Collibra

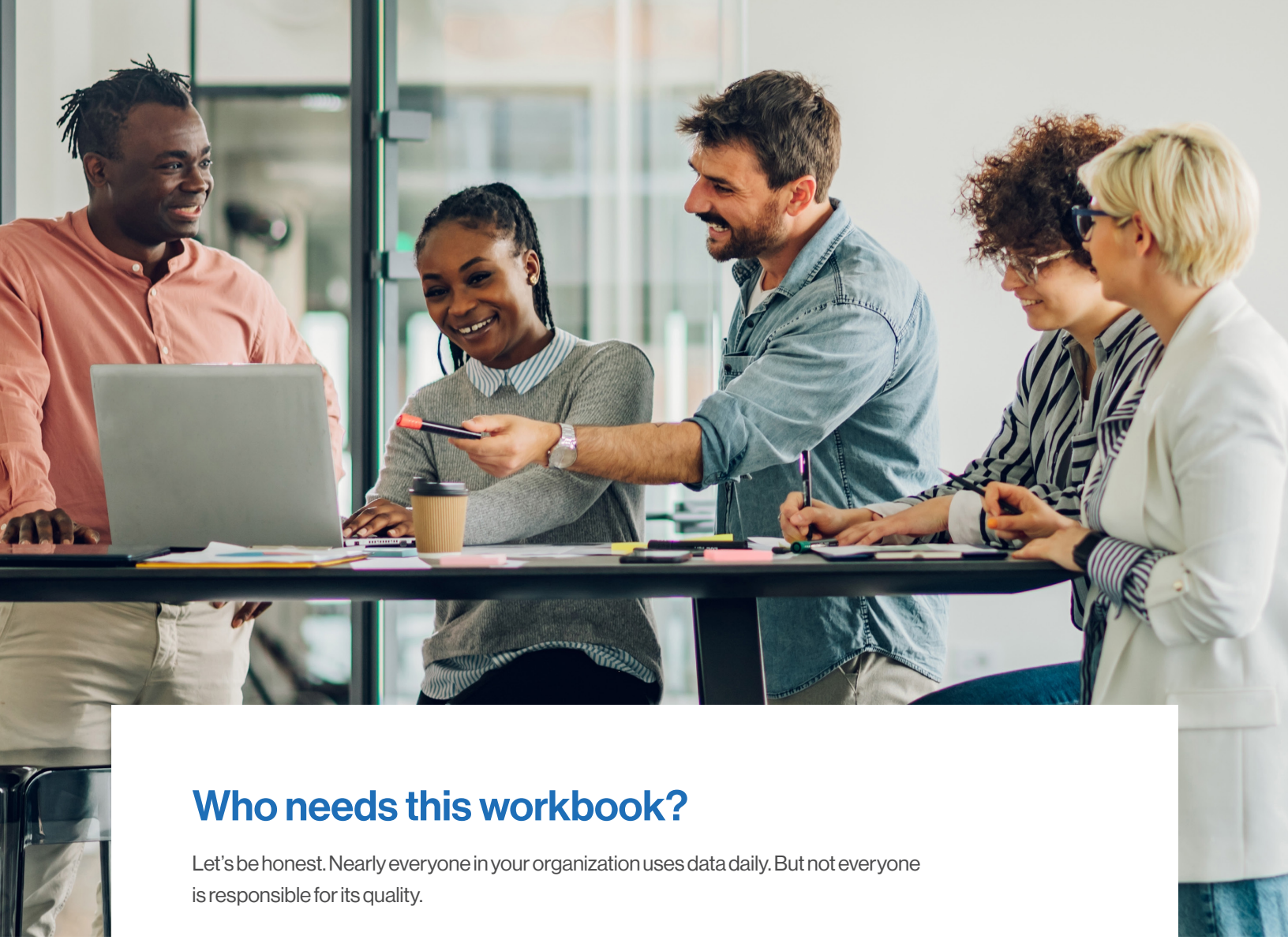
Workbook

How to become a data quality expert

A step-by-step workbook to ensure Data Confidence™ with Collibra Data Quality & Observability

Table of contents

Who needs this workbook?	3
Why data quality matters	4
Current and future state assessment	5
Steps to Data Confidence	7
Collibra Data Quality & Observability: Automated business and technical rulest	13



Who needs this workbook?

Let's be honest. Nearly everyone in your organization uses data daily. But not everyone is responsible for its quality.

If you analyze data sources, manage data pipelines, remediate data issues or champion data quality across your organization, then this workbook is your blueprint for success.. As you go through this workbook, you'll develop a confident approach to implementing a strategic data quality framework and discover how Collibra Data Quality & Observability unifies governance to provide trusted, reliable data across your entire organization.

Why data quality matters

Using high-quality data to make informed decisions isn't just a nice-to-have. Especially in our AI era, it's essential for organizational survival.. But your data is all over the place. Business users are disconnected from accessing the data they need. Data quality definitions vary from department to department. And siloed catalogs create multiple "sources of truth." It's a condition we call governance fragmentation. And it makes data quality nearly impossible. As organizations grow, it becomes difficult to ensure consistent data quality across all data sources. Processes that may have worked at the outset are not necessarily applicable or useful at scale.

So, how do you ensure that you have good data quality across the entire organization? How do you create unified governance?

It comes down to three key steps.

1. Establish a unified process for monitoring and managing data issues on a continuous basis
2. Get buy-in and support from key stakeholders across your organization
3. Create a culture where everyone participates — from data creators to processors to consumers

How is your organization approaching data quality today? If you're like most, the answer is "inconsistently at best."

Setting the stage | Current state assessment

Let's start out with a self-assessment on the current state of your data quality strategy. Check all the scenarios that apply or add your own inputs.

1 List your company's strategic initiatives and their expected outcomes.

Select and outline all that apply:

- ☐ Increase revenue and profit
- ☐ Reduce costs
- ☐ Increase productivity and efficiency
- ☐ Reduce regulatory risk
- ☐ Increase customer satisfaction and retention
- ☐
- ☐

2 Which of those initiatives are being negatively impacted by poor data quality?

Select and outline all that apply:

- ☐ Increase revenue and profit
- ☐ Reduce costs
- ☐ Increase productivity and efficiency
- ☐ Reduce regulatory risk
- ☐ Increase customer satisfaction and retention
- ☐
- ☐

3 What are your most pressing data quality concerns?

Select and outline all that apply:

- ☐ Accuracy
- ☐ Completeness
- ☐ Consistency
- ☐ Freshness
- ☐ Integrity
- ☐ Uniqueness
- ☐ Validity
- ☐
- ☐

4 Which activities around data quality take the most time for your and your organization?

Stack-rank from 1-5 with 1 being the most time-consuming:

- ☐ Discovering and profiling data across all sources
- ☐ Creating rules and mapping them to data assets
- ☐ Detecting data issues and assessing cause and impacts
- ☐ Prioritizing issue management based on business severity
- ☐ Identifying who is responsible for resolution and ensuring accountability
- ☐
- ☐

Setting the stage | Future state assessment

Let's start out with a self-assessment on the current state of your desired future state. Check all the scenarios that apply or add your own inputs.

1 What new data quality capabilities do you need to develop or acquire to meet your current or future needs.

Select and outline all that apply:

- ☐ Automated data profiling and classification
- ☐ Automated rule creation based on profiling and classification
- ☐ Automated rule creation based on governance policies
- ☐ GenAI text to SQL rule creation
- ☐ Automated mapping of rules to data assets and governance policies
- ☐ Shareable rule templates
- ☐ Automated monitoring across data sources and data pipelines
- ☐ Automated source to target data reconciliation
- ☐ Automated mapping of data quality scores to data catalog assets
- ☐ Technical and business lineage overlaid with quality monitors and controls
- ☐ Proactive notification of data issues, cause, impacts and policy violations
- ☐ Automated task assignment and workflows based on data ownership and issue management processes

2 What KPIs will you use to measure success?

Select and outline all that apply:

Business KPIs

- ☐ % increase revenue
- ☐ % decrease in costs
- ☐ $ reduction in regulatory penalties

Technical KPIs

- ☐ % increase in trust with respect to data
- ☐ % increase in data stewardship productivity
- ☐ % reduction in time to respond to data issues

3 Who are your key stakeholders across your organization?

- ☐
- ☐
- ☐
- ☐

4 What are the key use cases of those stakeholders and how will you prioritize?

- ☐
- ☐
- ☐
- ☐

5 What's your implementation timeline?

- ☐ Next 1-3 months
- ☐ Next 3-6 months
- ☐ Next 9 months
- ☐ Next 12-18 months
- ☐
- ☐

Steps to success: From fragmentation to Data Confidence

The core challenge: Fragmented governance

You're motivated to change your current state. But you're blocked. And the obstacle is fragmented governance.

Fragmented governance occurs when your visibility, control and policies are disjointed across different clouds, applications and data stores. It's the silent killer of data initiatives, manifesting when:

- Your business users are disconnected from accessing and stewarding the data they need
- Data quality definitions vary across departments and systems
- Policies are inconsistently applied across your data landscape
- Multiple siloed catalogs create contradicting "sources of truth"
- Data lineage breaks down as information moves between systems
- Knowledge about data context remains trapped in specific teams or tools

This fragmentation prevents organizations from scaling data and AI use cases safely, creating blind spots in what data exists, how it's used and who has access. As data estates grow more complex and AI multiplies the risk of unreliable use, the problem compounds exponentially.

The transformative goal: Data Confidence

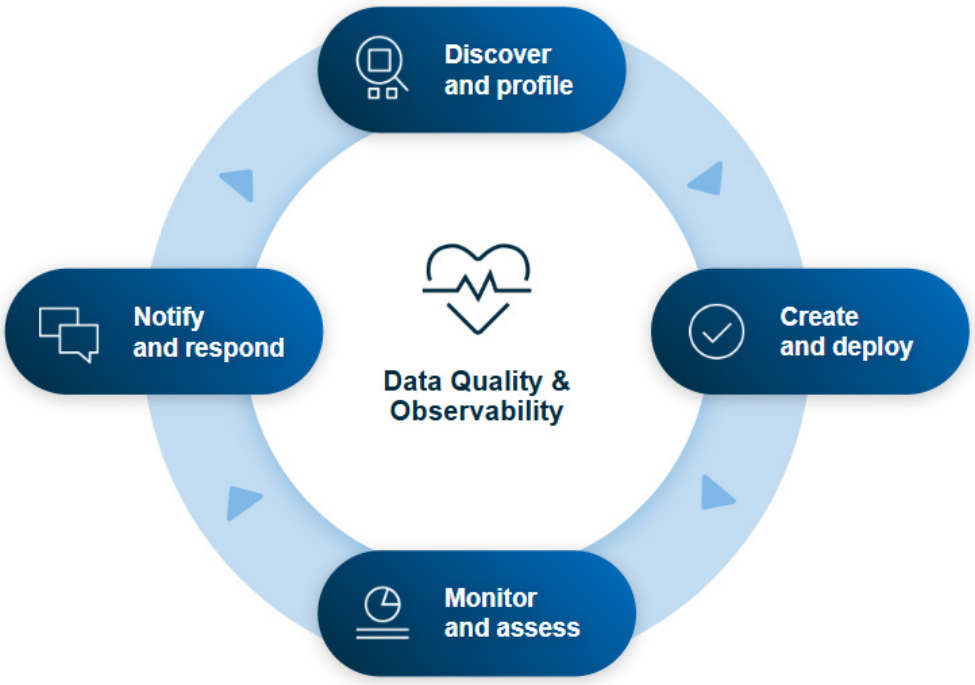
Data Confidence is the organizational state Collibra helps you achieve through unified governance. It represents:

- The feeling of knowing your people are using trusted, high-quality data
- The ability to accelerate all data-driven use cases, safely
- The capacity to observe, track, protect and democratize data across your entire ecosystem
- The foundation for reaping first-mover benefits from emerging AI opportunities.

Put simply, Data Confidence is what happens when everyone in your organization can trust, comply and consume data without friction or fear. To build this foundation, true collaboration across your entire data ecosystem is essential – but how do you secure continuous support from stakeholders across the organization?

The following four steps represent best practices that hold team members accountable and keep stakeholders informed. These create the foundation for Data Confidence and accelerate every data and AI use case.

Do these steps align to your existing data quality process?
If not, what additional steps would you add?



Step 01: Data discovery and profiling

Start by profiling your data to identify data structure, content, classification and sensitivity across sources. AI-powered data profiling enables you to automate discovery of data types in columns, and the classification of data in those columns based on categories such as name, credit card number and address, as well as identify domains of data like customer, employee and partner in database records. This foundational step is crucial for setting the stage for effective data quality and observability, especially as you increase your AI investments and the regulatory landscape evolves.

See it in action: Automated Profiling and Classification

Best Practices

Consider and acknowledge all descriptive statistics about your data set (uniques, count, mean) in order to build a comprehensive data profile.

Define all the data classes you need to monitor and manage so you can create the semantic tags to assess the quality of your data set.

How are you profiling data today?

What data classes are you currently monitoring?

How are you tagging data assets with those semantic classifications?

What new data classes would you like to monitor?

Step 02: Rule creation and deployment

After finding your data and understanding what type of data you have, you can define and deploy data quality rules to monitor and manage your data assets. Common rules include checks and thresholds for completeness, accuracy, consistency and timeliness. As well as more complex multistep rules like checking that the United States zip code field is not empty, the format is five digits, it only contains numeric characters, and it is a valid zip code based on the State and City fields in the record. Many rules for known data types and entities can be inferred using AI as part of the discovery and classification process. And AI can automate the mapping and execution of data quality rules to data assets throughout the sources in your enterprise.

See it in action: Adaptive and custom rules

> **Best Practices**
>
> Determine the impact of each dimension applied to your data to create a prioritized scoring mechanism for your data; a single dimension may not be sufficient to assess the quality of your data.
>
> Remember that multistep rules may be needed to ensure the fitness of data for a specific use case or context.

How are you creating and deploying rules today?

What data quality dimensions do you have rules for?

Do you currently use multistep rules?

What new multistep rules do you need for different use cases?

Step 03: Data quality monitoring and assessment

Creating and deploying data quality rules enables you to continuously monitor and assess the quality of your data. To ensure confidence and trust in data, monitoring and assessment needs to happen both with data at rest in sources,as well as data in motion in pipelines. For example, data and schema drift significantly impact AI model accuracy and reliability. As real-world data evolves or data structures change, models trained on outdated data degrade, leading to inaccurate predictions. Monitoring and mitigating drift ensures models remain effective and adapt to new realities, maintaining performance and trust in AI-driven decisions.

See it in action: Data drift and shift detection

Best Practices

Set up the frequency of your data quality checks to align with your policies for how often you want to validate your data quality, so that your monitoring is compliant with your standards.

Ensure you have end-to-end technical and business data lineage so you can identify the root cause of problems as well as the downstream impacts.

How are you performing data validation and anomaly detection today?

Are you monitoring and assessing quality in data sources and pipelines?

What governance policies are impacted by data quality?

How quickly can you determine data issue causes and impacts?

Step 04: Data issue notification and response

Once issues, causes and impacts are identified you must notify all relevant stakeholders, and initiate issue management workflow. Proactively notify stakeholders about any AI systems, reports, or business processes that rely on the compromised data. This helps prevent data issues from becoming business issues. Assign tasks based on data ownership and follow pre-defined issue management processes and workflows. This ensures clear accountability and speeds up the response. And prioritize response based on the severity of business impact and policy non-compliance.

Learn more: Assignment queue

> **Best Practices**
>
> Reuse the work you've done in data governance for assigning data owners to data sets, establishing SLAs for response, and defining issue management processes and workflows.
>
> Use pre-built reports that show data quality coverage across all technical systems and business units to provide stakeholders easy visibility into your data quality and issue status.

How are you performing notification and response today?

Who are the people accountable for issue management for different data sets?

Stakeholders need to be informed when data issues arise?

How easily can you prioritize responses based on business severity?

Collibra Data Quality & Observability: Find data issues before they become business issues

Collibra transforms data anomalies into actionable data quality and policy compliance insights, by providing automated visibility and business context across every source and system, using a complete platform that unifies quality, lineage, catalog, and governance.

Create more confidence in your data warehouses and lakes, business and regulatory reporting, and machine learning and AI agents.

Broad connectivity

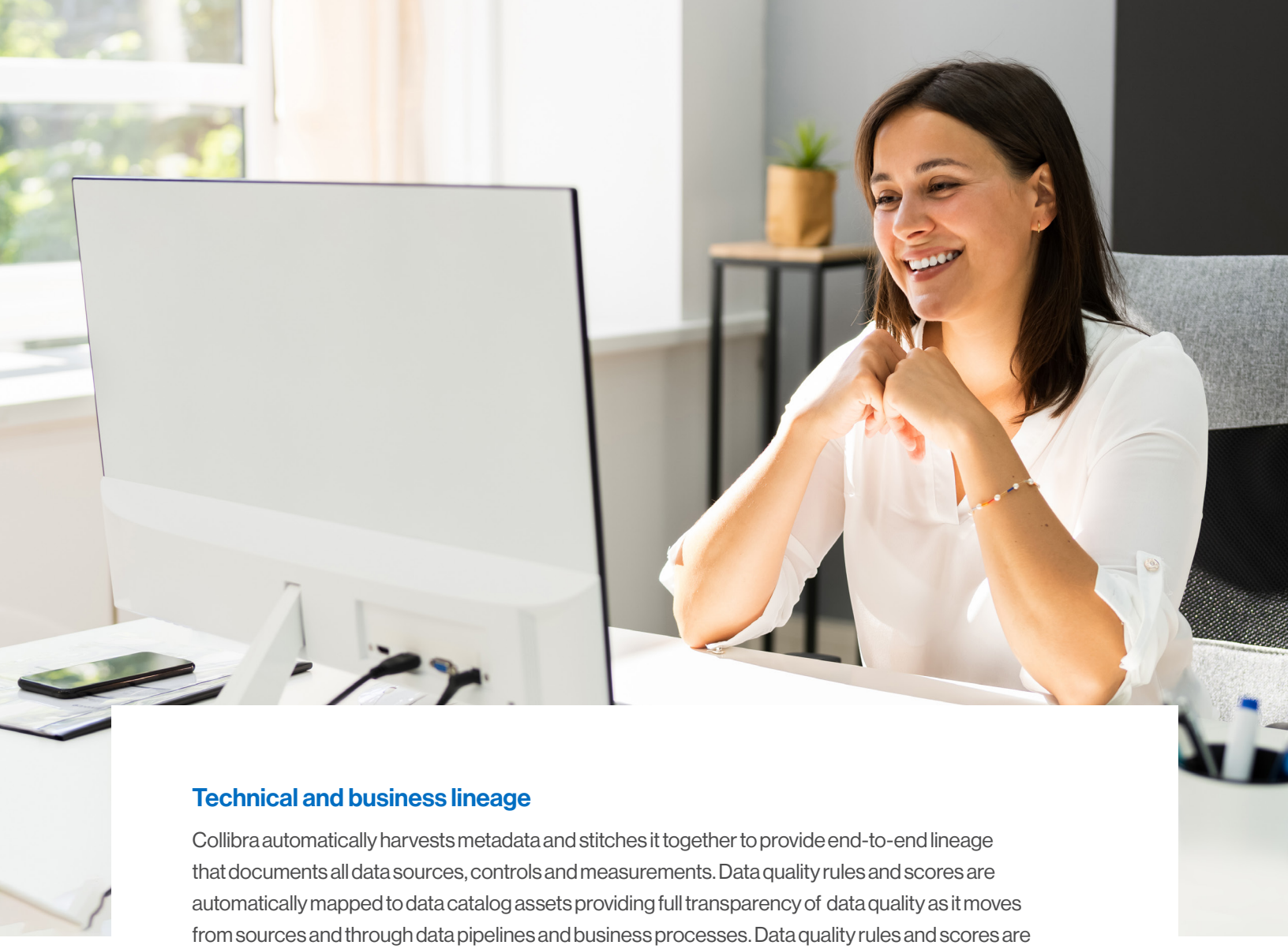
It is important to monitor and manage all your critical data elements. Collibra Data Quality and Observability provides a wide variety of connectors to ensure source systems, data stores, and files being ingested are managed consistently. Regardless of the data sources and how datasets are aggregated, joined or combined we help you use the identical set of monitors and controls to ensure quality.

Streamlined rule creation

Collibra Data Quality and Observability provides out of the box data classification and autogeneration of rules based on profiling results. There is a coding framework for developers, data engineers, and ETL designers that want to build real-time data quality into their broader data pipeline. A SQL assistant interface using Generative AI enables business people to use natural language prompts to turn business rules into technical rules.

Adaptive and advanced data quality monitors

Identifying accuracy typically requires context and time, to fully understand if the field value is correct. Collibra does this by observing data points over time while keeping a measurement of when each data point was correct or accurate. It can then automatically flag data points that become inaccurate.



Technical and business lineage

Collibra automatically harvests metadata and stitches it together to provide end-to-end lineage that documents all data sources, controls and measurements. Data quality rules and scores are automatically mapped to data catalog assets providing full transparency of data quality as it moves from sources and through data pipelines and business processes. Data quality rules and scores are also mapped to policies and business rules providing visibility of policy compliance.

Data quality scorecards and reports

Collibra Data Quality and Observability offers many ways to report on data quality. Scorecards let you visualize the health, consistency, and evolution of a dataset over a user-defined date range.The Pulse View dashboard provides a heat map of data quality jobs running in sources and pipelines. There are also a variety of out of the box reports and the ability to use BI tools like Tableau and PowerBI create custom reports.

Flexible and adaptable deployment

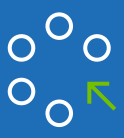
Collibra Data Quality and Observability enables you to monitor and measure data quality across systems, pipelines and sources in hybrid cloud and on-premises environments. You can deploy on a single virtual machine or large Kubernetes clusters, and jobs can be run in a spark compute layer or pushed down into the database layer. And our extensive APIs enable integration with notebooks and orchestration tools like Airflow.

Collibra

Turn your data into a competitive advantage

By reading this workbook, you've taken a critical step toward unified governance and achieving Data Confidence. With clear processes, consistent monitoring and automated insights, you can equip your organization to identify and address data quality proactively. And you can turn your data into a reliable asset that empowers informed decisions and fuels innovation.

As you continue your data journey, keep refining your approach, engaging stakeholders and embedding data quality into your organizational culture. At Collibra, we're here to help you accelerate all your data and AI use cases.



Learn more about Collibra Data Quality & Observability.
Ready to take a tour?

CO20-P0352-11