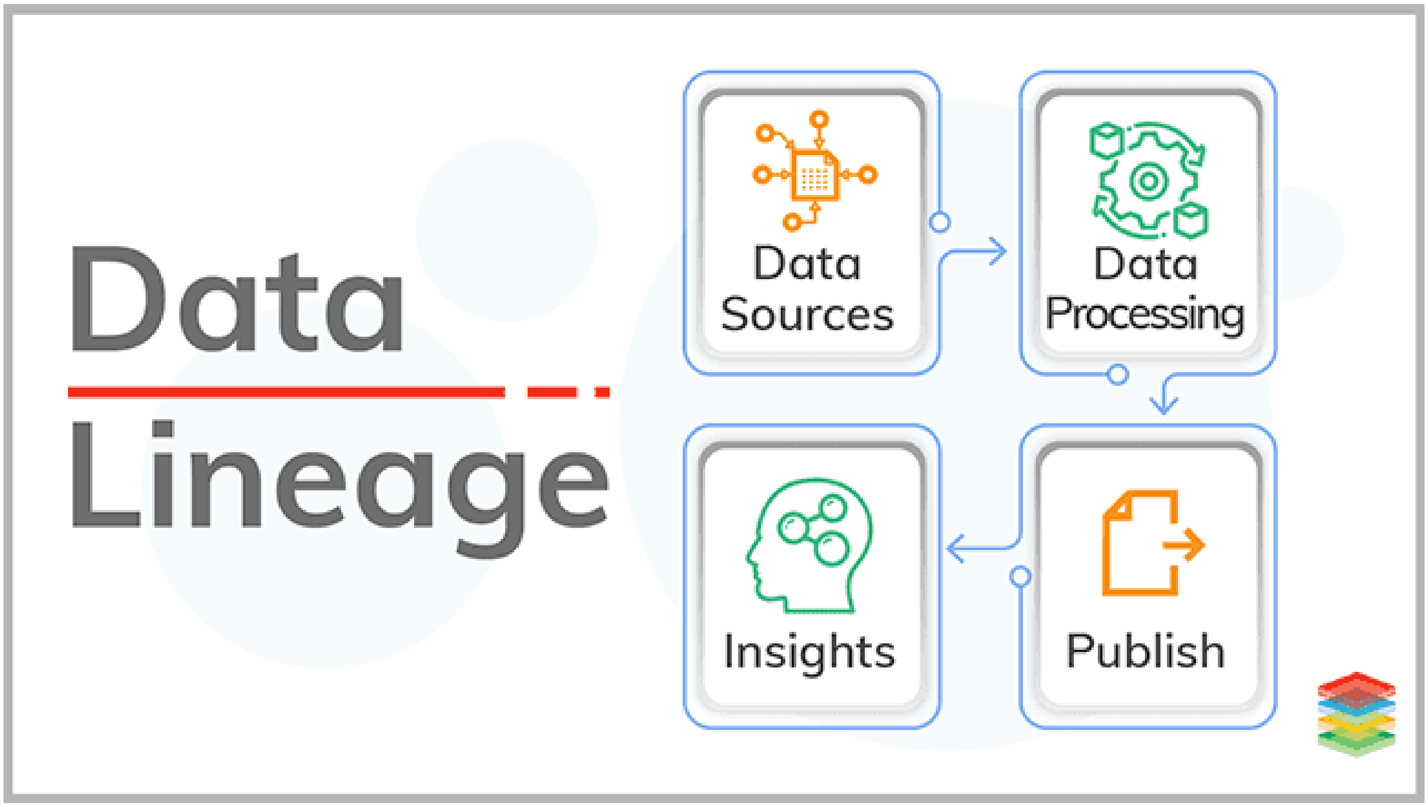
# What is Data Lineage, Best Practices and Techniques

Insights | 6 mins read | Apr 13, 2019



## What is Data Lineage?

Data Lineage is defined as the life cycle of the data. Data Lineage shows the complete data flow from origin to destination. Data lineage is the process of understanding, documenting, and visualizing the data from its origin to its consumption. This life cycle includes all the transformation done on the dataset from its origin to destination. Data lineage gives a better understanding to the user of what happened to the data throughout the life cycle also.

Data Lineage helps the user to make sure if the data is coming from a reliable data source, transformations are done appropriately and loaded correctly to the designated location. Data Lineage plays an important role where key decisions rely on accurate information. Without appropriate technology and processes in place tracking, data can be virtually impossible or at the very least a costly and time-consuming endeavor.

Data lineage enables the tracking of the data stream from both endpoints to ensure the data is accurate and consistent. It allows the user to look for the data in both directions (forward and backward) between origin to destination of the data.

Data Lineage provides us the answers for any specific dataset such as:

- Who created the data?
- What information does the data contain?
- Where is the data located?
- When was the data created?
- Why does the data exist?

We will discuss these questions in a later section.

## Why We Need Data Lineage

### For an ETL Developer

ETL stands for Extract, Transform, and Load. ETL job is a function where we need to extract data from any defined data source and put it into another location after applying some data transformation on the collected data.

### For a Data Steward

To play the role of a data steward, the person needs to know everything about the data which is being used in an organization. Data lineage helps the person to identify the least and most usable data assets in an ETL job. Data lineage provides transparency to the user who is responsible for that particular data asset.

### For a Business User

Data lineage helps a business user to find the reports based on any particular data fields or column. Example: there is some data source that includes data fields named sales and gender if the user needs to find the reports of the bases of these data fields. Data Lineage can help the business user to check whether the data is accurate or not.

### For a Troubleshooting Operator

When we need to troubleshoot for any of the wrong reports, lineage can help us to identify which process and jobs are involved in creating that particular report. In the case when we have some failed jobs, data lineage can help us to find the target tables and fields affected which are being used in the reports.

**Data Lineage is a data intelligence technique that makes it possible to determine the origin or lineage of a data object.**

Source- *Data Lineage: The What, Why & How*

## 5 W's of Data Lineage

### When the data is created/updated?

There is also some parameter which needs to define at the time of data creation. The data owner has the responsibility to store the data into the appropriate location and to grant access to the data. Know the owner of data is most important as it gives clarity that who is maintaining that data and to whom the user should contact in case of any problem with the correction.

### What information does it contain?

We always need to define some access policies to the data. And before that, it is also necessary to understand what information does the data contains. It helps in classification the data so that we can understand which data policies need to define against the data so that we can protect our sensitive data.

### How is it being used?

In an organization, the data is used to create several reports. These reports are used to make decisions for the growth of the organization. These reports are created by using several datasets that are generated within the organization. The data lineage diagram can show us which datasets are being used. So in case if we got some wrong reports this can help us to trace the source of the error if we have any.

### Why is it stored/used?

There is one more important question for the existence of data. Why does this data exist? This is one of the most important questions because if we don’t need any data it should be deleted. The data which is no longer required can lead to unnecessary time and money. So we should know about every dataset which is stored.

To capture the data lineage we need to collect the metadata after each of the data transformations. So metadata on each stage is collected and stored in the metadata store which can be used for lineage representation.

### Data Ingestion Lineage

Data Ingestion lineage can be used to track the complete data flow within the Data Ingestion Job. It can also be used for tracing any bug/error within the Data Ingestion job.

**Below we are going to discuss the data lineage of Apache NiFi by using Apache Atlas.**

Apache NiFi is a UI based platform where we need to define our source from where we want to collect data, processors for the conversion of the data, a destination where we want to store the data. Apache Atlas is the governance and metadata framework for Hadoop which can be used for data lineage. Apache NiFi also has a controller that can be configured to push metadata of the data flow to the Apache Atlas.

### Data Processing Lineage

Spark is very popular nowadays for Distributed Processing of Data. So, When we are working with the Apache Spark Lineage, the only thing which matters is RDDs. In spark, existing RDDs point towards their parent RDDs. Consider a simple job:

- First RDD: When we read a text file and make an RDD.
- Second RDD: When we apply map operation on the first RDD.
- Third RDD: When we apply filter operation on the second RDD.
- Fourth RDD: When we apply count operation on the third RDD.

### Query History Lineage

When users are querying Data Warehouse, Then they might keep on applying filters or joining the tables, etc. So, Query Lineage also becomes very necessary so that Data Engineers can observe what are the most frequent filters, joins used and They can accordingly optimize their partitioning keys or denormalize the tables, etc, and Other optimizations as well. Example: Uber Query Parser

### Data Lake and Warehouse Access Lineage

When proper Data Governance is applied on Data Lake and Data Warehouse like RBACs, Row Level & Column Level Permissions, Then Query Lineage along with MetaStore logs can help to visualize if some user is trying to access non-authorized data and accordingly administration team can take action on it. Example: Apache Atlas, Cloudera Navigator.

To know more about best data lineage consulting, Get in touch with us.

## Leave a Comment

Name *

Name

Email Address *

Email address

Comment *