

How to become a data quality expert

Technical Solution Workbook

A step-by-step workbook to accelerate time to value
with Collibra Data Quality & Observability

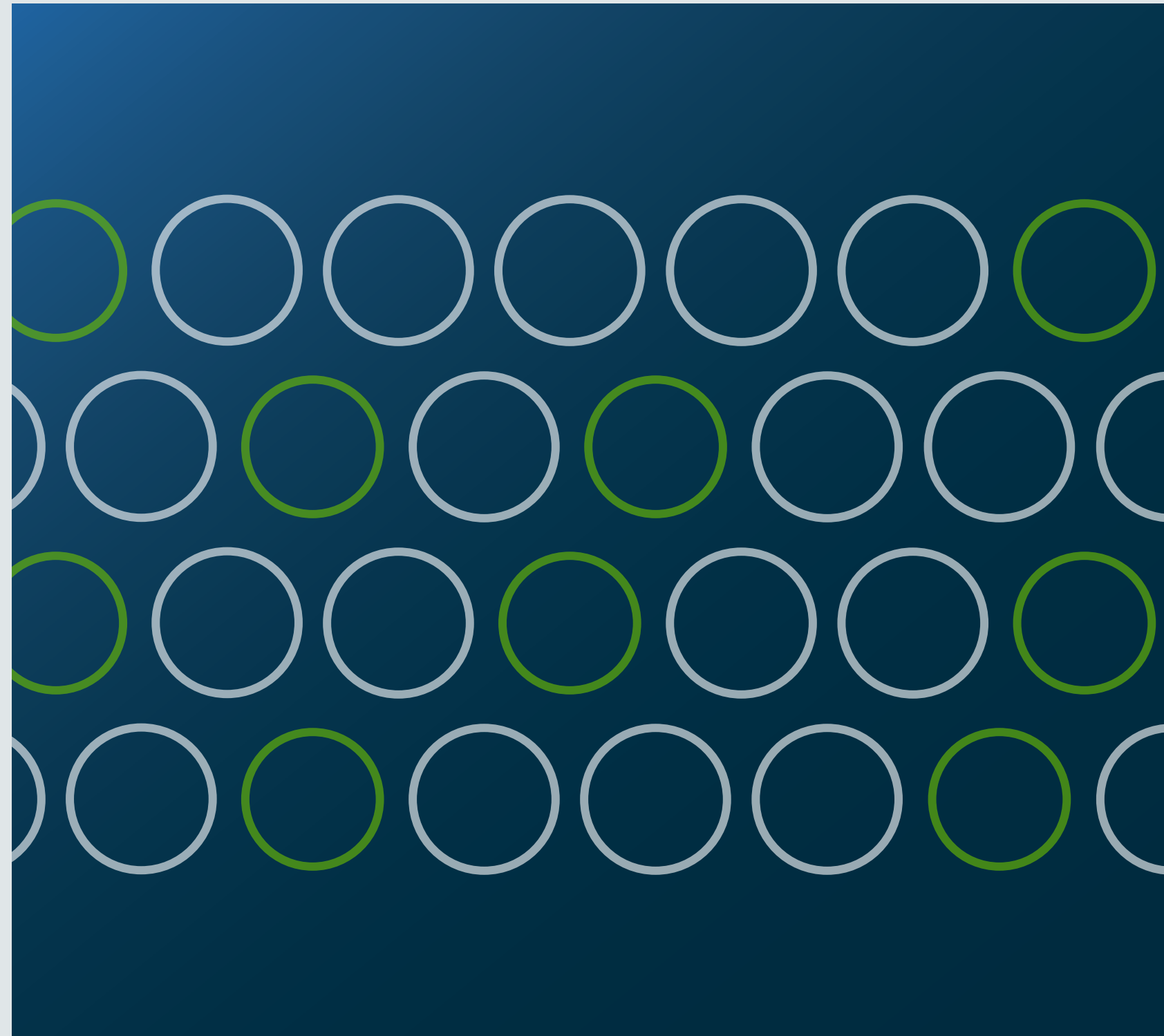


Table of Contents



- 01 Who is this workbook for?
- 02 Why data quality matters
- 03 Current and future state assessment
- 04 Steps to drive adoption and ensure success
- 05 Colibra Data Quality & Observability:
Automated business and technical rules
- 06 How Colibra Data Quality & Observability can help
- 07 Paving the way to faster time to value
- 08 Resources and further reading

Who is this workbook for?

Do you use data everyday to do your job? Most likely the answer is **yes.**

If you analyze data sources, data sets or data pipelines, act on data issues, and uphold data quality throughout the organization then this workbook is for you. As you go through this workbook you will learn how to confidently implement a data quality strategy and how Collibra Data Quality & Observability can help ensure the use of consistent, reliable, and trustworthy data across your entire organization.

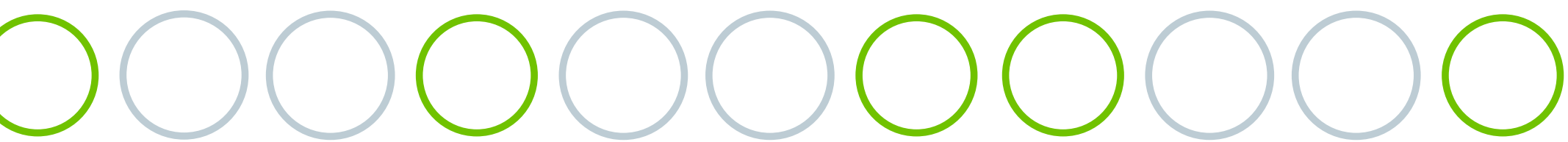
Why data quality matters

Using high-quality data to make informed decisions is essential for any business. But getting access to quality data is not easy. As any organization grows, 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?

It comes down to **three key steps**.

- 01** Create a clear process that allows you to proactively discover, acknowledge, and remediate your data issues on a continuous basis.
- 02** Get buy-in and support from key stakeholders across your organization.
- 03** Ensure that all the stakeholders creating, processing, and consuming the data contribute to these efforts.



How is your organization leveraging data quality today?

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.

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

- Select and outline all that apply:
- Increase operational efficiency
 - Increase productivity
 - Reduce regulatory risk
 - Reduce revenue loss
 - Increase data migration efficiency

02 List your top concerns or challenges with this approach

- Select and outline all that apply:
- Negatively impacting productivity
 - Negatively impacting regulatory compliance
 - Negatively impacting revenue
 - Negatively impacting customer experience
 - Negatively impacting business agility

03 What are your top data quality challenges?

- Select and outline all that apply:
- Error-prone, ad-hoc manual rule writing and management
 - Inability to scan all data sources
 - Siloed approaches to data quality across various departments
 - Lack of integrity of data as it moves across applications
 - Lack of business engagement and commitment

04 Flag those activities around data quality that take the most time for you and your organization.

- Stack-rank from 1-5 with 1 being the most time-consuming:
- Manual rule management and tasks
 - Comparing data across disparate sources
 - Chasing people to fix data issues
 - Fixing late-stage or downstream data issues
 - Convincing the business to trust models and reports

Setting the stage | Future state assessment

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

- 01

List the new data quality capabilities you believe you need to develop or acquire to meet your current or future needs.

Select and outline all that apply:

Autonomous rule management

Shareable rule templates

Horizontal and vertical scalability

A unified DQ scoring system

Remediation workflows

Data quality checks across data pipelines

Data masking

Lineage and time series analysis capabilities

Proactive anomaly detection

Data discovery and rule enforcement

- 02

What KPIs will you use to measure success?

Select and outline all that apply:

% increase in trust with respect to data

% increase in productivity

\$ reduction in regulatory penalties

\$ reduction in revenue losses

% reduction in risk and costs around data migrations

- 03

List key stakeholders across your organization

- 04

List key use cases you are looking to address

- 05

List timing for addressing these changes

Next 1-3 months

Next 3-6 months

Next 9 months

Next 12-18 months

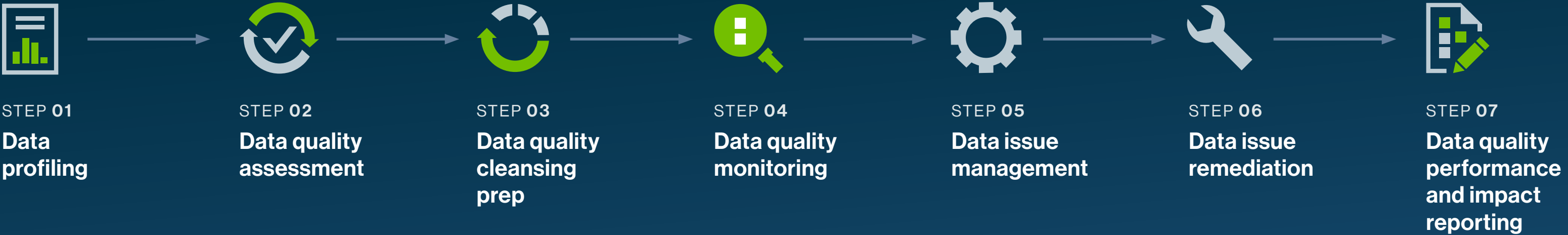
Steps to drive adoption and ensure success

Collaboration is essential to a successful data quality strategy. But how do you guarantee continual support from across the organization?

This workbook includes the key steps and best practices that you need to hold your team members accountable and your stakeholders informed. These steps encourage collaboration and ensure buy-in from the top down.

Do these steps align to your existing data quality process?

If not, list additional steps you take here:





STEP 01

Data profiling

We believe the first step to a successful data quality strategy is to acknowledge and classify issues in your data sets. But this isn't always easy.

See it in action

[Profiling and Scanning](#)

In a typical organization, data profiling is done manually using SQL and Excel. Every time you update or change your data source, you need to manually profile the data. This is time consuming and limits your ability to scale.

Collibra Data Quality & Observability is automatic and easy to use so you don't have to waste time. You simply point it to your data warehouse or data lake, and it will scan and profile your data to identify data quality issues.

You can also set up automatic profiling and customize the scope of your scans. Over time, Collibra Data Quality & Observability learns about your data and provides a wide variety of profile insights, so that you can automatically expand your awareness of the data without any manual effort. More specifically, Collibra Data Quality & Observability provides time series analysis and playback features that improve root cause analysis of data quality problems.

Best Practices



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



Define and classify all the data issue types in terms of severity so that you can create the building blocks to assess the quality of your data set.

How are you performing this activity today?

Stakeholders that need to **implement this new approach** *(List)*

Stakeholders that need to be **informed with this new approach** *(List)*



STEP 02

Data quality assessment

The next step is to set up your data assessment criteria. Data quality dimensions will help you assess if your data is good enough to use or if you need to make improvements. Once you identify the data quality issues, you need to write new rules that enforce quality across your organization.

See it in action

[Data Quality Dimensions](#)

We understand that when you profile your data, it takes a considerable effort to present the results of your profiling to your stakeholders. All stakeholders need to agree on a consistent definition of data quality and the rules to enforce quality, which takes time. You may also be spending considerable time with IT to make sure they create and implement these rules and definitions.

Collibra Data Quality & Observability provides an intuitive user experience out of the box, which can be used for sharing profiling results with various roles, including those who are responsible for the source data. You can also share these results directly with IT, and other data owners, so everyone is on the same page, which sets you up for a good governance model.

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 the aggregated scores of multiple dimensions represent data quality in your specific context and indicate the fitness of data for use.

How are you performing this activity today?

Stakeholders that need to *implement* this new approach (*List*)

Stakeholders that need to be *informed* with this new approach (*List*)



STEP 03

Data quality cleansing prep

Next, you need to prepare for the cleansing and standardization of the data. You must convert your data into the desired format so that data quality rules or checks can be applied.

See it in action
[Adaptive Rules](#)

We understand that this isn't easy. Data quality triage and clean up is a huge chunk of your daily routine. Plus, it can take immense discipline on your part to keep your rule library up to date and in line with your organizational growth. Over time, you may find it challenging to apply these rules consistently.

Collibra Data Quality & Observability has a robust data quality rule library, which cuts down on your efforts to write manual rules from scratch. It is also powered by a machine learning-enabled rules engine. In simple terms this means as your data grows, new rules are automatically generated, classified, and stored in your rule library. This reduces the effort to maintain your rule library by 50%. If you choose to still write your own custom rules, you have the flexibility to quickly build your rules in an intuitive no-code interface. This will allow you to implement these rules across your data landscape to promote consistency and standardization in your business data.

Best Practices



Create standardized rules that can evolve and align to your organization's growth. This ensures rules do not become obsolete over time.



Remember to always fix data quality issues at the source so that quality data flows downstream to all data consumers, not just a single data set.



Ensure that once you have identified the critical data issues, you codify the data cleansing into your ETL or MDM tool of choice.

How are you performing this activity today?

Stakeholders that need to **implement this new approach** *(List)*

Stakeholders that need to be **informed with this new approach** *(List)*



STEP 04

Data quality monitoring

You now have your data analyzed with clear quality rules in place. What's next? You have to continually perform quality checks on your data by setting up regular monitoring.

See it in action

[Data Quality Scorecards](#)

Depending on the number of data sources, setting up monitoring, profiling, and rule execution can be a challenge. Data quality evolves (and deteriorates) over time, so you may be required to constantly analyze every new profile scan and stay on top of all the new issues that arise with recent changes. You may also have different types of data stores, which can lead to different monitoring implementations every time.

Collibra Data Quality & Observability helps by providing you with a unified scoring system to report across all your data sources. Historical profiling helps you gather a baseline, and the built-in machine learning capabilities allow you to identify new data issues and generate new rules on the fly. You also have the ability to scan across all major database implementations, file systems, and streaming data, which provides horizontal and vertical scalability, all from a single platform interface.

Best Practices



Ensure that the frequency of your data quality checks aligns with how often you want to scan your source data, so that your data quality rules kick in as soon as a new scan is initiated.



Set up your data scoring mechanism to create scoring thresholds and automatically notify your team when the scores fall below this threshold to initiate timely action.

How are you performing this activity today?

Stakeholders that need to *implement* this new approach (*List*)

Stakeholders that need to be *informed* with this new approach (*List*)



STEP 05

Data issue management

You have your monitoring and rules. Now it is time to address individual data quality issues. But that is not always easy.

See it in action

[Data Issue Assignments](#)

There is no standard (and stress-free) way to communicate data quality issues to a large group of data owners and stakeholders. Data teams usually end up using a mix of issue management systems, emails, and meetings to drive the agenda, which wastes time and energy.

Collibra Data Quality & Observability speeds up the process by providing a set of business-friendly data quality scorecards. These reporting scorecards send out notifications to inform and assign data quality issues to data owners, so that you do not have to manually manage multiple systems for issue resolution.

Best Practices



Ensure that your data quality issues are visible and tracked in the same issue management system to guarantee visibility to the respective data teams (like your internal ticketing or incident platforms).



Plan to consistently report progress against these data quality issues to your business leadership to ensure their continued support.

How are you performing this activity today?

Stakeholders that need to **implement this new approach** *(List)*

Stakeholders that need to be **informed with this new approach** *(List)*



STEP 06

Data issue remediation

Once you have your monitoring and rules in place, it is time to address individual data quality issues. You can do this by assigning data quality issues to the respective data owner so they can be notified immediately when a problem occurs. This process helps ensure issues are quickly addressed and resolved.

See it in action

[Data Quality Investigation](#)
[Preventing Data Quality Issues](#)

But we know this is easier said than done. Without proper data lineage you cannot see the origin of the data quality issue, and therefore, cannot properly resolve any problem at its origin. You only see an anomaly in the data set, but you can't detect the root cause of the problem upstream. It's important to fix data issues in the source data by retracing lineage to the original data set.

Collibra Data Quality & Observability provides access to the unified data scoring system, which serves as a great starting point for your root-cause analysis. Through the configurable DQ workflows, you can inform multiple data owners across data sets to initiate remediation when data quality scores drop below the target threshold score.

Best Practice



Identify a list of all the data owners for each data source so that you can investigate lineage issues as a group across multiple data sets.

How are you performing this activity today?

Stakeholders that need to **implement** this new approach (*List*)

Stakeholders that need to be **informed** with this new approach (*List*)



STEP 07

Data quality performance and impact reporting

Last but not the least, you need to continually report on your data quality measures. You can do this through reports that include metrics e.g. the number of identified issues for each data asset or the average time from detection to resolution.

See it in action

[Data Quality Built-in Reports](#)

A successful data quality strategy ultimately comes down to your ability to present your efforts in a cohesive narrative. Unfortunately, the tools you use to enforce data quality may not be intuitive and flexible when it comes to reporting. You might be using an ad-hoc reporting solution to address this need, which could become costly to maintain.

Collibra Data Quality & Observability is set up to support the entire gamut of data quality-related activities for an enterprise. This means you have all your data quality results in a central location so you can easily report on the overall health of your data quality efforts across the organization.

Data quality performance and impact reporting is one of the most important steps in the process. This is the step where you show success and continue to garner support across the business.

Best Practice



Leverage pre-built reports that show data quality coverage across all technical systems and business units, so that you have complete visibility into your data quality landscape.

How are you performing this activity today?

Stakeholders that need to **implement this new approach (*List*)**

Stakeholders that need to be **informed with this new approach (*List*)**

Collibra Data Quality & Observability

Automated business and technical rules

At Collibra, we believe in empowering you to enforce data quality standards with the least effort, yet be flexible in your strategy to stay on top of anomalies in your data.



ML-enabled, explainable and adaptive data quality rules

Reduce manual rule writing and maintenance efforts by 50-70%.



Proactive monitoring and anomaly detection

Automatically uncover data drift, outliers, patterns and schema changes to mitigate risks and improve business decisions.



Foundational metadata management*

Capture, reconcile and manage metadata to quickly connect teams to improve data quality and usability.



Intuitive, configurable workflows*

Initiate remediation workflows with the right data owners when data quality scores drop, to quickly resolve issues.



Data masking

Identify and automatically mask sensitive information to maintain compliance while performing data quality checks.



Horizontal and vertical scalability

Scale data quality across large and diverse databases, files and streaming data as your business grows.



End-to-end, automated data lineage*

Trace data movement across the lifecycle. Help data quality teams narrow the focus of root cause investigations and prioritize issues.



Data ownership and stewardship*

Establish accountability for data and boost trustworthiness by ensuring governance of critical data elements.


How Collibra Data Quality & Observability can help



When you connect Collibra to your data on Amazon Redshift or Google BigQuery, Collibra Data Quality & Observability scans through your data for patterns and generates adaptive rules. You can apply these rules right away or modify as needed. You can also write your own SQL rules in the Rule Builder.

Following these scans, Collibra Data Quality & Observability generates a wide array of observations for your review. These checks stem from over 20 proven data science and ML algorithms that allow you to uncover insights hidden in your data.


Boost your KPIs



Improve your TCO ROI on data

484% 3-year ROI and **34%** improvement in staff time to address data errors

Source: The Business Value of Collibra, IDC 2022



Increase employee productivity

Eliminate up to **60%** of manual data quality workloads with autonomous data quality rules.


Source: Customer benchmarks



Reduce regulatory & compliance risks

Avoid seven-figure fines for non-compliance of BCBS 239, CCAR, HIPAA, GDPR and other regulations

Source: Customer benchmarks








Cut costs

Accelerate cloud data migrations. One company **saved 2000 hours** of effort with rule-based data integrity validation.

Source: Customer benchmarks

Paving the way to faster time to value

	 Scoping & Data Profiling	 Rule Building	 Data Verification	 Data Integration	 Data Monitoring & Reports
Activities	<ul style="list-style-type: none">• Decide the scope and critical use case• DQ architecture workshop• ML based auto-rule discovery• Validate quality of data across 9 dimensions• Rule suggestions	<ul style="list-style-type: none">• Build the rules with business consensus• Test the system• Run matching algorithm• Generate and verify test results	<ul style="list-style-type: none">• DQ metrics collection• Generate and verify test results• Identify data issues and relationships• Re-visit data collection, data analysis and rule building, if required	<ul style="list-style-type: none">• Show or advise on how to import existing rules• Advisory support on how to integrate Colibra Data Quality & Observability with other systems using APIs• Share best practices	<ul style="list-style-type: none">• Demonstrate how to report data inconsistencies to data stewards efficiently• Continuous system monitoring• Generate and verify test results
Deliverables	<ul style="list-style-type: none">• Project Plan• Installation and configuration• Initial analysis & profiling	<ul style="list-style-type: none">• Test results & logs• Documentation on incorporated rules		<ul style="list-style-type: none">• Knowledge transfer on APIs	<ul style="list-style-type: none">• Knowledge transfer• DQ scorecard to measure DQ metrics

How you can plan to engage with Collibra

Collibra	Customer Organization		
<div>Collibra Enterprise Architect Expert for data quality implementation program Full time on the project</div>	<div>DevOps or Infrastructure Technical point person for daily interaction with Collibra resources Week 1 Until able to demonstrate connectivity</div>	<div>Technical SME Support data quality requirements 2 weeks On call to answer questions</div>	<div>Executive Sponsor + Stakeholders Attend workshops and presentations, align on business objectives Approx. 3 days Planned by Data Management/DQ Program Director</div>
<div>Collibra Implementation Manager Project direction and best practice guidance Part-time</div>	<div>Data Architect SME for Collibra Data Quality & Observability in the wider enterprise architecture Part Time During initial week, available for questions a few hours a week thereafter</div>	<div>Data Quality Developer Responsible for developing and maintaining DQ rules 2 weeks Part-time</div>	<div>Data Steward Responsible for reviewing and validating DQ rules 2 weeks <i>(starting Week 2)</i> Part-time</div>

Resources and further reading

Resources

- [Collibra Data Quality Test Drive](#)
- [Collibra Data Quality Bootcamp](#)
- [Data Quality education for all: Free Self-paced Learning @ Collibra University](#)
- [Collibra Data Quality - Quick Demos](#)
- [Collibra Data Quality User Guide](#)
- [Collibra Data Quality Data Assessment](#)

Further reading

- [DQ Rule Cheat Sheet](#)
- [**Factsheet:** Required capabilities for enterprise-scale data quality and observability](#)
- [**Whitepaper:** Create an enterprise vision for data quality and observability](#)
- [**Blogs** on data quality and observability](#)

