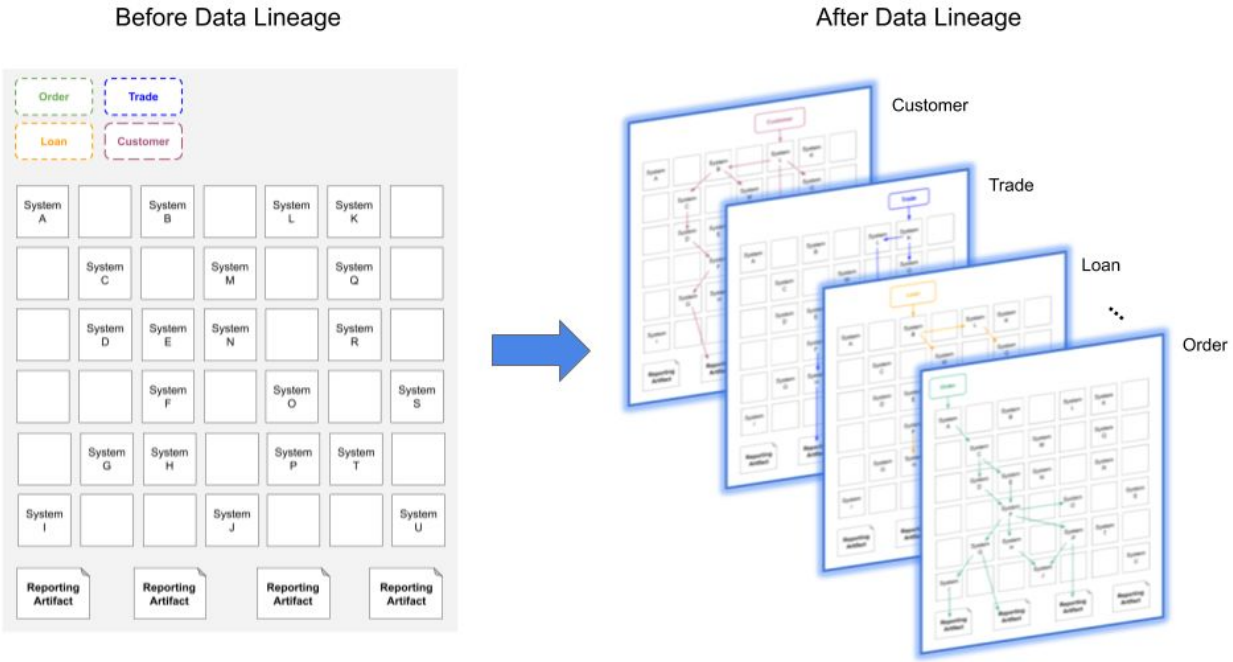
Executive Summary

A Pragmatic Approach to Enterprise-wide Data Lineage

A Global IDs Technical Whitepaper in Two Parts
November 21, 2019 (revised)

Arka Mukherjee, Ph.D., Kirk Kanzelberger, Ph.D., Bill Winkler, and Stefanos Damianakis, Ph.D.
info@globalids.com



KEY TAKEAWAYS

- Data lineage software traces *how information flows* through enterprise data ecosystems.
- To enable key business functions and satisfy regulatory requirements, the view of information flow must be *comprehensive, logical, granular,* and *objectively validated.*
- Manual approaches to lineage *do not scale* to large enterprises, and lack the objective validation provided by automation.
- The automation of data lineage is a *challenging* problem, requiring advanced algorithms for data classification and validation of flows down to the level of individual field values.

- *Automating data lineage is possible* given a small amount of high-level human input. Machine learning algorithms link logical business concepts to their physical representations in database tables. Software can then detect hypothetical data flows between systems and validate actual flow using automated sampling methods.

What do we mean by data lineage? For any data item, data lineage answers two key questions: *Where did the data come from?* and *What path did it take through the enterprise?* As depicted in the figure above, data lineage discovers how every core business object (Customer, Trade, Loan, Order, etc.) actually flows through the data ecosystem of the enterprise.

Data lineage enables key business functions. Data lineage enables key business functions in the following areas:

- Data lineage provides objective evidence of ***data provenance***, which is key to regulatory compliance.
- Data lineage underlies an objective and systematic approach to ***data quality***, since it uncovers all of the critical points in the flow of information and thus guides the most effective placement of data quality checks and controls.
- Data lineage also underlies an objective approach to ***impact analysis*** of changes to the data ecosystem owing to data harmonization, cloud migration, organization mergers, application rationalization, and so forth. Not understanding how information actually flows from end to end through the enterprise undercuts rational decision-making in all these areas.

Requirements for rigorous, actionable data lineage. In order to support these key functions, data lineage must provide a view of data flow that is:

- C*omprehensive* over the whole enterprise.
- *Granular* down to the object and attribute levels.
- *Logically framed* in terms of business concepts.
- *Objectively validated*.

The necessity of automation. Manual approaches to data lineage simply cannot scale to hundreds or thousands of systems and thousands or millions of attributes. Moreover, regulatory authorities are increasingly unwilling to accept human reporting that lacks objective verification using automated third-party methods.

Implementation of automated data lineage. In Part 2 of the whitepaper, we outline the implementation of data lineage:

- Humans define what kinds of entities are present in enterprise data (business concepts).
- Advanced algorithms intelligently link physical data attributes across many systems and applications to logical business concepts. This is the key step that enables a logical depiction of the flow of business concepts through the enterprise.
- Machine techniques for flow discovery generate a map of which business concepts are flowing between which systems across the whole enterprise. Each flow segment (i.e., every path followed by a specific type of data from one system to another) is qualified down to the particular attributes and values that are actually flowing between systems.
- Data is reconciled across each flow segment, and the flow of data through the enterprise may be visualized segment by segment, at the required level of granular detail.

Conclusion

A pragmatic approach to understanding and documenting enterprise-wide data lineage involves initial input from humans and large-scale machine automation. Humans follow a defined methodology to establish the high-level business context for information flow. Machines use advanced classification algorithms to link physical data on many systems to the business context. Machines then automatically generate and validate hypothetical flow segments, resulting in rigorous, system-generated data lineage documentation, including flow diagrams for each core business concept.

For more information on data lineage and its implementation in enterprise data ecosystems, please request the two-part Data Lineage technical whitepaper available from Global IDs.