Collibra

Workbook

Stop bad data before it stops you

How to ensure reliable data for AI and analytics success

The data observability workbook

Table of contents

Who this workbook is for

Welcome to the data observability workbook. As the volume and complexity of data landscapes grow, ensuring the reliability of your data is crucial for making informed decisions and driving business success. This workbook is designed to guide data leaders through the essential steps needed to implement a robust data observability program.

What will you learn?

By the end of this workbook, you'll not only have a comprehensive understanding of the steps involved in ensuring reliable data, but also practical insights into executing these steps effectively.

Get ready to gain the essentials you need to:

- Understand the importance of data observability and its impact on your organization
- Learn how to profile and classify your data to identify quality issues
- Develop and enforce data policies and quality rules to maintain data integrity
- Utilize machine learning to detect anomalies and ensure data consistency
- Monitor the impact of data quality issues and prioritize remediation efforts
- Create effective notification and escalation processes for data issues
- Continuously optimize your data observability practices for ongoing improvement

You'll also be ready to explore the benefits of leveraging Collibra Data Quality & Observability for your data observability program.

AI is raising the stakes on data reliability

There's no doubt. AI is disrupting every industry. And data and AI leaders are scrambling to stay ahead of the competition. It's an insight everyone can acknowledge. But building AI into your organization isn't easy. **If you have doubts about the decisions being made with your data now, you're not alone. In fact, more than half of all data leaders don't trust their organization's data.**[1]

The trust gap has consequences beyond just poor decision-making – it's blocking innovation and competitive advantage. As AI adoption accelerates and regulations evolve, manual data quality processes simply can't keep up.

Data observability is your foundation for scaling AI and analytics safely. It gives you automated visibility into data quality and reliability across your entire data ecosystem. With proper data observability in place, you can:

- Detect and prevent data quality issues before they impact decisions
- Build trust in your data across technical and business users
- Accelerate AI and analytics initiatives with confidence
- Meet compliance requirements through automated monitoring

Why data quality programs fail

The telltale sign of a data quality program that's faltering is lack of visibility into data issues and their causes until they impact decision making and downstream systems. This lack of transparency results in inaccurate AI and analytics that lead to poor decisions and compliance issues that can result in fines and damage to your company's reputation. While many organizations have robust programs to remediate data they lack tools and processes to observe data quality as data moves from ingestion and storage, to compute and consumption.

[1] Source: https://www.idc.com/getdoc.jsp?containerId=US51397423

Data observability breakdown

Data observability capabilities and principles can be a powerful accelerator to business success, offering significant benefits, including:

- **Quality and explainability**: Ensure that high-quality data is maintained across the organization, enabling clear understanding and use of data for decision-making
- **Compliance and trust:** Ensure compliance with relevant laws and regulations, avoiding legal consequences and reputational damage
- **ROI tracking:** Fully understand the ROI of data quality investments while identifying and mitigating risks associated with poor data quality

Data observability defined

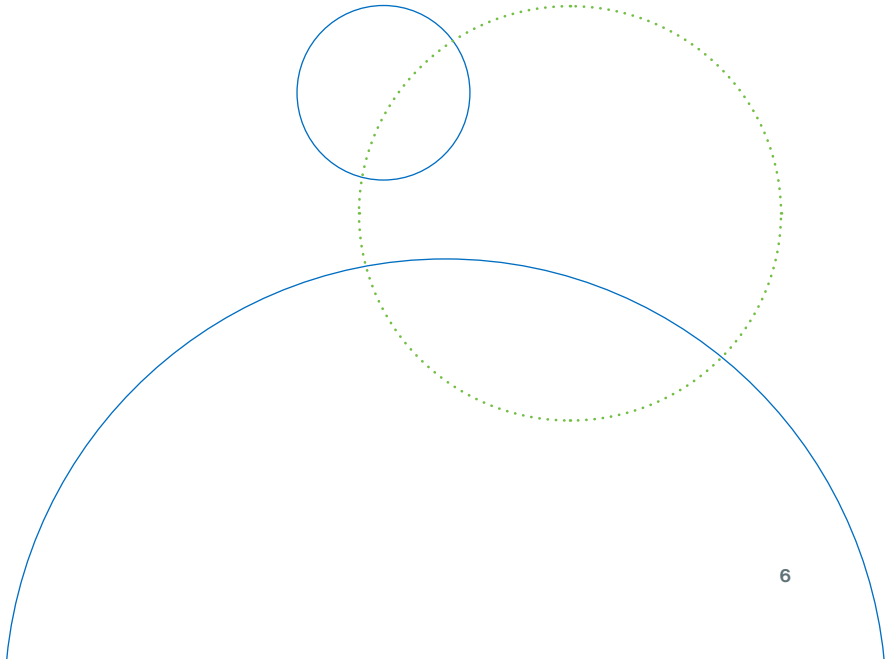
Data observability is software and processes for profiling, monitoring, and notifying stakeholders about data quality in sources, pipelines, AI, and analytics. It automatically collects metadata to determine data structure and content, suggest data quality rules, build historical baselines, detect anomalies, determine root cause and impact, and respond to issues to ensure reliable data.

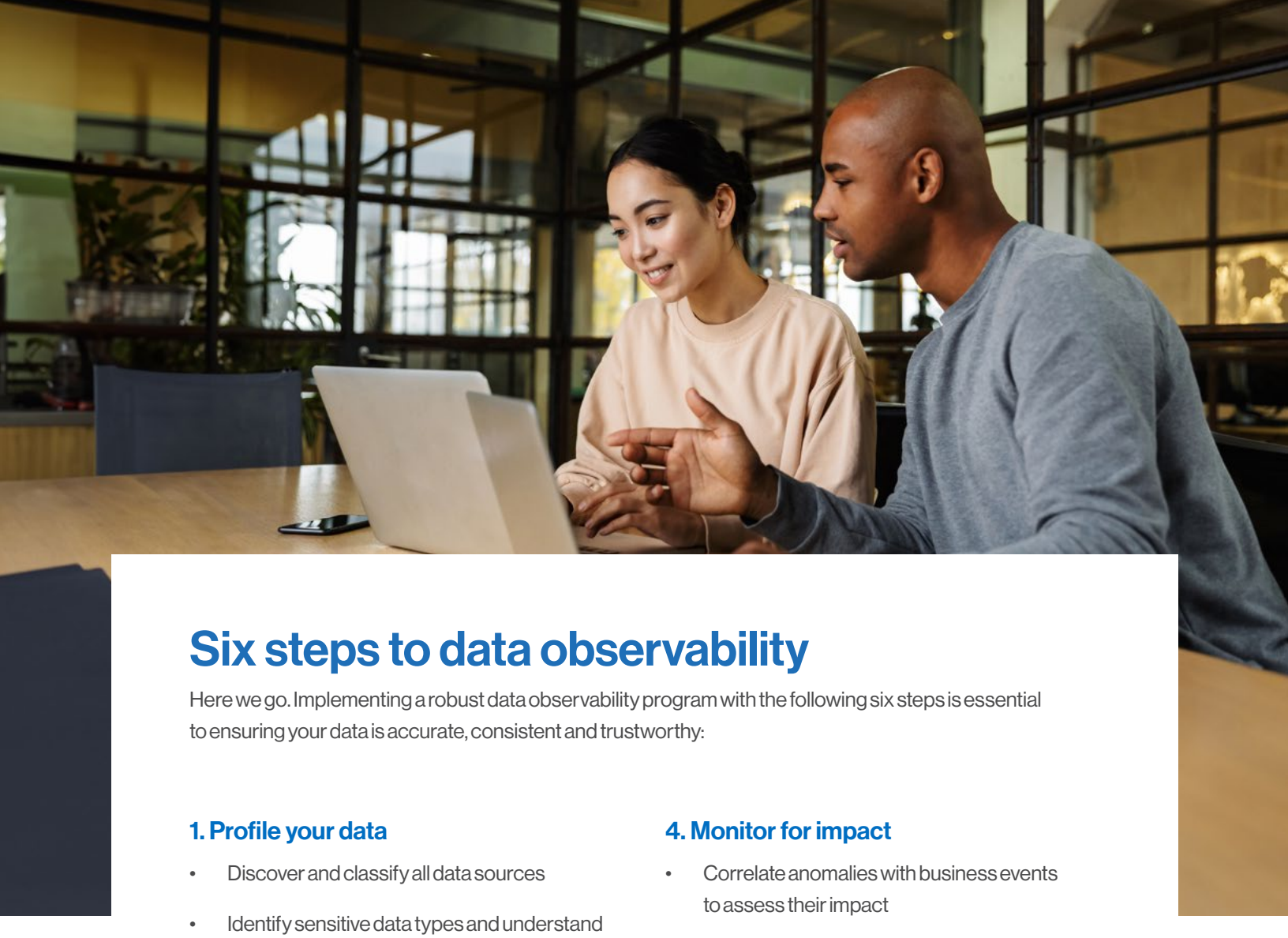
Ready for data observability?

There's no better time than right now. Start a data observability initiative by asking questions about these key topics:

1. **Understand data quality events:** Are data quality issues recognized and how often do they occur?
2. **Identify success metrics:** What key metrics (accuracy, completeness, freshness, etc.) reflect data quality in your organization?
3. **Ensure stakeholder visibility:** Who needs access to data quality insights (data engineers, data stewards, business analysts, etc.)?
4. **Evaluate technical expertise:** Do team members involved in data validation have the necessary technical backgrounds?
5. **Assess automation level:** How manual are your data quality processes? Can automation improve efficiency and accuracy?

These questions will help you get ready for implementing a robust data observability program, guiding you to evaluate your current readiness and pinpoint areas for improvement.





Six steps to data observability

Here we go. Implementing a robust data observability program with the following six steps is essential to ensuring your data is accurate, consistent and trustworthy:

1. Profile your data

- Discover and classify all data sources
- Identify sensitive data types and understand data structures

2. Define data policies and quality rules

- Establish clear data policies and quality rules
- Implement continuous testing and validation to maintain data integrity

3. Detect anomalies

- Use machine learning to establish baselines and detect deviations
- Identify potential data quality issues early

4. Monitor for impact

- Correlate anomalies with business events to assess their impact
- Prioritize remediation efforts based on severity

5. Notify key experts

- Alert relevant stakeholders about data issues
- Initiate and manage remediation processes effectively

6. Optimize continuously

- Evolve data policies and practices based on insights from monitoring
- Implement continuous improvement loops for ongoing optimization

These steps will guide you in building a reliable data infrastructure, ensuring high data quality and supporting informed decision-making.

1. Profile your data

Profiling your data is the first crucial step in ensuring reliable data. By thoroughly understanding and classifying all your data sources, you can gain a comprehensive view of your data landscape. This foundational step helps identify potential data quality issues and sets the stage for effective data governance and management practices.

OBJECTIVE

Understand and classify all data sources to gain a comprehensive view of your data landscape

Activities

- Data discovery and classification
- Identifying data structure, types, and sensitivity

Checklist

- ☐ Where is our data stored, and how is it structured?
- ☐ What types of data do we have, and which are sensitive or regulated?
- ☐ How complete and accurate is our data?
- ☐ What tools and processes do we need to efficiently profile our data?

Practical exercise: Data Profiling

Instructions

- Select a data source (for example, a customer database)
- Use a data profiling tool to analyze the data
- Document the types of data, data formats, and any sensitive data elements
- Identify any data quality issues such as missing values or inconsistencies

2. Define data policies and quality rules

Defining data policies and quality rules is critical to maintaining data integrity and ensuring reliable data across the organization. Establishing clear policies and rules helps prevent data quality issues and ensures that data handling aligns with regulatory requirements and organizational standards.

OBJECTIVE

Establish and enforce data policies and business rules to maintain data integrity

Activities

- Continuous testing and validation mechanisms
- Creating quality rules for data measurement and validation

Checklist

- ☐ What are the critical data policies and quality rules needed for our data types?
- ☐ How can we ensure continuous testing and validation of data quality?
- ☐ What are the potential risks if these rules are violated?
- ☐ Who is responsible for enforcing these policies, and how do we ensure accountability?

Practical exercise: Defining Data Policies

Instructions

- Choose a specific data set (for example, sales transactions)
- Define policies for data accuracy, completeness, and consistency
- Develop quality rules for data measurement and validation
- Implement continuous monitoring and testing mechanisms to ensure reliable data

3. Detect anomalies

Detecting anomalies is essential for identifying deviations from normal data behavior that could indicate data quality issues. Using machine learning to establish baselines and detect anomalies helps ensure that data remains accurate and reliable, thereby preventing potential problems before they escalate.

OBJECTIVE

Use machine learning to identify deviations from normal behavior and detect anomalies

Activities

- Establishing baselines for normal data behavior
- Using AI to automate anomaly detection

Checklist

- ☐ What constitutes normal behavior for our data?
- ☐ How do we automate our monitoring and detection processes?
- ☐ How do we balance sensitivity and specificity to minimize false positives?
- ☐ How often should we review and update our baseline metrics and monitoring processes?

Practical exercise: Detect anomalies

Instructions

- Select a data set with historical records (for example, financial transactions)
- Use a data observability tool to establish a baseline for normal behavior
- Run the anomaly detection process to identify outliers
- Document and analyze any anomalies detected

4. Assess impact

Monitoring for impact involves correlating detected anomalies with business events to assess their potential effects. This step is crucial for prioritizing remediation efforts based on the severity of the impact and ensuring that data quality issues are addressed promptly to prevent business disruptions.

OBJECTIVE

Correlate anomalies with changes and assess their impacts on business operations

Activities

- Root cause and impact analysis
- Prioritization of remediation efforts

Checklist

- ☐ How do we identify and document the root causes of anomalies?
- ☐ How do we determine the downstream AI systems and analytics that are impacted?
- ☐ How do we prioritize issues based on their potential impact?
- ☐ What metrics should we use to assess the impact of data quality issues?

Practical exercise: Impact assessment

Instructions

- Take the anomalies detected in the previous exercise
- Correlate these anomalies with business events or outcomes
- Assess the potential impact of these anomalies on business operations
- Prioritize remediation efforts based on the severity of impact

5. Notify key stakeholders

Effective communication is key to managing data quality issues. Notifying relevant stakeholders about data anomalies ensures that the right people are informed and can take immediate action. Establishing a clear notification and escalation process helps maintain data integrity and operational efficiency.

OBJECTIVE

Alert relevant stakeholders and initiate remediation processes for data quality issues

Activities

- Creating and managing notifications
- Setting up an escalation process for unresolved issues

Checklist

- ☐ Who needs to be notified when a data quality issue is detected?
- ☐ What information should be included in the alerts to make them actionable?
- ☐ How can we ensure timely responses to critical alerts?
- ☐ What is our process for escalating unresolved issues?

Practical exercise: Notification and escalation

Instructions

- Identify key stakeholders (for example, data stewards, data engineers and business analysts)
- Develop a notification process for data quality issues
- Create sample notification messages with necessary context
- Set up an escalation process for unresolved issues

6. Optimize continuously

Continuous optimization is vital for maintaining high data quality standards. By regularly reviewing and refining data policies and practices, organizations can adapt to new challenges and opportunities. This ongoing improvement process ensures that data observability remains effective and aligned with evolving AI, analytics and business needs.

OBJECTIVE

Evolve data observability practices based on insights gained from monitoring and anomaly detection

Activities

- Continuous improvement and feedback loops
- Updating data observability practices based on new insights

Checklist

- ☐ How do we gather and analyze feedback to improve our data policies and rules?
- ☐ What process do we have in place for updating our data observability practices?
- ☐ How do we stay informed about regulatory changes and ensure compliance?
- ☐ How can we measure the effectiveness of our optimizations over time?

Practical exercise: Continuous optimization

Instructions

- Review feedback from previous steps
- Identify areas for improvement in data policies and rules
- Update data observability practices based on new insights
- Implement a continuous improvement loop to keep practices up-to-date

Collibra

From data reliability to Data Confidence™

The path to reliable data requires more than just technical solutions—it demands a holistic approach that brings together governance, quality and observability. When you unify these capabilities, you create a foundation that lets you accelerate AI and analytics initiatives while ensuring compliance.

Success isn't just about implementing tools. It's about enabling your organization to trust, comply and consume data at scale. This means bringing technical and business users into the fold, automating quality controls and building a culture of data confidence.

You're ready to move beyond hoping your data is reliable to knowing it is.

With robust data observability powering your governance foundation, you can accelerate AI and analytics initiatives safely—and help your organization achieve true Data Confidence.

What is Data Confidence?

Data Confidence is the way you and your colleagues feel when your organization can accelerate every data and AI use case — without compromising on safety or quality.

It happens when governance becomes an enabler rather than a bottleneck. Your people can find, understand and use trusted data across every system. Business context flows alongside technical metadata. And policies apply consistently everywhere data lives.

Bottom line: When your people can trust, comply and consume data confidently, innovation accelerates.

That's Data Confidence.

Learn more about Collibra Data Quality & Observability