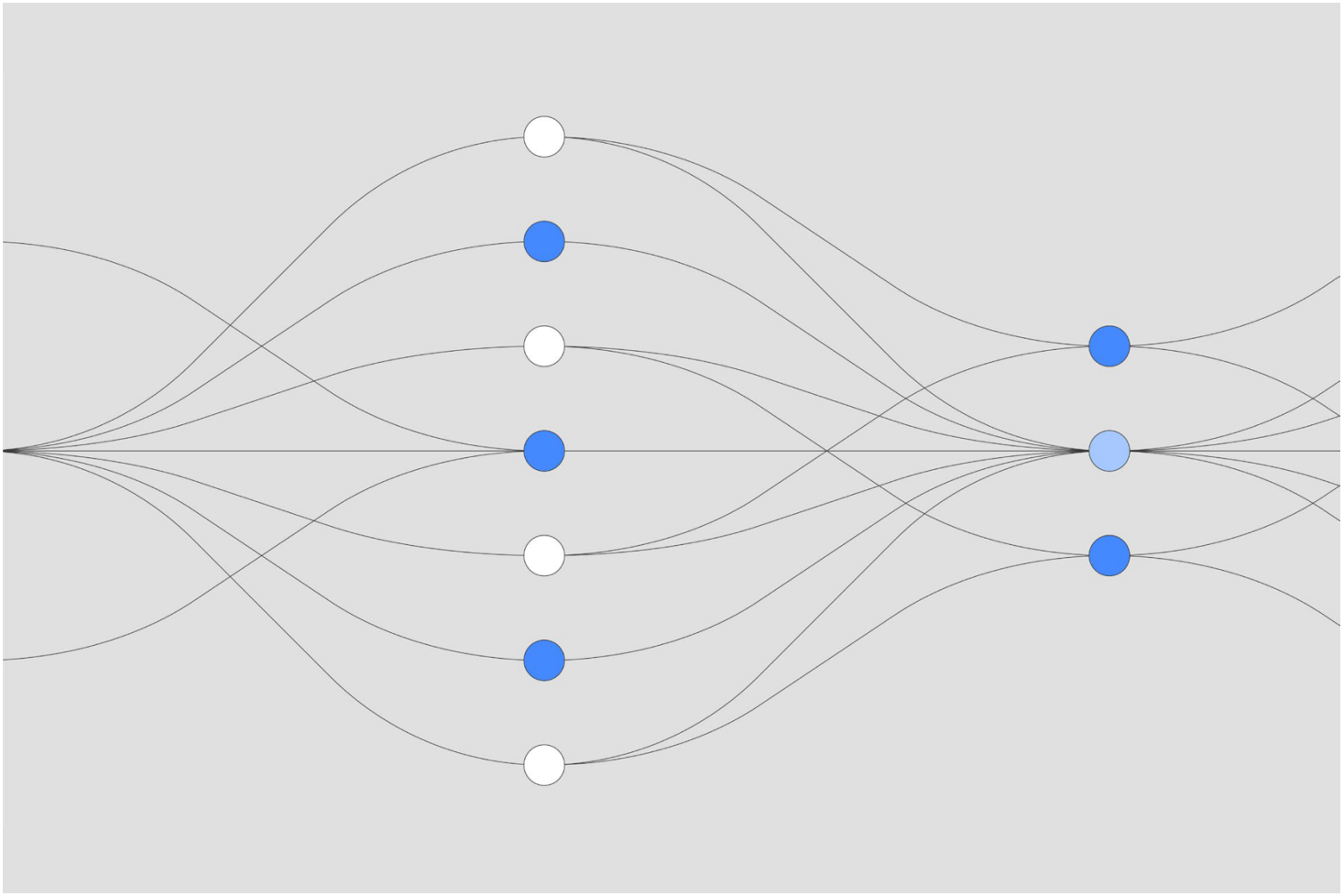
IBM watsonx.data intelligence
Data Lineage



A comprehensive guide to data lineage

Managing your evolving data infrastructure while retaining trust and visibility into pipelines

Table of contents

01
The drawbacks of not managing your data

Every minute of every day, new data flows throughout your organization. With the high volume of data you're collecting, you need to be able to act and make informed decisions—quickly. But the complexity of today's data systems can create blind spots in the data environment. And unfortunately, with limited visibility, you only have limited control.

When you have data from so many sources and systems, multiple applications and microservices, all interconnected but extremely diverse, it's a challenge to extract meaningful insights. It can be difficult for any one person or team to have complete visibility over your entire environment. And you can lose sight of your primary goal: to extract more value from data. These data blind spots and complexities can lead to challenges and consequences.

Slower delivery of predictive insights

In many companies, data remains inaccessible because:

- It's stored in an unusable format or inaccessible location.
- It cannot be traced and therefore cannot be trusted.
- It's difficult to determine what data is important.
- 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 the delivery of new features.

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. 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]

Current data observability tools are still primarily reactive—organizations find bugs after there's already been an incident instead of preventing them. This is concerning because even a single data incident can damage your organization financially or reputationally.

Decreased trust in reports and insights

Gartner predicts that 80% of data and analytics (D&A) governance initiatives will fail by 2027.[2] If you can't fully explain how data was collected or verify its origins, you can't answer basic questions about the data's authenticity or trust the data for better customer outcomes. Again, this can cause multiple business impacts and frustration.

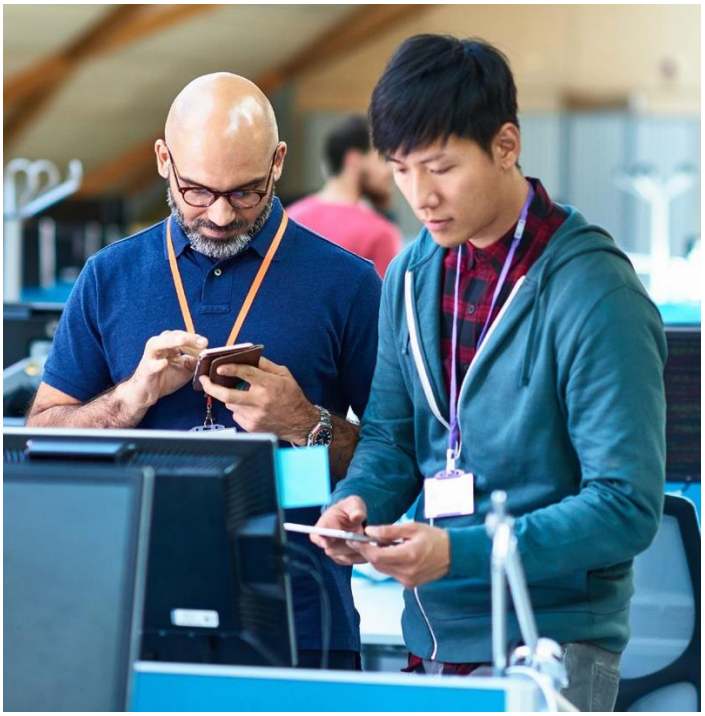
Shortage of data engineering talent

Engineering talent can hard to find. In a 2023 survey, 54% of tech team managers reported data engineering talent shortages. In fact, data engineers are the most sought-after skills with this talent gap ranking first before enterprise architects, software engineers and technical architects.[3]

Due to the growing complexity of the data stack, data engineers have become more critical than virtually any other business role because they're responsible for overseeing data pipelines and integrated data structures. These tasks require a larger skill set which makes good data engineers harder to find—and even harder to keep.

Increased risk of noncompliance

Regulators are intensifying enforcement 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 fact, in April 2022, the Digital Services Act threatened big tech with a 6% revenue penalty for having illegal content live on their sites. And the EU AI Act demands stringent AI model documentation, including information regarding the provenance of training data.



02

Overcome the complexity of the way data travels through your enterprise

There are seemingly endless opportunities when you're able to tap into the full potential of your data. To do that, activating metadata is key. Gartner's definition of active metadata in their "Market Guide for Active Metadata Management" touches on several key aspects of metadata:[4]

- **Continuous access:** Metadata isn't collected once per year or per month, it's a continuous process.
- **Connecting dots:** Metadata is constantly processed to distill information and knowledge from all the signals and noise. With the right feedback loop, your system can get smarter over time, collecting and learning.
- **Actionable:** All the metadata-derived insights are not locked into a silo but rather delivered in the form of recommendations, warnings and notifications to humans, systems or applications that may need it.
- **Embedded:** Actionable information and knowledge are integrated into the processes that humans and machines perform. Users aren't forced to search for the insights. Instead, active metadata comes to them—when and where they need it.

According to Gartner, active metadata capabilities will expand to include monitoring, evaluating, recommending design changes and orchestrating processes in third-party data management solutions. Gartner also predicted 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%.[4]

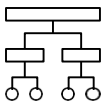
However, by activating existing metadata with automation and intelligence, data lineage can provide the visibility and control you need to help you become more aware of your data management and proactive in your data usage. But what is data lineage? Traditionally, it has been seen as a way of understanding 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. But data lineage is actually much more.

It represents a detailed map of all direct and indirect dependencies between the data entities in your environment. Why is this so important? This 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. It can also assist in providing ethical, compliant and efficient data management processes.

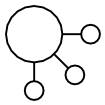
For example, a detailed dependency map can tell you:

How changing a bonus calculation algorithm in the sales data mart will affect your weekly financial forecast report

Where heavily regulated data is being used and for what purpose



What is the best subset of test cases to 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

In short, data lineage helps reduce complexity in your data environment by providing a complete picture of your data landscape, so you can tame data chaos and optimize data for valuable insights. It helps speed up the delivery of predictive insights by automatically generating comprehensive data flow visualizations. It also provides IT teams with high levels of observability without expending large amounts of manual intervention. This allows teams to be truly proactive so they can catch incidents before they happen.

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

Organizations that invest in data lineage can overcome the shortage of data engineering talent, allowing the IT team to refocus their efforts on tasks that can't be automated. Plus, whether you need to comply with the GDPR, HIPAA, Sarbanes-Oxley or the FDA, data lineage provides a complete overview of all regulated data processed by your organization. This information helps you prepare for audits and avoid penalties for noncompliance.

In conclusion, data lineage helps you leverage metadata to better manage, consume and optimize your data—even when working with overly complex data environments. These actions can lead to immediate and long-term business benefits.

03
The business benefits of data lineage

As data becomes increasingly decentralized and diverse, organizations need comprehensive visibility into its movement and usage. You can use this visibility to maintain control, help ensure compliance and boost overall operational efficiency. Data lineage is emerging as a game-changer in the realm of data management, empowering organizations to navigate the complexities of ever-growing datasets and derive maximum value from their investments. With data lineage, businesses can achieve higher value, which can lead to improved incident prevention, accelerated migrations, enhanced regulatory compliance and increased trust in data.

Automated impact analysis for improved incident prevention

Impact analysis is crucial. It helps illustrate the consequences of an application design or coding change that can affect downstream information assets. Decisions to make application changes can impact everyone from customers to stakeholders to salespeople. During impact analysis investigations, data lineage helps foster an environment where reports and data can be trusted. Data lineage provides that level of reliability.

Greater data pipeline visibility for faster incident resolution

As mentioned earlier, there are countless threats to your organization's bottom line. Whether it's a successful ransomware attack or a poorly planned cloud migration, identifying the problem before it can cause damage is always less expensive. That's why data pipeline visibility is so important. It not only helps protect your organization but also your customers.

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

Improved regulatory compliance support

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, CCAR and the EU AI Act just to name a few.

All of these regulations require accurate tracking of data. If called upon, your organization must be able to answer questions like these:

- Where does the data come from?
- How did the data get there?
- Can you prove it with up-to-date evidence whenever necessary?
- Do you need weeks or months to complete a report?
- Is the report you complete 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 if you face a regulatory compliance audit.

Faster and more efficient migrations

If you've ever been involved in the migration of a data system, you know how complex the process can be. Many enterprises are discouraged from adopting the cloud because of migration costs. The process can be so complex and expensive because every system consists of thousands or millions of interconnected parts, and it's impossible to migrate everything in a single step.

Dividing the system into smaller chunks of objects (such as reports, tables, workflows and so forth) can make it more manageable. But this tactic can pose 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 all dependencies are documented. Another benefit of lineage for migrations is identifying assets that are still being used so you only migrate those and avoid moving assets that are no longer necessary.

Retention of data engineering talent

Data engineers, developers and data scientists continue to be some of the fastest-growing and hardest-to-fill roles in tech. The shortage of data engineering talent is growing bigger, and so is the complexity of data systems. To retain talent, it's crucial that you eliminate as many manual tasks as possible—like chasing data incidents, assessing the impacts of planned changes or answering the same questions about the origins of data records.

Data lineage can help automate routine tasks and enable self-service in many situations, 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 allows faster onboarding of data engineers to integrate new or less-experienced engineers into the role without impacting the stability and reliability of your data environment.

Established trust in data

Data governance is a clear priority in many organizations. 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 data moves through your organization, to understand all touchpoints and recognize how they interact with one another. You can only completely trust your data when you have a total understanding of it.

Data lineage provides a comprehensive overview of all your data flows, sources, transformations and dependencies. With data lineage, you can expect accurate reporting, easily examine 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
- Misalignment due to miscommunication
- Juggling multiple simultaneous changes
- Lack of overall visibility

With data lineage, leaders gain greater visibility into the impact of proposed changes, benefit from greater pipeline visibility and achieve faster migrations. 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 identifying and detangling interconnected data dependencies.

04
Getting the most out of your data with data lineage

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 to do this, you need the right solution with the right tools. Here is where Data Lineage, as part of IBM watsonx.data intelligence, delivers most value.

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

Accurate and detailed metadata

Recognizing and capturing the dynamic aspects of data—the transformations, calculations and movements, all of which represent a type of dependency—are vitally important for successful data management. These attributes are best represented by data lineage. Without understanding and controlling data lineage, your data management objectives will likely remain unattainable.

Dependencies are everywhere in your data stream—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.

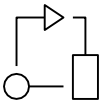
Another challenge is that dependencies must be mapped accurately and in extreme detail. Otherwise, the resulting map may contain too many false positives or will miss too many critical relationships among the data. Without detail and accuracy, any attempt to control dependencies is destined to fail.

Semantics and AI

To get the most out of your data and maximize insights, you need to add the power of AI.

Core information about the flow of data and the data journey must 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 can give you the ability to:



Differentiate between different types of dependencies (direct and indirect).

Understand the evolution of data lineage over time (time slicing and revisions).

Translate the real data processing code into more high-level, user-friendly expressions.

Activating integrations

Historically, metadata catalogs have focused on passively storing static metadata while 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 an 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?
- How can accurate, detailed, semantically rich metadata help with automation?
- Is there anything that would have a major impact that you're not doing today, but 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.

IBM brings intelligence to metadata management by providing an automated solution that helps you drive productivity, gain trust in your data and accelerate digital transformation. By creating a comprehensive view of data flows and dependencies, Data Lineage in IBM watsonx.data intelligence helps organizations identify potential issues early for smoother AI integration and a heightened focus on regulatory compliance. Embrace the future of data management and unlock the full potential of your AI initiatives with watsonx.data intelligence's data lineage capabilities.

05

What to consider when getting started with data lineage

To 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?

Automated approaches to data lineage

According to the way it's achieved, data lineage can be classified in the following ways:

- Pattern-based lineage
- Runtime lineage
- Code-based or reverse engineered lineage

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. However, the disadvantages of this approach are that you may miss important details, and it isn't always accurate. The impact on performance can be significant and can put data privacy at risk.

Runtime lineage
This technique relies on runtime information extracted from the data environment—log files, execution workflows exported by ETL/ELT tools, or any other source with sufficient runtime 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. Then it tracks each label from start to finish. The disadvantages of this technique are twofold: inaccurate data lineage because recently executed data flows can fail to capture scenarios that aren't executed with the same frequency, and the absence of transformation details.

Code-based or reverse engineered 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. However, this technique is a challenge because of the overwhelming variety of code. Parsing and reverse engineering all these types of code is far more difficult than parsing log files and requires specialized scanners for all of the supported technologies.

Understanding the process

There are three major process approaches for building your data lineage map:

1. Manual data lineage analysis
2. Self-contained data lineage analysis
3. External automated data lineage analysis

Manual data lineage analysis
Manually resolving lineage usually starts at the top with your people mapping and documenting the knowledge in their heads, or by tedious and labor-intensive human review of existing scripts and code. 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. The disadvantages of this process are the unreliability of data lineage (because of the human factor), its laborious and time-consuming nature, and the fact that it's unsustainable.

Intraplatform data lineage analysis
This approach uses a single tool's native platform as a source. It takes advantage of the tools that control the movement of your data, its changes and the entire data processing workflow to give you full insight. It's the preferred choice of ETL/ELT vendors. Here, the disadvantages are the fact that data lineage is self-contained (limited to the controlling platform), and it is also limiting for most of data engineering tasks.

Independent automated data lineage analysis
External automated data lineage analysis is designed for both diverse data system environments and diverse sources of lineage. 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 that can synthesize these sources into a single view of the environment. There are few if any disadvantages for this approach.

06

Thinking about the future of data management with data lineage

Poor data management can lead to numerous challenges for organizations, including inaccurate decision-making, increased risk of data breaches and wasted resources. Moreover, poor data management can hinder innovation and growth because it prevents organizations from leveraging their data effectively.

Data lineage is crucial in managing complex data environments because it allows organizations to understand the origin, movement and transformations of data throughout the entire lifecycle. By tracking data lineage, IT teams can identify bottlenecks, detect anomalies and help ensure proper data flow, ultimately improving overall data management and facilitating better decision-making.

IBM watsonx.data intelligence, through its data lineage capabilities, offers a powerful solution to these challenges. By providing a comprehensive map of all data flows, sources, transformations and dependencies, data lineage enables organizations to gain total visibility into their data infrastructure. This visibility allows organizations to identify and address issues related to data quality, trace the origins of sensitive data and help ensure compliance with regulatory requirements.

As a result, organizations can make informed decisions, minimize risks and optimize their data resources.

Beyond its functional benefits, the Data Lineage feature within IBM watsonx.data intelligence can also lead to significant business advantages. By improving data governance and trust, organizations can enhance their brand reputation, build greater customer confidence and foster long-term loyalty. Furthermore, the solution can help organizations unlock the full potential of their data by facilitating advanced analytics, machine learning and artificial intelligence initiatives.

In conclusion, addressing poor data management requires a robust approach that includes solutions like data lineage. By investing in high-quality data management solutions, organizations can overcome common challenges, unlock new opportunities, and ultimately drive sustainable growth and success.

Drive productivity, gain trust in your data and accelerate digital transformation with IBM watsonx.data intelligence's data lineage capabilities.

Learn more →

1. Number of user accounts exposed worldwide from 1st quarter 2020 to 4th quarter 2023, Statista, January 2024.
2. Gartner Predicts 80% of D&A Governance Initiatives Will Fail by 2027, Due to a Lack of a Real or Manufactured Crisis, Gartner, February 2024.
3. Looking Forward, Looking Back: The Digital Leadership Report 2023, Nash Squared, November 2023.
4. Market Guide for Active Metadata Management, Gartner, November 2022.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
April 2025