

Whitepaper

# Three case studies of Data Observability

How data engineers leverage Collibra to catch bad data



## Executive summary

Data engineers work on the infrastructure required to deliver reliable and high-quality data to data consumers. As the data quality fundamentals shift with data volume, sources, storage, and the desired state, data engineers find it challenging to deliver healthy data pipelines and products.

Data observability monitors data quality and reliability of data pipelines and rapidly helps remediate anomalies to deliver reliable and trusted data products. The three case studies in the paper describe how Collibra Data Quality & Observability helps deliver data health.



Case 1: Delivering trusted data with auto-generated rules

Case 2: Managing data lake health efficiently

Case 3: Accelerating cloud data migration

# What data engineers do

Data-driven organizations invest heavily in analytical tools and other resources for precise insights. Even though the analytical insights can only be as good as the input data, the behind-the-scenes data handling does not get the same attention. The large volumes of raw data arriving from various sources are messy. They are riddled with missing, duplicate, and inconsistent records. They also come in different formats or with mismatched attributes. Getting high-quality, relevant data from these chaotic silos is a mammoth task, which is usually overlooked and underappreciated.

Data engineers perform this complex task of making trusted data accessible to data analysts, data product owners, data scientists, and business analysts. They create platforms and pipelines to transform and transport data to data consumers. They routinely carry out data integrations in databases, data lakes, and data warehouses.

Data migration is another regular task for data engineers, along with optimizing the infrastructure for scalability. They also frequently perform feature engineering to improve the performance of data projects. In short, data engineers focus on the reliability and performance of the end-to-end data ecosystem. They are accountable for all the piping and plumbing needed to get data from its original state to the desired state.

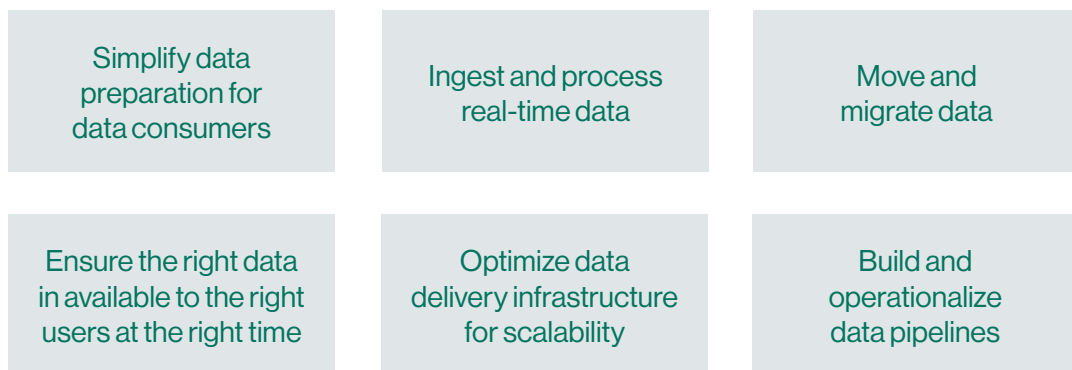


Figure 1. What does a data engineer do?

# Challenges with shifts in data quality fundamentals

Changes in the way data arrives have pushed organizations to rethink data quality. It is no longer a static measurement of data accuracy or validity. It is also not about finding and fixing data errors. The shifts in data quality fundamentals are seen in the following factors:

- Storage and applications moving to the cloud necessitate high-quality data pipelines with minimal data downtime
- High data volumes and increased speed of data arrival require identifying data issues and assuring data quality in near-real-time
- Diverse sources, formats, and shapes of data do not provide any static baseline for data quality rules, demanding quick rule adaptability
- Complex transformations, such as data enrichment and reverse ETL, can intensify quality issues
- Multiple delivery channels and expanded use cases add to the pressure for the timely delivery of reliable, high-quality data

It is quite clear that the traditional data quality approach does not align with these shifts. Gartner notes that a third of the organizations report difficulties in promoting data and analytics initiatives to production, despite massive investments in data delivery initiatives.<sup>1</sup>

**"Unfortunately, I spend 70% of my day identifying data issues and triaging pipeline failures..."**

A data engineer in a Fortune 500 company

---

<sup>1</sup>Gartner Research (Dec 2019): [Data Engineering Is Critical to Driving Data and Analytics Success](#)

Data engineers who create robust data pipelines know these difficulties are related to excessive manual tasks. They often spend 70% of their time identifying data issues and fixing broken pipelines. Though fixing at one place cannot stop the issues from reaching the downstream applications.

By 2025, 463 exabytes of data will be created each day globally, which means huge amounts of data in different formats and time frames to process, thereby resulting in a significant increase in the data engineering effort.

[Source: Forbes \(Jan 2021\): Will Data Engineering Efforts Reduce In The Future?](#)

While they use traditional data quality tools to automate quality checks and notify of failures, continuously pouring diverse data is still a challenge. Even with leveraging data governance for data ownership and automation, data engineers find it difficult to:

- Get real-time visibility into the health of enterprise data
- Proactively identify potential issues
- Enable fixing issues at the source
- Scale quickly for high volumes and faster arrival of data

The common factor here is monitoring data health in real-time to predict errors before they can happen, and prevent them from propagating downstream. A task that requires much more than the traditional rule-based find-and-fix approach.

# Data observability empowers data engineers to address the challenge of data trust

There is no use if data feeding the operational and analytical tools is unreliable or old. But building trust in data is not a one-time, isolated activity. It demands continuous monitoring of data, profiling of data, predicting errors, and proactively preventing them. For this, data engineers need to focus on assuring the quality of data at rest (datasets) as well as data in motion (pipelines) before the errors can affect operations.

Gartner notes that data observability empowers data engineers to provide accurate and reliable data to consumers and applications within expected time frames. It helps IT and business leaders to have a degree of control over data usage and capacity planning.

What they need is data observability to ensure the quality of data as it moves through the enterprise systems. It broadens the focus to include data lineage, context, business impact, and quality to track the health of enterprise data systems. The visibility into data movement from ingestion to consumption across applications and infrastructure provides enormous opportunities to improve data trust.

Source: [Gartner Quick Answer: What Is Data Observability?](#)

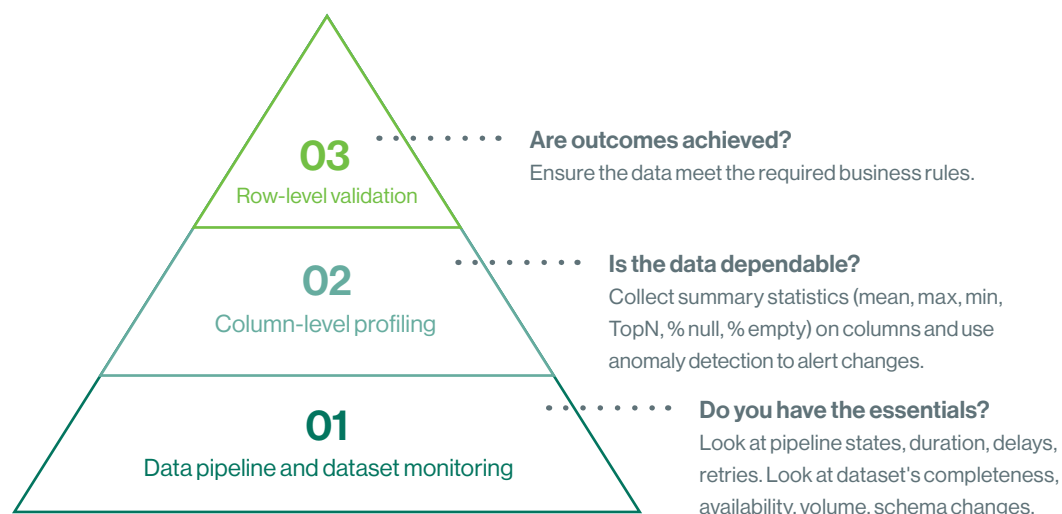


Figure 2. Data observability: three levels of granularity

Data observability leverages sophisticated ML technology to help data engineers monitor datasets and data pipelines, detect anomalies, and find and fix the root cause quickly. The column-level profiling checks if data is consistent, while the row-level validation confirms if data conforms to the business rule requirements. Using ML ensures rule adaptability for frequently changing data coming from diverse sources.

Data observability (i.e. data pipeline and dataset monitoring) empowers data engineers to follow the path of data upstream from the point of failure and help fix it at the source. Data stewards, on the other side, ensure high-quality, error-free datasets (via column-level profiling and row-level validation) for downstream operations. Both personas complement each other to deliver the best approach to healthy, trusted data.

**Data observability is a set of tools to track the health of enterprise data systems and identify and troubleshoot problems when things go wrong.**

[Source: Forbes \(2020\): Data Observability Ushers In A New Era Enabling Golden Age Of Data](#)

# Case studies: Driving data health with data observability

Here are three case studies that illustrate how data engineers leverage Collibra Data Quality & Observability (Collibra DQ&O) for catching bad data before it can hurt.

## Case 1: Delivering trusted data with auto-generated rules

"Is the data consistent and valid?  
Is data in the correct format?  
Has the schema changed?"



Writing rules is the most time-consuming part of managing data quality. Many data quality solutions automate the use of these rules, but writing and maintaining them is still a huge task.

Writing rules manually puts a lot of constraints on the team. For example, rule writers who may be proficient in one language need time to learn new languages. They also need to understand the business aspect to write efficient rules. When 70% of the rules degrade or become obsolete within weeks, rule writers need to start all over again. This grind takes its toll, and the rules become practically unmanageable.

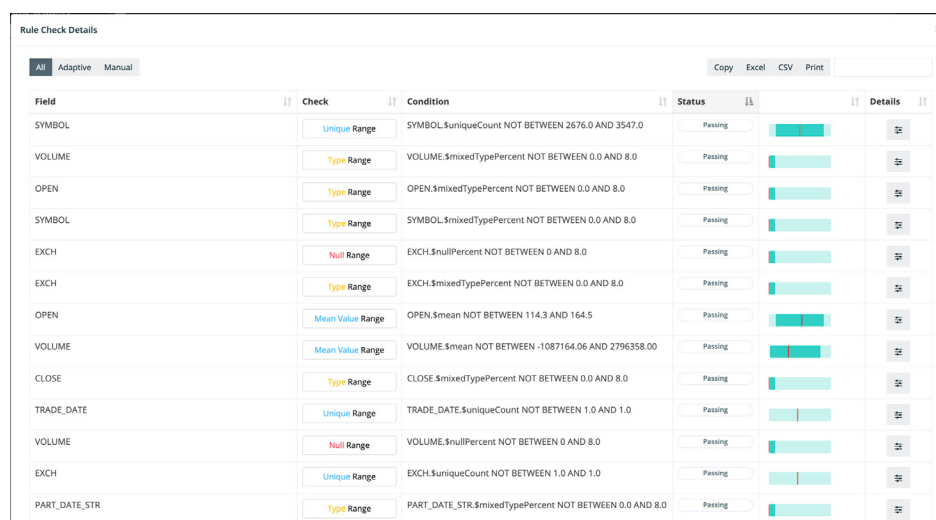
### The challenge:

A top global investment bank relied on manual rule writing. Naturally, it involved considerable efforts, and the associated costs skyrocketed. The company also faced delays in updating the rules manually. They could not manage to ensure data quality with the changing nature of large volumes of data. Ultimately, they failed to deliver healthy data to business users.



## The solution:

Collibra DQ&O leverages ML to auto-generate rules, saving several weeks of manual work and assuring ease of maintenance. The team could see the immediate result of higher data quality with less manual effort.



Field	Check	Condition	Status	Details
SYMBOL	Unique Range	SYMBOL.\$uniqueCount NOT BETWEEN 2676.0 AND 3547.0	Passing	
VOLUME	Type Range	VOLUME.\$mixedTypePercent NOT BETWEEN 0.0 AND 8.0	Passing	
OPEN	Type Range	OPEN.\$mixedTypePercent NOT BETWEEN 0.0 AND 8.0	Passing	
SYMBOL	Type Range	SYMBOL.\$mixedTypePercent NOT BETWEEN 0.0 AND 8.0	Passing	
EXCH	Null Range	EXCH.\$nullPercent NOT BETWEEN 0 AND 8.0	Passing	
EXCH	Type Range	EXCH.\$mixedTypePercent NOT BETWEEN 0.0 AND 8.0	Passing	
OPEN	Mean Value Range	OPEN.\$mean NOT BETWEEN 114.3 AND 164.5	Passing	
VOLUME	Mean Value Range	VOLUME.\$mean NOT BETWEEN -1087164.06 AND 2796358.00	Passing	
CLOSE	Type Range	CLOSE.\$mixedTypePercent NOT BETWEEN 0.0 AND 8.0	Passing	
TRADE_DATE	Unique Range	TRADE_DATE.\$uniqueCount NOT BETWEEN 1.0 AND 1.0	Passing	
VOLUME	Null Range	VOLUME.\$nullPercent NOT BETWEEN 0 AND 8.0	Passing	
EXCH	Unique Range	EXCH.\$uniqueCount NOT BETWEEN 1.0 AND 1.0	Passing	
PART_DATE_STR	Type Range	PART_DATE_STR.\$mixedTypePercent NOT BETWEEN 0.0 AND 8.0	Passing	

Figure 3. Auto profiles with replay and trend analysis

The auto-generated rules learn from data and adapt to do away with the massive rule update efforts. The adaptive rules adjust independently to data variance over time, resulting in a 50-70% reduction in the total number of rules. Explainable and shareable rules can be easily ported across different systems, freeing the team to focus on critical issues and creative use of their skills.

The team also benefited by using the advanced data profiling feature of Collibra DQ&O. Profiling datasets, tables, or columns helps quickly discover common data quality issues of availability, validity, conformity, and schema changes.

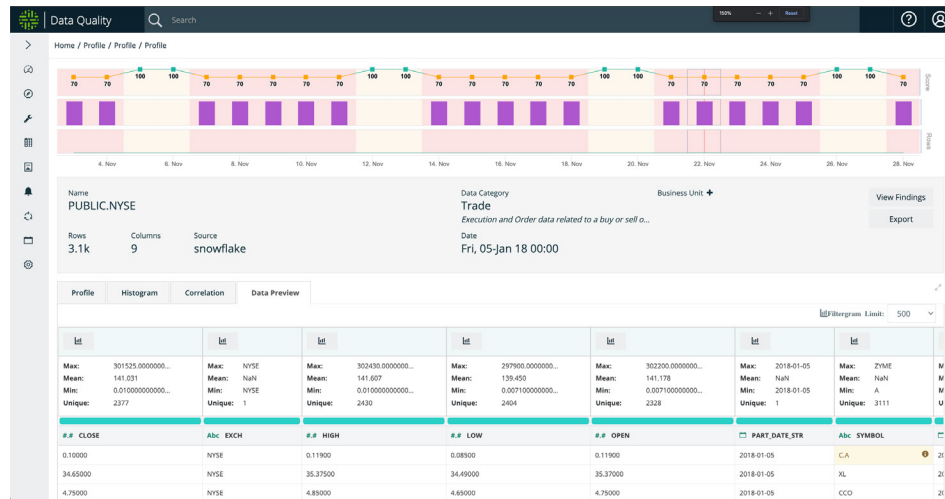


Figure 4. Auto Profiles with Replay and Trend Analysis

**The result:**

With the auto-generated rules and data profiling, the top global investment bank eliminated 60% of manual data quality workload and saved \$1.7M+ in costs.

Collibra DQ&O is built on Spark, providing the fastest performance and highest quality analytics on DataFrames. By simply passing in a Spark DataFrame to Collibra DQ&O, users can automatically add Profiles, Duplicates, Outliers, Relationships, Patterns, and more to their DataFrames. This approach enables real-time data monitoring to deliver healthier data to consumers.

Collibra DQ&O provides correlation across columns to discover hidden relationships and measure the strength of those relationships.

The advanced profiling techniques improve root cause analysis of data quality problems. For example, time series analysis and the playback feature help narrow down the time frame of the issue. They also enable drill-down into profiling to see broken records that violate the user-defined or auto-generated rules.



## Case 2: Managing data lake health efficiently

Data engineers managing data lakes are most concerned with identifying missing or incomplete data. The downstream consumers suffer from poor quality data if data load jobs fail to load the complete number of rows and columns. No data-driven organization can work with uncertainty about data completeness. But often, the scattered and siloed data is the starting point of this problem.

### The challenge:

Every time a top financial services company did a data loading job they faced the issue of data completeness. As a result, the data engineering team manually checked whether the tables were complete after loading. Data scientists and decision-makers were equally concerned about data health across the organization.

"Did all my jobs run on time?  
Did each database table load completely?  
Is my data complete?"

### The solution:

To overcome this challenge, the team hooked Collibra DQ&O into each client database via JDBC connection options. Collibra's out-of-the-box drivers and plug-ins can connect to a wide range of databases and data lakes. The Collibra DQ&O scheduler can automatically crawl and add database tables programmatically without human intervention.

### The result:

Immediately the team saved time because the scheduler takes approximately one second to add a database table, instead of the typical one hour. View the [full list of connectors](#).

Once everything is connected and the information is unified, Collibra DQ&O automatically creates consistent profiles. It observes data continuously and alerts data engineers if data loads run late, incomplete, or suddenly divert from the past behavior patterns.

Use of ML helps identify missing records and pattern breaks proactively, building trust in the data lake. The proactive, continuous monitoring ensures that the data engineering team can take immediate action in case of issues and deliver data completeness quickly.

Collibra also offers a pulse view to provide a clear overview of operational health with every dataset broken down by business units. This at-a-glance view helps data engineers see all the jobs and select any missing runs or quality failures in a single heatmap. They can see inside data pipelines to reveal blind spots in their data operations and distinguish one-off incidents from system-wide hiccups.

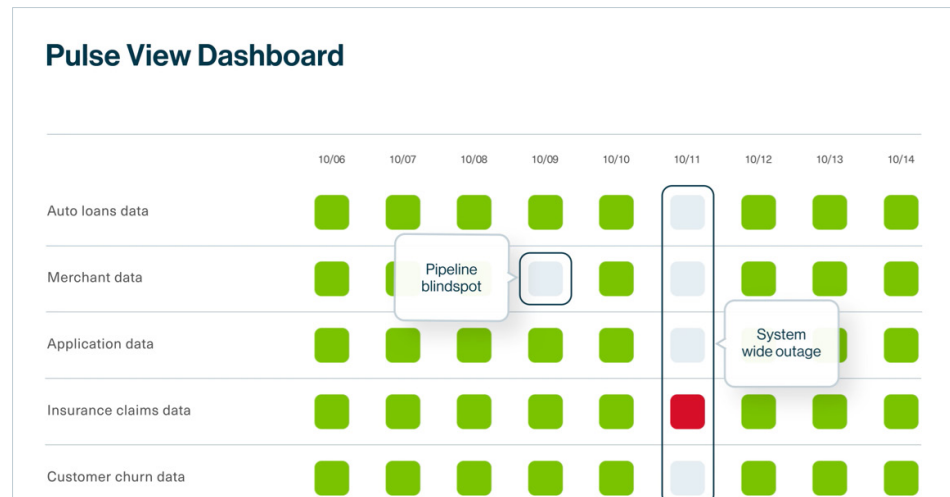


Figure 5. Data pipeline monitoring

With Collibra DQ&O, the top financial services company now manages the data lake efficiently and keeps up with changes.



### Case 3: Accelerating cloud data migration

In most organizations, data engineers need to extract, transform, and load third-party files into the data lake or warehouse. They also move data from the existing database storage to optimized cloud storage. They may consolidate storage systems to a single data lake or warehouse. And may copy the same data between Dev, QA, and Prod environments for multiple projects. Though a routine task, assuring the quality of the migrated data is challenging.

**"Did my cloud migration complete on time?  
Does my data lake reconcile with  
the source system?"**

#### The challenge:

A top healthcare insurance company was struggling with lengthy cloud data migration. The manual process of validating data integrity between the source and the target required a great deal of time. And despite their best efforts, it did not ensure trust in data. The team was almost on the verge of giving up.

Data migration is perhaps the most demanding activity that tests the patience of data engineers. Matching the target data to the source data can take hours, and can still end up problematic. Validating data integrity between distinct storage systems requires identifying missing records, values, and broken relationships across tables or systems. Source-to-target mapping depends on the types of systems, and the mapping can easily fall out of sync.

#### The solution:

Collibra DQ&O provided the data engineering team with the key capability of automated rule-based data integrity validation. With it, they could automatically check if every record in every cell matches between copies. This level of granularity, besides the standard row count, column, and conformity checks, helped the team build complete trust in the data.

## Assuring high data quality during cloud data migrations

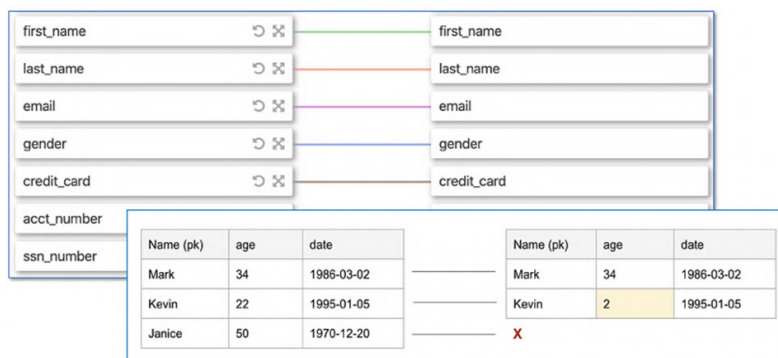


Figure 6. Data reconciliation between the source and the target

### The result:

The accelerated cloud data migrations with Collibra DQ&O saved the top healthcare insurance company 2000 hours of effort while delivering healthier data.

The data observability feature is especially useful when a data engineer loads in third-party data files during cloud migrations and after moving data to persistent storage. Collibra DQ&O is source-agnostic and is compatible with any source and target, making life easy for data engineers.

Collibra DQ&O assures that quality data can move between systems by five key functions:

- Data profiling and cataloging of the source systems to understand the quality of data on the source system
- Data validation rules and duplicates detection
- Source-to-Target data reconciliation
- Periodic data monitoring
- Inbuilt workflow to proactively resolve the data issues

## Conclusion

The shifts in data quality are urging data engineers to rethink their approach to delivering high-quality data pipelines. Data observability empowers them to track the health of enterprise data systems and predict issues before they happen.

The complete stack of data quality and observability helps build trust in data. Diverse use cases illustrate how data engineers can leverage Collibra Data Quality & Observability for increased speed and accuracy while driving data health.



### Tour the product →

Test drive Collibra Data Quality & Observability through an interactive guided tour for 14 days at no cost.

### Start a free trial →

Install Collibra Data Quality & Observability in your own environment and try it with your own data for 20 days at no cost.

### Request a demo →

Speak one-on-one with a Collibra expert and get a personalized demo of Collibra Data Quality & Observability.