Collibra

Whitepaper

Unlocking business opportunities through data lineage

Introduction

Data lineage can seem like an insignificant challenge to those unfamiliar with the complexity of enterprise data architectures. But that is far from reality. Maintaining an accurate and current map of data lineage across an enterprise is anything but simple. Yet the need to do so has become a business imperative for a variety of reasons.

This paper examines how data lineage can support key business objectives. It also looks at the challenges and hurdles that organizations face in maintaining an active map of key data as it flows across an organization.

What do we mean by data lineage?

Data lineage means not only being able to trace the source of data elements, but also charting their journey through an enterprise as they are aggregated, transformed, processed and analyzed by various applications and reporting tools. And it is important to note that this data journey can and should be viewed at different levels of granularity in order to provide different information to different audiences. Two important levels of granularity are:

- Summary business lineage, which shows a summary view of how data used by business users is transformed and aggregated as it flows from source to endpoint, such as a report
- Detailed technical lineage, which tracks data flows and transformations at the lowest levels of granularity, by table, columns, and queries

Data lineage needs to be solved at both of these levels of granularity. Solutions that focus exclusively on technical lineage risk missing key business benefits and vice versa.

Why do we need data lineage?

Understanding use cases and benefits

Data lineage may initially seem to support technical use cases. It enables organizations to ascertain the physical source of data elements and track the flow of data as it is processed by various systems. This technical capability has clear benefits to data engineers, data architects and technology teams, helping them to understand dependencies, usage patterns and determine the impact of changes to an organization's data architecture.

However, data lineage also offers clear business benefits. Data-driven organizations need to know that their analysis is based on trusted data. To do so means being able to confirm not only the physical source of data elements underpinning a report or AI model, for instance, but also to trust in the accuracy of that data and understand how it has been brought together. Data lineage can also help to address regulatory obligations relating to data privacy, which require organizations to know where personally identifiable information (PII) is stored and track how it is processed.

Below we explore the benefits offered by data lineage in more detail:

1. Speeding time to insight

Most professionals have been in a situation where they view a report that contains impactful conclusions but question the validity of the underlying data. Data lineage can speed time to insights by addressing doubts over data quality. If executives can trust that the data underlying a report is certified, sourced from trusted systems, and well-governed, then they can make informed decisions without wasting time arguing over the validity of conclusions.

2. Improving data quality

If a certain piece of analysis leads you to suspect there is an error in the data, how do you check and correct it at the source? You first need to know where that data has come from. Maintaining current and accurate data lineage fosters better data quality. It helps ensure all records can be corrected at the source more efficiently. It can also identify problems where source data may be accurate, but an error was introduced in the way the data was processed.

3. Complying with relevant regulations

Data lineage is key to supporting a variety of regulatory obligations. For example, GDPR Article 30 requires organizations to keep records of their data processing activities, including transfers of personal data. Privacy regulations such as GDPR and CCPA also afford consumers the right to have their personal data deleted, which requires organizations subject to those regulations to know exactly where that data resides. Other industry-specific regulations such as BCBS 239 (which governs the way banks aggregate risk data) require lineage to demonstrate that data has been sourced from the right systems and aggregated correctly. Additionally, emerging regulations are driving the need for increased transparency in AI models, including comprehensive lineage of the data powering the models.

4. System rationalization

Large organizations experience a natural ebb and flow between the proliferation of new systems and data, and the need to rationalize and simplify data architectures. There are many reasons why disparate systems and data architectures proliferate. Mergers and acquisitions are perhaps the most obvious, but organic sprawl also contributes, with distinct business and operating units often arguing their case for autonomy over IT procurement decisions. Rationalizing and harmonizing systems is a perennial project for most organizations — one that helps reduce costs and improve operational efficiency. Maintaining an accurate map of data lineage is crucial in supporting rationalization projects.

5. Cloud migration

The economic benefits of cloud services, not only in terms of cost but also the ability to innovate more quickly, has led many organizations to adopt cloud-first IT strategies. However, in migrating applications and processes to the cloud, there are numerous challenges to overcome. Mapping out logical data flows is a vital first step in any such effort, while understanding lineage at a technical level — the precise fields, table joins and transformations required to support individual processes — is also key to any migration project.

6. Asset management

Asset management can be viewed as a by-product of cloud migration and rationalization projects. As organizations look to simplify their data architectures, they should identify which data and sources are most used, and conversely, which are no longer required and can be decommissioned. Having an automated map of technical data lineage can identify exactly this kind of information.

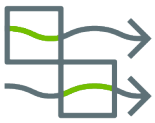
7. Impact analysis / Change management

Technology and data architectures can be strewn with complex interdependencies. When it comes to making changes to systems, the knock-on implications can be very difficult to model. What may seem like a trivial change from the perspective of a database administrator could trigger the failure of a key business process. The only way to mitigate against such risks is by maintaining detailed and accurate data lineage. Knowing how data flows through an organization is crucial when it comes to managing the impact of change.

Challenges to achieving data lineage

Given the clear benefits of building and maintaining data lineage, both from a business and technical perspective, one may think that organizations would prioritize this. However, that is not necessarily the case. This chapter explains some of the challenges that make the task of mapping and maintaining data lineage a complex and moving target.

Factors contributing to this complexity include both macro trends and operational challenges, which are detailed below:



Macro trends

Rapidly evolving regulations

Rules and regulations relating to data privacy and data aggregation are evolving rapidly. The fact that many pieces of legislation are supranational makes it even harder for organizations to manage and track obligations across multiple jurisdictions. The ability to manage consent, usage and retention policies requires detailed knowledge of where and how data flows within an enterprise.

Growing complexity of analytical requirements

Gaining insights into modern business operations is a complex challenge. How do I optimize my supply chain? How do I get more timely insights into my finances? How do I better understand my customers and better serve their needs in a timely manner? How do I use my data to address challenging or uncertain times? These questions require analysis of varied data sets sourced from numerous physical data stores. Modern data-driven enterprises require ever more diverse insights. This makes it all the more important to maintain accurate data lineage, to ensure your analysis is pointed at the right sources of data and one can trust in its conclusions.

Operational challenges

Time-consuming manual processes

With more data to manage, along with complex operating structures, data architectures and analytical requirements to support, automation is crucial when building and maintaining data lineage at an enterprise scale. Key business concepts and logical flows can be mapped out by data architects, but automated solutions are required to gather and keep track of technical lineage. This saves time spent building data lineage manually, allowing developers and data engineers to focus on more critical business initiatives.

Disparate data architectures

Data lineage would be relatively straightforward if organizations were unencumbered by legacy technologies and maintained simple data architectures that were not subject to change. Unfortunately, those caveats simply don't hold true. Instead, data resides across a multitude of applications and physical data stores. New sources are constantly being brought on-board. Technology and data architectures are also subject to continuous change, which makes the task of maintaining lineage a moving target. Any enterprise-wide solution will therefore require automated tools to keep track of technical lineage.

Fragmented business operations

Solving data lineage is not simply a technical challenge. It requires an understanding of business terminology and logical data flows. Whether you are a financial institution striving for a 360-degree view of your customer or a retailer piecing together your supply chain and distribution network, your analysis will require aggregating and tracking data across multiple operating units and business entities. This challenge becomes even more complex when an organization has grown through mergers and acquisitions. Organizations that are fragmented and/or siloed have greater challenges in maintaining accurate data lineage, yet also stand to yield greater benefits by doing so.

Poor visibility into data quality

Data quality and data lineage are two concepts that are intertwined. Ensuring data quality means knowing more than simply where data is located and which columns it contains. It also means answering pertinent questions like: is the source still current and valid? How is each field defined? Are those business definitions consistent? Has the data set undergone any transformations? Who is responsible for its quality? In order to answer those questions, organizations need insights into technical data lineage. They also need to understand lineage from a business perspective and to tie data lineage together with data governance — ensuring that they not only know where data comes from, but also who is responsible for its accuracy.

How can Collibra help?

Collibra provides native, automated lineage via an integrated, enterprise-wide platform, with embedded governance and privacy. We enable organizations to manage data lineage in a business and technical context — to drive trust in their data, derive meaningful business insights, and comply with relevant regulations.

Collibra Data Lineage offers the following benefits:

1. Improve efficiency

Collibra Data Lineage allows organizations to track data across disparate systems by capturing end-to-end business and technical lineage automatically — saving a significant amount of time to achieve the same outcome manually. These capabilities span many commonly used source systems, with support for a wide range of SQL dialects, ETL tools and business intelligence platforms

2. Comply with relevant regulations

Collibra Data Lineage helps organizations keep track of data assets subject to regulatory policies. Whether that means keeping a tight handle on personal identifiable information (PII) to comply with privacy regulations like GDPR and CCPA, or tracking the way that data is aggregated to meet BCBS 239 principles.

3. Better understand your data

To make better business decisions, you need to understand the full context of your data. With Collibra's data lineage capabilities you can see how a data set is built and aggregated, check whether it's from a trusted source, see who the data owner is, and see if it's been certified as reliable. With this enriched context, you can help ensure that accurate, complete and trustworthy data is used to drive business decisions.

4. Enhance IT and data operations

Our data lineage solution helps support a range of technical functions relevant to both IT and data architects. From mitigating the risks associated with system changes via impact analysis, to helping support system rationalization and cloud migration initiatives, Collibra Data Lineage will help drive better intelligence across all operations.

Practical examples of Collibra Data Lineage

The previous sections analyzed the high-level benefits and use cases related to data lineage. Now, we'll look at a few practical examples of exactly how Collibra Data Lineage can be used. Below we walk through a few scenarios that help to bring those conceptual benefits into real-life practical examples:

1

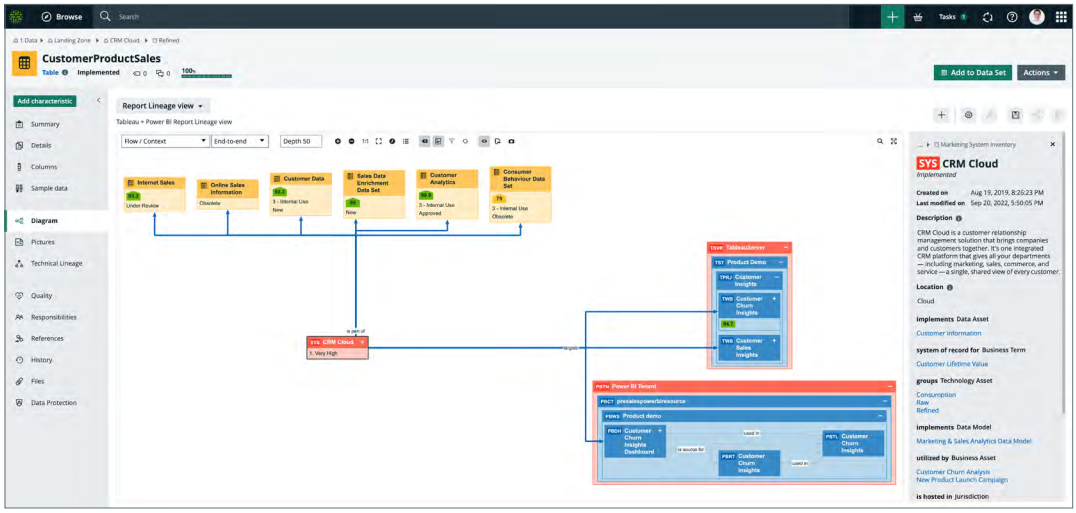
Scenario 1:

Your organization has grown through M&A and now supports several different CRM systems, with overlap between customers. You would like to analyze your aggregated customer base through a range of different reports, but you know that one source is the most accurate and complete and would prefer all your analysis to use data from that source where possible.

How Collibra Data Lineage helps

With business lineage, you can check whether reports requiring customer data use your preferred data source. If not, you can update the reports to point to the preferred data source, ensuring the best quality data is used.

Detailed technical lineage diagrams can help you confirm that the preferred source is in fact pulling customer data from the right databases.



Summary business lineage from source to report

2

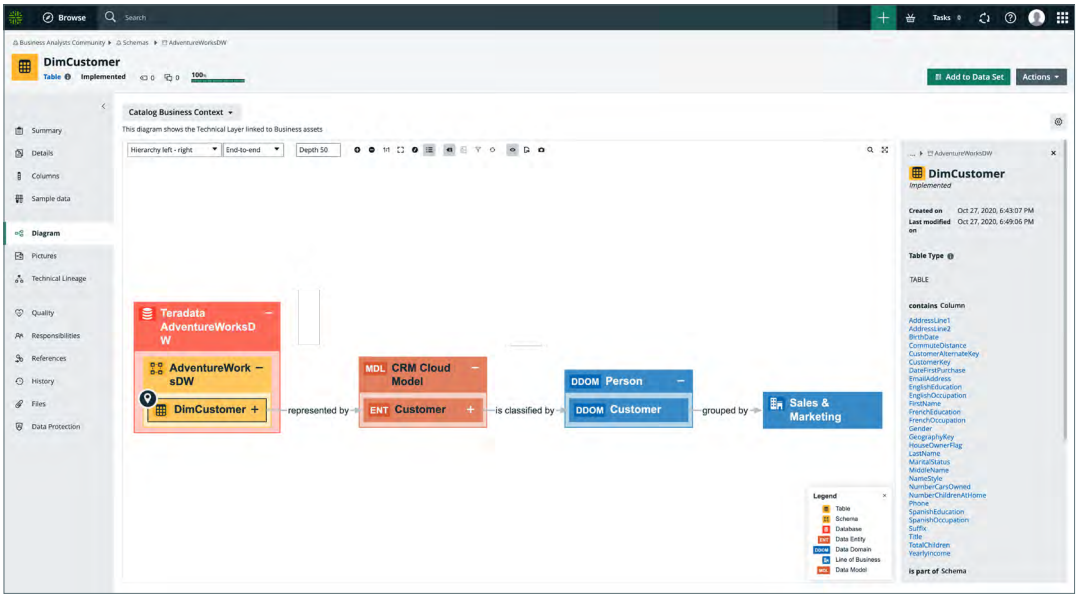
Scenario 2:

Your organization has embarked on a project to rationalize CRM systems and migrate to a preferred cloud-based CRM provider. This project requires understanding key business terminology and processes, as well as system and data dependencies.

How Collibra Data Lineage helps

Business lineage provides an overview of the data architecture relevant to this project. It shows how customer data from a particular data source maps into the logical layer (data attributes and entities) and the conceptual layer (data domains) — akin to a whiteboard sketch. This helps organizations understand what data is relevant and should be moved. Additionally, business lineage can show what data is being used in reports so that users can be alerted to use another data source if necessary.

With our detailed technical lineage diagrams, you can see the mapping of all of the system processes (e.g. table joins, queries and calculations) required to process CRM data and ensure that the impact of any change will be minimized.



Summary business lineage diagram showing how a data source is connected to the logical and conceptual layers of the data model

3

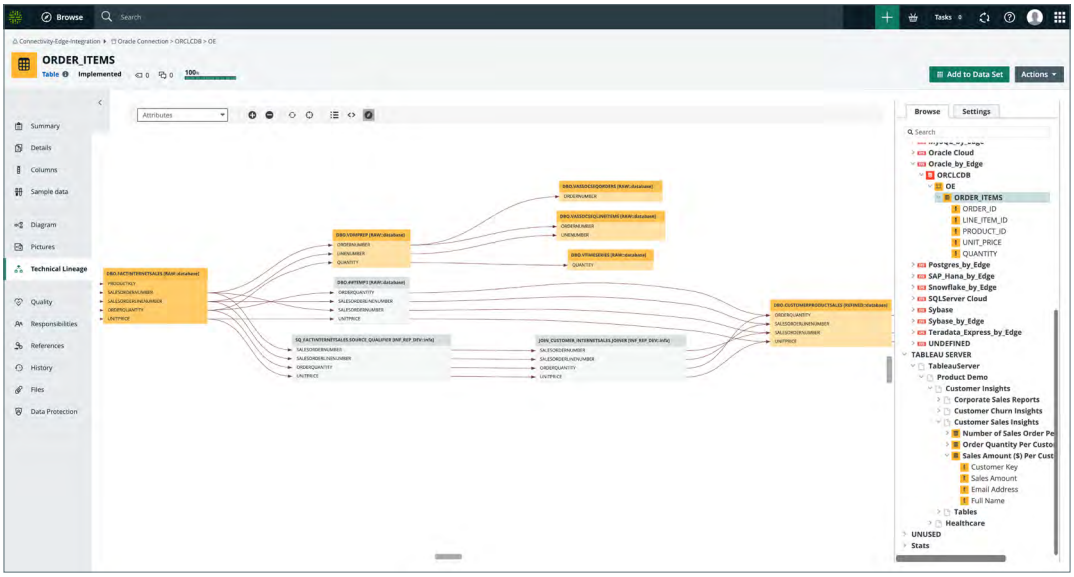
Scenario 3:

Your organization is seeking to migrate to a new SQL database as part of a strategic decision taken by the CTO.

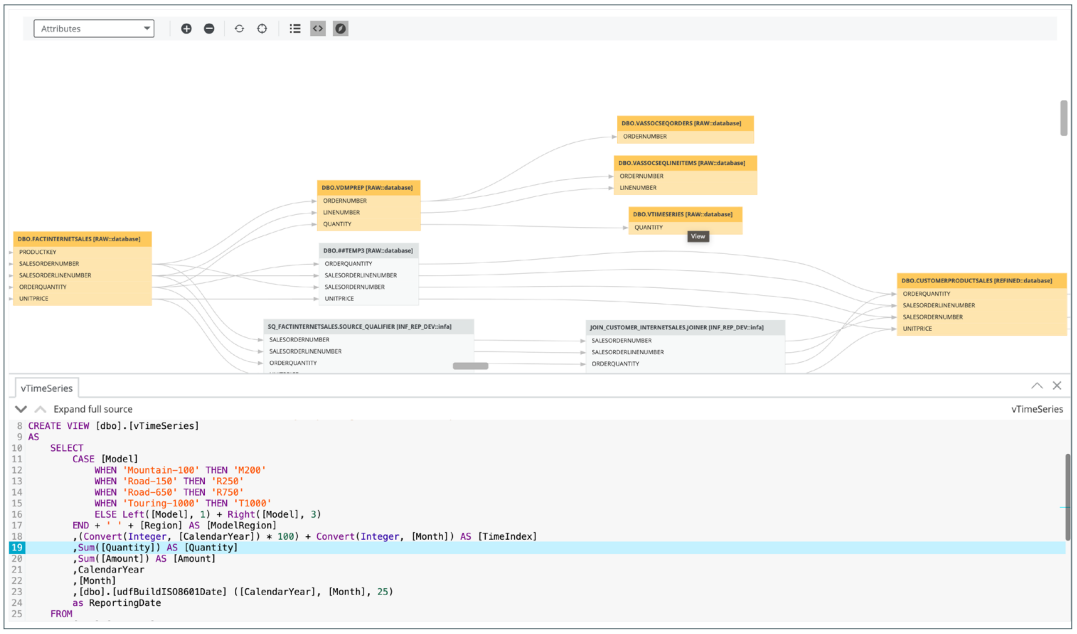
How Collibra Data Lineage helps

Business lineage can help by identifying all of the downstream applications and reports that rely on the old database. All data and report owners can easily be identified and notified as part of the project plan.

Detailed technical lineage can help by showing specific query-level interactions that will need to be reviewed given the slight nuance in SQL dialect resulting from the change.



Detailed technical lineage diagram



Drill down into SQL code within the technical lineage diagram

Conclusion

Data lineage is a crucial step in the journey to data intelligence enabling organizations to better understand and trust their data. Data lineage provides a roadmap of data consistency, completeness and accuracy, which ensures business users employ the right data to make impactful business decisions. With Collibra Data Lineage, business and IT come together with the combination of business-friendly summary lineage views and detailed technical lineage views, helping them to successfully make data-driven decisions that reduce costs, improve operational efficiency and enable innovation.

