OCTOPAI

Enhance Data Value with Multilayer Data Lineage

David Loshin
President, Knowledge Integrity, Inc.

Accessible and Visible Data Informs Strategic Decision-Making

In today's data-driven world, it would be unusual to find organizations that are not leveraging a wide variety of data sets for reporting, business intelligence (BI), and analytics to inform many business decision processes. Most organizations extract data from existing transaction or operational systems and integrate that data using standardizations, transformations, and loading into data warehouses, marts, and other analytical platforms enabling enterprise reporting and analytics.

Conventional data warehouses that are populated through batch data extracts from on-premises transaction and operational systems stream data through staged transformations in preparation for loading have become commonplace. Reports derived from organizational data funneled into a traditional data warehouse architecture inform business leaders about the current state of operations, while data transformed into actionable information drives strategic decisions.

However, enterprise data strategies have evolved beyond the limitations of extracted structured data sets originating within on-premises systems. The staged, lock-step batch processing associated with traditional data warehouses is perceived to be increasingly brittle and lacking in supporting more sophisticated analysts and data scientists. And with growing BI, reporting, and analytics demands, the existing (and future!) needs of a growing array of business data consumers of varying skills and analytical sophistication are increasingly difficult to meet. This becomes particularly concerning when business analytics and strategic decision-making relies on available, trustworthy, current, and accessible data.

Accessible and Visible Data Informs Strategic Decision-Making

In today's data-driven world, it would be unusual to find organizations that are not leveraging a wide variety of data sets for reporting, business intelligence (BI), and analytics to inform many business decision processes. Most organizations extract data from existing transaction or operational systems and integrate that data using standardizations, transformations, and loading into data warehouses, marts, and other analytical platforms enabling enterprise reporting and analytics.

Conventional data warehouses that are populated through batch data extracts from on-premises transaction and operational systems stream data through staged transformations in preparation for loading have become commonplace. Reports derived from organizational data funneled into a traditional data warehouse architecture inform business leaders about the current state of operations, while data transformed into actionable information drives strategic decisions.

However, enterprise data strategies have evolved beyond the limitations of extracted structured data sets originating within on-premises systems. The staged, lock-step batch processing associated with traditional data warehouses is perceived to be increasingly brittle and lacking in supporting more sophisticated analysts and data scientists. And with growing BI, reporting, and analytics demands, the existing (and future!) needs of a growing array of business data consumers of varying skills and analytical sophistication are increasingly difficult to meet. This becomes particularly concerning when business analytics and strategic decision-making relies on available, trustworthy, current, and accessible data.

The challenge is that the conventional data warehouse architecture is predicated on presumptions that:

- All data originate from known sources.
- There are well-defined processing pipelinesthat transform data in its original format to one usable for reporting and business intelligence (BI).
- There is some control over the data production processes.

In fact, the opposite is increasingly true. Data warehouses are rapidly being expanded to incorporate data from a much wider variety of structured, semistructured, and unstructured data sources. Organizations are ingesting data sets originating from a wide variety of external sources. Increased enterprise data awareness inspires different kinds of data consumers to reuse both source data as well as processed data. And that processed data is likely to be fed back many times across the environment to power corporate information intelligence offerings such as recommendation engines, prediction models, and other machine learning and artificial intelligence (AI/ML)- driven capabilities that rely on accurate and reliable data supplies.

Organizations are also expanding their enterprise data architectures across an increasingly hybrid environment composed of both existing on-premises systems and cloud-based platforms. Data sets are expected to flow across these platforms to support end-user business intelligence and analytics. Yet at the same time, data consumers want to continue to use their favorite end-user BI tools and analytics techniques.

In other words, there are more and more end-user analytics environments with their own data pipelines. This distributed and decentralized approach to analytics leads to diminishing centralized oversight and control of how data flows across the environment. Because enterprise data environments are increasing in complexity, overall data awareness is eroding.

With an increase in distributed authority and a growing number of sophisticated BI and analytics data consumers who exercise control over their own data pipelines, there is a need for additional insight to ensure that business decision-making can be enabled by available, trustworthy, and current data. In this paper we look at how increasing data visibility through a combination of methods for data lineage can, among other things, provide insight into data dependencies across the enterprise, simplify the analysis of root causes for data issues, and reduce risks associated with auditing and compliance.

Data Lineage: Addressing Challenges That Impact Information Value

A key to addressing some these challenges requires understanding the nature of data pipelines. Simply put, a data pipeline is a sequence of data processing steps, in which each step consists of a source, one or more processing stages or transformations, and delivery to a destination. Collectively, the business data consumers rely on information produced as a result of data pipelines to meet a business objective.

There are different types of data consumers, and all are becoming more sophisticated in the way they assemble their reports and analyses. Data analysts tasked with producing reports and their business analyst counterparts rely on self-service analytics to configure and produce their reports. Data scientists are afforded opportunities to create their own data pipelines using data preparation tools. And as more analytics consumers develop their reports and analyses, there is a corresponding step-up in the creation of data pipelines.

There are inherent risks associated with the rapid proliferation of data pipelines across the enterprise. Flaws introduced into data at the source may unknowingly impact multiple reporting and analytics applications downstream. Changes to one or more data sources may require additional tinkering at target data repositories. The legal and compliance department may insist on an assessment of impacted system related to changes in externally directed policies (such as industry standards or compliance with data privacy laws).

And as enterprise data architectures become more complex, the lack of visibility into the data pipelines and the corresponding information production processes compounds the issues. Not knowing what information is available for use, not knowing which data assets to use, and a lack of knowledge about data element values' contents, accuracy, relevance, and timeliness all can impact the downstream data consumers. These effects are multiplied with growth in the number, depth, and breadth of data pipelines.

The remedy to the opacity of these intertwined data pipelines is enabling full visibility into the data and information production pipelines across the enterprise. That visibility is provided by data lineage. Data lineage is a capability allowing data practitioners to infer the different stages of data pipelines and produce a collection of end-to-end mappings for the collections of data pipelines in the enterprise. Producing these mappings allows data professionals to gather intelligence about how data flows throughout the organization and understand systemic dependencies related to information production. Data lineage provides the much-needed visibility into data production and pipeline processes to empower data consumers to amplify organizational information value.

Enterprise Data Visibility Enhances Data Value

Data lineage is a tool that supports multiple operational use cases that depend on visibility across the multiple dimensions, such as:

- **Speeding up the time-to-value of data**: The ability to rapidly configure and produce an analysis or a report is critical for taking the greatest advantage of enterprise knowledge. Data lineage provides data analysts and data scientists with insight into the available data landscape and can speed the development of analytical results that drive decision-making.
- **Simplify and automate compliance**: Most people immediately equate data "compliance" with observing data privacy laws, but compliance actually subsumes a much broader spectrum of conforming to defined data policies. By providing a visual map of the ways that data flow across the enterprise, data lineage allows compliance analysts to determine whether there are any vulnerabilities that introduce risks of non-compliance. At the same time, the lineage map can suggest where integrated compliance monitors can be inserted to enable integrated audit reporting.
- **Capture and preserve corporate data architecture**: Assembling a comprehensive map of the data pipelines that drive corporate applications (particularly reporting and analytics) documents key aspects of the organization's data architecture. Not only is this valuable when considering technical modernization and application renovation, it also provides a foundation for training newly onboarded technical staff, who can be briefed on the corporate data landscape by touring the data lineage maps.
- **Impact analysis**: As a corollary to documenting data architecture, data lineage provides an operational perspective of organizational data flows that can be used to assess the impact of a modification to a data source or some aspect of code within the dependent data pipelines.
- **Optimization**: With the proliferation of data pipelines, there are bound to be situations in which data is extracted from the same data sources multiple times or that the same data transformations are replicated in segments of multiple data pipelines. Data lineage provides visibility into those pipelines in which redundant or replicated processes can be identified and then optimized to speed data delivery.
- **Speed up and optimize root cause analysis**: Data incidents are often the result of data flaws whose impacts are manifested at the end of one or more data pipelines. In many cases, those individuals that discover (and are most affected by) data flaws are the least likely to have control over the processing stage that introduced flaws. By providing multilayered visibility into the data lineage, a data practitioner is empowered to rapidly reduce the search space for where the data flaws are introduced, speeding root cause analysis as well as informing remediation of discovered issues.

Comprehensive Data Lineage to Support Enterprise Use Cases

While there are different approaches to mapping data lineage, many are limited in their ability to visualize both a global perspective displaying end-to-end data dependencies as well as drillable perspectives presenting the details of how each pipeline works and, correspondingly, how data element values are produced. Therefore, for a data lineage tool to be most effective, it must provide a cross-dimensional view of the collection of data pipelines to support a variety of enterprise use cases:

- **Cross-System Data Lineage**: Cross-system data lineage maps how data flows across systems and provides linear visibility at the systemic level. Cross-system data lineage produces a view that differentiates between the different components of all data pipelines at a system level. This perspective enumerates the different types of data sources such as database systems, source files, or incoming data streams, the various data transformation process steps, and the different types of delivery points such as produced reports, analytics applications, or end-user analytics tools. This allows data consumers to see the full trajectory of information from an origination point within the environment, across data integration stages, analytical processing, and delivery to one or more final destinations.
- **Inner-System Data Lineage**: Inner-system lineage analyzes the actual data integration procedures and, through reading procedural directives including metadata, integrated SQL queries, code within SQL scripts and stored procedures, and additional application resources, infer the specific transformations applied to column data within each data pipeline's processing stages. The result is that inner-system data lineage provides the most detailed visibility into column-to-column data flows and transformations. This detail maps each column's heritage from originating sources to any target within a specific process or report, and highlights dependencies within all the layers and transformation of the reporting system (physical, semantic, and presentation).
- **End-to-End Column Lineage**: Provides details about between-system dependencies for specific columns. This visualizes the sequence of processes applied to a collection of source data elements and shows the lineage of column relationships across the pipeline landscape from entry point into the BI landscape all the way through to final destinations, even if the column changes its name along the data flow from system to system and from source to target.

The presentation of comprehensive lineage must align naturally with the expectations of the data consumers. The user interface must display lineage consistent with a step-by-step manner: for any data element in any data artifact, first survey the nearby lineage, and then incrementally expand the view along the different dimensions to inform the users but not overwhelm and confuse them with complex diagrams and maps.

The Need for Automation

Visualizing the different dimensions of data lineage empower the data consumers and the data practitioners. However, the challenge lies in accumulating, collating, and managing the information that is necessary to produce a coherent view of the different dimensions of data lineage. Earlier manual attempts at capturing data lineage were hampered by two critical barriers to success: comprehensiveness and consistency.

Comprehensive data lineage must be informed by all types of data integration, business intelligence, reporting, and presentation tools and technologies. That suggests, though, that comprehensiveness is bound to be a challenge – as the enterprise data landscape continues to expand, the effort required to survey the breadth of all critical collections of data pipelines and data flows is beyond the abilities of a small team manually reviewing code, data streams, and data element metadata. And even if that team were able to assemble a reasonable snapshot of the corporate data lineage, the dynamic nature of continuous application development and deployment suggests that the snapshot will rapidly become out of date and inconsistent.

Manual attempts at capturing data lineage are bound to be insufficient – it is not feasible to attempt to manually update and manage the different dimensions of data lineage. Only an automated mechanism can surface all the lineage relationships, maintain the representations of that lineage, and be able to visualize the different dimensions.

Automated data lineage discovery addresses both challenges. An automated discovery process surveys the data integration, ETL, data preparation, report-building, end-user query and reporting, and data visualization processes and the associated code and data artifacts. And suffice it to say, for comprehensive data lineage to provide a truly holistic view of the way information flows across the environments, the automated discovery processes must maintain an agnostic perspective about specific methods of data integration, preparation, and delivery. In other words, all the dimensions of data lineage types must be captured through visibility into all the main BI and reporting ecosystem vendor products. Not only that, the discovery process must provide visibility across the hybrid platform composed of both on-premises and cloud-based data sources, data pipelines, data delivery, and report/BI presentation and visualization tools.

That discovery process will register the relevant metadata, assess in-process and cross-process data dependencies, document the transformations applied, and chronicle the collections of data flows. When performed on a regular cadence, automated discovery will assemble a complete, comprehensive, and up-to-date data lineage map documenting the different dimensional perspectives.

Considerations: What to Look for in a Data Lineage Solution

Finally, it is important to differentiate between conventional approaches to data lineage that provide limited visibility into the various data flows from source to reports or end-user applications and a multilayered data lineage solution that provides a comprehensive perspective on data lineage across the enterprise's data ecosystem.

When considering and comparing data lineage technologies, look for a complete solution supporting the development, management, and visual presentation of data lineage mappings. Consider these six key questions when comparing tools:

- **Breadth**: Does the tool provide all perspectives of data lineage showing the flow of information across system boundaries, column-to-column lineage between systems, as well as the logic and data flow that composes the columns within a report or analytical artifact?
- **Integration**: Can the tool provide a viewpoint of how cross-system and inner-system data lineage results in integrated end-to-end mappings on a column-by-column basis?
- **Hybrid support**: Is the tool capable of surveying and capturing the lineage of data pipelines that cross on-premises and cloud boundaries and still provide a holistic view of all the dimensions of data lineage?
- **Automation**: Does the tool automatically survey data integration, ETL, and report-building processes, data integration, reporting, and BI tools, and the associated code and data artifacts data streams, and data element metadata to accumulate the information required for comprehensive and consistent data lineage?
- **Coherence and timeliness**: Does the tool refresh its view of the data lineage on a regular basis to provide the most consistent and up-to-date view of the different dimensions?
- **Visualization**: Does the tool have a user interface designed to visualize the different dimensions of data lineage that lets the data consumer switch between the different dimensional perspectives and drill down and across accordingly? Does the tool present information in a way consistent with how the data consumers can absorb the scope of the different dimensions of data lineage?

The answers to these questions will not only enable you to differentiate between conventional data lineage tools and those with bleeding-edge capabilities, but it will also help distinguish those multilayered data lineage solutions that empower the different data consumers to amplify information value.