Collibra

eBook

Four steps for ensuring reliable data

How CDAOs can improve decision making

Table of contents

Why you need to rethink data quality for AI and analytics success

Nearly ninety percent of C-suite executives plan to increase their spending in data, analytics, and AI.

Accenture: Forge the Link Between Technology and Business Value

Companies in every industry around the world are in a race with competitors to extract value from data. Data is the foundation for generating new products and services, improving the productivity of the workforce, and making business operations run smarter and faster. The last decade has brought tremendous growth in the collection and use of data for a wide range of Artificial Intelligence (AI) and analytics applications. And Generative Artificial Intelligence (GenAI) has created even more excitement and expectations for being able to use data to drive performance and innovation.

However, the exponential growth of data sources and volume of data in those sources creates silos of inaccurate, incomplete and inconsistent data. And changes to data sources, whether intentional or unintentional, break data pipelines feeding downstream AI and analytics applications. The rise in the volume and frequency of change in data has become overwhelming for data teams to manage without automation and intelligence. Continuing to rely on manual data quality and observability processes will result in poor data, poor decisions and poor outcomes. Not only will leadership lose enthusiasm for and confidence in AI and analytics initiatives, the legal and brand reputation impacts can be disastrous.

In this ebook you will learn about four key steps CDAOs can take to overcome data reliability issues. By embracing an AI-powered approach to data quality and observability, you can deliver the reliable data your organization needs to ensure AI and analytics success. Let's see how you can deliver reliable and trusted data to the people who need it, when they need it.

1. Discover and classify

The first step to managing the quality and reliability of your data is to determine where it's located in the complex landscape of on-premises and multi-cloud data sources. Once data sources are discovered, the data types and business entities contained in those sources can be parsed and understood.

> **The amount of data created in 2025 is projected to be 181 Zetabytes.**
>
> Statistica: Worldwide Data Growth

Locate data

Manual processes for locating data cannot keep up with the exponential growth of data. On average companies use over 200 applications[1] and have hundreds or thousands of databases, data warehouses, data lakes, and lakehouses spread across on-premises and multiple cloud environments. To automate the discovery process you need broad connectivity to these diverse data sources. Broad connectivity enables AI to continuously scan metadata in applications, databases, and Continuous Integration/Continuous Deployment (CI/CD) pipelines and immediately detect and notify you of new data sources added to your IT landscape.

Profile data

Manual processes for inspecting and evaluating data can't keep up either. Large enterprise applications like SAP have millions of columns spread across thousands of database tables. 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. Lineage and relationships between data can also be inferred using AI to further speed the discovery and stewardship process.

You can't improve the reliability of data you don't know about. Discovering all your data sources and classifying the types of data they contain is essential for managing data quality. AI-powered approaches to discovery and classification plays a vital role in automating and scaling your efforts, enabling you to keep up with the ever-growing and evolving data landscape.

1. Okta: Business at Work Study

2. Define and associate

After finding your data and understanding what type of data you have, you can define and associate appropriate policies and rules to your data. Policies and rules provide the what and how of ensuring high-quality data fit for AI and analytics.

> **67% of executives cite potential errors as the top risk of AI adoption.**
>
> CFO Dive: Flawed Data Ranked as Top AI Risk

Quality policies

Well-structured policies provide a clear framework for what data quality dimensions should be managed, service level agreements (SLAs) for quality, and accountability for delivering on the quality SLAs. Policies address common dimensions of quality including completeness, accuracy, validity, consistency and timeliness, as well as the thresholds to measure quality against and the processes for monitoring and reporting issues. Defining the data types, systems and processes the policy applies to as well as the roles and responsibilities necessary for maintaining data quality are also part of policy definitions. For example, a policy for customer data might require customer addresses to be complete, accurate, valid, up to date and consistent across systems. The AI techniques used to find and classify data also enable you to automate mapping of data policies with data assets and enforcement of those policies.

Quality rules

Rules operationalize the principles laid out in policies by providing specific instructions and actions to ensure data meets organizational quality standards. For example, a rule for the United States zip code field in customer records might mandate the field cannot be empty, the format is five digits, it only contains numeric characters, and is verified as a valid zip code. 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.

Defining and associating policies and rules with your data assets provides the necessary standards and controls for reliable data. AI-powered approaches to definition and association of policies and rules plays a vital role in automating and scaling your efforts, enabling you to govern and manage your data quality effectively and efficiently.

3. Monitor and assess

Associating policies and rules with your data assets enables you to continuously monitor and assess the quality of your data. To ensure reliable and trusted data, monitoring and assessment needs to happen both with data at rest in sources, as well as data in motion in pipelines.

> **Data scientists spend 60% of their time cleaning data.**
>
> ResearchGate: What Data Scientists Spend Their Most Time Doing

Pattern analysis and anomaly detection

Continuous data monitoring allows you to establish a historical baseline of expected patterns and ranges for data within sources and pipelines. Once this baseline of normal behavior is defined, you can set thresholds based on it to detect anomalies. AI can automate the calculation of normal ranges for data columns and perform data quality checks for issues such as incorrect formats, types and missing values. Additionally, AI can automate schema change detection, identifying when columns are added, removed or modified. By automating the identification of patterns, trends and outliers, you can significantly improve the productivity of your data quality efforts.

Cause and impact analysis

Once anomalies are detected, AI can be used to determine the likelihood that a particular event or change caused the data quality issue. For example, if monitoring detects a sudden spike in the number of null values in the "CustomerID" column of sales transaction data, AI can correlate this anomaly with timestamp data. The AI identifies that the issue began when a new data source was added to the ingestion process. By using data lineage to trace the flow of data through the pipeline and performing dependency analysis, AI determines that the lakehouse and sales reports will be impacted by this error.

Monitoring and assessing your data, both at rest in sources and in motion through pipelines, enables you to quickly identify data quality issues before they become major problems. AI-powered approaches to monitoring and assessment plays a vital role in automating and scaling your efforts, providing the visibility and speed you need to ensure reliable data for decision making.

4.

Respond and adapt

Once issues, causes and impacts are identified you must respond to resolve current problems. Monitoring and assessment also enable you to continuously adapt your data quality processes to ensure reliable and trusted data in the future.

> **Just 32% of organizations say they are adequately mitigating inaccuracy.**
>
> McKinsey: The state of AI in 2023

Notification and workflow

When an issue is detected relevant stakeholders can be automatically notified about anomaly details, suspected causes, and impacted systems and processes. Workflow in tracking systems like Jira or ServiceNow, can be initiated based on the type, cause, and severity of the anomaly with assignment of tasks to responsible team members, and setting of priority levels for resolution. For instance, in the case of the null values in the "CustomerID" column example mentioned earlier, the data steward responsible for the source system receives a high-priority task for remediation. Simultaneously, the data engineer overseeing the pipeline and the developer managing the sales report are also notified of the issue and the assignment given to the data steward.

Continuous optimization

The goal of data quality and observability is to continuously adapt to your data landscape, ensuring reliable and trusted data. As new data is constantly added, AI can automate the application of existing rules to data and create new rules based on profiling and classification of the new data. Policies and rules are not static either and must continually adapt to evolving needs. By analyzing historical data, AI can define what is "normal" for each data set and recommend updates to adapt existing policies, rules and thresholds. For example, AI may suggest stricter validation thresholds for data types or sources that frequently cause issues. Beyond automatically tuning policies and rules, AI can also help provide alerts, reports, and other insights to data stewards so they can be more effective and efficient in managing and governing data quality.

Responding and adapting to data quality issues is essential for managing current problems and preventing future ones. AI plays a vital role in automating and scaling your efforts, ensuring reliable and trusted data.

CDAO checklist for reliable data

The four key steps explained in this ebook can help CDAOs provide reliable and trusted data for AI and analytics. Using them as a checklist helps you empower everyone in your organization to be confident in using data to make decisions.

1. Discover and classify

Locating and profiling data are the first steps in creating a continuous process for managing and governing data quality. AI plays a vital role in automating and scaling your efforts to keep up with the ever-growing and evolving data landscape.

- ☐ Locate data
- ☐ Profile data

2. Define and associate

Defining and associating policies and rules with your data assets provides the necessary standards and controls for reliable and trusted data. AI plays a vital role in automating and scaling your efforts to govern and manage your data quality effectively and efficiently.

- ☐ Quality policies
- ☐ Quality rules

3. Monitor and assess

Monitoring and assessing your data, both in sources and in pipelines, provides the necessary visibility and speed to ensure reliable and trusted data. AI plays a vital role in automating and scaling your efforts to quickly identify issues before they become major problems.

- ☐ Pattern and anomaly analysis
- ☐ Cause and impact analysis

4. Respond and adapt

Responding and adapting to data quality issues is essential for managing current problems and preventing future ones. AI plays a vital role in automating and scaling your efforts to create a continuous process for managing and governing data quality.

- ☐ Notification and workflow
- ☐ Continuous optimization

Data is more important than ever to business success.

It fuels innovation and growth, enhances customer and employee experiences, increases efficiency and productivity, and improves risk and cost management. The ever growing volume of data means manual processes for managing and governing data quality will never provide the scale and agility required for today's business needs. Chief Data and Analytics Officers who embrace an automated and intelligent approach to data quality and observability can empower employees across functional areas of the business with reliable and trusted data that drives better decision making and business performance.

Using an AI-powered, continuous cycle of discovering and classifying data, defining and association policies and rules, monitoring and assessing quality, and responding and adapting to issues you can deliver the reliable data your organization needs to ensure AI and analytics success.

Want to learn more about how CDAOs can close the AI trust gap? Check out this ebook