FUJITSU

The Importance of Data Lineage and the Business Benefits

DATA LINEAGE:

Data Lineage is the transparency and visibility over the journey of data as it flows throughout the business. Good data lineage provides the traceability and ability to answer where you got the data from, the transformations that happened to the data and what else has happened to the data on its journey to a report, dashboard or connected system.

HOW CAN DATA LINEAGE HELP:

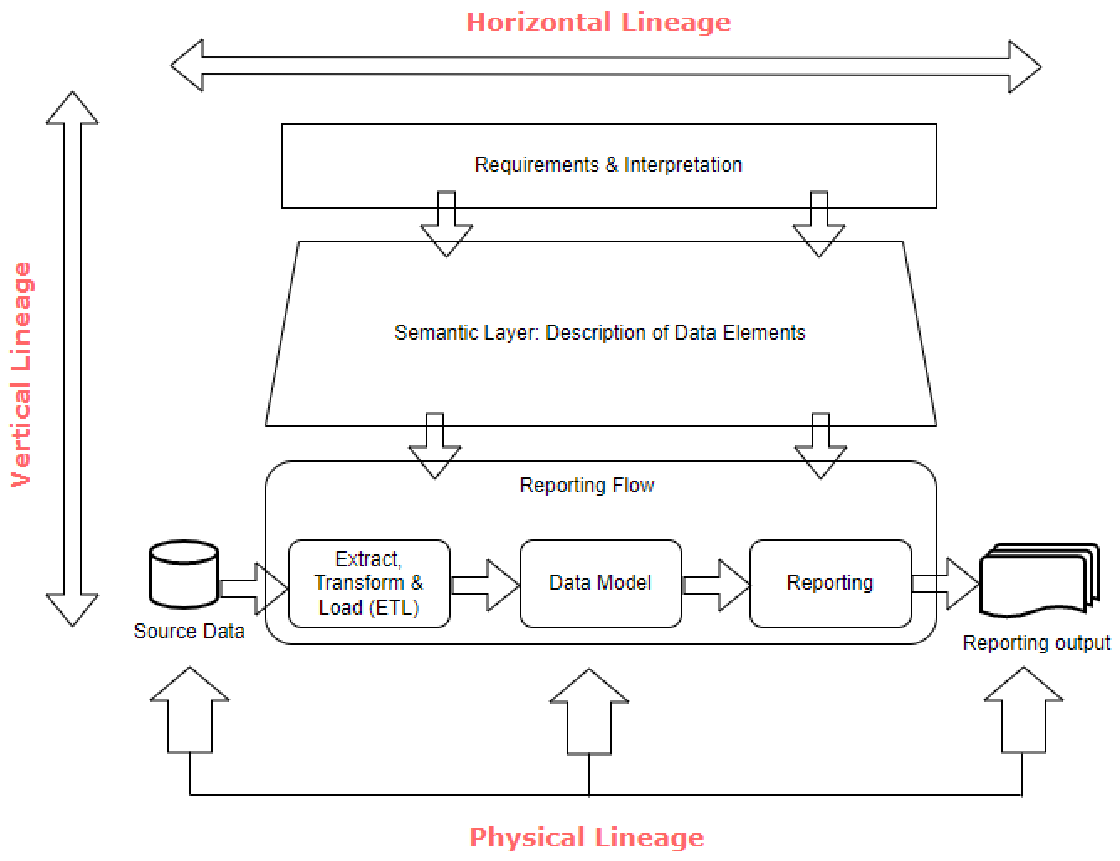
Many companies understand what data lineage is, but struggle to understand the possible uses and the value that good data lineage bring. When enriched by the knowledge of what good looks like, organisations can answer data lineage fundamentals of where, when, how, what, and why?

Good data lineage allows data professionals and consumers of data answers to questions such as:

- Why do we need this data?
- Where did the data originate?
- When was the data created (is it still accurate)?
- How did the data flow from source to target?
- What happened to the data as it moved?

Data Lineage can be split into 3 different types:

1. **Vertical lineages** demonstrate the origin of a data from regulations towards deployment in a data model on a metadata level.
2. **Horizontal lineages** show the mapping of source data to target output on a metadata level. It demonstrates the functional logic of how source data is transformed towards a target end state.
3. **Physical lineage** demonstrates the actual data flow from source system to reporting solution, supporting the metadata architecture of the data driven reporting environment.



WHY DO YOU NEED A DATA LINEAGE TOOL?

1. Enhanced data governance and regulatory compliance.

For modern enterprises that operate with big volumes of information, the governance policies and practices help control and clarify for managing sensitive data.

GDPR and other data privacy laws can be extremely hard to achieve if you do not have data lineage software to log records through all the touchpoints.

2. Improved data quality.

The interpretability of data lineage software brings higher accuracy and consistency in decision-making, and increased reliability of the data analytics to drive business operations. This also results in faster time to decisions.

3. Detailed impact analysis.

Issues with data discovered from downstream or upstream data processes to map where the error is produced. This expedites error removal and delivers faster and higher levels of data quality.

4. **Easier migrations.**

Successful migrations need to be able to understand touchpoints between different systems. Lineage tools which map all the workflows of how data travels make this easier.

USE CASES OF DATA LINEAGE:

1. **Cross-system lineage** has data flows and dependencies to provide extensive cross-system views of the entire data landscape.

o Predicting the impacts of a process change
o Analyses the broken processes
o Discovers parallel processes performing the same tasks
o High-level visualisations of data flow

2. **End-to-end column lineage** details column-to-column-level lineage between systems.

o Impact analysis of a change in a column in the source systems
o Root cause analysis to uncover the sources of reporting errors
o Column-level visualisation of data flows.

3. **Inner-System Lineage** is when you want to dive even deeper into the details of a system. Understanding the logic and data flow for each column provides visibility at the column level no matter how complex the report.

o Visualising the logic of a report, ETL, or database object data flow.
o Dependencies within a report.

DATA LINEAGE IN MODERN DATA PLATFORMS:

Data Lineage in AWS:

1. Build data lineage for data lakes using AWS Glue, Amazon Neptune, and Spline.
2. Spline agents capture runtime lineage information from Spark jobs, powered by AWS Glue.
3. Use Amazon Neptune, a purpose-built graph database optimized for storing and querying highly connected datasets, to model lineage data for analysis and visualisation.

Data Lineage in Azure:

1. Data lineage is built using Microsoft Purview Data Catalog
2. Microsoft Purview Data Catalog connects with other data processing, storage, and analytics systems to extract lineage information.
3. The information is combined to represent a generic, scenario-specific lineage experience in the Catalog.
4. Where there is data moving across multiple systems the Data Catalog can connect to each of the systems for lineage.

Data Lineage in Databricks:

1. Capture and view data lineage with Databricks Unity Catalog
2. Databricks Unity data lineage supports all languages and is captured down to the column level.
3. The lineage can be visualised in Data Explorer in near real-time and retrieved with the Databricks REST API.
4. In this scenario lineage is aggregated across all workspaces attached to a Unity Catalog metastore.
5. Data lineage captured in one workspace is visible in any other workspace that is sharing that metastore.

Data Lineage Tools:

This section discusses metadata management and data lineage specialists, such as Alex Solutions, Manta and Octopai; and vendors of data Catalog tools, such as Alation, Atlan, Data.world and OvalEdge.

By using the above tools, the metadata is maintained and triggered automatically from the data, so you do not have to do any additional work.

These tools also track and audits along the data flow such as data governance policies, data stewardship, master data management (MDM), reference data management (RDM), etc.

These tools can be used to map the data flows and the relationships, and dependencies between different data elements both forward and backward and work particular well alongside open-source platforms such as Databricks.

Conclusion:

Transparency and visualisation of data flows enables reporting specialists, business analysts and data consumers to work together and understand each other's task. It enables an efficient "search for data" in any organisation, resulting in cost reductions and a shorter lead time of projects, change requests and investigations.

When implemented and used correctly, data lineage capabilities and tools enhance the control on the data transformation and reporting processes. Which in turn results in improved quality of the data and reports.

If you need help implementing data lineage tools within your organisation then please contact our Fujitsu Data & AI specialist now.

Contact

Fujitsu Data & AI
+61 3 9924 3000