

The Ultimate Guide to Data Lineage



manta

Table of Contents

Table of Contents	2
Why Are We Talking About Data Lineage?	3
What Is Data Lineage?	5
What Is Metadata?	6
How Data Lineage Activates Metadata	7
Business Benefits of Data Lineage	9
How to Create Data Lineage and Keep It Up to Date	12
What to Look for in a Data Lineage Solution	17
How Manta Can Help with Data Lineage	19

Introduction

New data flows throughout your organization every minute of every day. With the high volume of data you're collecting, you need to be able to act and make informed decisions—and quickly.

But today's data systems are deeply complex, and this complexity creates blind spots in the data environment. **With limited visibility, you have limited control.** How can you overcome these complexities and get the full picture of your data landscape?

Data lineage combats complexity in your data environment by providing a complete picture of your data landscape, allowing you to tame data chaos and optimize data for valuable insights.

In our ultimate guide to data lineage, we will discuss:

- What data lineage is and why it matters in a variety of industries
- How to activate metadata
- How to create data lineage
- What to look for in a data lineage solution
- Why Manta should be part of your data toolkit

Why Are We Talking About Data Lineage?

Data management has undergone a massive transformation in the past decade. Data infrastructure is constantly growing in complexity, evolving into data ecosystems with thousands of components all aimed at one goal: to derive more value from data.

Today's data ecosystems are too much for the human brain to handle. They are too diverse and interconnected, rife with directly and indirectly connected applications, microservices, and infrastructures, with countless dependencies defining how all these touchpoints interact with one another.



When you have data from this many sources and systems, you can't extract meaningful insights. It can be difficult for any one person or team to have complete visibility over your entire environment.

In complex environments, it's not a question of if your governance team will miss something, but when.

These data blind spots and complexities can lead to the following challenges:

Slower Delivery of Predictive Insights

MIT and Hewlett Packard Enterprise reported that **data-driven companies are 58% more likely to beat revenue goals than non-data-driven companies, and 162% more likely to significantly outperform them.**¹ This is the power of predictive insight. But due to an overabundance of data, much of it goes to waste. One report found that as much as **68% of data goes unleveraged.**²

In most cases this data remains inaccessible because:

1. It is stored in an unusable format or inaccessible location
2. It cannot be traced and therefore cannot be trusted.
3. It's difficult to determine what data is important
4. The data is sensitive and needs to be protected

When data engineering resources are spent on unproductive impact analysis (such as assessing the impact of new development requirements), it distracts developers and slows down the delivery of new features. Data lineage helps speed up the delivery of predictive insights by automatically generating comprehensive data flow visualizations.

Growing Number of Data Incidents

During the third quarter of 2022, nearly 15 million data records were exposed worldwide through data breaches—a 37% increase compared to the previous quarter.³ Due to the limited visibility of complex data systems, assessing the end-to-end impact of data dependencies and changes is demanding and sometimes impossible.

1 MIT Technology Review; Getting the most from your data-driven transformation: 10 key principles; Janice Zdankus; Anthony Delli Colli; 14 Oct. 2021

2 Seagate; Rethink Data' Report Reveals That 68% Of Data Available To Businesses Goes Unleveraged; 15 July 2020

3 Statista; [Number of data records exposed worldwide from 1st quarter 2020 to 3rd quarter 2022](#)

Current data observability tools are still primarily reactive—meaning organizations find bugs after there has already been an incident, rather than preventing them. This is concerning because even a single data incident can cause severe damage to your organization. IBM reports a single data breach costs an average of \$4.3 million.⁴

Data lineage grants IT teams high levels of observability without large amounts of manual intervention. This allows teams to be truly proactive—to catch incidents before they happen.

Decreased Trust in Reports & Insights

In 2020, 90% of companies reported that their data governance projects had failed.⁵ Years later, companies are still struggling to build trust.

If you can't fully explain how data was collected or verify its origins, you can't answer basic questions or leverage data for better customer outcomes, and you're going to experience severe business impacts and frustration.

Data lineage is a powerful tool in the fight for building trust in data. Detailed lineage creates an added semantic layer for more accurate and timely reports that lay the foundation for more informed decisions and better forecasting without second-guessing.

Automated lineage also puts an end to the costly, lengthy, manual processes of lineage collection and updating.

Manta customers have saved between \$5-15M in the initial phase of implementation alone.

Shortage of Data Engineering Talent

Engineering talent is hard to find, especially in the competitive post-COVID-19 environment. The last thing an organization should do is waste their team's time on manual, routine tasks. This can increase their frustration and likelihood of leaving.

Due to the growing complexity of the data stack, data engineers have become more critical than any other role, as they oversee data pipelines and integrated data structures. This requires a larger skill set, which makes good data engineers harder to find and even harder to keep.

Organizations that invest in data lineage remove this burden from their IT team, allowing them to refocus their efforts on tasks that can't be automated.

Increased Risk of Non-Compliance

Regulators are cracking down on organizations in every industry. In the United States, we've already seen a rapid expansion of data regulations, like PCI, FERPA, and FISMA.

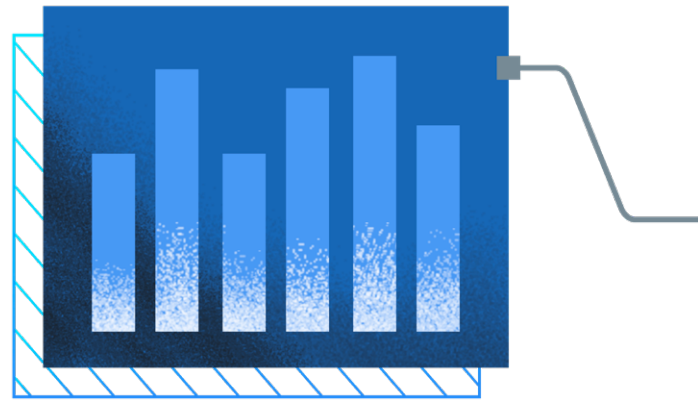
The EU is experiencing similar regulatory challenges. In April of 2022, the Digital Services Act threatened Big Tech with a 6% revenue penalty for having illegal content live on their sites.

⁴ IBM; [Cost of a data breach 2022: A million-dollar race to detect and respond](#)

⁵ Gartner; [The State of Data and Analytics Governance Is Worse Than You Think](#); Saul Judah, Andrew White; 19 June 2020

Whether you need to comply with the GDPR, HIPAA, the Sarbanes-Oxley Act, or the FDA, your organization is at risk without lineage. Data lineage provides a complete overview of all regulated data that is processed by your organization. This helps you prepare for audits and avoid hefty penalties for non-compliance.

You need to get your data systems under control and find a way to stay efficient despite this skyrocketing complexity. You need data lineage.



What Is Data Lineage?

Traditionally, data lineage has been seen as a way of understanding how your data flows through all your processing systems—where the data comes from, where it's flowing to, and what happens to it along the way.

In reality, data lineage is so much more.

Data lineage represents a detailed map of all direct and indirect dependencies between the data entities in your environment.

Why is this so important?

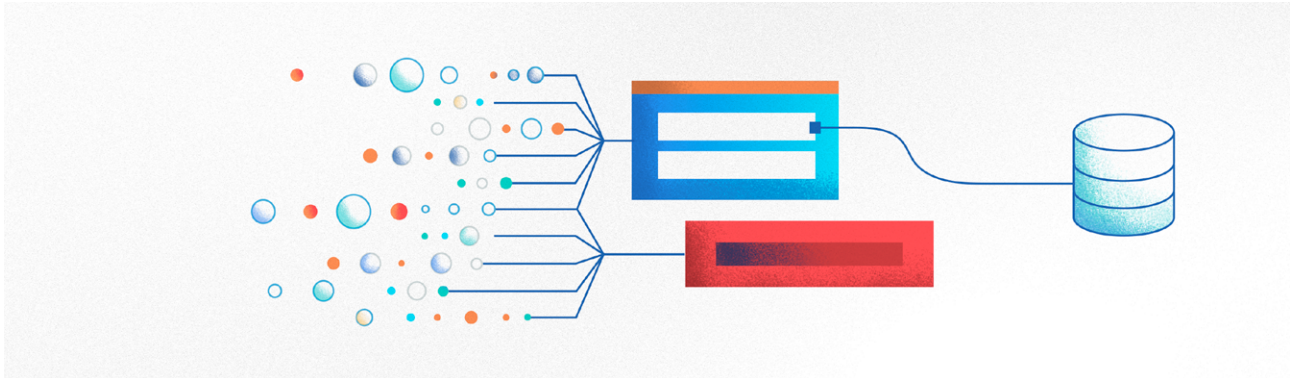
A detailed dependency map is the core component of a modern data stack. It allows you to gain complete visibility and a clear line of sight to uncover data blind spots throughout your data systems, while also helping ensure ethical, compliant, and efficient data management processes.

A detailed dependency map can tell you:

- How changing a bonus calculation algorithm in the sales data mart will affect your weekly financial forecast report
- Where data that is heavily regulated is being used and for what purpose
- What is the best subset of test cases that will cover the majority of data flow scenarios for your newly released pricing database app
- How to divide a data system into smaller chunks that can be migrated to the cloud independently without breaking other parts of the system

There are endless opportunities when you tap into the full potential of your data. To do that, activating metadata is key.

What Is Metadata?



You may have heard metadata described simply as data about data—essentially being information that describes or catalogs what the data is.

As defined by [Dr. Irina Steenbeek](#): The terms “data” and “metadata” have a complicated relationship. Data is the physical or electronic representation of facts or signals “in a manner suitable for communication, interpretation, or processing by human beings or by automatic means.”

Metadata puts data in a particular context. A **data model is an example of metadata**. For example, if you had a dataset with customer financial

information, you couldn’t use the dataset without a detailed description. The description of a dataset is metadata that defines data and explains the context in which data can be used.

The challenging aspect of defining metadata is that the same data can be recognized as either data or metadata, depending on the context. For example, data models are metadata for business users. For data modelers, on the other hand, data models can be considered data that will in turn require other metadata to describe data models. Different sources contain different approaches to classifying metadata.



Business metadata

Business metadata focuses largely on the content and condition of the data and includes details related to data governance.

Note: “data governance” has different meanings in different contexts.



Technical metadata

Technical Metadata provides information about technical details of data, the systems that store data, and the processes that move it within and between systems.

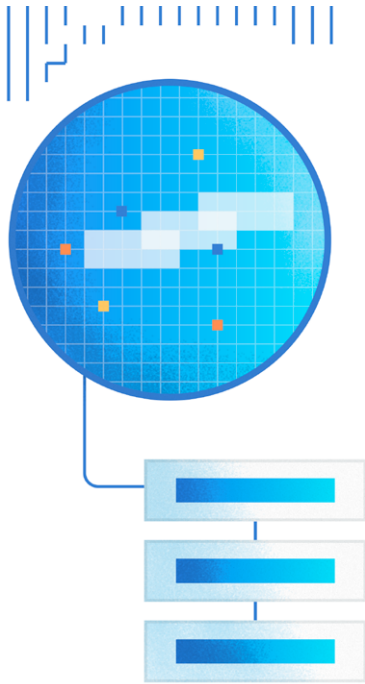


Operational metadata

Operational Metadata describes details of the processing and accessing of data.

Data lineage is crucial for understanding and utilizing dynamic metadata—how it moves across the systems, where it originated, how it’s interconnected, and how it transforms.

How Data Lineage Activates Metadata



We typically use data in the following ways.

- **We query data to get answers to our questions.** That is the very basic use case we usually think of first. It can be good old pre-built reports, ad-hoc queries, or smart AI/ML algorithms digging insights from the data we have. By performing queries, we turn data into information and eventually knowledge.
- **We embed data into places where people or machines naturally need them.** We do not force a sales representative to log into our reporting platform and write ad-hoc SQL queries or use pre-built reports. No. Rather, we prepare all the data they may need about a prospect or a customer, turn it into information, and deliver it to their workspace (as a dashboard in an application like the CRM they use daily). On top of that, we also enrich internal data with valuable external data to provide an even more complex view of the customer.
- **We also use data to automate tasks and processes.** Instead of waiting for the sales representative to open their workspace and search for customers who may be a good fit for a new product offering, we have an algorithm running in the background that scores existing customers and sends proactive notifications that suggest who to call and what (or even how) to offer. Or, for an even simpler example, we automatically send a reminder to a sales representative in case they take no action (even if they should).

Obviously, there are more ways that we interact with data; the above are the most traditional examples. The first “search” case represents a very “static” experience. Everything is sitting in a silo (e.g., a data warehouse or data lake), and we expect people to come, find what they need, and ask questions they need to ask. Do not get me wrong - it is awesome for some use cases and, when compared to a case with no data available, a huge jump forward.

However, we see that data is put to much better use in the other two examples, actively supporting users with limited data engineering skills and dramatically increasing their productivity. Compared to the first example, the latter are more “active” and thus more useful and accessible to a broader audience. And that is what we want to achieve with active metadata too.

Gartner’s definition of metadata in their most recent [Market Guide for Active Metadata Management](#) touches on several key aspects of metadata:

- Continuous access - metadata is continuously collected. It is not something you do once per month or once per year, as we want to collect every change and every signal and respond to it.
- Connecting dots - metadata is not just collected; it is constantly processed to distill information (and knowledge) from all the signals and noise. And with the right feedback loop, your system gets smarter over time, collecting and learning.
- Actionable - all the intelligence and insights derived from metadata are not locked into a silo, but rather delivered in the form of recommendations, warnings, and notifications to humans and systems/applications that may need it.
- Embedded - actionable information / knowledge is integrated into processes humans and machines perform, embedded into their workspace. People are not forced to go in and look for the insights. Instead, active metadata comes to them - when and where they need it.

Merely collecting and cataloging metadata fails to maximize its potential. This is one reason why metadata has not historically been practical. We've spent decades focused on metadata collection, but data sitting in a repository is not useful and data without understanding can be a liability. This is changing now with the [focus on active metadata](#).

According to Gartner, active metadata capabilities will expand to include monitoring, evaluating, recommending design changes, and orchestrating processes in third-party data management solutions.⁶

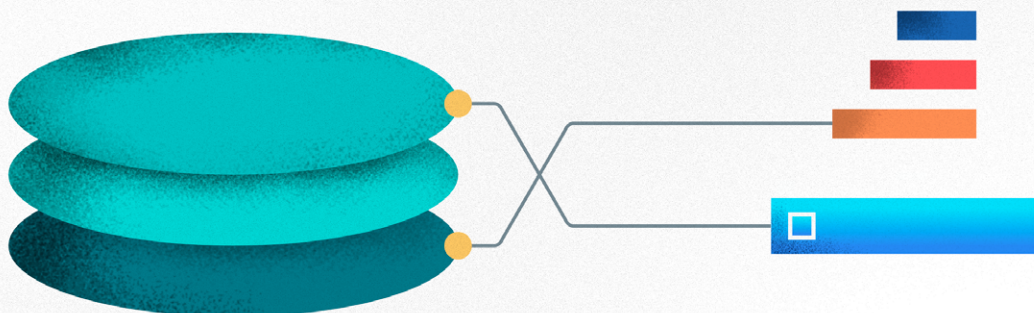
Gartner also predicts that organizations that adopt aggressive metadata analysis across their complete data management environment will decrease the time to delivery of new data assets to users by as much as 70%.⁷

By activating existing metadata with automation and intelligence, data lineage can provide the needed visibility and control to help you become more aware of your data management and proactive in your data usage.

This approach can empower you with:

- Continuous detection of “dead tables” where potentially sensitive information is stored but not accessed or used
- Instant alerts if a change negatively impacts tactical management reports or key data features used by the data science team
- Notifications about overly complex parts of data pipelines where refactoring or redesign would help to reduce the risk of failure
- Warnings if the design of a data pipeline moves data between locations where no data should ever be

As you can see with these examples, data lineage allows you to leverage metadata to better manage, utilize, and optimize your data—even when working with overly complex data environments. That can lead to immediate and long-term business benefits.



⁶ Gartner, Market Guide for Active Metadata Management; Guido De Simoni, Mark Beyer; 14 Nov 2022

⁷ Ibid

Business Benefits of Data Lineage

Automated Impact Analysis for Improved Incident Prevention

In business, every decision contributes to the bottom line. That's why impact analysis is crucial—it predicts the consequences of a decision. How will one decision affect customers? Stakeholders? Sales?

Data lineage helps during these investigations. Because lineage creates an environment where reports and data can be trusted, teams can make more informed decisions. Data lineage provides that reliability—and more.

One often-overlooked area of impact analysis is IT resilience. This blind spot became apparent in March of 2021 when CNA Financial was hit by a ransomware attack that caused widespread network disruption. The company's email was hacked, consumers panicked, and CNA Financial was forced to pay a record-breaking \$40 million in ransom.⁸ This is where lineage-supported impact analysis is needed. If you experience a threat, you will want to be prepared to combat it, and know exactly how much of your business will be affected.

IT resilience is also threatened by natural disasters, user error, infrastructure failure, cloud transitions, and more. In fact, according to McKinsey, **76% of organizations experienced an incident during the past two years that required an IT disaster-recovery plan.**⁹

Most organizations struggle with impact analysis as it requires significant resources when done manually. But with automated lineage from Manta, customers have seen as much as a **40% increase in engineering teams' productivity after adopting lineage.**

Greater Data Pipeline Observability for Faster Incident Resolution

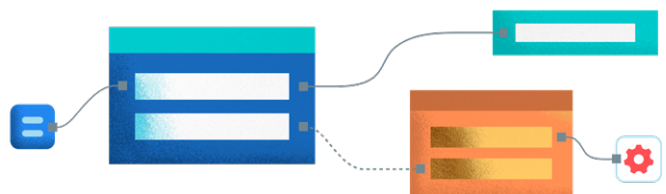
As discussed above, there are countless threats to your organization's bottom line. Whether it is a successful ransomware attack or a poorly planned cloud migration, catching the problem before it can wreak havoc is always less expensive.

That's why data pipeline observability is so important. It not only protects your organization but also your customers.

Data lineage expands the scope of your data observability to include data processing infrastructure or data pipelines, in addition to the data itself. With this expanded observability, incidents can be prevented in the design phase or identified in the implementation and testing phase to reduce maintenance costs and achieve higher productivity.

Manta customers who have created complete lineage have been able to trace data-related issues back to the source 90% faster compared to their previous manual approach.

This means the teams responsible for particular systems can fix any issue in a matter of minutes, according to Manta research.



⁸ Bloomberg; CNA Financial Paid \$40 Million in Ransom After March Cyberattack; Kartikay Mehrotra and William Turton; 20 May 2021

⁹ McKinsey Digital; IT resilience for the digital age; Arun Gundurao, Jorge Machado, Rut Patel, and Yanwing Wong; 11 May 2021

Improved Regulatory Compliance

Depending on your industry, you have to ensure you're in compliance with a host of regulatory bodies and policies—BASEL, HIPAA, GDPR, CCPA /CPRA, and CCAR, just to name a few.

All of these regulations require accurate tracking of data. Your organization must be able to answer:

1. Where does the data come from?
2. How did the data get there?
3. Are we capable of proving it with up-to-date evidence whenever necessary?
4. Do we need weeks or months to complete a report?
5. Is that report even entirely reliable?

Data lineage helps you answer these questions by creating highly detailed visualizations of your data flows. You can use these reports to accurately track and report your data to ensure regulatory compliance.

Faster and More Efficient Migrations

McKinsey predicts that \$8 out of every \$10 for IT hosting will go toward the cloud by 2024.¹⁰ However, if you have ever been involved in the migration of a data system, you know how complex the process is.

Approximately \$100 billion of cloud funding is expected to be wasted over the next three years—and most enterprises cite the costs around migration as a major inhibitor to adopting the cloud.¹¹ The process is so complex (and expensive) because every system consists of thousands or millions of interconnected parts, and it is impossible to migrate everything in a single step.

Dividing the system into smaller chunks of objects (reports, tables, workflows, etc.) can make it more manageable, but poses another challenge—how to migrate one part without breaking another. How do you know what pieces can be grouped to minimize the number of external dependencies?

With data lineage, every object in the migrated system is mapped and dependencies are documented.

Manta customers have used data lineage to complete their migration projects 40% faster with 30% fewer resources.

Retention of Data Engineering Talent

Data engineers, developers, and data scientists continue to be fast-growing and hard-to-fill roles in tech. The shortage of data engineering talent has ballooned from a problem to a crisis, made worse by the increasing complexity of data systems. The last thing you want is to continually overstretch your valuable data engineers with routine, manual (and frustrating) tasks like chasing data incidents, assessing the impacts of planned changes, or answering the same questions about the origins of data records again and again.

Data lineage can help to automate routine tasks and enable self-service wherever possible, allowing data scientists and other stakeholders to retrieve up-to-date lineage and data origin information on their own, whenever they need it. A detailed data lineage map also enables faster onboarding of data engineers to integrate new or less-experienced engineers into the role without impacting the stability and reliability of the data environment.

¹⁰ McKinsey; Cloud-migration opportunity: Business value grows, but missteps abound; Tara Balakrishnan, Chandra Gnanasambandam, Leandro Santos, and Bhargs Srivathsan; 12 Oct 2021

¹¹ Ibid

Established Trust in Data

Data governance is a clear priority in almost every organization, regardless of industry. **In one survey, 60% or respondents planned to spend more than \$49K on data governance technology and tools in the next one to two years.**¹²

Report developers, data scientists, and data citizens need data they can trust for accurate, timely, and confident decision-making. But in today's complex data environment, you must contend with dispersed servers and infrastructure, resulting in disparate sources of data and countless data dependencies. You need a complete overview of all your data sources to see how it moves through your organization, understand all touchpoints, and how they interact with one another. You can only completely trust your data when you have a complete understanding of it.

Data lineage provides a comprehensive overview of all your data flows, sources, transformations, and dependencies. With data lineage, you will ensure accurate reporting, see how crucial calculations were derived, and gain confidence in your data management framework and strategy.

Improved Change Management

One of the most critical processes for every business, regardless of size, is change management. Organizations face a variety of change management challenges or obstacles, including:

- A lack of executive support or buy-in
- Mis-alignment due to miscommunication
- Juggling multiple simultaneous changes
- Lack of overall visibility

Nearly all of the benefits mentioned above address these challenges directly. With data lineage, leaders will gain greater visibility into the impact of proposed changes, greater pipeline visibility, and faster migrations (to name a few). With greater trust in the data, getting executive support and communication alignment is easier, and through greater visibility you'll be better equipped to manage multiple simultaneous changes without the pressure of detangling interconnected data dependencies.



¹² Zaloni & Dataversity; The 2022 State of Cloud Data Governance; Michelle P. Knight, Annie Bishop

How to Create Data Lineage and Keep It Up to Date

Now that you know what data lineage is and the business benefits it provides, it is important to understand how data lineage is created and delivered.

We've defined static metadata (the information about tables, fields, columns, business terms, and their data types, locations, quality attributes, tags, etc.) and dynamic metadata (the information about the data's journey from source to target, and all the changes, transformations, and calculations that happen along the way).

In activating this metadata, data lineage creates a map to understand its movements, connections, and dependencies.

Lineage metadata is about logic—instructions, stored procedures, or code in any form. It can be an SQL script, a database stored procedure, a job in a transformation tool, a Java API call, or a complex macro in an Excel spreadsheet. It's essentially anything that moves your data from one place to another, transforms it, or modifies it.

To understand this logic and then build your data lineage map, you need to be able to answer two questions:

1. What is the source of information for building the data lineage map?
2. What is the process for building the data lineage map?

Question 1: Determining the Source of Data Lineage

Data lineage information can be derived from three major sources for three types of data lineage:

1. Data as a source for pattern-based lineage
2. Logs as a source for run-time lineage
3. Code as a source for design lineage

1. Data: Pattern-Based Lineage

This technique reads metadata about tables and columns and uses information about data profiles to create links representing possible data flows based on common patterns or similarities. This could be something like a table or column with similar names and data values. When these similarities are found between columns, they can be linked together in the data lineage diagram.

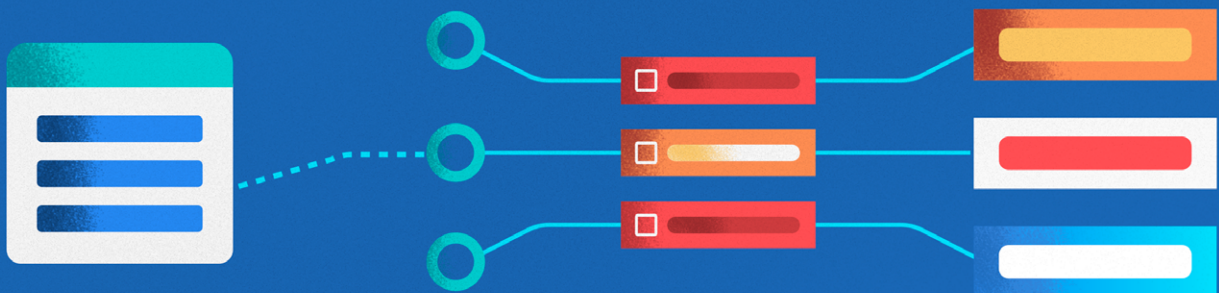
 Advantages	 Disadvantages
It's the best approach for identifying manual data flows happening outside of the system—like copying data to a flash drive, modifying it on another computer, or storing it on a different part of the system.	You may miss important details. Because you're only watching data, this lineage is limited to the database—you're not seeing the application side of your environment or the so-called "transformation logic" of how and where data is being modified.
You don't have to worry about the integration of different system technologies because you're watching the data as the source, rather than algorithms.	The approach is not always accurate. The impact on performance can be significant, and data privacy is at risk.
It's the best approach in cases when it is impossible to read the logic hidden in your programming code because the code is unavailable or proprietary and cannot be accessed.	

2. Logs: Run-Time Lineage

This technique relies on run-time information extracted from the data environment—log files, execution workflows exported by ETL/ELT tools, or any other source with sufficient run-time details. Some data processing engines use a trick called data tagging, where each piece of data being moved or transformed is tagged or labeled by a transformation engine, which then tracks that label all the way from start to finish.

 Advantages	 Disadvantages
It has an operational nature, which is valuable for incident resolution because it provides accurate information about the flow of a specific data element that has been identified as erroneous.	Inaccurate data lineage. Run-time lineage only captures information about recently executed data flows and may fail to capture data calculations and scenarios that are not executed equally or with the same frequency. This can lead to inaccurate or inconsistent lineage, as some parts are either missing or are no longer valid.
It considers different technologies in the data stack (unlike pattern-based lineage), as the format and structure of the logging information vary significantly.	The absence of transformation details. Not everything is or can be logged, especially in the case of more complex algorithms or processing done outside the database/ETL/ELT world. As a result, run-time lineage can often capture only very high-level and generic table-to-table mappings.
Regular expressions, rules, or AI/ML can be deployed to identify relevant parts of log files and derive data flow information.	

Blindly using such metadata poses a big risk for an organization. If used by a data engineer to run impact analysis, it leads to a high probability of incidents when designing and implementing changes in the system and new requirements. If used by a risk analyst to prepare a regulatory report, it leads to inaccuracies in the report and increased risk of (public) incidents and penalties. If used by a data scientist to analyze and prepare data to train a new model, it leads to inherent inequality encoded into the AI/ML algorithm.



3. Code: Design Lineage

This technique looks directly into the code that processes and transforms data records to identify data flows. This is “code” in the broadest sense—such as an SQL script, a PL/SQL stored procedure, an ETL/ELT workflow encoded in a proprietary XML format, a macro in an Excel spreadsheet, a mapping between a field in a report and a database column or table, a Java API, a Kafka stream definition, an XSLT transformation, or a Python algorithm in a Jupyter notebook.

 Advantages	 Disadvantages
The variety of code. The functionality to work with this variety gives design lineage the advantage as the best approach for gaining detailed visibility into your data environment to identify and eliminate data blind spots.	The variety of code. It’s a challenge because parsing and reverse engineering the code is much tougher than parsing log files, and it requires specialized scanners for all supported technologies.
It is the most accurate approach to lineage, with very few false positives. This is critical for incident management as it narrows down the scope of the investigation and makes change management and impact analysis more efficient.	
It accurately detects all data flows, with a close to zero chance of missing any, even those rarely used or not used at all. This is critical for change management processes and impact analysis, as well as for migration projects, privacy programs, and regulatory reporting.	
It can reliably detect indirect data flows—where one data element influences another, even without a direct data lineage connection. This is essential for change management, impact analysis, incident management, migration projects, and regulatory reporting.	
It records details about transformations and calculations used to process data, which is especially important for compliance and regulatory reporting.	

These advantages make design lineage the preferred approach for the most successful vendors and organizations.

Question 2: Understanding the Process

Now that you know the potential sources of your information and lineage techniques, let's look back at question number two: What is the process for building your data lineage map?

There are three major process approaches:

1. Manual Data Lineage Analysis
2. Self-Contained Data Lineage Analysis
3. External Automated Data Lineage Analysis

1. Manual Data Lineage Analysis

Manually resolving lineage usually starts at the top with your people, by mapping and documenting the knowledge in their heads. This process involves interviewing application owners, data stewards, and data integration specialists for information about data movement within your organization. Then, you must begin inputting that information into spreadsheets or other mapping mechanisms so the lineage can be defined.

 Advantages	 Disadvantages
It's the starting point. Manual data lineage analysis is where a lineage project needs to start to be able to gain insight into what is going on across the entire environment. It may be that there isn't any code at all or any permissions to access and profile data directly (especially with legacy systems). In these cases, domain experts—your people—are your only source of lineage.	The lineage cannot be trusted. You're relying on what people are telling you. Their information may be contradictory, missing important details, or simply wrong. This can lead to a situation where you have lineage, but you're unable to use it because it cannot be trusted.
	It's tedious. Manual data lineage analysis uses code as a source, where the code is analyzed by its authors or external resources. This means manually examining the code, comparing column names and reviewing tables and file extracts by hand. Unless you have team members with the requisite skills and expertise in the programs and modules you need to map, manual data lineage analysis may not even be worth attempting.
	It's unsustainable. Due to code volumes, complexity, and the rate of change, manually managed lineage will fall out of sync with the actual data transfers in the environment, and you're back to having data lineage that cannot be trusted.

2. Self-Contained Data Lineage Analysis

This approach uses logs as a source. This approach uses a tool that fully controls your data's movement, its changes, and the entire data processing workflow to give you full insight. It's the preferred choice of ETL/ELT vendors.

 Advantages	 Disadvantages
It's fully automated, so no tedious manual analysis is needed.	The data lineage is limited to the controlling platform—it's self-contained. Anything that happens outside the controlled environment is invisible. More complex components within the environment can be missed. The result is incomplete lineage.
Complete lineage of the entire data processing platform. It provides full insight, control, unlimited access to internal logs, details about executed workflows, and processing instructions.	It is limiting for the majority of data engineering tasks. Organizations using this approach enforce a single data processing platform or prohibit the use of its more complex components, as they'll likely be missed. However, this slows down new development and is limiting and frustrating for data engineers.

3. External Automated Data Lineage Analysis

External automated data lineage analysis is designed with the diversity of the data system environment in mind. It does not require all data processing to happen in one tool or platform. As the name indicates, this approach also offers fully automated data lineage analysis.

 Advantages	 Disadvantages
It doesn't require all the data processing to be on one platform. Unlike self-contained data lineage analysis, external automated data lineage analysis can be done across system platforms, components, and tools.	None
It can use any of the three sources. Using either logs or code as a source for data lineage discovery is most common, but data as a source can be used too. It's also versatile enough to combine sources and approaches.	
Its versatility allows for flexibility. It can be adjusted based on the user's level of understanding and needs.	

External automated data lineage is a powerful tool for gaining full visibility of the data environment, overcoming data blind spots, and taking informed, timely action from your data.

What to Look for in a Data Lineage Solution

Tapping into the true potential of data lineage means automating manual processes, enabling trust in data, and increasing the productivity of your organization for better business outcomes. But in order to do this, you need the right solution with the right tools.

To achieve your goals, the following key data lineage elements must be present:

1. Accurate and Detailed Metadata
2. Semantics and AI
3. Activating Integrations

1. Accurate and Detailed Metadata

We've emphasized the importance of recognizing and capturing the dynamic aspects of data—the transformations, calculations, and movements, all of which represent a type of dependency. These are best represented by data lineage, but without understanding and controlling data lineage, your data management will remain inaccessible.

Dependencies are everywhere and are usually well hidden. There are even indirect dependencies like filtering conditions. Automated discovery is non-negotiable—it's the only thing that can uncover these hidden dependencies.

Another challenge is that dependencies must be mapped very accurately, in detail. Otherwise, the resulting map will contain too many false positives or will miss several critical relations among the data. Without detail and accuracy, any attempt to control dependencies is destined to fail.

2. Semantics and AI

Just mapping dependencies is not enough. To get the most out of your data and maximize insights, you need AI.

Core information about the flow of data and the data journey has to be enriched by its meaning—what does a specific transformation mean, and how does it affect the data?

The ability to answer such questions provides more power and control over dependencies and allows for the deployment of more advanced techniques for automation. To fully deploy AI and other advanced techniques, semantics is key.

The semantic layer of data lineage provides various capabilities:

- The ability to differentiate between different types of dependencies (direct and indirect)
- The ability to understand the evolution of data lineage over a period of time (time slicing and revisions)
- The ability to translate the real data processing code into more high-level, user-friendly expressions



3. Activating Integrations

Historically, metadata catalogs have focused on passively storing static metadata, overlooking its dynamic properties.

In activating metadata, the ultimate task is integrating it into all data management processes, so you can proactively use this knowledge to speed up processes and reduce manual tasks. A data catalog, data privacy, or ETL/ELT tool that has access to detailed, accurate, semantically rich data lineage opens new doors for activating additional metadata.

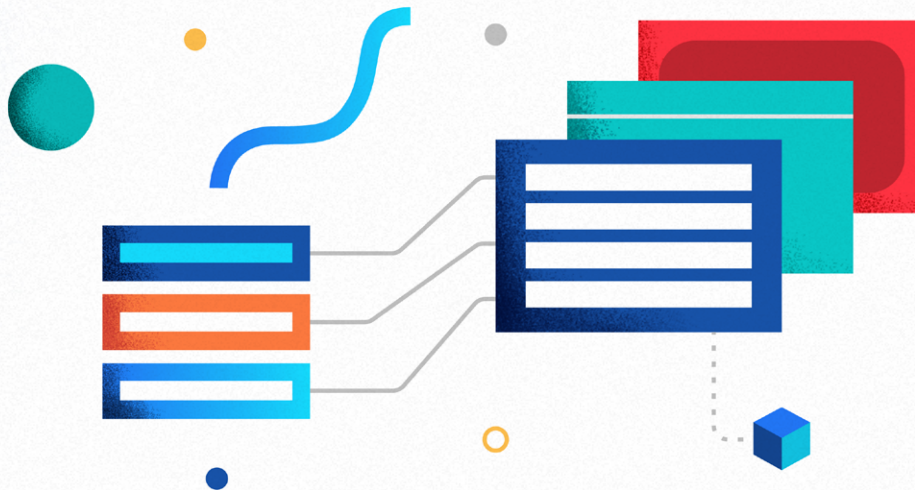
Activating integrations saves time. You won't have to spend hours manually analyzing and extracting data.

Strategies for activating data lineage metadata can differ based on the domain it's being integrated into, but for every domain, you want to ask the same set of questions.

- What processes and tools are currently in use?
- What is still being done manually? Why hasn't it been automated yet?
- How can accurate, detailed, semantically rich metadata help with automation?
- Is there anything that would have a major impact that we are not doing today but we could do if it were automated?
- Is there a way to use automation to redesign and improve an existing process?

This ability to automate is why so many successful organizations deploy enterprise-wide data lineage platforms—to integrate them with other parts of their data infrastructure.

How Manta Can Help with Data Lineage



As a modern organization, you process high volumes of data. Your IT environment will only increase in complexity, and your IT team is struggling to keep up. You need to get your data systems under control and find a way to stay efficient despite this skyrocketing complexity. **You need data lineage, and the bare minimum ‘good enough’ data lineage that comes with your data catalog can lead to costly updates later.**

Manta has helped nearly one hundred organizations realize the benefits of data lineage. We bring intelligence to metadata management by providing an automated solution that helps you drive productivity, gain trust in your data, and accelerate digital transformation.

The Manta platform includes unique features to make the most value out of your lineage, with more than **50 out-of-the-box, fully [automated scanners](#)**. In addition, **Manta works alongside the most popular data catalogs**; our platform integrates with catalogs like [Collibra](#), [Informatica](#), [Alation](#), and more.

Don't wait. Realize the benefits of automated data lineage today.

➤ **Schedule a demo**
with a Manta engineer to learn more.