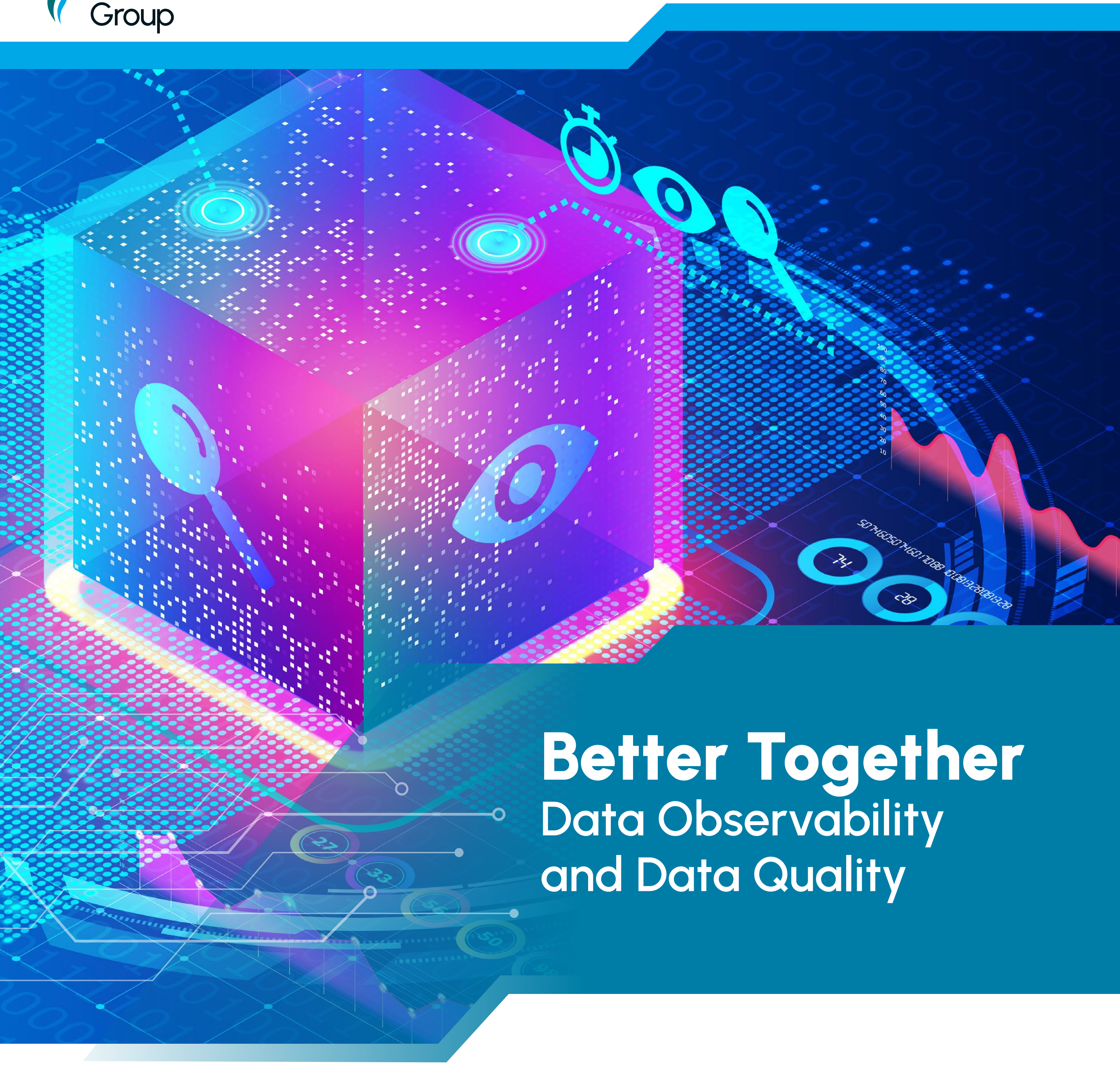
The Futurum Group



Better Together

Data Observability and Data Quality

AUTHOR

Steven Dickens
Vice President and Practice Lead | The Futurum Group

Dave Raffo
Senior Analyst | The Futurum Group

OCTOBER 2023

IN PARTNERSHIP WITH

IBM®

Executive Summary

Exponential data growth across all industries puts pressure on organizations to guarantee their data is accurate and accessible. Because data is more important than ever, data quality has never been more important than it is today. As companies strive to derive value from their data, they need to know they have the right data in the right format.

This emphasis on good data brings great demands for IT vendors who sell data management products and services. To guarantee data quality, they need to incorporate features such as data observability to uncover issues with corporate data. Observability adds a new dimension to data quality for the enterprise, yet the concept is not always understood by people who work with data in their everyday roles.

This paper shows how data observability can help prevent data quality problems for specific data use cases. Although it is not designed to address all data quality use cases, data observability complements other data quality tools such as data catalogs that discover data. When used correctly by data engineers in a data fabric, data observability can help guarantee an organization's data is of the highest quality.

Introduction: What Do We Mean by Data Quality and Data Observability?

Making sure data is accurate and accessible gets more difficult as the amount of data grows. This growth will only accelerate, especially with new AI tools rapidly requiring even more data. Large language models (LLMs) require reliable data to feed their models.

Along with data growth, companies must also deal with data trust – how do they know if they're working with good data? Data reliability ensures data is complete, accurate, and consistent wherever it is stored. Only consistent and accurate data is truly reliable.

Data quality is a measure of the condition of data. It determines accuracy, completeness, consistency, reliability, and whether it is up to date – factors that separate good data from bad data. Data quality metrics help organizations find data errors that must be resolved and determine whether the data in their IT systems can serve its intended purpose.

That is where data observability comes in. Data observability tools help companies understand the state and health of their data. These tools help identify, troubleshoot, and resolve data issues in real time. Data observability is used by data platform, data analytics, and data engineering teams responsible for integrating and automating data flows between managers and consumers inside an organization.

An effective data observability tool should:

- Collect and process telemetry data across data sources.
- Review through comprehensive monitoring of networks, servers, databases, other on-premises and cloud applications, and storage the health of data sources
- Alert when abnormal behavior is detected
- Provide a centralized repository to support data retention and fast data access.
- Provide all necessary information to fix and debug any data issuespractices.

Critical Components of Data Observability for Data Fabric and Data Mesh

A well-designed data architecture is a foundational element for ensuring high data quality. An organization's data architecture is the key to how its data assets and management resources are structured. It is the way organizations enforce data governance through the use of metadata and AI technologies, data quality monitoring, data modeling lineage for root cause analysis of data quality, dataset value determination, and proper data hygiene. Ensuring data quality requires an effective data architecture that provides a framework for collecting, transporting, storing, security, sharing, and using data for data science use cases.

Data architectures have evolved from monolithic, centralized approaches such as data warehouses, business intelligence, and big data platforms. These architectures involved data ingestion, processing, cleansing, aggregation, and serving. Some of these systems could be useful only to data engineers, and others created technology hurdles that affected scalability. They could not keep up with modern data growth challenges, such as constant change and growth of sources, processing, and number of consumers. Organizations could not respond fast enough to gain maximum value from their data.

To guarantee data quality, enterprises need a data architecture that quickly connects all their data from on-premises, hybrid cloud, or multi-cloud sources. The best way to do that is through modern data architectures such as data fabric and data mesh.

Data Fabric

A data fabric uses intelligent and automated systems to integrate data pipelines and cloud environments. This setup improves visibility of these diverse data resources while embedding governance and adding data accessibility, security, and privacy. A data fabric can bring together data platforms used for different lines of business, such as human resources (HR), supply chain management, and customer records into a single management platform.

A data fabric allows data scientists to build holistic views of customers, and common uses include customer profile building, fraud detection, preventive maintenance analysis, and risk model builds for return-to-work and other initiatives.

Data Mesh

A data mesh is decentralized architecture organized by separate business domains such as sales, marketing, and customer service. Each domain is managed by its business and data owners, who set policies according to their specific requirements. This approach provides each line of business with flexibility and allows them to create self-service data products for their unique needs.

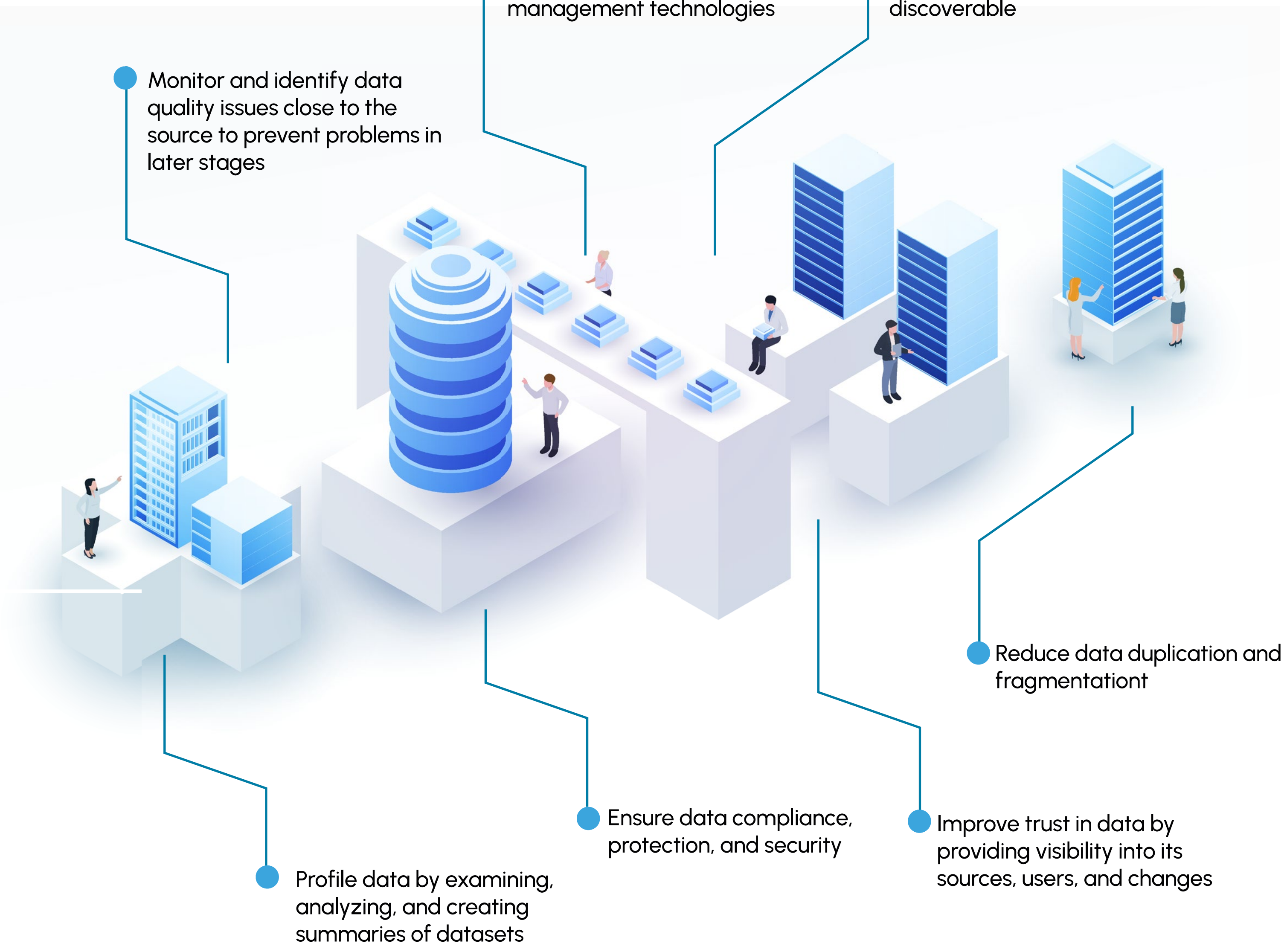
Data mesh common use cases include creating business intelligence dashboards and automated virtual assistants, gaining customer insights, and reducing data processing time for machine learning (ML) projects. Traditional storage systems such as data lakes and data warehouses can be used in a data mesh

Differences Between Data Fabric and Data Mesh

A data fabric is more technology-focused than is a data mesh, with an emphasis on using AI/ML to discover metadata patterns. Data mesh is focused more on individual team ownership and governance. However, both architectures are designed to connect new data sources and accelerate development of data pipelines across on-premises, hybrid, and multicloud environments. They both play a central role in a company's data strategy, and a data fabric's automation features can be incorporated into a mesh for some use cases.

Whether an organization uses a data fabric, data mesh, or a combination of both, the benefits and strategies for building a modern data architecture are identical.

A modern data architecture should:



- Allow efficient adoption of data platforms and new data management technologies
- Apply metadata to make data searchable and discoverable
- Monitor and identify data quality issues close to the source to prevent problems in later stages
- Reduce data duplication and fragmentatiton
- Ensure data compliance, protection, and security
- Improve trust in data by providing visibility into its sources, users, and changes
- Profile data by examining, analyzing, and creating summaries of datasets

How Data Observability and Data Catalogs Help Data Quality

Data observability tools are one way to improve data quality. Data catalogs are another. Data observability and data catalog products are complementary and can be used in combination.

A data catalog provides a detailed inventory of all data assets in an organization. It is designed to help data professionals such as data analysts, data scientists, and data stewards quickly find the right data for their analytical or business goals.

A data catalog uses metadata to create a searchable inventory of all data assets in an organization. These assets can include:

- Structured data
- Unstructured data, including documents, web pages, email, social media content, mobile data, images, audio, and video
- Reports and query results
- Data visualizations and dashboards
- ML models
- Connections between databases

Metadata is data that provides information about one of the above data types to make it easier to find and understand. A data catalog should include ways to apply tags, associations, ratings, annotations, and other information to metadata that helps users find it faster.

A complete data catalog system should also:

- Enable data discovery through search.
- Use ML to automate compliance.
- Deploy wherever data resides – on-premises or in clouds – to connect enterprise assets.
- Integrate with other data quality and governance tools, including those that manage AI models.

As data warehouses grow to thousands of documents, these processes must be automated. If not, data users will spend more time searching for data than performing analysis on the data.

Although cataloging organizes reams of data, data observability shows the live changes and impacts of an organization's daily processes. A data observability tool allows an organization to track errors before they degrade the quality of data.

Data observability tools show data engineers the health and quality of data. Observability tools monitor data behavior so that data problems can be identified and prevented before they negatively affect analytics. Observability includes anomaly detection, root cause analysis, and workflow automation to find and fix data problems.

To perform these capabilities, data observability tools must monitor data at rest and in motion. Data at rest is in a fixed location, and monitoring tells data engineers whether the dataset arrived on time, if it is being updated frequently enough, and if the expected volume of data is included. Monitoring data in motion covers data while it moves through pipelines. It shows if pipeline performance affects data quality, the conditions for success, and what operations impact the dataset before it reaches its destination. Monitoring data at rest and data in motion is connected and should be handled by the same tool.

Customer Challenges: Steps to Ensuring Reliable Data

Ensuring you have data you can trust is a multi-step process. Some organizations are already well down the path, while others are still in the planning phase.

Phase 1

Characterize Your Data Use Cases

Divide data use cases into analytic data used for business in a business intelligence (BI) dashboard, operational data that may be used in real time (streaming, for example), or customer-facing data. Perform a data quality assessment of each type of data. At this stage, determine the changes you will need to make and settle on an optimal data architecture for your business.

Phase 2

Build Internal Support

Show risks and consequences of areas where you have bad data and make a case for adding a data quality product or service. Make sure to inform all stakeholders of the state of their key data. Show them how you can identify business use cases for their data to make better decisions and reduce costs where applicable.

Phase 3

Classify Your Data

Suppose you have 50 data tables in a data warehouse. Some of these tables are more critical than others. Your monitoring tool should be able to treat these tables differently depending on how critical they are, and its alerting features must reflect the differences.

Proper classification makes it easier for data engineers to define and build pipelines for specific data types. When you classify data, you can build pipelines around use cases and give data the correct levels of latency, security, storage, monitoring, etc.

Phase 4

Implement Data Quality Policies

Effective data quality management requires an organization to set and enforce data quality policies that outline standards, roles, responsibilities, and processes of data management. These should be set to guarantee an organization will remain in compliance with data regulations as well as improve customer relationships and reduce costs. Data policy guidelines should cover how information is collected, stored, processed, and shared.

Data observability is necessary to enforce data quality policies by monitoring changes in data quality, availability, and flow. Effective data observability must include these features as part of its end-to-end data operations workflow:

- Monitoring dashboard that provides an operational view of a pipeline or system, including comparisons over time
- Alerting for expected events and anomalies
- Event tracking
- Automated issue detection
- Event logging
- Service level agreement (SLA) tracking to measure data quality and metadata against pre-defined standards.

Phase 5

Create a Roadmap

Develop a plan for going forward that includes implementing the data architecture and proper governance. This step includes creating key performance indicators (KPIs) to measure success and the processes that will be used to implement data observability after the architecture is in place.

Recommendations and Solutions to Consider

An effective data quality platform should include a comprehensive set of features, such as data profiling, cleansing, monitoring, and validation. It should also be easy to use, intuitive, and scalable as your company and its data grows. You also want a platform that can integrate with other tools and processes in your company.

IBM provides tools that combine data quality and governance with data observability. The two main offerings are IBM Knowledge Catalog and IBM Databand.

IBM KC is a cloud-based enterprise metadata repository that helps customers access, curate, categorize, and share data, knowledge assets, and their relationships across an organization. It is available as an IBM Cloud-hosted software as a service (SaaS) solution and as part of IBM Cloud Pak for Data. In Cloud Pak for Data, KC is part of an integrated modular set of software components for data analysis, organization, and management. Cloud Pak for Data is available as an on-premises software build on the Red Hat OpenShift container platform or as a managed version on IBM Cloud.

KC use cases focus on improving data quality, remediating issues, and managing workflows. KC SLAs are related to data, setting specific SLA targets for areas such as data freshness, data completeness, and data validity.

KC helps business users measure and improve data quality scores at various phases: raw data (i.e., Kafka streams), standardized and cleansed data, gold production data, and master data records. It computes data quality scores based on quality dimensions for each individual column in the data asset, and then calculates a combined quality score for the entire data asset. If the KPIs used to measure data in a column fall below expectations, KC can be used to remedy the data quality issues.

IBM's Databand observability platform is a central place for defining and receiving alerts of data incidents. It enables rapid detection and alerting of incidents, debugging, and lifecycle management. Databand continuously monitors data in transit and at rest, captures error logs and root causes from metadata, and alerts stakeholders of problems. Databand's debugging shows specific errors within specific tasks for faster resolution of issues.

A key use case for Databand is to ensure data flows in a data platform perform reliably so that the data can be trusted. Databand can detect anything suspicious or abnormal about the data and send alerts to the data engineering team. Databand can tell users, for example, if a data build tool (dbt) job is performing as expected, or if anomalies in a data profile for data at rest indicates an dbt job will malfunction upstream.

The SLAs for Databand are related to the data platform, measuring uptime based on data performance and reliability targets. Databand SLAs help increase uptime of the data platform by monitoring reliability of dataflows.

The major distinctions between KC and Databand are as follows.

IBM KC Augmented Data Quality

OVERALL GOAL	PRIMARY USERS	KEY USE CASES	CAPABILITIES	DATA QUALITY GOALS
Activate business-ready data for AI and analytics from high-quality data.	Data stewards, data engineers, data analysts, data scientists	Allow businesses to measure and improve data scores at phases such as raw data, standardized and cleansed data, gold production data, and master data records; show if the data quality score in the silver zone of Databricks lakehouse is acceptable before it is autoloaded to the gold zone for consumption; display the quality score trend on a Kafka stream for a specific time period.	Augmented data quality; ability to measure and improve data quality scores	Completeness, accuracy, validity, freshness, schema, lineage when used with Databand

Databand Data Observability

OVERALL GOAL	PRIMARY USERS	KEY USE CASES	CAPABILITIES	DATA QUALITY GOALS
Detect data incidents earlier and resolve them faster before they cause quality issues.	Data, platform and analytics engineering teams	Prevent engineering operations from processing large volumes of data with high frequency from rolling back corrupted data; report if dbt jobs are performing as expected without anomalies	Incident management, SLA alerting, data pipeline monitoring, and impact analysis	Completeness, accuracy, validity, freshness, schema, lineage

Conclusion

Data quality, observability, and reliability tools should always deliver trusted data because applications such as AI are only as good as the data they use. You want solutions that automate management of data lifecycles with governance, security, and lineage for self-service data consumption.

Together, IBM KC and IBM Databand:

- Automate management of data lifecycles with governance, security, and lineage for self-service data consumption
- Provide readily consumable and properly governed data to your teams anytime and anywhere
- Deliver reliable data by detecting data incidents early and resolving them quickly with continuous data observability
- Drive faster and scalable insights by delivering a single and comprehensive view of an entity's data across an enterprise

Although they are separate products, KC and Databand can be used in combination to solve problems for data engineers and business users. A customer using KC might want to proactively monitor data flows with Databand, or a Databand customer might expand beyond monitoring and add a complete data catalog with KC as the next step in improving data quality.

Important Information About this Report

CONTRIBUTORS

Steven Dickens
Vice President and Practice Lead | The Futurum Group

Dave Raffo
Senior Analyst | The Futurum Group

PUBLISHER

Daniel Newman
CEO | The Futurum Group

INQUIRIES

Contact us if you would like to discuss this report and The Futurum Group will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "The Futurum Group." Non-press and non-analysts must receive prior written permission by The Futurum Group for any citations.

LICENSING

This document, including any supporting materials, is owned by The Futurum Group. This publication may not be reproduced, distributed, or shared in any form without the prior written permission of The Futurum Group.

DISCLOSURES

The Futurum Group provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

ABOUT IBM

IBM is a leading provider of global hybrid cloud and AI, and consulting expertise. We help clients in more than 175 countries capitalize on insights from their data, streamline business processes, reduce costs and gain the competitive edge in their industries Learn more here.

ABOUT THE FUTURUM GROUP

The Futurum Group is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

CONTACT INFORMATION

The Futurum Group LLC I futurumgroup.com I (833) 722-5337 I