Collibra

Whitepaper

Seven data quality features to ensure AI success

Written by Sanjeev Mohan, Principal SanjMo

Let's face it. The shine is off artificial intelligence (AI) and large language models (LLMs) are fast becoming commodities. But AI development is not slowing down although the focus has shifted to building data and AI solutions that have demonstrable ROI and deliver value.

The data teams have learned the lesson that the most important 'moat' for any business is their data. However, it has to be high quality, reliable, contextual, and secure. Data isn't static. While workflows move data from system to system, pipelines deteriorate multiple times during their lifecycle. This deterioration may be caused by upstream data changes, newer fields being added or logic change. Tracking and governing changes over time can provide insights into your data's journey over a longer duration at an individual data pipeline and dataset levels.

This paper explores what should comprise a comprehensive data quality and reliability framework and how it enables organizations to deliver analytics successfully. The framework underscores why data and AI solutions will fail to deliver the corporate mandate if data quality issues aren't tackled in a timely and comprehensive manner.

Comprehensive Data Quality and Reliability Framework

Data quality is a cornerstone of successful data-driven decision making because the outcomes are only as good as the data they are based on. Hence, it is essential to ensure that data is accurate, complete, consistent, relevant, timely, and accessible. The outcome of comprehensive data quality ensures organizations can trust the data they use, leading to more reliable and impactful business decisions.

Besides building trust and credibility, effective data quality reduces costly mistakes, faulty forecasts and other operational inefficiencies. Clean and reliable data minimizes the need for rework, mitigates risks, and saves resources.

Maintaining data quality is essential for meeting regulatory standards and compliance requirements, which are becoming increasingly stringent across industries. Clean and reliable data helps avoid legal penalties and safeguards the organization against regulatory breaches.

High-quality data serves as a foundation for exploring new opportunities, optimizing processes, and driving innovation. It also enhances customer satisfaction when customer preferences, behaviors, and interactions are correctly captured and utilized. This can lead to higher customer loyalty and stickiness. In other words, data quality is not just a technical requirement but a business imperative and truly a differentiator between successful and also-ran companies.

But can data teams ensure they have adequately handled all aspects of data quality and reliability? That requires a comprehensive framework as shown in Figure 1.

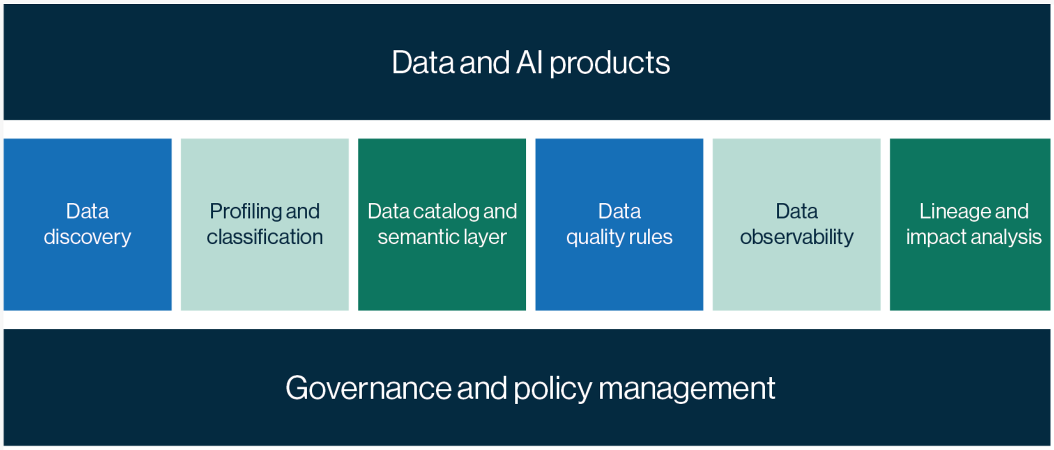

Data and AI products
Data discovery
Profiling and classification
Data catalog and semantic layer
Data quality rules
Data observability
Lineage and impact analysis
Governance and policy management

Figure 1: A structured and comprehensive approach to data quality will lead to reliable and trusted data

This integrated approach leverages metadata to ensure users can derive real-time, trusted insights. Some vendors offer capabilities that are focused on specific areas but most businesses seek an integrated and comprehensive set of capabilities to holistically govern end-to-end data pipelines. Such an approach unifies data management practices across the organization, ensuring consistency, accuracy, and reliability. By exposing data quality across diverse sources and systems, uniform data quality standards, policies, and processes can be established. When data sources are spread across internal and external locations, data quality issues are often exacerbated by siloed teams. An integrated approach fosters collaboration across departments (e.g., IT, data management, marketing, finance) by breaking down silos and encouraging a shared responsibility for data quality.

Let’s examine each layer.

Data Discovery

Data discovery is the first foundational element to deliver reliable data and AI products because it provides a comprehensive overview of all available data assets, including their sources, formats, quality, and relationships. By understanding the full data landscape, organizations can identify the most relevant data for their needs, ensuring that only accurate, reliable, and high-quality data is used in data and AI initiatives. Knowing what data exists and its current state helps avoid the use of outdated, irrelevant, or erroneous data, which could compromise the quality of data and AI products.

As organizations increasingly rely on data to drive business decisions, the ability to discover, understand, and utilize data effectively has become more advanced and essential.

Data discovery, once a simple task of locating data sources, has now expanded into a sophisticated process that leverages the power of large language models to find relationships that may not have been as evident. These hidden patterns and insights help in assessing whether the data is suitable for the intended purpose and set up the subsequent steps of data profiling, cataloging, and data quality and observability. Traditionally, the data discovery layer has connected to the structured and unstructured data sources and business applications to create technical or operational metadata. In the modern world of AI, synthetic data should also be included in the scope of data discovery because this artificially generated data mimics the characteristics of real-world data while not revealing any actual user information or removing any biases. In many cases, real-world data is incomplete, missing key records or features necessary for building robust analytical models. Synthetic data can help fill these gaps by creating representative data points, ensuring datasets are complete and can be effectively used for analysis or machine learning.

A data discovery tool should be able to connect to the data sources using optimized native connectors or using open standards like ODBC/JDBC, and APIs, including RESTful services, SOAP, and GraphQL. Native connectors are designed to consistently meet high-throughput, performance, security, and reliability needs, which enhances the efficiency of data discovery efforts. Often, the connectors help with real-time extraction of metadata for newly created or updated data in the sources. This is achieved through the change data capture (CDC) capabilities.

Data Profiling

Organizational data is often messy and has issues, such as quality (e.g., missing values, duplicates, inconsistencies), lack of understanding, and integration challenges across disparate sources. Hence, the need to uncover hidden relationships and identify data drift over time. This transparency can help optimize data processing workflows, improve the reliability of analytics, and assist in meeting compliance requirements. In other words, without this clarity, organizations risk making decisions based on flawed data, leading to inefficiencies and potential compliance risks.

Profiling enables organizations to gain a clear view of their data landscape by ensuring data is accurate, relevant, and well-understood. It enriches the data team's knowledge of the discovered data by helping them understand its characteristics, like uniqueness, cardinality, range of values, and sensitivity. It involves collecting statistics, metadata, and other information about the data that is then used for data management, data quality assurance, regulatory compliance, and analytics initiatives.

Data profiling uses a sophisticated set of algorithms to assess the context and quality of data across various dimensions. This process is often automated but should be customizable to an organization's specific needs. These needs may pertain to merging of data from different sources by identifying dependencies and relationships between columns or identifying unnecessary duplicate information or highly correlated columns for data optimization and storage efficiency. Other needs may pertain to data preparation, anomaly detection, data migration, business rule management or mitigating risks because of poor data quality.

Profiling of source data can be compute intensive and can potentially slow down the operational system. In order to minimize load on source systems, profiling may be done on a user-defined sample of data. Another option is to extract the data into an external cluster using, say, Apache Spark. Each option has its own trade-offs and businesses should choose the option that best meets their requirements.

Data profiling can be run on-demand or be scheduled to run at certain intervals. Additionally, workflows help automate the process. These workflows should have the ability to integrate with an off-the-shelf orchestration engine, like Apache Airflow for efficiency, reliability and scalability reasons. Automated data profiling reduces manual intervention, minimizes errors, and ensures that workflows run smoothly.

Data Classification

Once the data is profiled, it must be tagged or classified into a structured format to improve data management, usage, governance, and usability. Without classification, data can become disorganized, leading to challenges in ensuring data quality, maintaining compliance with regulations, protecting sensitive information, and optimizing data integration and analysis. Unclassified data can increase the risk of security breaches, regulatory violations, and operational inefficiencies, as it hinders the ability to apply appropriate controls and governance.

Data classification tags act as metadata descriptors that make it easier for users to search, find, and access relevant data and to categorize data according to its sensitivity, usage, and ownership. This metadata can provide context to AI and machine learning models, thereby reducing hallucinations.

This classification process of creating tags or labels is also used in the later stages of the data quality and reliability framework, such as applying quality rules and access policies.

Once again, ML algorithms are used to automate creation of the labels, but large language models (LLMs) are now being used to leverage their understanding of semantics which significantly improves traditional classification which only used keywords and static taxonomies. For example, LLMs use their understanding of language patterns and context to automatically classify text into predefined categories, such as sentiment analysis (positive, negative, neutral). Finally, LLMs can also be used to classify unstructured data, such as emails, chat messages, social media posts, and other non-traditional data formats.

To ensure high standards in data classification, two measurements are used:

- **Precision:** Measure of accuracy of the positive predictions made by the model. Higher precision shows that classification is relevant and accurate. A high precision indicates that the model makes very few false positive errors. When there are false positives, tickets are raised to request data owners to perform the necessary remediation.
- **Recall:** Measure of the ability of the model to identify all relevant positive instances or its completeness. Higher recall indicates that classification did not miss tagging data elements.

In summary, automated profiling and classification can proactively and cost effectively detect anomalies, inconsistencies, and errors in datasets, alerting data engineers and stewards to potential issues before they impact downstream processes.

Data Catalog and Semantic Layer

Organizational data is often siloed, difficult to locate, poorly documented, and inconsistently managed, leading to inefficiencies, errors, and compliance risks. Users struggle to find the right data, understand its context, and trust its quality, which hampers data-driven decision-making and collaboration.

A data catalog addresses these issues by centralizing data discovery, enhancing data governance, and promoting data literacy, ultimately enabling more effective and compliant use of data across the organization. It helps users understand what data is available, where it comes from, how it can be used, and its quality and governance status.

Metadata forms the foundation of a data catalog to provide a comprehensive view of the data assets within an organization. The metadata that has been generated in the discovery, profiling, and classification phases is stored in the data catalog so that it can be searched and utilized in the decision-making process. The metadata in the catalog is of three types:

- **Technical:** Describes the technical aspects of data, including its structure, storage, and processing details, such as schema information, column data types, data distribution histogram, indexes, etc.

  For files, data catalogs show data formats, like CSV, JSON, and XML etc. and infer schemas if they are not explicitly defined.

- **Operational:** Focuses on the data's usage, performance, and lifecycle, such as ownership, data retention policies, data refresh schedule frequency, and access policies.

  The data access and security policy information helps enforce data governance policies to ensure compliance with regulations (like GDPR or CCPA), and provides an audit trail for how data is used, transformed, and shared.

It also helps in observability and lineage (discussed later) as it includes usage statistics (e.g., frequency of access, most queries tables, most active users) and performance metrics (e.g., query response times, system load.)

- **Business:** Provides a business context to the data, making it understandable and relevant to business users. It is also called the semantic layer and includes business glossary and terms (e.g., KPIs, metrics, dimensions), business rules, contextual details of how data is used in business processes.

Descriptive and contextual business descriptions are now being automatically generated with the help of LLMs' capabilities and stored in the data catalogs. This further eases the ability to do natural language Q&A of metadata. Traditionally, catalogs allowed consumers to search using keywords, but with the advent of integration with LLMs, catalogs now support semantic searches using natural language.

The terms in the business glossary are mapped to the underlying technical metadata. Domain experts can define business-friendly and intuitive terms that are more relevant to performing analytics. The glossary can also organize terms in a hierarchy and provide an audit log of changes for full transparency and history. Several standards, including BPMN, OMG SBVR, Object-Role Modeling, Fact-oriented Modeling, RDF/OWL and SKOS are available to manage business glossaries but are outside the scope of this document.

Recent years have seen an expansion of data catalog's scope by housing all types of data assets, including data products, advanced analytical models, reports, rules, and KPIs, etc. Although data catalogs, as the name suggests, were originally launched to discover and search metadata, they are now being used to develop new assets.

For example, a user can search for a data product and then combine it with other assets to build and publish a new data product. In this scenario, the data catalog becomes a marketplace offering shareable data assets. These marketplaces may simply allow data sharing or have the ability to calculate usage and do chargebacks, thereby enabling monetization of data. This is an exciting future for data catalogs as they evolve into strategic products to create new revenue streams.

Data catalogs, in essence, act as an excellent collaboration workspace for data producers and engineers and data consumers and business stakeholders. Users can annotate, rate, and rank data assets, allowing consumers to effortlessly 'shop' for the right product and trust it. Data contracts is an emerging concept that defines the attributes of data assets stored in the data catalog so consumers can build service level agreements (SLAs) around their offerings and solutions.

In summary, data catalogs significantly enhance the ability to unlock the value of all types of data and analytics assets to improve decision-making. Data becomes a strategic asset that is well-documented, enriched with metadata, trusted, and easily accessible. By increasing data utilization, companies can improve efficiency, generate new insights, and optimize operations.

Data Quality Rules

Data quality refers to ensuring the data set is accurate, consistent, complete, and reliable for decision-making. High-quality data is foundational for any organization aiming to derive actionable insights from their data assets and to reduce hallucinations from LLMs.

Traditionally, data quality was assessed using custom rules built with manual queries. However, as data transforms rapidly and as new types of data are created, manual processes are not sustainable due to the dynamic nature. This led to statistical analysis of the profiled data using standard deviation and Z-Score to determine how data is clustered around the mean and thereby identify outliers. Random Forest is particularly good at identifying potential relationships among certain data sets in the profiled data. Levenshtein Distance uses fuzzy matches to detect potential duplicates. Other ML algorithms may be tuned to pinpoint frequency of data values as low frequency values may indicate anomalies.

These ML algorithms help predict potential failures or bottlenecks in the pipeline, allowing for proactive remediation. They automatically detect deviations from expected patterns, signaling potential data quality issues or pipeline failures.

Data Observability

Data observability is the ability to understand and monitor the health of data pipelines, ensuring that data flows smoothly from source to destination without degradation in quality, reliability or cost performance. As organizations scale their data operations, maintaining visibility into the health of these pipelines becomes increasingly complex. Hence, continuous monitoring of data as it moves through the pipeline provides visibility into anomalies, failures, performance degradations, cost overruns and accuracy issues at every stage.

Figure 2 shows the key components of a data observability workflow which helps identify and troubleshoot issues faster than the cases where no data observability product is in use.

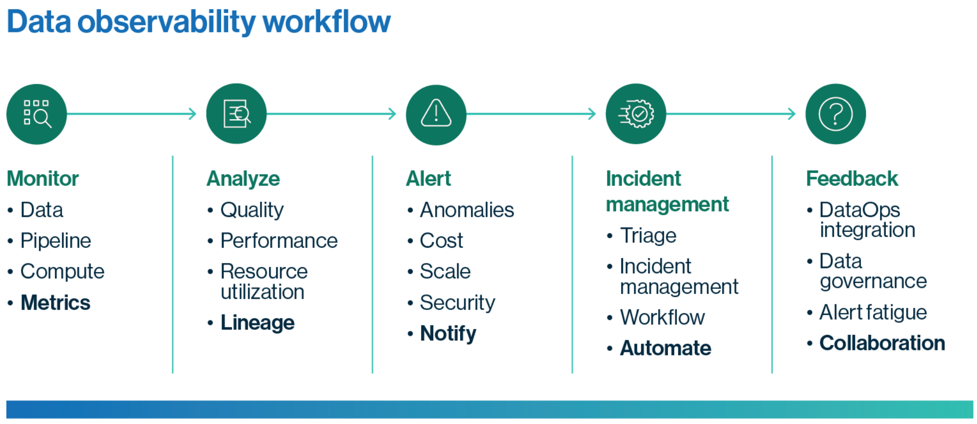


Figure 2: Data observability workflow

The various components of the data observability workflow includes:

- **Monitor**:

  Continuous monitoring of data and metadata enables detection of patterns and anomalies as soon as the problem occurs. Organizations should prioritize critical data elements and relevant data sources based on strategic imperatives in order to maintain an effective focus and reduce unnecessary alerts. They should monitor key metrics like data drift, volume, quality, SLAs, and resource usage to ensure comprehensive oversight.

- **Analyze**:

  Analysis of data and metadata helps identify hidden patterns, failures, and anomalies, enabling timely, and even proactive, interventions to prevent downstream impacts. Effective observability tools dynamically detect drifts, optimize resource use, and continually retrain models to maintain system efficiency and accuracy.

- **Alert**:

  When many fine-grained alerts are generated the responders stop paying attention. This situation is called 'alert fatigue'. Hence, it is crucial that the data observability tool manages alerts intelligently and escalates the most critical ones.

  Data observability tools proactively alert teams about anomalies and manage alert fatigue by intelligently adjusting thresholds based on normal ranges, as well as categorizing or customizing notifications to reduce unnecessary notifications. This approach helps ensure critical alerts are attended to, improving pipeline uptime and issue resolution speed.

- **Incident Management:**

  Incident management enables root cause analysis, preventing technical debt by addressing issues at their source rather than downstream. This process supports collaboration across business units to initiate remediation steps, improving overall system reliability.

  Remediation of anomalies is often a manual step because the mission-critical source systems may have their own stringent operational processes to update data. Often, the data steward teams receive alerts when data quality or reliability threshold are breached and then take the necessary action based on the priority of the notifications.

- **Feedback**:

  Feedback loops in data observability ensure that the system continuously evolves and meets SLAs. Operational feedback, like latency or missing data, drives immediate improvements, while business feedback fosters adoption by demonstrating value through enhanced data quality checks and deployment transparency.

AI is further advancing the data quality and observability space. LLMs are adept at understanding the semantics and using Euclidean distance to find similarities. In addition, if copilots can write a full-featured code for us, why not the rules? The idea is to leverage AI to infer hidden relationships and context patterns to detect and write rules automatically and then apply them.

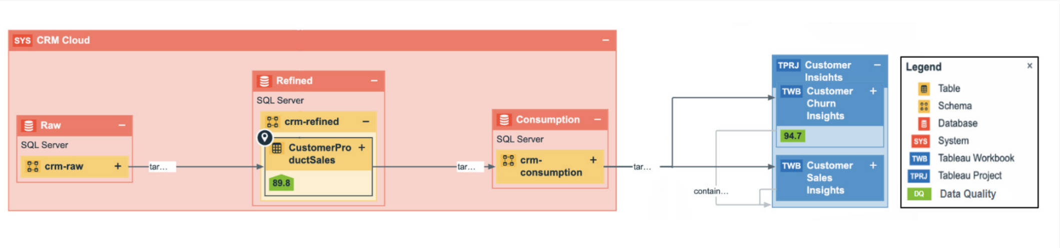
The data observability capability can pinpoint the impact of changes in the source systems on the downstream systems. This process is called impact analysis and relies on the understanding of the data pipeline lineage. This is the last step in the framework and is covered next.

Lineage and Impact Analysis

Understanding data lineage helps organizations trace the path of data from its source to its final destination, providing insights into how data is used, the level of accuracy, and making it easier to identify and resolve issues, such as errors or inconsistencies.

Lineage tracks and documents the data's origin, movement, and transformation as it flows through various systems and processes within an organization. It's like a map that shows where the data comes from, how it's processed or altered, and where it ultimately ends up.

Its use cases include regulatory compliance, data quality management, or operational efficiency. As a result, various personas consume lineage output. For example, non-technical stakeholders understand how data flows align with business operations, decisions, and reporting requirements. Data engineers gather insights into internal processes within the technology stack, such as how data is transformed from raw inputs to processed outputs. Data scientists use the lineage of datasets and models for managing reproducibility, compliance, and model integrity.



In this representative lineage example the raw CRM data in SQL Server is refined, and made available in the consumption zone of a cloud data lake. A Tableau workbook providing customer sales and churn insights consumes the data in this example. Data quality score (89.8%) is also visible at the Customer Product Sales table, and at consumption (94.7) in the Tableau project.

Lineage should be available at any physical layer, like schema and tables to the lowest level of granularity, like columns for more precise impact analysis and debugging. Additionally, lineage should include the transformation logic between the assets. This requires extraction of transformations metadata from ETL/ELT, SQL, and BI tools. If data elements or transformations change, a lineage diagram should make it easy-to-understand the impacts in both upstream and downstream systems.

To help improve data reliability the data quality rules, dimensions, metrics and scores can be overlaid onto the lineage, which enables the business users to identify where the relevant quality controls are being implemented.

AI can automatically infer and predict data lineage using similar data sets. This enables organizations to quickly identify the origin of data, understand its journey, and assess the impact of any changes or errors even for new datasets. By simplifying these complex processes, AI helps maintain data integrity, support compliance efforts, and enhance decision-making by providing clear insights into data dependencies and potential risks.

A new open standard called Open Lineage is in beta currently. Once this standard is available and widely accepted by data governance, ETL and BI products, bi-directional sharing of metadata will become much easier.

Governance and Policy Management

So far, the focus of this document has been on discovering, profiling, classifying, storing, and sharing of metadata and creating derived data products. But organizations need to ensure that data consumers adhere to agreed-upon usage and governance policies. The governance and policy management function supports and ties all the other pieces of the framework together. The policies pertain to definitions, rules, metrics, roles, responsibilities, workflow and processes:

- **Data policies:** pertain to classification, quality, usage/privacy, security, etc
- **Business terms policies:** glossaries, classification, metrics, etc.
- **Stakeholder management policies:** who does what, who needs to be notified, etc
- **Process policies:** what are the processes for issue management, what are the processes for policy, rule, metric creation and approval
- **Data access policies:** ensure that data consumers can only access data they have been authorized to see to protect privacy, manage data breach risks, and meet compliance guidelines.

Organizations enforce policies for various reasons, like the right to be forgotten (technically, right to erasure), data retention, access control, and usage. The data governance platform acts as a single pane of glass for managing security policies across all the underlying technical platforms, thereby ensuring consistency.

The data access policies start by detecting where all personal, sensitive data exists in the pipeline and then classifying it according to the security, privacy, legal and compliance requirements. This visibility is crucial for managing data privacy risks for data flow across various systems, applications, and storage locations. It can also detect if redundant copies of data exist to introduce processes to reduce the attack surface by limiting unnecessary data proliferation. The concept of data minimization is mandated by many compliance regulations, like the EU GDPR. Other common regulations with specific guidelines include PCI DSS 4.0, California Consumer Privacy Act (CCPA), and HIPAA, etc. Discussion of these regulations is outside the scope of this document.

Next comes the policy management stage where access and usage policies are first defined and then enforced. Organizations should prioritize protection measures based on the sensitivity and criticality of the data, ensuring that the most at-risk data is secured first for the identified use cases.

Ideally, policies can be written in natural language using drop down options as administrators are not always adept at writing complicated logic in languages like SQL or Python. The data observability tool applies the policies to the relevant tags created during the classification stage and validated by the respective owners. For example, if the social security number is tagged as sensitive, then the policy may say this data should be encrypted, tokenized or redacted except the last four digits for most of the data consumers.

While a data catalog serves as a central repository where data governance policies are defined, managed and enforced, the underlying systems handle their execution. These systems may be on-premises or across different cloud providers. Access policies are enforced using role-based access control (RBAC), attribute based access control (ABAC), masking, tokenization, anonymization, pseudonymization and various other newer approaches, like differential noise. These approaches are used to adhere to various security and compliance regulations, especially for sensitive data like personally identifiable identity (PPI), Payment Card Information (PCI) and Protected Health Information (PHI).

In this section of the data quality and reliability framework, the focus is primarily on protecting data assets, but the space of security is much wider. It includes multi-factor authentication (MFA), firewalls, intrusion detection and prevention systems (IDPS), antivirus software, endpoint detection and response (EDR), Data Loss Prevention (DLP), and Security Information and Event Management (SIEM,)etc. These are used to monitor and control users, network traffic, and devices.

Closing Thoughts

So, why is the need for good governance - including quality, reliability, access control, lineage, observability, semantic layer etc. more important than any time in the past?

Because more people are accessing more data for more business use cases than ever before. Without trusted and reliable data for AI and analytics, the outcomes will be poor, time and money will be wasted, and business leadership will lose enthusiasm for and confidence in AI and analytics. A structured and comprehensive approach to governing data will enable your organization to deliver the high quality and reliable data needed for AI and analytics success.

Some ways a disciplined approach can accelerate AI and analytics development and deployment, and improve the accuracy and performance of solutions include.

- **Increasing Data Quality Transparency:** By providing visibility into different dimensions of data quality like accuracy, completeness and consistency for AI development and operations. Data quality transparency reduces the risk of errors, biases, and unreliable outputs in AI applications.
- **Enabling Retrieval Augmented Generation:** By enabling access to trusted and reliable data that ensures the accuracy of Generative AI model outputs. This approach eliminates hallucinations and enables fact checking and validation of generative AI output.
- **Building Trust in AI Output:** By providing visibility into data pipelines, how data is being processed, and the behavior of AI systems in real-time. This gives customers, employees, and regulators confidence that AI outputs are within defined operating ranges and meet expected standards.

Establishing your organization at the forefront of using AI and analytics to improve business outcomes requires immediate action on trusted and reliable data to fuel the AI and analytics engine. High-quality data to train and augment your AI models leads to high-quality model outputs and better business results. Failure to act decisively on this foundation for AI and analytics adoption and use will put you at risk of falling behind.

Sanjeev Mohan
Principal, SanjMo

Sanjeev Mohan is an established thought leader in the areas of cloud, modern data architectures, analytics, and AI. He researches and advises on changing trends and technologies and is the author of Data Product for Dummies. Until recently, he was a Gartner vice president known for his prolific and detailed research, while directing the research direction for data and analytics. He has been a principal at SanjMo for over two years where he provides technical advisory to elevate category and brand awareness.