Collibra

eBook

Getting to data quality: Data reliability in the AI era

Discover the six steps to ensuring data quality in a rapidly evolving data-driven world

Table of contents

Data reliability: The health of your data (and your business) depends on it.

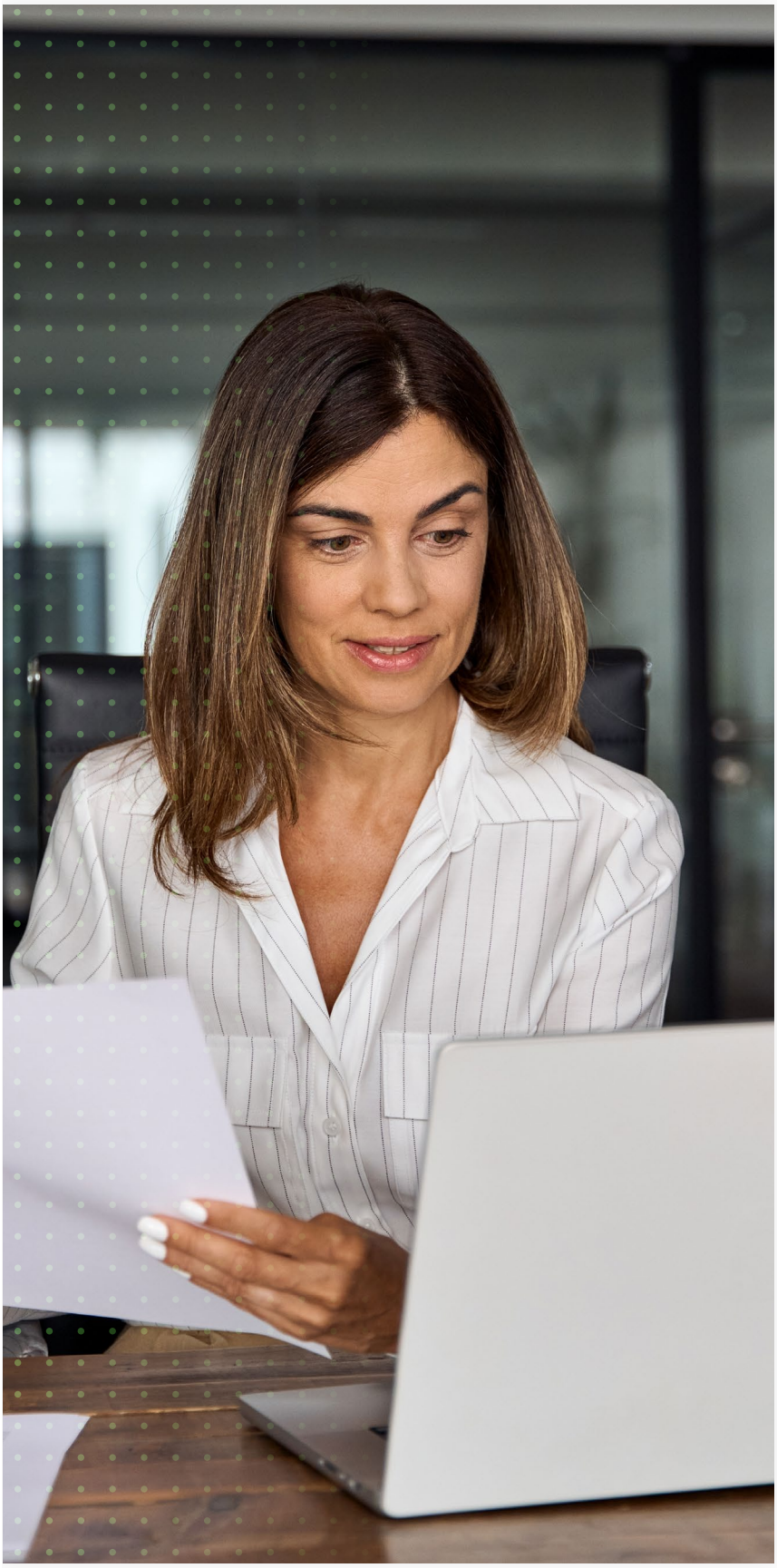
55% of those responsible for data don't trust it.[1]

The recognition that data is the essential enterprise asset has been commonplace for at least a decade. With the ascendancy of advanced AI, however, the importance of data — and, by extension, the critical importance of data quality — has only become greater. But the problem is that the people responsible for data still don't trust it.

The impact of incorrect data is substantial. Organizations are struggling to find scalable and continuous solutions to reduce direct and indirect costs of poor data quality in the face of many challenges, including:

- Exponential interest in AI
- Increasing data volumes
- Expanding regulatory landscape

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



The truth is if more than half of all data leaders don't trust their own data, then you can't expect your colleagues in Marketing, Sales, and Finance to trust it either. The lack of trust is costly. Moreover, the proliferation of data regulations have significantly increased the cost of poor data quality. On average, poor data quality costs organizations $12.9M annually.[2] It's why 50% of organizations will adopt modern data quality solutions to support their initiatives, according to Gartner.[3]

What should you do? Relying on manual data quality and observability processes is no longer sustainable and will lead to poor data quality, resulting in poor decisions and outcomes. Leveraging the power of machine learning (ML) and automating these processes is essential to ensure organizational data integrity and enable effective decision-making. That's why data leaders like you are seeking data quality and observability solutions with flexible rules for safe, trusted business decisions.

Unreliable data can be costly to your organization, including:

- **Inaccurate risk evaluations:** Using the wrong data can mean the cost of improper balancing of financial, operational and security issues
- **Unsustainable resource allocations:** Unreliable data can mean spending too much on budgeting, staffing or inventory. It can also mean missed opportunities
- **Erroneous performance evaluations:** Missing or unreliable data can result in a lack of accountability, losses to competition and outcomes that were never in line with stakeholders expectations in the first place

2, 3. https://www.gartner.com/en/newsroom/press-releases/2023-05-22-gartner-identifies-12-actions-to-improve-data-quality

1. Getting prepared for data reliability

You can improve the health of your data with an effective data quality solution. To get started, you'll want to ask yourself and your colleagues a few fundamental questions.

Question 1.
Are data quality and reliability "events" understood when they occur? How commonplace are they?

Understanding the frequency and nature of data quality issues is crucial, especially as AI investments expand and regulatory landscapes evolve. This insight helps determine the effectiveness of your current data management practices and highlights areas needing attention.

Question 2.
What data quality dimensions or metrics show success for you?

Identify the key metrics that reflect data quality in your organization. This could include accuracy, completeness, consistency, timeliness and validity. Knowing these will help you set benchmarks and track improvements over time, ensuring your data meets the standards required for AI and compliance purposes.

Question 3.
Who across your organization must have visibility on the reliability of data?

It's essential to identify the stakeholders who need access to data quality insights. This could include data engineers, analysts, business leaders and compliance officers. Ensuring that the right people have visibility into data reliability helps in making informed decisions and maintaining accountability, which is increasingly important in a regulated environment.

Question 4.
Do those involved with business and rule validation have technical backgrounds?

Understanding the technical expertise of your team members involved in data governance is crucial. This will help in designing training programs and support systems to bridge any knowledge gaps and ensure effective rule validation, which is vital as AI integration becomes more complex.

Question 5.
How manual is the process of generating data quality rules today?

Assess the current level of automation in your data quality management processes. Manual processes can be error-prone and time-consuming. Identifying opportunities for automation can enhance efficiency and accuracy, which is essential for keeping pace with the rapid changes in data volume and regulatory requirements.

These questions will guide you in evaluating your current data quality and observability readiness. And, they'll help you pinpoint where you need focus and improvements, setting a solid foundation for robust data governance and AI implementation.

The benefits of Collibra Data Quality:

- Monitor data quality issues across many sources with automated and targeted rule writing
- Remediate faster with no-code, business ready DQ dimensions and self-service rules for accurate and consistent data products
- Visualize data health and certify data for trusted business decisions

2.

How to use ML to solve your data observability challenges

As data volumes continue to grow exponentially and the complexity of data sources increases, ensuring data quality and observability has become more challenging than ever. Traditional manual processes are no longer sufficient to manage the vast amounts of data and the rapid pace of change. This is where machine learning (ML) can play a transformative role.

By leveraging ML, organizations can automate and enhance their data observability practices, ensuring reliable data quality, regulatory compliance, and improved decision-making.

However, it's important to take a structured approach when implementing machine learning to tackle data observability challenges.

Here are the six key steps to ensure your team's success:

1. **Profile the data:** Discover and classify your data sources, focusing on sensitive data types
2. **Define data policies and business rules:** Implement continuous testing and validation mechanisms to maintain data integrity
3. **Detect anomalies:** Use machine learning to monitor data and identify deviations from normal behavior
4. **Monitor for impact:** Correlate anomalies with changes and events to understand their impact
5. **Notify key experts:** Alert relevant stakeholders to initiate remediation processes
6. **Optimize continuously:** Evolve policies, rules and reports to enhance data quality goals

Ready to move forward on your journey to data reliability? In the pages to follow, we'll expand on each of these steps, identify common obstacles and provide four questions you can ask to clear your path to data reliability.

3.

The 6 steps to data reliability

Step 1: Profile the data

Start by discovering all your data sources and classifying them, with a particular focus on identifying sensitive data types. Profiling your data helps in understanding the nature, structure, and quality of your data assets. This foundational step is crucial for setting the stage for effective data governance and observability, especially as you increase your AI investments and the regulatory landscape evolves.

Common obstacles

- **Data silos:** Data spread across different departments or systems can be challenging to consolidate
- **Data volume:** Large volumes of data can make profiling and classification time-consuming and resource-intensive
- **Data complexity:** Different data formats and structures add to the complexity of profiling

Four questions to ask

1. Where is our data stored, and how can we access it?
2. What types of data do we have, and which are sensitive or regulated?
3. How complete and accurate is our data?
4. What tools and processes do we need to efficiently profile our data?

Step 2: Define data policies and business rules

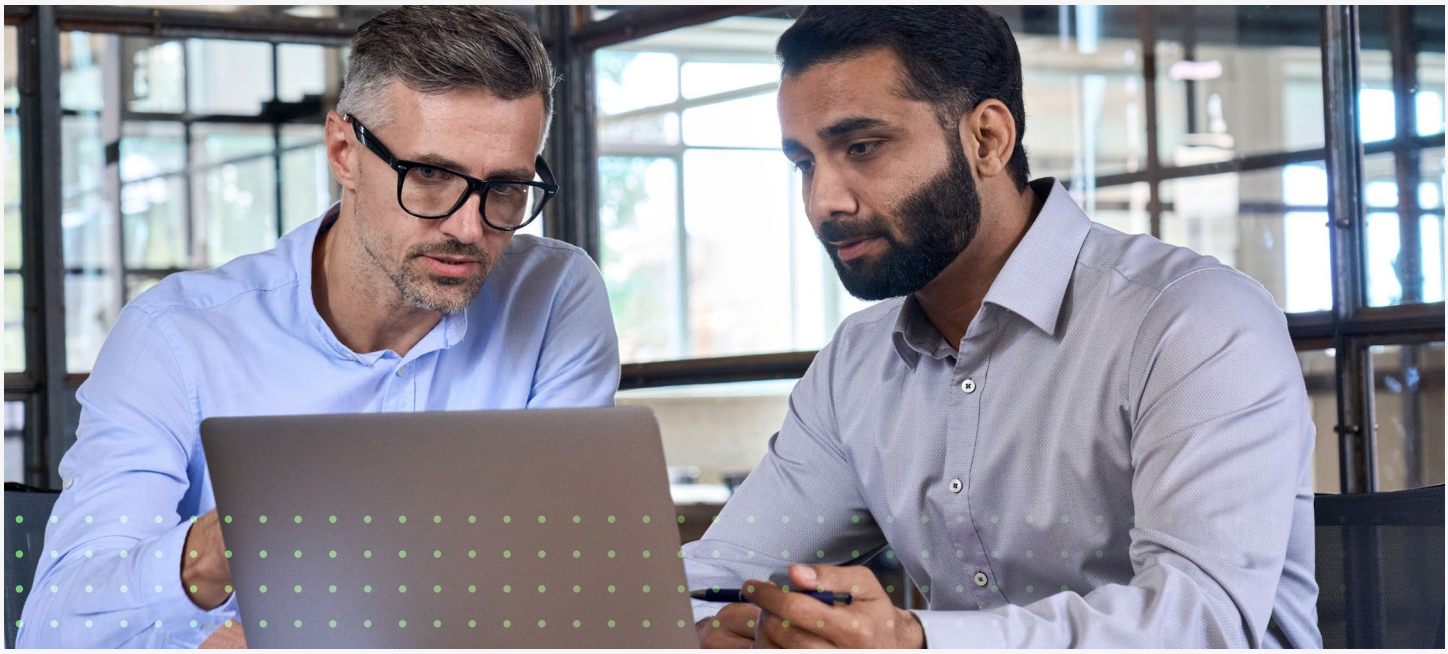
Based on the types of data you have, define policies and business rules that guide how data should be handled. Implement continuous testing and validation mechanisms to ensure data pipelines do not contain data that violates these policies. This proactive approach helps your organization maintain data integrity, while AI becomes more integral to decision-making.

Common obstacles

- **Lack of standardization:** Inconsistent data policies across the organization
- **Regulatory complexity:** Navigating the myriad of evolving data regulations
- **Resistance to change:** Difficulty in getting buy-in from all stakeholders

Four questions to ask

1. What are the critical data policies and business rules needed for our data types?
2. How can we ensure continuous compliance with these policies?
3. What are the potential risks if these rules are violated?
4. Who is responsible for enforcing these policies, and how do we ensure accountability?



Step 3: Detect anomalies

Establish a baseline for normal behavior within your data. Use machine learning algorithms to monitor data continuously and detect deviations from this baseline. Anomaly detection is vital for identifying potential data quality issues before they escalate into bigger problems, which is increasingly important as AI applications depend on high-quality data to function correctly and comply with regulations.

Common obstacles

- **Defining normal behavior:** Establishing accurate baselines can be complex
- **Algorithm selection:** Choosing the right ML algorithms for effective anomaly detection
- **False positives:** High rates of false positives can overwhelm teams and reduce trust in the system

Four questions to ask

1. What constitutes normal behavior for our data?
2. Which ML algorithms are best suited for our anomaly detection needs?
3. How do we balance sensitivity and specificity to minimize false positives?
4. How often should we review and update our baseline metrics?

Step 4: Monitor for impact

Once anomalies are detected, correlate these deviations with unintended changes and other events to identify the root cause and assess the potential impact. Understanding the impact of data quality issues on your operations helps in prioritizing remediation efforts effectively, ensuring your AI models and analytics applications continue to perform accurately and reliably.

Common obstacles

- **Correlation complexity:** Difficulties in accurately correlating anomalies with their impacts
- **Resource allocation:** Limited resources to investigate and resolve detected issues
- **Impact assessment:** Challenges in quantifying the business impact of data quality issues

Four questions to ask

1. How do we identify and document the root causes of anomalies?
2. What tools can help us correlate anomalies with their impacts efficiently?
3. How do we prioritize issues based on their potential impact?
4. What metrics should we use to assess the impact of data quality issues?

Step 5: Notify key experts

Provide contextual alerts to relevant stakeholders—including data engineers, analysts, business leaders and compliance officers—to initiate remediation processes. Timely and informed notifications ensure that the right people can take swift action to resolve data quality issues, maintaining both operational efficiency and regulatory compliance.

Common obstacles

- **Alert fatigue:** Too many alerts can lead to important notifications being ignored
- **Contextual information:** Ensuring alerts contain enough context for quick action
- **Stakeholder engagement:** Ensuring all relevant stakeholders are engaged and responsive

Four questions to ask

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

Step 6: Optimize continuously

Evolve your data policies, rules, and reports based on insights gained from monitoring and anomaly detection. Continuous optimization ensures that your data governance practices remain effective and aligned with evolving data quality goals, supporting the scalability of AI initiatives and adherence to changing regulatory standards.

Common obstacles

- **Continuous improvement:** Keeping up with the need for ongoing adjustments and optimizations
- **Feedback loop:** Establishing effective feedback mechanisms to inform improvements
- **Regulatory changes:** Adapting to new and changing regulations in a timely manner

Four questions to ask

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

The path to data reliability is here

By following these steps, your organization can leverage machine learning to enhance data observability, ensuring high data quality and compliance. This structured approach leads to better decision-making, reduced costs associated with poor data quality, and improved overall data management.

4.

Ensuring data reliability in the AI era

As organizations face the challenges of exponential data growth and increasingly complex regulatory landscapes, ensuring data quality and observability has never been more critical. By implementing machine learning to automate and enhance data observability practices, you can maintain high data quality, ensure compliance, and make better business decisions.

Collibra is uniquely positioned to support your data quality and observability initiatives. Our solutions empower business users to write data quality rules in natural language, reducing reliance on technical expertise and speeding up the validation and testing processes. With our GenAI-based SQL assistant, you can automatically generate checks and convert them into data quality rules without writing SQL, enabling a self-service approach to data governance.

Collibra Data Quality and Observability (DQ&O) leverages Adaptive Rules for intelligent monitoring, providing complete visibility into technical metrics like null checks, row counts and outliers. Our machine learning algorithms ensure these rules are always up-to-date by learning from past and new data. This comprehensive approach connects data issues to their root causes, linking data ownership, lineage and detailed reliability analysis within our platform.

We offer direct integration for data quality across the data catalog, a unique feature among vendors. Our anomaly detection capabilities provide thorough outlier detection, and our health reporting tools offer business leaders like you clear insights into data quality dimensions specific to your needs, without overwhelming you and your data professionals with information.

Collibra's DQ Pushdown enhances efficiency, cost management, security and reduces time to value. Our break record storage and quarantine options for exception records ensure safe and secure handling of data remediation issues.

Collibra

eBook

Ready to ensure data reliability in your organization? Discover how Collibra can transform your data quality and observability practices. Learn more about Collibra Data Quality and Observability.

CO10-P0742-11