IBM Data and AI /
Data Governance

IBM data lineage

Gain deeper visibility into the provenance of your data and its journey from source to end use

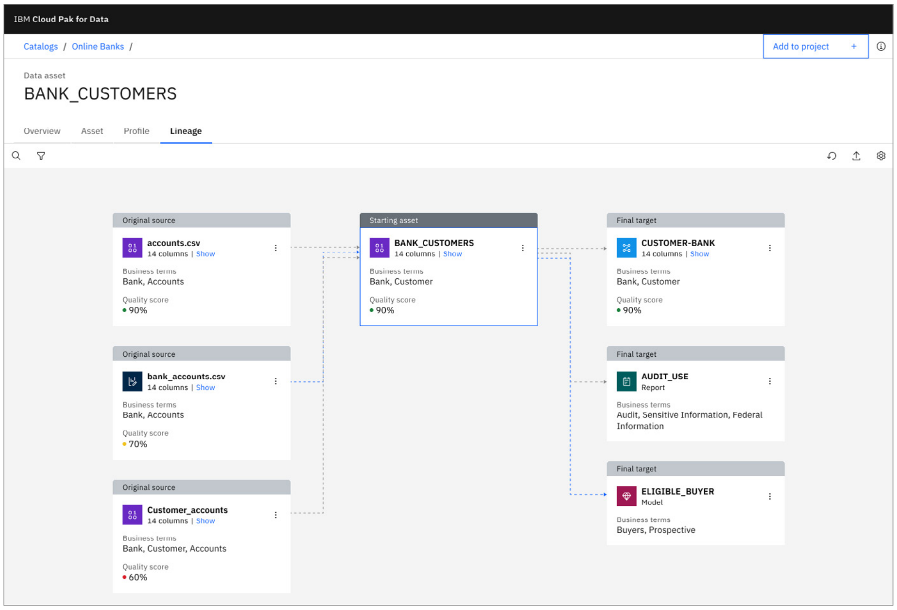
Highlights
Enable regulatory compliance, conduct impact analysis, build trust in data

Deliver deeper data lineage and faster time to value

Simplify data lineage with automated scanning of third-party data flows

Data proliferation has made it challenging to trust your data. With so much data, how can you know where it came from, how it has changed, and who's using it? Poor understanding of the data journey from source to end use can have considerable consequences. As you integrate AI into your workflows, a lack of transparency around the journey of data fed into AI models can hinder auditability. A lack of visibility into where sensitive data resides can increase risks of non-compliance with data privacy and industry regulations. Data engineers could spend disproportionate time to analyze impact of planned data changes.

To overcome these challenges, you need a map that simplifies the understanding of a dataset's journey from its origin to its end use, with specific details on how data is transformed, and by whom, along the way. Data lineage is a visual representation of a dataset's journey from its origin to end use. It has evolved into the primary enterprise tool for understanding the flow of data and the contribution of each person and program throughout that data's lifecycle.



Data asset
BANK_CUSTOMERS

Original source
accounts.csv
14 columns | Show
Business terms
Bank, Accounts
Quality score
● 90%

Original source
bank_accounts.csv
14 columns | Show
Business terms
Bank, Accounts
Quality score
● 70%

Original source
Customer_accounts
14 columns | Show
Business terms
Bank, Customer, Accounts
Quality score
● 60%

Starting asset
BANK_CUSTOMERS
14 columns | Show
Business terms
Bank, Customer
Quality score
● 90%

Final target
CUSTOMER-BANK
14 columns | Show
Business terms
Bank, Customer
Quality score
● 90%

Final target
AUDIT_USE
Report
Business terms
Audit, Sensitive Information, Federal Information

Final target
ELIGIBLE_BUYER
Model
Business terms
Buyers, Prospective

Gain visibility into your data's journey from source to end-use

Enable regulatory compliance, conduct impact analysis, build trust in data
Data lineage is essential for modern data management and has a wide range of use cases. It's a required aspect of regulatory compliance and helps identify the origins of sensitive data, the various locations where it's stored, who can access it and which data should be anonymized. Due to new regulations like Europe's AI Act, you need to keep detailed notes on what data is used within AI models. Data lineage allows auditors to understand what training was used to train a model. The 2023 CEO study conducted by IBM® found the number one barrier to generative AI adoption is concerns from leadership about the lineage or provenance of data.[1]

Enterprises are constantly implementing changes to their data architecture and pipelines. Without a data lineage, it can be difficult or impossible for you to assess the impact of planned changes. Research from IBM shows that fixing a bug in production is 15 times more expensive than fixing it during the implementation phase.[2] Data lineage gives you insight into the downstream impacts of these changes before potentially costly bugs are introduced.

Data lineage can also enable analysts and data consumers to conduct root cause analysis by diagnosing issues and discrepancies in data and reports, offering you the power to speed up migration processes while undergoing digital transformation. Engineers can also gain visibility into which architectural components must be migrated at once and which don't need to be migrated at all.

Only when analysts and data scientists have a complete understanding of data can they rely on it for confident decision-making. Data lineage is a critical capability of modern data governance to deliver trust in the data used for analytics and AI by providing visibility into your data's provenance and its end-to-end journey.

Deliver deeper data lineage and faster time to value

IBM Knowledge Catalog with Manta, an IBM Company delivers a business-friendly data lineage native within the data catalog while allowing users to drill deep into the technical details data engineers need. View historical versioning of lineage to understand how pipelines change over time and quickly resolve issues. IBM can provide quicker time to value not only through the automation of previously manual processes, but also through the ability to more rapidly answer questions about whether certain data is trustworthy. IBM data lineage enables you to create a complete end-to-end data lineage for full understanding, observability and control of your data.

Simplify data lineage with automated scanning of data flows in third-party tools

With data lineage capabilities bolstered by the acquisition of Manta, IBM can help organizations ease the amount of manual effort necessary for robust data lineage. This is achieved by providing scanners for the automated discovery of data flows in third-party tools like Power BI, Tableau and Snowflake. This information is then automatically scanned into IBM Knowledge Catalog's Data Lineage UI and becomes visible alongside data quality, business terms and other metadata previously available to IBM Knowledge Catalog users.

Automated data lineage helps you avoid the need for the time-consuming, manual creation of a data lineage. This kind of manual process can often be tedious, contain contradictory or missing information, and lead to teams relying on unsound lineages to make critical decisions.

Conclusion

The acquisition of Manta will allow IBM to provide quick time to value not only through the automation of previously manual processes, but also through the ability to answer questions more rapidly. This will empower you to build organizational trust in your data.

Manta, an IBM company is available as an integrated solution with IBM Knowledge Catalog or as a standalone tool today. When integrated with IBM Knowledge Catalog, Manta, an IBM Company allows users to understand technical data lineage information alongside critical business metadata created in the data catalog. Examples include business terms and data quality scores.

The capabilities delivered by Manta, an IBM company will be available as an add-on to IBM WatsonX™ to display lineages for data stored in watsonx.data and AI models created in watsonx.ai.

Why IBM?

IBM has been named a leader in multiple analyst assessments, including the Gartner Magic Quadrant for Data Quality solutions, the Forrester Wave: Enterprise Data Fabric Q2 2022 and the Gartner Magic Quadrant for Data Integration Tools.[3, 4, 5]

IBM offers integrated data governance and data integration alongside automated data lineage, data quality, data privacy and entity resolution—all parts of IBM Cloud Pak® for Data, which is an open and extensible data and AI platform that can be deployed on any cloud. The integration of these capabilities within a unified environment helps streamline data governance tasks to accelerate understanding of what data means, where it comes from and how it relates to other assets.

For more information

To learn more about IBM data lineage, contact your IBM representative or IBM Business Partner. Watch this webinar to see data lineage capabilities from Manta, an IBM company in action. Visit the IBM data governance page to explore how IBM enables the creation of a governed, compliance-ready data foundation.

1. Be a creator, not a consumer, IBM.
2. IBM System Science Institute Relative Cost of Fixing Defects, IBM.
3. Gartner Magic Quadrant for Data Quality Solutions, Gartner, 2022.
4. Forrester Wave: Enterprise Data Fabric Q2 2022, Forrester, 2022.
5. Gartner Magic Quadrant for Data Integration Tools, Gartner, 2022.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
November 2023

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

It is the user’s responsibility to evaluate and verify the operation of any other products or programs with IBM products and programs.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON INFRINGEMENT.

Statement of Good Security Practices: No IT system or product should be considered completely secure, and no single product, service or security measure can be completely effective in preventing improper use or access. IBM does not warrant that any systems, products or services are immune from, or will make your enterprise immune from, the malicious or illegal conduct of any party.

The client is responsible for ensuring compliance with all applicable laws and regulations. IBM does not provide legal advice nor represent or warrant that its services or products will ensure that the client is compliant with any law or regulation.