embedding the data controls. Using data lineage to automate the integration of data controls that proactively monitor for data errors supports the data steward's job of continuous data quality management.

ENFORCING REGULATORY COMPLIANCE

In response to the growing number of global laws and regulations about protection of personal and private information, there is a need to be able to operationalize enforcement of regulatory compliance associated with data protection. This includes identifying sensitive data elements, monitoring how those data elements are processed and disseminated, identifying where there are risks of exposure, as well as complying with requests associated with consumer data protection rights, such as the "right to erasure" from the European Union's General Data Protection Regulation (GDPR) or the California Consumer Privacy Act's (CCPA) "right to deletion."

Yet many organizations are not aware of how complex compliance may be, especially in cases like the CCPA, where the definition of personal information includes "inferences drawn from any of the" enumerated types of personal information. This means that a consumer's request for deletion might encompass a variety of dependent data objects that generally would be difficult to pinpoint.

Data lineage provides visibility into that data dependency chain. And not only will it support the operational aspects of privacy regulation compliance, the data lineage mappings can be used to produce compliance reports through automated auditing of data pipelines to demonstrate that proper data protection precautions are instituted.

CONSIDERATIONS

As this paper has suggested, data lineage augments the organization's toolkit for empowering data stewards to implement data governance. It helps in analyzing data dependency, validating semantic consistency, and facilitates the analysis of impacts of changing data sources and requirements. Data lineage supports operational stewardship tasks such as data quality root cause analysis and the integration of data controls. Finally, it supports enforcing and reporting on data privacy regulatory compliance.

When considering technologies that provide data lineage capabilities, look for products that not only support capturing and managing data lineage, but also provide these value-added capabilities:

- The ability to enable users to see the flow of data through the data production lifecycle.
- A mechanism for enumerating the data sources for the different data pipelines.
- The ability to identify data elements and link them to data models and to metadata for data element concepts and business glossaries.
- A method of documenting data transformations and allowing data professionals to review those transformations across a variety of data pipelines.
- The capability of interoperating with existing ETL/data integration tools to import data pipelines along with their collected transformations.
- A means for collaboration around data pipelines and associated metadata.
- The ability to display a visual presentation allowing data stewards to review the data lineage.

ABOUT THE AUTHOR

David Loshin is the President of Knowledge Integrity, Inc., a consulting and development company focusing on customized information management solutions including information quality solutions consulting, information quality training and business rules solutions. David is a recognized thought leader and expert consultant in the areas of analytics, big data, data governance, data quality, master data management, and business intelligence. Along with consulting on numerous data management projects over the past 20 years, David is also a prolific author regarding business intelligence best practices, with numerous books and papers on data management.

ER/STUDIO ENTERPRISE TEAM EDITION

ER/Studio Enterrpise Team Edition provides the collaborative framework for managing business glossaries, metadata, reference data, rules, and policies. This helps to drive stakeholder collaboration through context-based links and management. Administrators can configure stewardship roles and permissions to enforce security policies. Common and consistent reference data can be linked across models and reconciled across operational systems for standardized reporting and analytics. Advanced change management, audit trails, and optional JIRA integration are enabled by Team Server.

Request a Demo