MANTA

WHITEPAPER

Automated Data Lineage: The Cornerstone of Effective DataOps

Automated Data Lineage: The Cornerstone of Effective DataOps

ABSTRACT

Data Lineage, the DataOps Cornerstone

Today's enterprises are equipped with hundreds of specialized tools and technologies to enable data experts and business intelligence operators to collect, integrate, and analyze record volumes of data. So why does the divide between data creators and data consumers continue to widen? Why does it feel like the more we can do with our data, the less we understand what should be done with it?

The answer is that we are being challenged by the complexity that all these integrations, tools, and processes have introduced to the data environment, wreaking havoc on our ability to effectively manage data pipelines and trust the data in our reports.

We have recruited DataOps teams to deal with the spinning plates by aligning the people, pipelines, and processes to manage the increasingly complex data environment.

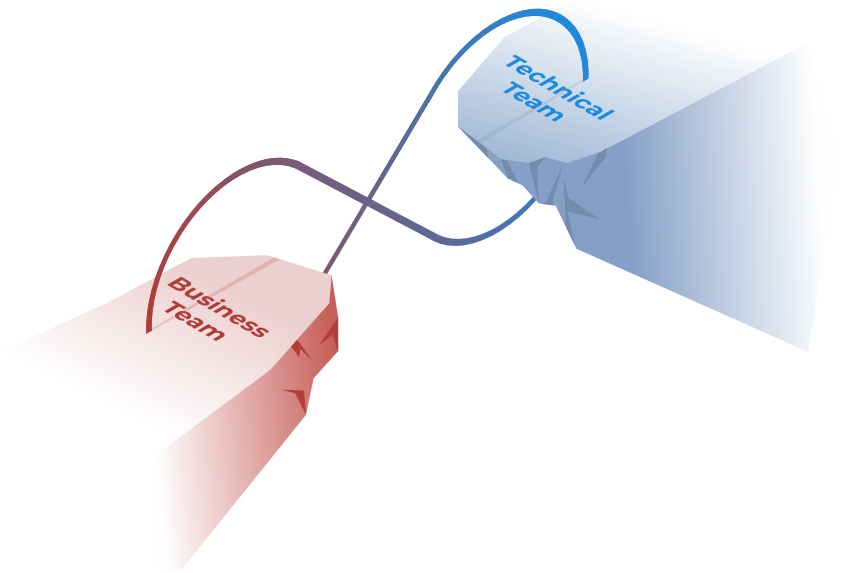
A fundamental component of that high-stakes goal is an effective cross-team collaboration between data creators and data consumers. But teams are struggling to deliver what's expected of them, wasting hours on manual data processing just to figure out what's going on in the pipeline and where to start troubleshooting.

Automated Data lineage can provide the visibility teams need to effectively identify dependencies, evaluate the impact of transformations, find the point of failure, and isolate data incidents. Lineage illustrates how data flows through the environment and transforms over the entire data lifecycle.

There's nothing new in our need for reliable and accurate data lineage, but the incomplete and platform-locked lineage of the past is no match for DataOps.

DataOps needs next-generation lineage solutions that can provide historical revisions, ingest custom metadata, and provide observability to stakeholders across the entire pipeline.

This paper focuses on the imperative role of automated data lineage in bridging the gap between technical and business stakeholders and outlines what leadership should look for in a lineage solution to fuel DataOps success.



INTRODUCTION

The Mounting Complexity of the Enterprise Data Environment

Over the past five years, most companies have adopted new technologies for data streaming, big data, cloud data storage, and modern artificial intelligence and machine learning (AI/ML). Our data landscape has evolved from a simple architecture with basic extract transform load/extract load transform (ETL/ELT) and a data warehouse into a living and breathing ecosystem with tens or hundreds of different technologies.

As the number of data transformations increases, our data environments grow increasingly complex. Today, most enterprises are struggling with data pipelines because of that complexity.

Our pipelines are misunderstood by non-technical users and poorly managed across multiple stakeholders and teams. Data privacy and compliance are no longer the main issues.

We're able to do "more" with our data than ever before, but these capabilities come at a high cost. The price we pay for limited visibility and observability of data pipelines, the price we pay for "not knowing," is enormous. Wrong and/or late business decisions cost companies millions in revenue, while data engineering resources are wasted on frustrating and repetitive manual tasks.

Complicated transformations, more touchpoints in the pipelines, and the growing divide between data creators and data consumers bring a confounding need for DataOps to tackle the encroaching problem of complexity.

Key Challenges for DataOps Teams

More data and more requirements lead to more complex data pipelines and limited visibility. Complexity with no visibility translates to data incidents, zero trust in data, and decreased performance and agility of data teams. This is why DataOps initiatives call for powerful data lineage.

Data comes from many disparate sources across the enterprise, including on-premises and multi-cloud. An enterprise may have hundreds or thousands of data sources and tens of millions of data objects. Given the volume and scale of data today, it's virtually impossible for organizations to trace data's journey through their systems' infrastructure using manual processes. That's why DataOps should take a lineage-first approach and build an effective layer of metadata before any other tools such as data catalogs are considered.

Many tools and technologies offer limited lineage capabilities, but they provide a siloed view of data within an organization. For instance, the lineage you may have for Hadoop provides visibility only as far as your Hadoop clusters. For these reasons, enterprises are turning to comprehensive lineage solutions that can provide a view across or **a high-level summary for business analysts** and a complete overview of the entire pipeline.

Companies go to great lengths to monitor data quality to limit downtime and speed up incident resolution, but monitoring what flows through your pipeline is only half the battle.

Here's why.

If we think of data as the water we drink, we know that ensuring it is clean and safe for drinking needs to happen at the receiving end, coming out of the tap. Rusty or damaged pipes will contaminate the water coming through them, rendering even the cleanest water at best low-quality and at worst completely undrinkable.

DataOps emerged when more data and more requirements led to a mounting complexity of data pipelines and limited visibility. As a result, we have all the data incidents, zero trust in data, and decreased performance and agility of data teams. Complexity with no visibility is a productivity killer.

KEY CHALLENGES

- Complexity and continuously shifting requirements prevent data teams from establishing a sustainable pace and project continuity.
- Inconsistent coordination and a lack of clear communication amongst stakeholders make building, deploying, and maintaining data pipelines unnecessarily difficult.
- Teams struggle with Increasing delays in operationalizing models due to a lack of quality data lineage.
- Without automated lineage data, analysts cannot scale data qualification procedures and spend hours manually cleaning and preparing data instead.
- Data lineage solutions still require technical proficiency to leverage, making self-service impossible for business users.
- Without lineage to provide a verified source of truth, the lack of trust in data keeps efficient data availability and data democratization out of reach.

Where Existing Data Lineage Solutions Fall Short

There's nothing new about the need for lineage. Metadata harvesting is a basic industry practice. Data lineage empowers business and technical users with a deeper understanding of their data and reinforces their trust. Additionally, lineage helps improve collaboration by linking business views of data with underlying logical and detailed information.

However, the demands DataOps place on lineage tools are relatively new and also discipline-specific. The demand for data lineage comes from both technical and business users. Both rely on support for custom views with varying levels of detail, including business, logical and physical levels in order to meet their specific needs.

The issue is that most lineage tools also either lack sufficient **granularity** like the **column-level lineage or attribute-level level lineage** needed for root cause and impact analyses, or they don't provide the **historical lineage** and **transformation logic that's so critical for proactively resolving errors**.

Missing Data Processing Logic

Every data lineage solution starts with automated scanning. The less you do manually, the better for everyone. But there is a catch. In many cases, automation is too simplistic, and scanning is only done for data structures (like tables, columns, fields, etc.), not for data processing logic. But DataOps need to capture the dependencies between the **datasets** being produced and the business logic producing and transforming them. Without lineage derived from transformation rules and code, teams don't have real lineage or visibility of the pipeline. They have, at best, an unreliable interpretation and, at worst, a critical data incident on the horizon.

When assessing a metadata management technology, make sure the solution you are considering either can deliver lineage by analyzing

programs or procedures created outside the ETL tool or has a mechanism that allows for manual activities to “fill in the blanks.” Without knowing what’s missing from the lineage map, mitigating data incidents and performing root and impact analyses becomes impossible.

Revision History and Transformations

Just as the column-level granularity is key for understanding how a data citizen arrived at a given calculation, revision history provides a key snapshot of what your environment looked like in a specified time frame.

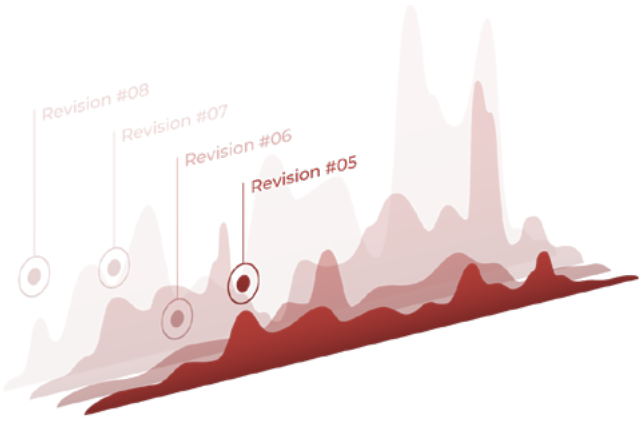
The ability to compare “states” of the data environment in different time frames helps accelerate incident and root cause analysis and cut out over 90% of manual data parsing/processing.

DataOps teams can analyze changes in total number of governed data assets across the environment or review trends in code changes or the number of touchpoints that access a particular element or table.

Tracking revisions allows teams to identify programmed changes in tables and columns, records, and column attributes, including transformation logic.

Revisions also solve collaboration issues. What used to take hours of reconstructing to compare past data flows can now be done in a few clicks.

By allowing DataOps to understand changes in context, historical lineage improves collaboration among pipeline stakeholders, speeds up root cause and impact analysis and eliminates the need for manual data wrangling.



Limitations of Runtime Lineage

While data lineage is a map of all possible data flows and is typically derived from the processing logic itself (by analyzing and decoding it), runtime lineage represents information about data flows executed recently and is usually derived from log files and execution plans generated data processing tools.

One major disadvantage of using only the runtime approach is that it only gives you high-level information about data flows, not the details of the calculations.

If you only need to see recently executed data flows, runtime lineage will be useful. However, for effective **incident prevention and impact analyses**, teams need to see all possible data flows, not just recent flows.

Even if collected over an extended time frame, runtime lineage remains incomplete because of the continuous changes happening in your data environment. Runtime metadata is a critical resource for managing data pipelines but cannot be your only source of lineage.

Conclusion

Data and analytics teams need automated lineage to tackle the complexity that has become the hallmark of enterprise data environments. Without a complete, detailed lineage map of the data flows happening across countless databases, warehouses, processes, and BI tools, any DataOps initiative becomes a losing battle.

Having lineage limits the scope of "not knowing" and keeps the associated costs to an absolute minimum while maintaining a reliable and seamless flow of data across every touchpoint.

Equipped with enhanced observability across data sources, transformations, and dependencies, DataOps teams can easily and rapidly identify the impact and origin of changes in the pipeline and prevent data incidents. By supporting efficient data processing, and comprehensive observability, lineage becomes a key business driver, enabling trust in data across the organization and beyond.

About MANTA

MANTA automates data lineage mapping for DataOps to drive significant and measurable impact for metadata management, data governance, and master and self-service interaction.

MANTA's united lineage platform provides a path to a DataOps practice that can tackle the complexity of the modern enterprise data environment and provide easy-to-understand, automated metadata that both business and technical users can leverage to tackle any data challenge.

Want to see how MANTA can support your organization's DataOps? Get in touch with us at manta@getmanta.com or book a call. We will be happy to walk you through a demo of the critical lineage capabilities you've just read about.