Collibra

The beginner's guide to

predictive data quality and observability

Poor data quality can have catastrophic consequences and monumental impact

For example, the Mars Climate Orbiter, which was a $125 million investment, crashed and burned after a 10 month mission because an engineer calculated the acceleration and distance to Mars using the International System of Units instead of the Imperial System of Units.

In business, without healthy data, every decision is at risk. And poor decisions immediately hurt critical business objectives. This leads to:

- Disengaged customers
- Missed revenue targets
- Increased operational risks
- Lack of compliance

19% of businesses have lost a customer by using incomplete or inaccurate data about them.

Source: Dun & Bradstreet Report (Jun 2019): The Past, Present, and Future of Data

47% of the recently created data records had at least one critical or business-impacting error.

64% assessments had completeness errors.

53% assessments had accuracy errors.

Source: Harvard Business Review (May 2020): Assessing Data Quality - A Managerial Call to Action

Trusted business decisions begin with trusted data

Trusted data drives accurate insights that lead to better and faster business decisions. But how do you ensure the use of trusted data across your organization?

The answer is data quality which indicates if data is fit for use.

By 2022, 70% of organizations will rigorously track data quality levels via metrics, improving it by 60% to significantly reduce operational risks and costs.

Source: Gartner: How to Improve Your Data Quality, June 2021

Barriers to achieving high data quality

There are numerous barriers that you may face when implementing a data quality strategy that slow down or complicate the process. These barriers include:

- Data quality professionals trained in technical skills need time to learn the business (what rules to write), while those familiar with the business take time to gain the technical competency (how to write rules).
- When you have disparate data sources and systems, it can be difficult to achieve consistent quality without extensive rewriting of the rule logic in different codes.
- Duplicate or ambiguous data arriving from multiple sources takes a much longer time to resolve, and results in wasted efforts at the enterprise level.
- Too much data means excessive time spent weeding out unusable data and a lesser focus on extracting value from the useful data.

Breaking these barriers means rethinking the approach to data quality.

Why a traditional approach to data quality doesn’t work?

The traditional approach to data quality is reactive, which works by finding and fixing data issues. In the process, the errors may get addressed at one location, but may continue at other locations.

More specifically, the traditional approach to data quality includes:

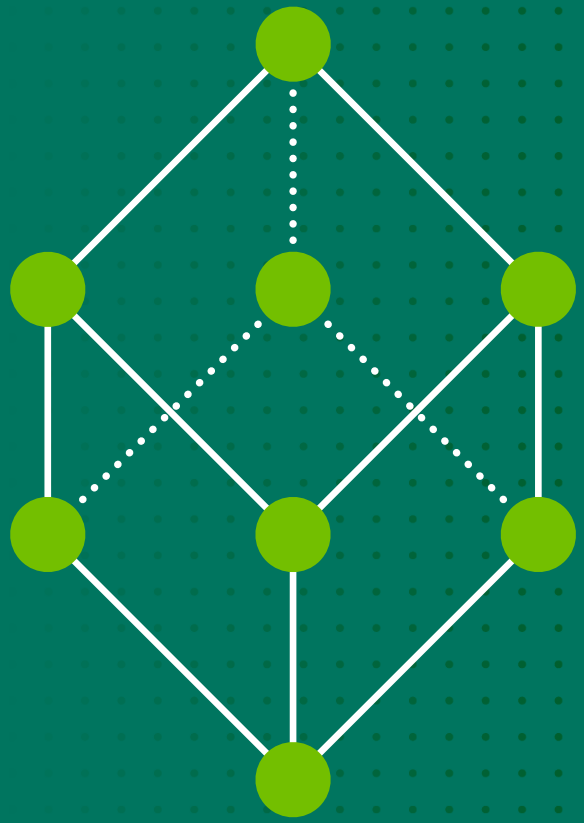
- Manual rule writing which is typically filled with errors, is labor-intensive, and takes a long time to develop. Naturally, manual approach cannot cope with the large volumes of rapidly arriving data.
- Discrete or homegrown automation tools which fail to easily integrate into a larger solution architecture and often force legacy technologies, hurting cloud adoption and digital transformation.

Why you need a modern approach to data quality

Data quality is not a one-time activity. You need a proactive and continuous approach for detecting and fixing quality issues before they affect the downstream operations.

A modern approach to data quality leverages the advancements in data science and machine learning to evolve rules on the fly. The rules can adapt to the changing data landscape while making the intelligence behind them transparent. A unified data quality scoring system across all data sources along with personal alerts and interactive dashboards put business users at the forefront of data quality.

The auto-discovered, adaptive, explainable rules power a predictive, enterprise-scale, self-service approach to data quality.



The 5-step process to predictive data quality and observability

Step 1
Connect data
Admins

Connect and scan a wide range of heterogeneous data sources and pipelines, including files and streaming data.

Step 2
Gain awareness
Any stakeholder

Display profiling statistics for each data set, table, and column, including hidden relationships and time series analysis.

Step 3
Automate controls
DataOps

Build generic DataOps and statistical controls with automated technical rules to detect unknown issues and scale your data quality operations.

Step 4
Define conditions
Data stewards

Build domain-specific controls with automated and custom business rules that are adaptive, non-proprietary, explainable, and shareable.

Step 5
Take action
Any stakeholder

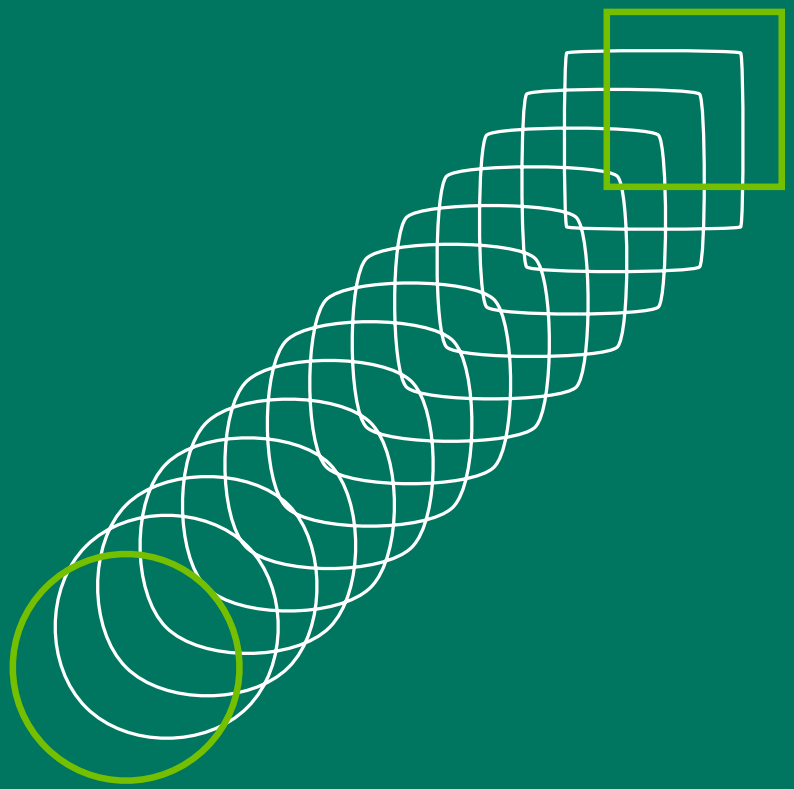
Embed data quality processes into critical business workflows. Initiate alerts with the right data owners when data quality scores drop, to resolve issues quickly.

Benefits of Collibra Data Quality & Observability

Collibra uses a predictive, enterprise-scale, and self-service approach to data quality. It brings advanced machine learning to address every data quality problem, from profiling, rules, and outliers to data reconciliation and discovering hidden relationships.

Collibra Data Quality & Observability helps you:

1. Get real time visibility into the health of all your data.
2. Discover sensitive data and enforce data quality rules.
3. Reduce the risk and cost of data migrations.



You can eliminate up to 60% of manual data quality workloads with autonomous data quality rules.

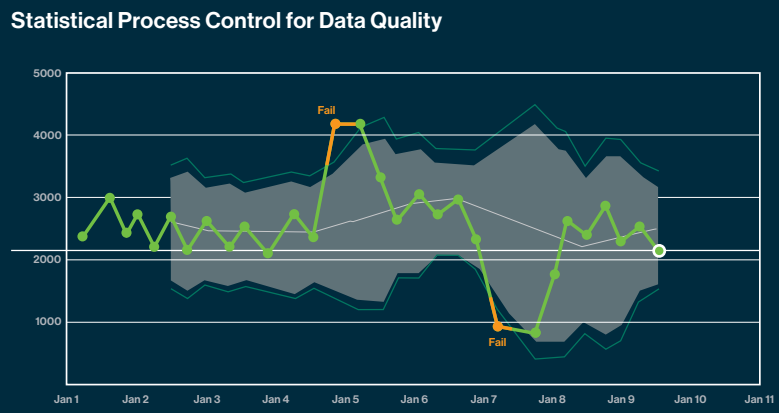
Source: Customer benchmarks

1. Get real time visibility into the health of all your data

Is your data often delayed, missing, or incomplete? With Collibra Data Quality & Observability, you can automatically generate rules to get complete visibility on metrics including null checks, row counts and outliers. More importantly, you can proactively detect and resolve data issues before they start impacting downstream applications.

How we deliver

- Automate rule management with unsupervised machine learning
- Detect and resolve data issues upstream to mitigate data loss and downtime
- Standardize data quality checks across data pipelines so you can build reliable data products

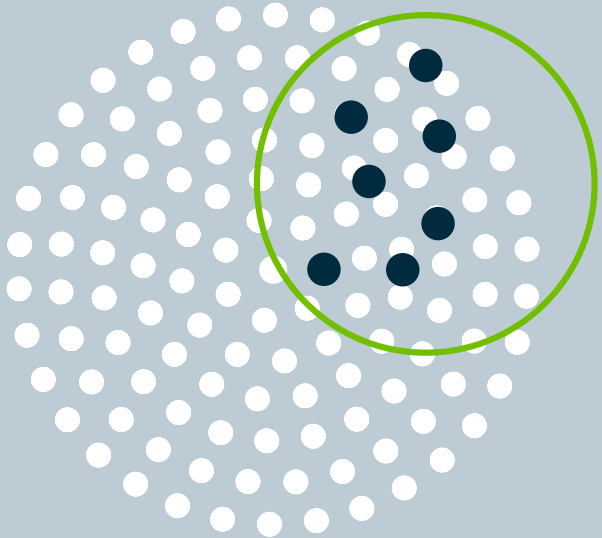


2. Discover sensitive data and enforce data quality rules

Having trouble discovering sensitive data and enforcing rules across all data types and sources? With Collibra Data Quality & Observability, you get access to an out-of-the-box repository of industry-specific, auto-validation rules that helps you automatically discover sensitive data, enforce rules and take action on broken records.

How we deliver

- Automatically classify sensitive and non-sensitive data with advanced meta tagging and auto-validation rules
- Mitigate compliance risk by ensuring your sensitive data is always timely identified, accurate, valid and complete
- Continuously monitor data for violations, generate alerts and initiate remediation workflows to deliver trusted data

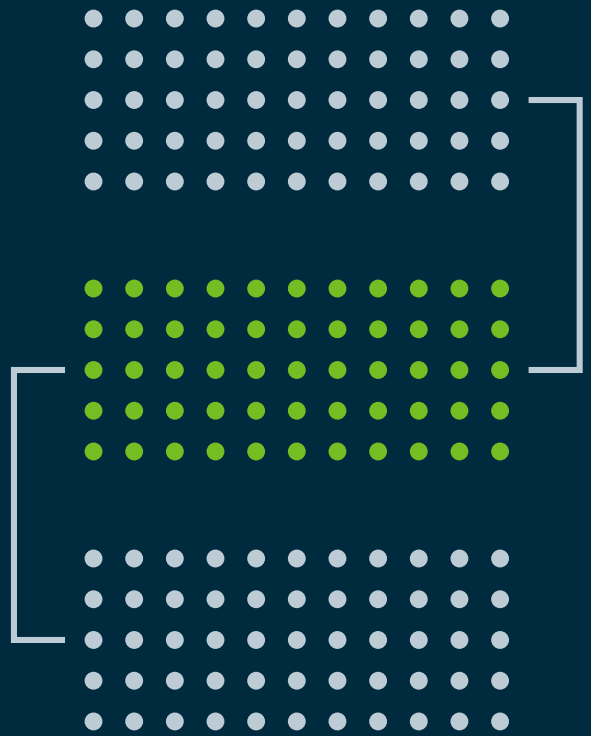


3. Reduce the risk and cost of data migrations

Do you struggle to customize rules as you migrate data to different systems and environments? With Collibra Data Quality & Observability, you get out-of-the-box rule templates that are global, explainable, and shareable. You can also quickly write your own rules in SQL and avoid being locked into proprietary languages.

How we deliver

- Eliminate silos across any data type, source and environment with portable and explainable rule templates
- Eliminate the need to rewrite rule logic in different codes



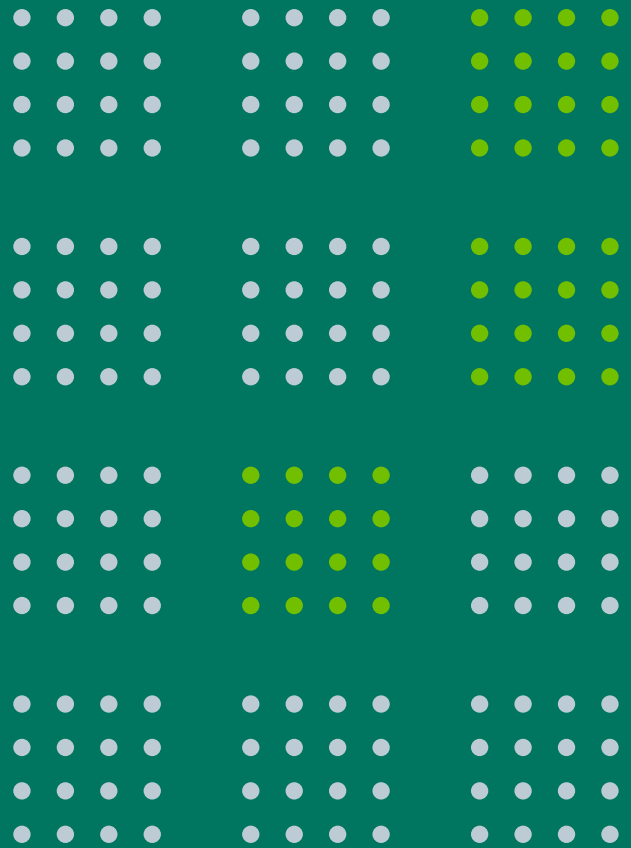
Data governance foundation for trusted data

The robust data governance foundation helps you focus on business-critical data and foster a culture where everyone in the organization contributes to trusted data:

- Data governance promotes a shared understanding of data across the enterprise for a collaborative approach to trusted data.
- A robust data ownership model helps assign data quality issues to the right owners for a quick resolution.
- Business-driven, collaborative workflows help easily identify, escalate and resolve data quality issues.

By 2025, 60% of data quality processes will be autonomously embedded and integrated into critical business workflows.

Source: Gartner Mar 2022 report - The State of Data Quality Solutions: Augment, Automate and Simplify



A unified data quality management approach with data catalog, governance and lineage

A unified data management approach works with data quality, observability, catalog, governance, and lineage together. It helps you centralize and automate data quality workflows to support a holistic view of managing data and get the best out of your data and analytics investments.

	Gain proactive data intelligence about...
Data quality + observability + data catalog	What datasets or columns you should scan first.
Data quality + observability + data lineage	What data issues you should resolve first to address the root cause.
Data quality + observability + data governance	Whom to reach out to resolve data issues.

Customer stories

Improve your ROI on data

484% 3-year ROI

34% improvement in staff time to address data errors

Source: The Business Value of Collibra, IDC 2022

A leading global financial services company

Expediting time to compliance with autonomous data quality rules

- Eliminated 60% of manual data quality workload
- Reported $1.7M+ cost savings
- Avoided seven-figure fines for non-compliance of BCBS 239 and GDPR

A top healthcare insurance company

Reducing the risk and cost of data migration by rule-based data integrity validation

- Accelerated cloud data migration
- Saved 2000 hours of efforts
- Ensured continued HIPAA compliance by discovering PHI (protected health information) and enforcing data quality rules

A top global health care services company

Modernizing data quality management via adaptive and explainable rules

- Reduced complexity, repetition, and guesswork in data quality management
- Maximized data quality coverage across all data sources and pipelines

Start building trust in your data.

Although data quality isn't easy, it is necessary for all data-led businesses. You can't make an informed decision unless you have high quality data that you can trust. Collibra's predictive data quality and observability solution ensures continuous, scalable and self-service data quality across your organization.

Test drive Collibra Data Quality & Observability

About Collibra — Since 2008, Collibra has been uniting organizations by delivering trusted data for every use, for every user, and across every source. Our Data Intelligence Cloud brings flexible governance, continuous quality and built-in privacy to all types of data. The Global 2000 relies on Collibra to create the critical alignment that accelerates workflows and delivers better results faster. We have a diverse global footprint, with offices in the U.S., Belgium, Australia, Czech Republic, France, Poland and the U.K. To learn more, visit **collibra.com**, follow **@Collibra** on Twitter or follow us on **LinkedIn**.

If you are interested in learning more, please visit collibra.com.