**Data Lineage**

Camilo Hoyos

DAT-375 Data Analysis Techniques

Jacqueline U. Clauss, M.S.

February 7, 2021

Data lineage is a document that describes many explicit details of data including the 5 W’s, the source of data, data enrichment, the characteristics of data, and the target of data. (*Data Lineage Overview, Best Practices and Techniques* 2019) By having this document several stakeholders who need to track and understand the data have a single source of truth for review and analysis. This also ensures that whenever departments have to communicate with one another there is a source of truth to discuss and elaborate. You know you have the correct lineage when it can be validated against the first entry point, likely an ETL job, the data warehouse, such as a database, and the exit point, such as another ETL job or an API. If the document is aligned with the code then the lineage is accurate.

Data lineages serve as a control document and currently have no valid reasons for not establishing the document. Data lineage can provide data governance. By understanding the 5 W’s a proper data governance team can build processes around ingestion, modification, such as enrichment, hashing, and masking, and as well as extraction. This can ensure data is protected and only accessed by authorized individuals. Additionally, data lineage can assist with audit or compliance. Software can be audited for its black-box processing and having a data lineage document can provide auditors with the evidence they need to ensure proper controls are enacted on data. Lastly, data quality can be ensured. In many cases data comes from third parties that aggregate and transform data. When the data is ingested it is not possible to assume the data as ingested is of total accuracy and completeness. By documenting source data, how it is enriched, aggregated, or segregated ensures issues with the data set can be evaluated and resolved. For instance, if two sources of data utilize a primary key but with different meanings, the data lineage would document that they should be not used to join the data. (Knight, 2017) The best possible scenario to track data lineage is upon its creation. In some cases it may not be possible to get all detail, especially if it is third party data, but tracking as much of it at the point of generation is paramount. On the other side of lineage you should do everything possible to track lineage outside of your own software as well if you provide the data as a product. Ensuring you understand how your clients use data ensures you will not disrupt your product whenever changes are made.

Talend, a leading data software company, defines data integrity as “... the overall accuracy, completeness, and consistency of data. Data integrity also refers to the safety of data in regards to regulatory compliance — such as GDPR compliance — and security. It is maintained by a collection of processes, rules, and standards implemented during the design phase.” (*What is Data Integrity and Why Is It Important? - Talend* 2020) Data lineage validates accuracy, completeness, and consistency through a few means. First, it is a process to execute and maintain data lineage, by having an individual responsible for updating the document this individual becomes the officer of data. They can alert all stakeholders if there is inconsistency in any transformations or enrichment through documentation. The data lineage document also defines what the data is and its standards such as length and data type. (Knight, 2017) If any other departments need to utilize data they know what to expect. Additionally, by documenting any enrichment processes anyone who utilizes the data lineage will understand any up-stream process that enhances the accuracy or completeness of the data set.

Not establishing a data lineage can have severe consequences. Whenever teams have to collaborate and discuss data it is possible to miscommunicate data processes or simply not even know what occurs outside of the teams’ immediate knowledge. In this lack of clarity, improper use of data attributes may affect data quality, and improper use of enrichment will in some cases make the data downright unusable. Without establishing a lineage there is too much risk in damaging data quality, integrity, and may simply invalidate the data for business use.

References

Data Lineage Overview, Best Practices and Techniques. (2019, April 13). Retrieved January 29, 2021, from https://www.xenonstack.com/insights/data-lineage/

Knight, M. (2017, April 20). Data Lineage Demystified: The What, Why, and How. Retrieved January 29, 2021, from https://www.dataversity.net/data-lineage-demystified/

What is Data Integrity and Why Is It Important? - Talend. (2020, August 10). Retrieved January 29, 2021, from https://www.talend.com/resources/what-is-data-integrity/