Box 21: Advanced Data Lineage – The #1 Key to Removing the Chaos in Modern Analytical Environments, Imhoff, C 2018

Sponsored by OCTOPAI

We have built Business Intelligence (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. [The] paper focuses on an important component of metadata, advanced data lineage, [which] consists of:

Octopai Horizontal data lineage diagram

Documents where the data came from, what happened to it in terms of data transformations as it travelled 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. Without a common store of horizontal data lineage, developers, analysts, data scientists, and others must repeatedly recreate or reengineer their own horizontal data lineage information before they can be comfortable with using the data it describes.

Octopai Verticaldata lineage diagram

Describes the individual extract, transform and load (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. 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?"

Seven Use Cases

1 Changing analytical environments

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 Data from mergers and acquisitions

Accelerate the process of mergers and acquisitions. Companies acquiring or merging with another entity struggle to determine the true value of proposed transactions. 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

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 visualisation, 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 utilisation of the analytics assets is of huge benefit to the organisation.

4 Support data governance

Support data governance. Data governance initiatives have been started many times in organisations 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 analysed. 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

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 Data for new reports and analyses

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 / compliance

Supporting regulatory reporting and compliance. Organisations have multiple stakeholders – executives, employees, customers, suppliers, even auditors – who must trust reported data and analytics. Regulatory compliance {and in a European context, 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.

A special type of master data used to categorise other data or used to relate data to information beyond the boundaries of the enterprise. Reference data can be shared across master or transactional data objects (e.g. countries, currencies, …etc.).

Source: https://www.octopai.com/paper/advanced-data-lineage-1-key-removing-chaos-modern-analytical-environments.
Precis of Imhoff, c Ph.D., 'Advanced Data Lineage: The #1 Key to Removing the Chaos in Modern Analytical Environments', October 2018,