EDM Webinar

Business-first Data Lineage: A New Approach

Live Date: May 11, 2022

Featuring:

Masood Khatri, Head of Digital and Product, Xoriant CDi

Apnav Agrawal, Direct, Data Governance Practice, Xoriant CDi

Mike Meriton, Moderator, Co-Founder & COO, EDM Council

Recording: View webinar

Presentation: View slide deck

EDM Council Homepage: edmcouncil.org

Xoriant CDi Homepage: cdi.xoriant.com

WEBINAR Q&A:

What is the distinction between Mapping and Matching?

Mapping: Mapping is defined as building a correspondence from source data towards the target attribute. Therefore setting up the relation between the attribute from the source with the attribute from of the destination data (like, a golden source repository)

Matching: Matching is the process used for identification of the records from the target universe (like, a golden source repository) to create/update/enrich from the source data. Fuzzy match is one of the examples where legal name match is used to match source and destination data for create/update/enrich data.

The system to system lineage that we (Capital Market Data Office) provide are not very meaningful for our business. What should we do?

Business users always prefer a simple lineage that will help them understand the end-to-end data flow. We recommend someone should liaise with business users to understand their requirements on a business prioritized basis and create lineage as per their specification. System to System lineage can be used as a reference and would be useful for technical teams while performing Root Cause Analysis.

Can you please define "functional" lineage? What is the difference with other "technical" lineages?

Functional Lineage is the traceability of data footprints throughout data journey or life cycle. It is a simplified view of lineage that highlights the transformation and aggregation of data that is needed and understood by a business user.

Technical Lineage shows the flow of physical data through underlying applications, services, data stores toward developing, updating and maintaining a Data Architecture.

Is it possible to provide a brief definition of data lineage?

Data lineage is the process of understanding, recording, and visualizing data as it flows from data sources to consumption. Lineage corresponds to completely tracking data changes from different sources and getting the details on how, what and where the data changed.

Do you recommend linking your data lineage tools to Data Catalog and Glossary tools?

Yes

Do you include Metric definitions (KPIs) in your data lineage solutions?

Yes

With organizations having legacy systems involved in data flow as source systems, what is the approach you recommend for an end to end data lineage?

In this case, it's important that the data should be tracked from the source to provide an end to end view of lineage.

How do you tie business lineage to technical lineage?

Business Lineage can be viewed as the subset of technical lineage presented with elements and terminologies that the business cares about and can understand independently.

What are some of the best practices to include transformations in data lineage flows?

Best Practices:

1. Automation of tracking the lineage
2. Frequent review of the meta data sources to identify the anomalies in the data
3. Progressive extraction of data will keep the metadata and lineage in sync. This will help reading lineage more productively
4. Track all sources of metadata
5. Periodic review of the rule and governance policies of data transformation and manipulation

What are some of the main advantages you see about data lineage?

Some of the advantages of having data lineage are:

1. Identification of data issue
2. Traceability of the data flow
3. Performing root cause analysis to drive data quality improvements and ROI
4. Performing Impact analysis
5. Data migration
6. Detailed and transparency for the audit trail of data management
7. Identification of KPI's
8. Identification of potential risk

Any advice for how best to capture/preserve historical (temporal) lineage?

Temporal lineage, the best would be to rely on the technical tool and database capabilities to preserve the historical data.

Best practices for keeping lineage current and accurate?

1. Build a hybrid model of managing data flow. Technology will provide the framework of the flow where the manual intervention help keep tracking the effectiveness of the flow
2. Periodic review of the data points and the rules of data transformations
3. Review of the result of RCA (Root cause analysis) to keep a track of the effectiveness of the lineage flow
4. Incorporate the latest AI and ML techniques to keep it updated

How do you gather the underlying metadata that produces these pictures? Who is providing the flows / transformations / rules information?

The process showcased as an industry illustration the tool integrated with our platform's ETL. The process outlined the activities we perform for maintaining our legal entity reference data. Over the period of 20+ years we have built reusable rules, transformations and matching/mapping algorithms that help in managing about 14+ million entities in our database.

Can you talk about the types of events you have and what is captured with an event to populate this information?

Our solution is an open framework, where you can define the events and metadata that you would like to capture based on where and what stages that your data is changing. The architecture provides full flexibility in terms of defining the events and the meta data structure as per your need.

Can you use Database logs as a source of the events?

No. The events need to be configured right when the data is ingested into the pipeline.

Do you think this solution assumes a level of data literacy that may not be realistic?

Using this approach, we believe that it can provide your business users with data lineage transparency which support's their data literacy. The API driven engine runs through your entire data journey and can unveil many new scenarios that were not found explicitly before.

What are the key reasons why Business Users are not able to use the lineage created by Tech teams. What is the "troubleshooting" that they are doing that needs to be eliminated?

Current tools are designed to be more at a technology level, like Schema, PL/SQL, etc.
Tech teams have to work with business in order to perform root cause analysis.

Is there a tool you can suggest for data lineage that is more business user friendly?

Based on what we have seen, the current tools in the market provide capabilities to build custom dashboards for business users, as they themselves are too technical in nature.

Is there a difference between data lineage and data provenance?

Data Lineage and Data provenance are similar as both track data lifecycle from the source to destination. The key goal of a data lineage tool is data lifecycle management right from the data source to the destination of the data. The key goal of the data provenance is specifically to track data sources for data in motion, data in process, and data in rest.

How are you solving the maintenance aspect of the Lineage i.e. how do you keep it up to date and accurate?

Best Practices for keeping data lineage current and accurate:

1. Build a hybrid model of managing data flow. Technology will provide the framework of the flow where the manual intervention helps keep tracking the effectiveness of the flow
2. Periodic review of the data points and the rules of data transformations
3. Review of the result of RCA to keep a track of the effectiveness of the lineage flow
4. Incorporate the latest AI and ML techniques - provides insights on the metadata and improves data quality

Are you expecting people to modify their data pipelines to produce specific events at every stage of the processing?

No. As an industry illustration framework is a single line of code that you need to insert in the pipeline for it to start tracking.

What kind of code needs to be integrated to generate events? Is it a library ?

We are developing an API based framework, where the integration would be as easy as calling an API with the desired metadata where it's ingested or changed.

Don't you think Business users will want to look at things from a higher level than feed or attribute. e.g. where does the Cost of Goods sold value come from?

Absolutely. Business users may want to look at the lineage information for the derived attributes as well, like the example provided, handling derived fields is something we have in our roadmap and we are looking forward to addressing the same in the near future.

Data quality rules can differ by business unit and their context. How do you incorporate multiple data quality rules for a single attribute?

Agreed, the data attribute can have a different journey based on the target systems or hoops it goes through. We recommend that the data quality rules be in sync at multiple levels to avoid any duplication or unnecessary exceptions when trying to execute multiple rules that contradict each other.