Data lineage

Definition

Data Lineage is metadata that identifies the sources of data and the transformations through which it has passed up to the point of applying. (DAMA-NL, 2020)

Notes 1

Other definitions:

- Data Lineage is a description of the pathway from the data source to their current location and thealterations made to the data along that pathway. (Brackett 2011).
- Data lineage is the description of data movements and transformations at various abstraction levels along data chains and of the relationships between data at these levels (Steenbeek, 2023).
- Backward DL is data lineage that describes where data come from.
- Forward DL is data lineage that describes where data are used.
- Horizontal DL is what is generally seen as data lineage.
- Vertical DL is data lineage that describes the relationship between the concept data model, logical data model and application data model.

Notes 2

Data lineage answers the 5 W's of data:

1. Where does the data come from or where does it go?
2. Who uses it?
3. When was it created?
4. What information does it contain? What transformations are executed?
5. Why does it exist?

Synomym

Data chain

Purposes

- To be able to conduct an impact analysis when making changes to data structures, data flows, or data processing.
- To identify opportunities for improvement in the existing data flow.
- To be able to investigate root causes of data issues.
- To be able to determine the reliability of data, based on its origin.
- To identify personal information (GPDS).
- To support migration of applications and to identify "dead-ends".

- To be compliant with standards that require DL.
- To enable Data Lifecycle Management

Life cycle

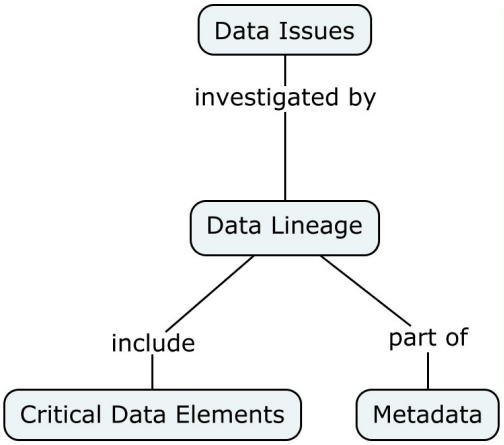
Phase	Activity
Plan	* Define the scope * Select a way to store the DL information, e.g., by an editor or in a DL tool * Collect the relevant metadata * Enter, change, or delete the metadata * “Stitch the nodes”
Do	* Use the DL for its purpose
Check	* Evaluate the effectiveness of the DL
Act	* Adapt the DL * Maintain the DL

Characteristics and requirments

Characteristic	Requirement
Completeness	DL is complete regarding the scope.
Maintainability	DL can be maintained efficiently.
Clarity	DL can be interpreted easily (zooming, filtering)

Relations

Data lineage	is supertype of	backward data lineage, forward data lineage, horizontal data lineage, and vertical data lineage
Data lineage	is an element of a	data quality management system
Data lineage	is part of the	business or technical metadata
Data lineage	includes a set of	data elements but especially critical data elements
Data lineage	is a method to assess the root cause of	data issues



Example(s)

Example 1: Horizonal data lineage

Example 2: Horizontal data lineage

Example 3: Horizontal and vertical data lineage

Story

Legislation requires the Valencia bank to report monthly to its regulator, the central bank. The regulator, however, also wants to know how these reports have been produced and where the data comes from. This is to assess the quality of the data.

Because the reports are generated by complex data flows, the bank decides to apply data lineage to map these flows and make them visible. It soon turned out that fields with the same meaning had different names in the systems involved.

Nevertheless, it was possible to link the fields and it became clear where the reported data came from. The bank can now satisfactorily inform the supervisor about the origin of the reported data. A data steward is made responsible for the maintenance of the data lineage in the tool, so that the metadata is kept up to date.

Data lineage also proves to be useful when making changes to the systems. The impact of changes in the systems downstream can be understood more quickly.

Reference(s)

- Achieve data lineage in data vault 2.0. (2017, July 18). Scalefree Blog. https://blog.scalefree.com/2017/07/18/achieve-data-lineage-in-data-vault-2-0/
- Colibra. (n.d.).
- The complete guide to Data Lineage. https://www.collibra.com/wp-content/uploads/Ebook-DataLineage-20200113.pdf.
- DAMA (2017). DAMA-DMBOK. Data Management Body of Knowledge. 2nd Edition. Technics Publications Llc. August 2017.
- DAMA Dictionary of Data Management. 2nd Edition 2011. Technics Publications, LLC, New Jersey.
- DAMA-NL (2020). Data Concept System for Data Quality Dimensions (DCS). Research Paper.
- Data lineage - Solidatus simplified data lineage solution. (n.d.). Solidatus - An Award-Winning Data Lineage Solution. https://www.solidatus.com/data-lineage/
- Data lineage 103: Legislative requirements. (2021, April 11). Data Crossroads. https://datacrossroads.nl/2019/03/17/data-lineage-103/
- Data lineage 104: Documenting data lineage. (2020, July 9). Data Crossroads. https://datacrossroads.nl/2019/03/20/data-lineage-104/
- Data lineage and metadata management: An innovative approach. (2019, December 18). DATAVERSITY. https://www.dataversity.net/data-lineage-and-metadata-management-an-innovative-approach/

- Data lineage. (2014, December 20). Wikipedia, the free encyclopedia. Retrieved May 21, 2021, from https://en.wikipedia.org/wiki/Data_lineage
- New vision on data lineage/ flow in DAMA-DMBOK2. (2021, April 13). Data Crossroads. https://datacrossroads.nl/2017/09/10/new-vision-on-data-lineage-flow-in-dama-dm-bok-2/
- Steenbeek, Irina (2023). Data Lineage: the Needs of and Benefits to Various Stakeholders