MANTA

WHITEPAPER

Supporting Data Quality Processes with Automated Data Lineage

WHITEPAPER

Supporting Data Quality Processes with Automated Data Lineage

Abstract

Data quality is an age-old problem for medium and large enterprises. With enormous swaths of data flowing between numerous complex systems, it is easy for quality issues to go unnoticed. On top of that, the way businesses transform, interpret, select, and move data can introduce new quality issues to datasets. Despite this, many organizations continue to rely on inefficient and error-prone manual tasks to support data quality processes.

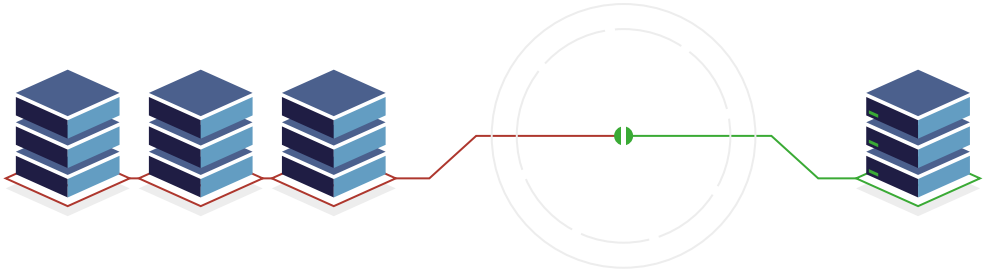
Fortunately, there is a better way to ensure data quality: automated data lineage. Automated data lineage maps out organizations' complete data flows from start to finish, eliminating the negative consequences of using poor-quality data to drive business decisions. With MANTA's automated data lineage, businesses can develop a high level of trust in the quality of their data.

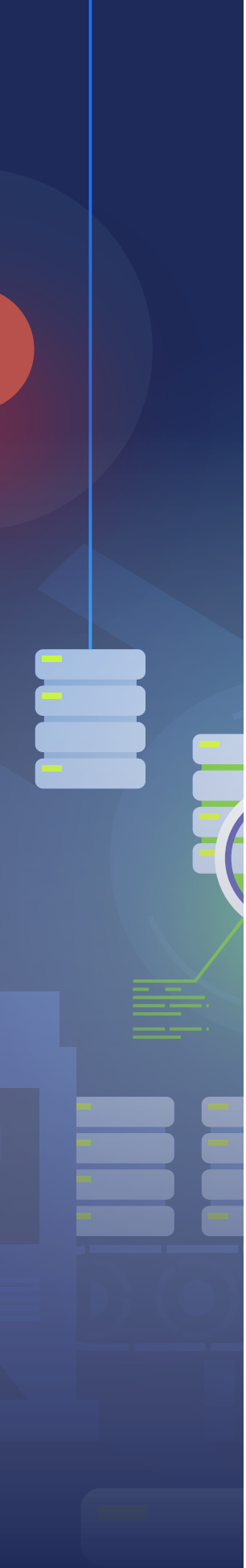
Introduction

Data quality measures how fit data is for its intended purpose. Several attributes, including data accuracy, completeness, consistency, reliability, timeliness, and integrity, contribute to whether data is high or poor quality. Poor data quality can lead to bad business intelligence, delayed cloud migrations, inaccurate consumer insights, regulatory compliance issues, and other negative outcomes.

For example, a highly regulated organization with poor data quality may struggle to prove to regulators that their data is trustworthy—a serious oversight that could lead to fines, investigations, shutdowns, and reputational damage. Alternatively, a supply chain management team relying on inaccurate sales data may underestimate future customer demand and order insufficient inventory.

If either of those organizations had traced their data's lineage, they could have flagged data quality issues before they became a problem. This rings true for all data quality projects, which depend on traceability to ensure proper data usage and storage.





The Business Applications of High-Quality Data

Data quality serves many purposes within a business. The longer poor data is allowed to flow throughout systems without intervention, the more problems it can cause. The following are just some of the business use cases of high-quality data.

Data Integration

Data integration is necessary whenever businesses come together in the form of a merger or acquisition. If either organization is suffering from poor data quality, this integration process suffers and becomes more difficult to manage. Data movement can also result in data being manipulated or changed in a way that impacts quality.

Many mergers and acquisitions occur in highly-regulated industries, making it even more important for organizations to ensure data quality to avoid compliance failures. If the acquiring company does not do due diligence, they can become liable for poor quality data that fails to meet regulatory requirements.

Data Migration

Successful data migration away from legacy systems also depends on quality. Assuring data quality before migration is underway can help prevent migration project failures and delays. On the flip side, poor data quality can cause migration issues such as poor business intelligence or business process disruptions.

Supply Chain Management

High-quality data can help businesses manage their supply chains by reliably predicting future demand, enabling businesses to respond accordingly with a business plan. When quality is missing, data-driven decisions become inaccurate and hinder these efforts.

Marketing Initiatives

First-party data is increasingly becoming the go-to method of gathering consumer insights. The accuracy of first-party data can greatly impact the outcome of marketing campaigns. In fact, Forrester found that 26% of marketing campaigns suffered from poor data quality in 2019. It also found that 87% of marketing decision makers said that high-quality data was important to the success of their brand's marketing performance.

Regulatory Compliance

Governments around the world are more frequently establishing regulations pertaining to data usage and storage, with two examples being the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). If data is disorganized, poorly maintained, or otherwise untrustworthy, it becomes more difficult to demonstrate compliance with data regulations.

Roadblocks to High Data Quality

Despite the critical importance of high data quality, businesses often fail to achieve it. In fact, most business leaders don't know the true cost of poor data quality for their organization. However, they are aware that a data quality problem exists. According to Experian, over half (55%) of business leaders lack trust in their organization's data assets. This is particularly true in the wake of the pandemic, with 93% of businesses experiencing data management issues as a result of COVID-19.

What's the holdup when it comes to quality? There are several roadblocks to achieving data quality, including data movement, data transformation, data interpretation, data selection, and data inaccuracies.

Key Challenges

1. **Data movement**
   Whenever data moves, something can go wrong. Data quality issues can become exacerbated if the new system it is moving to has more complex features and a stricter data model.

2. **Data transformation**
   Aggregation, manipulation, and ETL change the nature of data. This can cause inaccuracies that affect data quality.

3. **Data interpretation**
   Even if the data itself is accurate, getting the wrong results during data interpretation can make it impossible to understand what it is actually saying.

4. **Data selection**
   Not all data is equal, and not all data subsets carry the same weight. Determining which data matters and how to categorize it influences data quality.

5. **Data inaccuracies**
   Broken, missing, incorrect, or redundant data result in poor data quality.

Automated Data Lineage Eliminates Data Quality Challenges

Data lineage is a detailed map of data's journey throughout an organization's data processing systems. This encompasses where data comes from, where it is flowing, and what happens to it along the way. With data lineage, businesses gain a big-picture understanding of how data moves across systems and teams as well as concentrated visibility into smaller components and details.

Data lineage reveals the path of data through pipelines, databases, models, and analytical tools, revealing where quality issues might arise. With this visibility, quality concerns can be addressed before they become problematic. Yet despite the role data lineage plays in achieving data quality, most businesses are not actively using data lineage. In fact, a 2020 O'Reilly survey found that nearly 80% of organizations do not manage data lineage.

Businesses that do track data lineage have historically done so manually. Manual data lineage tracking generally begins with documenting what the people within an organization know. This involves interviewing data specialists and others with knowledge of how data moves within the business. The lineage is then defined using spreadsheets or other mapping mechanisms that map out the information uncovered in these interviews.

There are a few reasons why this is not the best approach. For starters, the process can take days, weeks, or even months, depending on the amount of data that needs to be mapped out. Manual lineage is also more prone to errors—data might be missing or contradictory, making the map incomplete or misleading.

Automation is redefining the lineage process. With automated data lineage, visibility into how data flows becomes instantaneous and accurate. The result is that it becomes significantly easier for businesses to locate where reporting errors originated and whether data quality has impacted data flows historically.

Supporting Data Quality Processes with MANTA

Data quality has historically been a pain point for businesses, but it doesn't have to be. With automated data lineage, error-prone and time-consuming manual data quality processes can be eliminated for good.

MANTA is the most accurate automated data lineage solution on the market for achieving high data quality. The world-class data lineage platform automatically scans mid-level company and enterprise data environments and builds a powerful map of all data flows and direct and indirect dependencies that can be used by technical and non-technical teams. With MANTA's self-service platform, organizations can trust that their data, information, and reports are accurate.