Module 5: Data Lineage

DAT-375

Brandon Hobbs

March 27, 2023

Data lineage according to Gaur (2022), “…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”. This process is important because “…[without] data lineage, big data becomes synonymous with the last phrase in a game of telephone. The original data from the first person (e.g., ‘a guppy swims in a shark tank) changes to something completely different when it ends with the last person (e.g., ‘The puppy that spins and barks, stank’)” (Knight, 2023).

**How is data lineage established?**

To establish data lineage, document the *where* and *how* of your data. The first step would be to identify the source, the systems generating, and storing the data. From here a data map could be generated – which is a visual representation of how the data flows through the system’s pieces. Highlight any transformations used as the data moves through the system, e.g., birthdate entered as a string and then converted to a DateTime for storage. These transformations may be purposeful or accidental, e.g., float to integer conversions, both should be analyzed and understood.

Once the analysis of transformations and the data map is built this needs to be documented for future reference. This reference can establish how things are actually done, i.e., root cause analysis after a failure, or how they were done at some point, i.e., regression testing.

At each of the transformation/storage steps it is important to establish who is using the data, when was the data created/modified, what information is available. At the data steward level, it would also be important to know how the data is being used and why is it being stored (particularly important for PII or sensitive data).

Finally, once this documentation is established data should be monitored through the pipeline. This can help to alert, in real-time, when data is modified or moved. All of this should then be logged.

**Why should we establish data lineage**?

As already mentioned, lineage helps to establish data provenance (where it comes from and the methods with which it was produced). Joshi (2021) says lineage is important because it can “…help you fix issues or perform system migrations, it also enables you to ensure the confidentiality and integrity of data by tracking changes, how they were performed, and who made them”.

It may not always be appropriate for all systems dealing with data. For example:

* In small teams or where resources are limited this could cause an undue burden on the team
* The data is low risk and small mistakes, or inaccuracies will have limited impact on the business
* The data is temporary, e.g., data collected for a one-time analysis

Lineage should be established as early as feasible – ideally from the moment data starts to be created, stored, and used by an organization.

**How can data lineage validate data integrity?**

Once lineage has been established it can be used in a few ways to ensure integrity. Firstly, it can help in finding root causes if an error is detected. Engineers should be able to use the lineage to trace the data back to the source and determine where the error occurred, e.g., integer overflow at storage or conversion of a float to an integer at the dashboard level.

Any potential error or planned change can have their impact analyzed. That is, if during lineage determination data consumers are identified, where they are collecting data, and how they use it disruptions can be planned or mitigated.

During a system upgrade or during migration ensuring the data is still intact or that data flows are re-established will be easier if documentation is available.

Documenting how “things actually work and not how we think they work” is always invaluable. During the data mapping exercise, it may be found that the same dataset is being stored multiple times with slight modifications. This is increasing the cost to the company for storage and I/O and could introduce some stale data to teams if they are pulling from a different databse.

Lastly, some data is controlled by law. Any business operating the EU needs to be compliant with GDPR and would need to survive any audits, e.g., data security, privacy, or retention.

**Ah oh, you didn’t establish lineage, what might happen**?

So, now the organization is creating, storing, and using data but lineage isn’t understood, and provenance cannot be ensured – now what? This could cause loss of data due to “due to technical issues, security breaches or simply human error”.

Decision making could be negatively impacted. As mentioned before, the data could be stale because the dashboard is pulling from the wrong database. Or two different teams are using “the same data” but from different sources that are not identical.

Data quality could suffer as errors and inconsistencies may go unnoticed. This could lead to inefficient data analysis as consumers may feel the need to validate the data at each intermediate step instead of turning the data into knowledge.

Federal agencies could levy fines or impose sanctions if legal compliance is not adequate, e.g., GDPR.

**Citations:**

EarlyFriday. (2023, February 23). *How correct Data Lineage can save your life (or at least your business).* LinkedIn. https://www.linkedin.com/pulse/how-correct-data-lineage-can-save-your-life-least-business

Gaur, C. (2022, October 26). *Data lineage best practices and techniques | compete guide*. Xenonstack Inc. Retrieved from https://www.xenonstack.com/insights/data-lineage

Joshi, S. (2021, September 28). *What is data lineage? Why it's important to track data flow*. https://learn.g2.com/data-lineage

Knight, M. (2023, January 5). *Data lineage demystified*. DATAVERSITY. https://www.dataversity.net/data-lineage-demystified/#