DATA LINEAGE

What is Data Lineage?

Data lineage uncovers the life cycle of data—it aims to show the complete data flow, from start to finish.

Data lineage means not only being able to trace the source of data elements, but also charting their journey through an enterprise as they are aggregated, transformed, processed and analysed by various applications and reporting tools. And it is important to note that this data journey can and should be viewed at different levels of granularity in order to provide different information to different audiences.

Two important levels of granularity are:

• Summary business lineage, which shows a summary view of how data used by business users is transformed and aggregated as it flows from source to endpoint, such as a report

• Detailed technical lineage, which tracks data flows and transformations at the lowest levels of granularity, by table, columns, and queries

Data lineage needs to be solved at both of these levels of granularity. Solutions that focus exclusively on technical lineage risk missing key business benefits and vice versa.

~ Data lineage is the process of understanding, recording, and visualizing data as it flows from data sources to consumption. This includes all transformations the data underwent along the way—how the data was transformed, what changed, and why.

Data lineage allows companies to:

* Track errors in data processes
* Implement process changes with lower risk
* Perform system migrations with confidence
* Combine data discovery with a comprehensive view of metadata, to create a data mapping framework

Data lineage helps users make sure their data is coming from a trusted source, has been transformed correctly, and loaded to the specified location. Data lineage plays an important role when strategic decisions rely on accurate information. If data processes aren’t tracked correctly, data becomes almost impossible, or at least very costly and time-consuming, to verify.

Data lineage focuses on validating data accuracy and consistency, by allowing users to search upstream and downstream, from source to destination, to discover anomalies and correct them.

Why is Data Lineage Important?

Data lineage may initially seem to support technical use cases. It enables organizations to ascertain the physical source of data elements and track the flow of data as it is processed by various systems. This technical capability has clear benefits to data engineers, data architects and technology teams, helping them to understand dependencies, usage patterns and determine the impact of changes to an organization’s data architecture.

Just knowing the source of a particular data set is not always enough to understand its importance, perform error resolution, understand process changes, and perform system migrations and updates.

Knowing who made the change, how it was updated, and the process used, improves data quality. It allows data custodians to ensure the integrity and confidentiality of data is protected throughout its lifecycle.

Data lineage can have a large impact in the following areas:

Strategic reliance on data — good data keeps businesses running. All departments, including marketing, manufacturing, management and sales, rely on data. Information gathered from research, from the field, and from operational systems helps optimize organizational systems improve products and services. Detailed information provided through data lineage helps better understand the meaning and validity of this data.

Data in flux — data changes over time. New methods of collecting and accumulating data must be combined and analysed, and used by management to create business value. Data lineage provides tracking capabilities that make it possible to reconcile and make the best use of old and new datasets.

Data migrations — when IT needs to move data to new storage equipment or new software systems, they need to understand the location and lifecycle of data sources. Data lineage provides this information quickly and easily, making migration projects easier and less risky.

Data governance — the details tracked in data lineage are a good way to provide compliance auditing, improve risk management, and ensure data is stored and processed in line with organizational policies and regulatory standards.

Data Lineage and Data Classification

Data classification is the process of classifying data into categories based on user-configured characteristics.

Data classification is an important part of an information security and compliance program, especially when organizations store large amounts of data. It provides a solid foundation for data security strategies by helping understand where sensitive and regulated data is stored, both locally and in the cloud.

In addition, data classification can improve user productivity and decision making, remove unnecessary data, and reduce storage and maintenance costs.

Data classification is especially powerful when combined with data lineage:

Data classification helps locate data that is sensitive, confidential, business-critical, or subject to compliance requirements.

For each dataset of this nature, data lineage tools can be used to investigate its complete lifecycle, discover integrity and security issues, and resolve them.

Data Lineage Techniques and Examples -

Here are a few common techniques used to perform data lineage on strategic datasets.

Pattern-Based Lineage -

This technique performs lineage without dealing with the code used to generate or transform the data. It involves evaluation of metadata for tables, columns, and business reports. Using this metadata, it investigates lineage by looking for patterns. For example, if two datasets contain a column with a similar name and very data values, it is very likely that this is the same data in two stages of its lifecycle. Those two columns are then linked together in a data lineage chart.

The major advantage of pattern-based lineage is that it only monitors data, not data processing algorithms, and so it is technology agnostic. It can be used in the same way across any database technology, whether it is Oracle, MySQL, or Spark.

The downside is that this method is not always accurate. In some cases, it can miss connections between datasets, especially if the data processing logic is hidden in the programming code and is not apparent in human-readable metadata.

Lineage by Data Tagging -

This technique is based on the assumption that a transformation engine tags or marks data in some way. In order to discover lineage, it tracks the tag from start to finish. This method is only effective if you have a consistent transformation tool that controls all data movement, and you are aware of the tagging structure used by the tool.

Even if such a tool exists, lineage via data tagging cannot be applied to any data generated or transformed without the tool. In that sense, it is only suitable for performing data lineage on closed data systems.

Self-Contained Lineage -

Some organizations have a data environment that provides storage, processing logic, and master data management (MDM) for central control over metadata. In many cases, these environments contain a data lake that stores all data in all stages of its lifecycle.

This type of self-contained system can inherently provide lineage, without the need for external tools. However, as with the data tagging approach, lineage will be unaware of anything that happens outside this controlled environment.

Lineage by Parsing -

This is the most advanced form of lineage, which relies on automatically reading logic used to process data. This technique reverse engineers data transformation logic to perform comprehensive, end-to-end tracing.

This solution is complex to deploy because it needs to understand all the programming languages and tools used to transform and move the data. This might include extract-transform-load (ETL) logic, SQL-based solutions, JAVA solutions, legacy data formats, XML based solutions, and so on.

Data Lineage for Data Processing, Ingestion, and Querying

When building a data linkage system, you need to keep track of every process in the system that transforms or processes the data. Data needs to be mapped at each stage of data transformation. You need to keep track of tables, views, columns, and reports across databases and ETL jobs.

To facilitate this, collect metadata from each step, and store it in a metadata repository that can be used for lineage analysis.

Here is how lineage is performed across different stages of the data pipeline:

Data ingestion — tracking data flow within data ingestion jobs, and checking for errors in data transfer or mapping between source and destination systems.

Data processing — tracking specific operations performed on the data and their results. For example, the data system reads a text file, applies a filter, counts values from a specific column, and writes to another table. Each data processing stage is analysed separately to identify errors or security/compliance violations.

Query history — tracking user queries or automated reports generated from systems like databases and data warehouses. Users may perform operations like filters, joins, and so on, effectively creating new datasets. This makes it critical to perform data lineage on important queries and reports, to validate the process data goes through. Lineage data can also help users optimize their queries.

Data lakes — tracking user access to different types of objects, or different data fields, and identifying security or governance issues. These issues can be complex to enforce in large data lakes due to the huge amount of unstructured data.

Benefits offered by data lineage -

1. Speeding time to insight - Data lineage can speed time to insights by addressing doubts over data quality.

2. Improving data quality - Maintaining current and accurate data lineage fosters better data quality. It helps ensure all records can be corrected at the source more efficiently. It can also identify problems where source data may be accurate, but an error was introduced in the way the data was processed.

3. Complying with relevant regulations - Data lineage is key to supporting a variety of regulatory obligations.

4. System rationalization - . Rationalizing and harmonizing systems is a perennial project for most organizations — one that helps reduce costs and improve operational efficiency. Maintaining an accurate map of data lineage is crucial in supporting rationalization projects

5. Cloud migration - The economic benefits of cloud services, not only in terms of cost but also the ability to innovate more quickly, has led many organizations to adopt cloud-first IT strategies.

6. Asset management - Asset management can be viewed as a by-product of cloud migration and rationalization projects.

7. Impact analysis / Change management - Technology and data architectures can be strewn with complex interdependencies. When it comes to making changes to systems, the knock-on implications can be very difficult to model. What may seem like a trivial change from the perspective of a database administrator could trigger the failure of a key business process. The only way to mitigate against such risks is by maintaining detailed and accurate data lineage. Knowing how data flows through an organization is crucial when it comes to managing the impact of change.

Data Lineage Tools - Alation, Atlan, Collibra, CloverDX, Datameer, Dremio, Immuta, Informatica, Keboola, MANTA, Octopai, OvalEdge, Talend, Trifacta.

Resources -

<https://www.imperva.com/learn/data-security/data-lineage/#:~:text=Data%20lineage%20is%20the%20process,Data%20lineage%20process>

<https://www.collibra.com/content/dam/documents/en/resources/Unlocking-business-opportunities-through-data-lineage.pdf>