October/2018

# Advanced Data Lineage:

# The #1 Key to Removing the Chaos

# in Modern Analytical Environments

Claudia Imhoff, Ph.D.

Sponsored By:

OCTOPAI

# Table of Contents

# Executive Summary

**"Any enterprise CEO really ought to be able to ask a question that involves connecting data across the organization, to be able to run a company effectively, and especially to be able to respond to unexpected events. Most organizations are missing this ability to connect all the data together…"[1]**

We have built BI and analytics environments for almost 3 decades. Implementation teams have gotten very good at gathering data, integrating it, keeping it in specialized storage technologies, and making it accessible to specific analytically-inclined personnel. So why is it so difficult for companies to get huge benefits from all these analytical capabilities? The answer is that today's enterprises have *several* of these environments, making searching for and analyzing data across these many instances a difficult, if not impossible, task.

The key solution is a comprehensive, easily created and accessed collection of metadata – an overarching "brain" that describes all aspects of the data found in these analytical stores, giving all users a comprehensive understanding of where the data resides, along with all its history. As a level set, most people define metadata as "data about data", but it is so much more than that! Donna Burbank, Managing Director of Global Data Strategies, defines metadata as "Data in Context – the Who, What, Where, Why, When, and How of Data."

This paper focuses on an important component of metadata, advanced data lineage. We discuss the many use cases for data lineage and stress the characteristics one should look for in mature data lineage technologies. The conclusion describes how organizations should begin their journey into solving their chaotic analytics environment by choosing a modern metadata management technology.

# Data Lineage Introduction

We are fortunate to have many articles, white papers, and treatises on the topic of metadata and data lineage in particular. This paper does not go over that well-worn turf, but instead refers readers who need

[1] Tim Berners-Lee – inventor of the World Wide Web

foundational information on data lineage to a blog written by the sponsor of this paper.[2]

Briefly, data lineage is defined as the journey data takes as it moves from its originating data source to the ultimate BI and analytic products created from it. This "lifecycle" consists of both horizontal and vertical lineage.

- *Horizontal data lineage* documents where the data came from, what happened to it in terms of data transformations as it traveled from its source to its ultimate data stores, what views and joins use it, to finally, what reports, visualizations, or analyses use it. Broadly, it refers to the system-to-system lineage of the data used for BI and analytics. (Figure 1: Horizontal Data Lineage from Source to Targets)

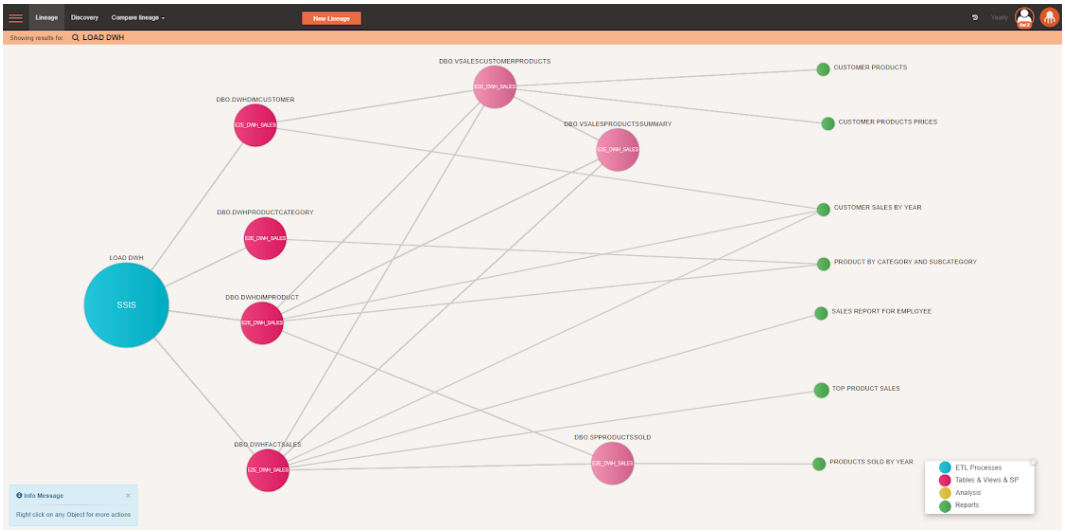


Figure 1: Horizontal Data Lineage from Source to Targets (example from Octopai)

Without a common store of horizontal data lineage, developers, analysts, data scientists, and others must repeatedly recreate or re-engineer their own horizontal data lineage information before they can be comfortable with using the data it describes.

- *Vertical data lineage* describes the individual ETL processes and analytic products themselves to provide an understanding of how each was created along with cross-relational impact analysis. Again broadly, it provides the column-to-column lineage within ETL and reporting systems. (Figure 2: Vertical Data Lineage Understanding the Content of Columns of Data)

[2] For a good primer on data lineage, please read https://www.octopai.com/what-is-data-lineage/

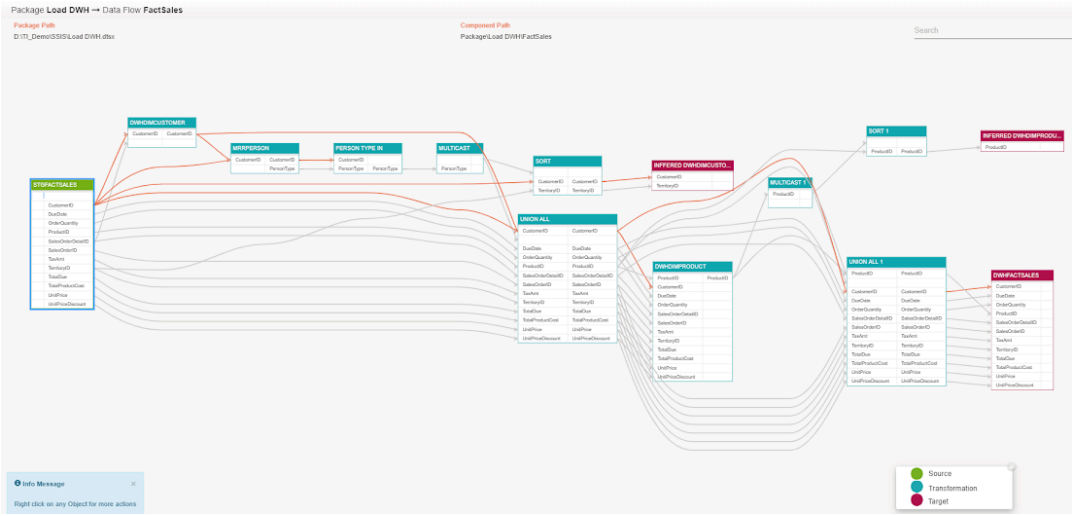


Figure 2: Vertical Data Lineage Understanding the Content of Columns of Data (example from Octopai)

Without a common store of vertical data lineage, those using BI and analytic tools cannot determine whether the products are suitable for their purposes, unless they go on the "hunt" for it. This is such a waste of time for both of these very critical resources. Vertical data lineage eliminates this wasted effort by quickly and clearly answering questions like "What is the source of an attribute in my report?", "How was this KPI calculated?", "Why do these two 'identical' fields have different values?"

Fortunately, metadata is what makes data lineage available, and demand for both is on the rise. For more clarification, let's look at the popular ways that advanced data lineage is being used today.

## Advanced Data Lineage Use Cases

Data lineage started as a means to simply understand why two reports of similar activities did not match. The ability to detect errors and then trace the error back to its source certainly gave data lineage a front row seat at the design table for most forward-thinking data analysts, BI developers, data architects, and data scientists. In today's modern analytics environment, advanced data lineage has become a mandatory component. The critical need for this advanced data lineage is best demonstrated by these seven use cases.

1. **Determine the impacts of changes on analytical environments**. All analytical environments constantly evolve; they are in a constant cycle of new development, testing and deployment. That is a good thing, but it also can cause severe problems for downstream consumers or producers of the analytical results. Changes that occur without thought as to their effect on downstream reports, analytics, and visualizations run the risk of breaking the system or, worse, changing the meaning of a data attribute or calculation with no notice to the consumers or producers of that data or result. Obviously, this can lead to erroneous or misleading outcomes. Horizontal data lineage is mandatory to eliminate these problems.

2. **Accelerate the process of mergers and acquisitions.** Companies acquiring or merging with another entity struggle to determine the true value of proposed transaction. Without well-constructed horizontal and vertical data lineage, it takes massive effort and risks serious errors in calculations to answer critical questions like: "How many joint customers do we have?" "What are projected combined P&L and Balance Sheet values?" "How accurate are predictions for growth and market share for each company?" Being able to study the lineage of the data speeds up the overall process – the analysts can quickly determine what data they need, where to find it, and how "reliable" it is for the crucial calculations. Companies can then base the soundness of these business opportunities with much higher accuracy and success.

3. **Discovery of data, reports, and analyses needed by the business community.** Another big benefit to consumers and producers using analytics environments comes from their ability to rapidly find the data and analytical results they need for their business decision-making. Unfortunately, many business people have a very difficult time locating the data, analytic, or visualization, and they have an even harder time confirming its appropriateness for their usage, determining the access mechanism, even getting approval for access. Vertical data lineage provides the information needed to quickly improve the productivity of these valuable resources. This increase in their utilization of the analytics assets is of huge benefit to the organization.

4. **Support data governance.** Data governance initiatives have been started many times in organizations only to falter due to a lack of technological support. However, the need for data governance in analytic environments has never lessened; in fact, it is needed more than ever due to the collection of unusual sources and increasing volumes of data now being analyzed. Fortunately, there have been

great advances not only in data lineage but in metadata management in general that successfully support all facets of data governance. Horizontal data lineage supplies the information that supports the governance of data as it moves through the system. And vertical data lineage provides information about where governed data should reside, who should have access to it, and how it relates to other sets of data.

5. **Reduce duplication of data and analytics.** Quickly finding appropriate data and analytical assets is a significant time-saver, but, perhaps more importantly, advanced data lineage (both horizontal and vertical) can reduce the likelihood of creating redundant reports, analytics, dashboard components, etc. Discovering that something you need already exists eliminates the risk of creating something over and over, wasting the valuable analyst time and cluttering up the environment unnecessarily. Reducing unnecessary, redundant, and possibly erroneous analytical components decreases the maintenance overhead and streamlines a complex set of processes.

6. **Determine the data flows needed for new reports and analyses.** Perhaps one of the more exciting new uses of data lineage information is its ability to give analysts and developers a fast and complete definition of what data, data feeds, data repositories, data views, and existing analytics are available to create new analytical components. Automating these data sources greatly reduces the time it takes to create these new components while ensuring the appropriateness of the data assets being used. You can understand how both horizontal and vertical data lineage would be quite useful for these users.

7. **Supporting regulatory reporting and GDPR compliance.** Organizations have multiple stakeholders – executives, employees, customers, suppliers, even auditors – who must trust reported data and analytics. Regulatory compliance and GDPR (General Data Protection Regulation) specifically require that companies track and understand how personal data flows through business processes and applications – including analytic ones. While most companies may have business process models as part of their enterprise architecture, it is rare that they have them for their analytic environments. Horizontal data lineage can locate any privacy sensitive data quickly and track how and where it flows from data access to data integration and data quality processes on to analytic applications. This information provides a clear picture of whether a company is following all regulatory mandates in its analytical activities.

# Characteristics of Advanced Data Lineage Solutions

Now that we have seen the many ways that advanced data lineage can be used, the next logical question is what does it take to get this end-to-end component of metadata in place? For many organizations, this meant laborious manual efforts to discover, collect, store and make data lineage information available. This is also the reason many of these initiatives failed or lost support. Manual processes are not efficient, error-prone, and get out of sync with the real environments rapidly. And they are very costly.

Fortunately, we now have well thought-out and designed solutions that remove much of the manual effort with ground-breaking techniques like machine learning and artificial intelligence. Here is a list of characteristics to look for when shopping for a metadata management tool.

- Top of the list of needed characteristics is *data lineage automation*. Being able to replace most of the manual activities with a tool that can automatically and quickly track much of the data lineage is an obvious requirement. Data lineage automation and the ability to generate the visual trace or mapping of the data as it flows through the analytical environments mean the developers and business community can easily discover the data and analytical assets they need. All seven of the advanced use cases discussed above are enabled by this innovation in metadata management.
- Second on the list of must-haves is a *clean and easy-to-use interface*. The clarity and simplicity of the interface is what makes data lineage usable by all analytical developers, as well as business consumers and producers. After all, if you can't locate the data or analytical asset you need quickly, then what is the point? Such an interface allows users to locate the data quickly and see everything related to the data – every process, table, or view – as it goes through its journey from start to finish. Good visibility ensures that all users have the appropriate data for their needs, and reduces the time required to find, analyze, and fix errors and other discrepancies.
- The next characteristic for a data lineage solution is *agile and rapid deployment*. The creation of horizontal and vertical data lineage is not a simple process for many metadata management tools due to an onerous and manual collection process. Look for a solution that automates as much as possible – including metadata collection. Once extraction is completed, automation continues for the rest of

the data lineage tracking. Look for a solution that can be up and running immediately, not after a long manual process of collection is completed.

- Since most organizations have multiple analytic environments, the data lineage solution must be a *cross-platform metadata management technology*. Metadata must be collected from all analytic platforms and stored in a single, central repository and made available to everyone. An automated, cross-platform metadata management solution should be able to locate all data regardless of where it resides; it should be able to use the names of fields, any tags, and of course, the actual content of the data (both structured and variably structured data). A central metadata repository has the added benefit of improving data governance and regulatory compliance across the enterprise.
- We are currently in a time of great innovation in analytics – machine learning, artificial intelligence, data science, and so on. For modern metadata management solutions, these sophisticated techniques, especially *artificial intelligence and machine learning, must be an integral part of the solution*. Machine learning is used to discover, define relationships and dependencies within and between the analytic environments, and retrieve the metadata to establish the data lineage. This characteristic enables the automatic creation of the much-needed central repository of metadata. Among other things, machine learning creates the modeling and indexing of the metadata and determines the existing dependencies between all data elements (even if the data elements have aliases – e.g., state, state code, state ID, etc. – even across different languages).
- Finally, as good as machine learning and automation is, there may be cases where programming code or stored procedures have been used. These can be difficult for many automated metadata solutions to decipher. In assessing a metadata management technology, make sure *the solution can either analyze programs or procedures created outside of an ETL tool or they have a mechanism that allows for manual activities to "fill in the blanks"*.

# Summary

**"It's difficult to imagine the power that you're going to have when so many different sorts of data are available."[3]**

[3] Tim Berners-Lee, Inventor of the World Wide Web

Metadata has been defined as "a love note to the future." If so, then data lineage is its heart! Advanced data lineage tools, such as Octopai, make the benefits discussed in this paper feasible by giving enterprises the ability to quickly and reliably capture and document all forms of metadata, including the much-needed advanced data lineage information – both horizontal and vertical. Through their intuitive and easy-to-use interfaces, these technologies give most consumers and producers of analytical assets the access they need to metadata, regardless of their skill sets.

The bottom line is that we all must work smarter, not harder. This is only possible by implementing a sophisticated, modern metadata management solution, the real intelligence behind business intelligence.