MCKNIGHT
CONSULTING GROUP

The Data Observability Advantage

Unlocking the Secrets to Reliable, High-Quality Big Data

Presented by: William McKnight

President, McKnight Consulting Group

3 x Inc 5000

in /in/wmcknight

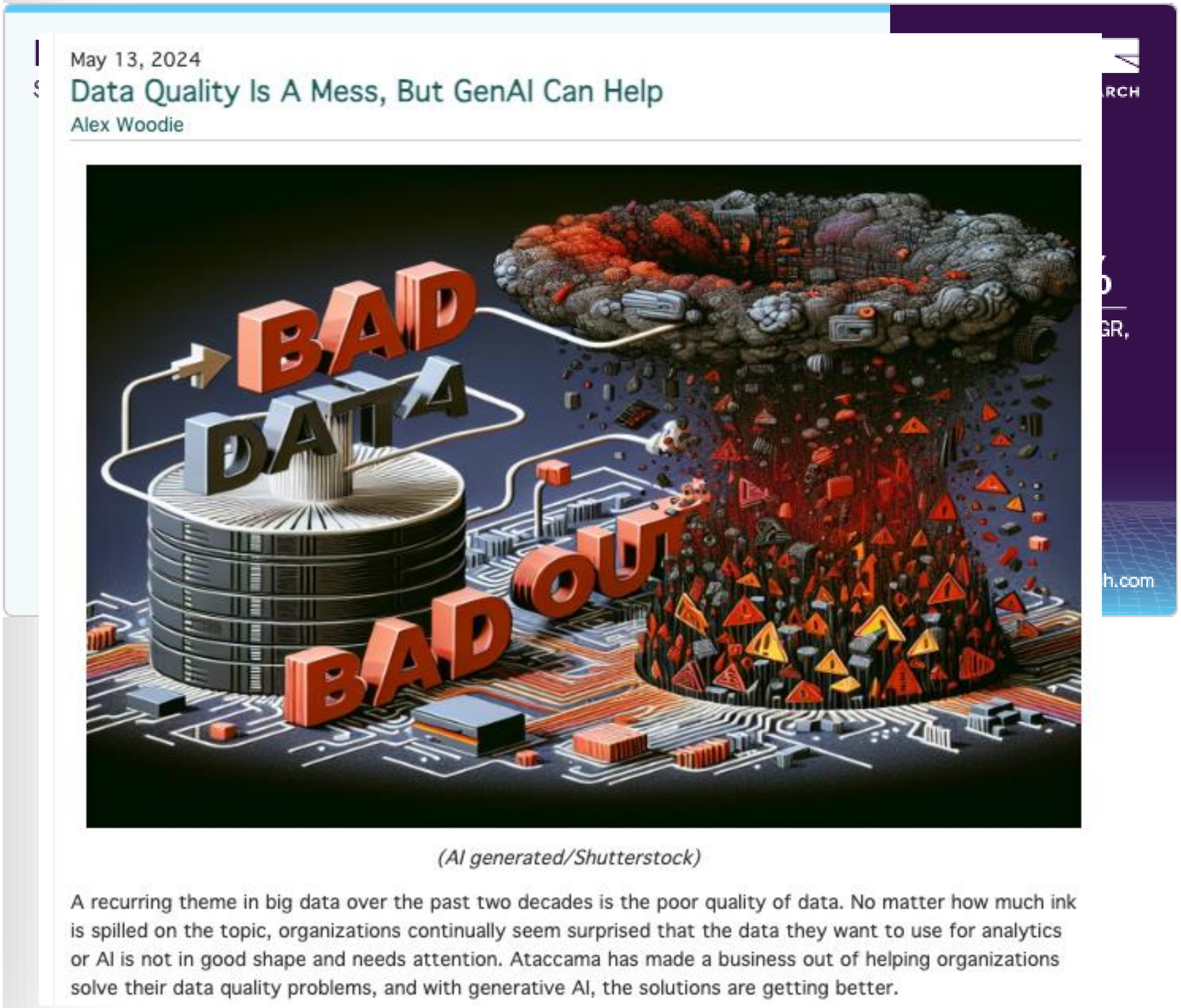
www.mcknightcg.com
(214) 514-1444

TOP VOICE
★ 2024 ★
thinkers 360
OVERALL

I'VE BEEN FEATURED
IN THE 2024
dataIQ™ 100 USA
THE MOST INFLUENTIAL PEOPLE IN DATA

Enterprise Data is Still a Mess

- The proliferation of data sources
- The complexity of data formats
- The lack of data governance
- The push into AI



May 13, 2024

Data Quality Is A Mess, But GenAI Can Help

Alex Woodie

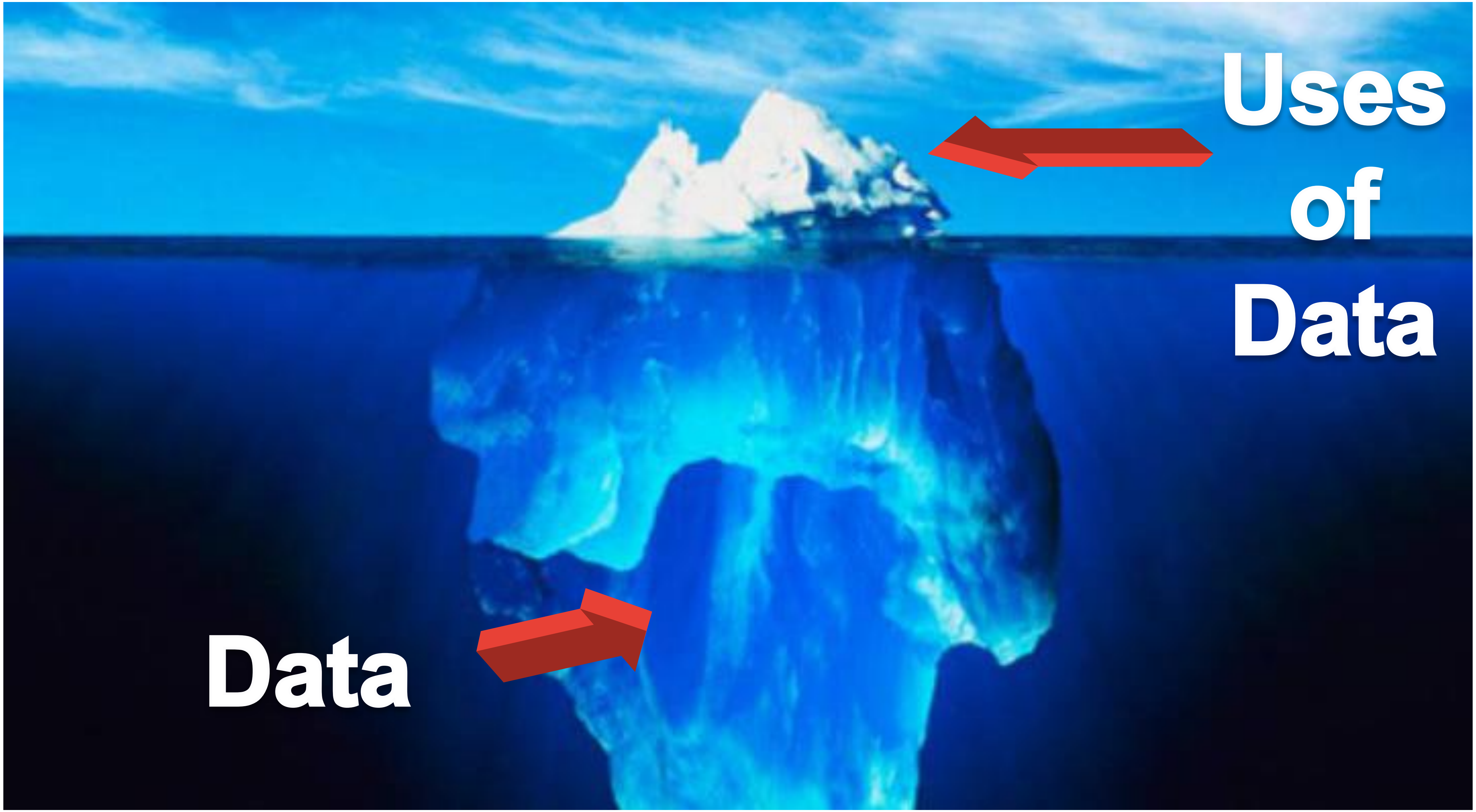
(AI generated/Shutterstock)

A recurring theme in big data over the past two decades is the poor quality of data. No matter how much ink is spilled on the topic, organizations continually seem surprised that the data they want to use for analytics or AI is not in good shape and needs attention. Ataccama has made a business out of helping organizations solve their data quality problems, and with generative AI, the solutions are getting better.

We Divide the Data Observability Vendors Into Two Camps

- Group 1 are platforms that specialize in data and pipeline health observability (henceforth ‘data observability’) and are **the subject of this presentation**. This group deals with data quality, consistency, freshness, distribution, volume, schema, lineage, and sometimes spend.

- Group 2 are platforms that primarily specialize in infrastructure and data traffic MELT (metrics, events, logs, and traces) observability. Due to the different focus, these will not be covered alongside the data and pipeline health observability vendors in this presentation.


Uses
of
Data
Data

Data Quality is Essential to Business Success

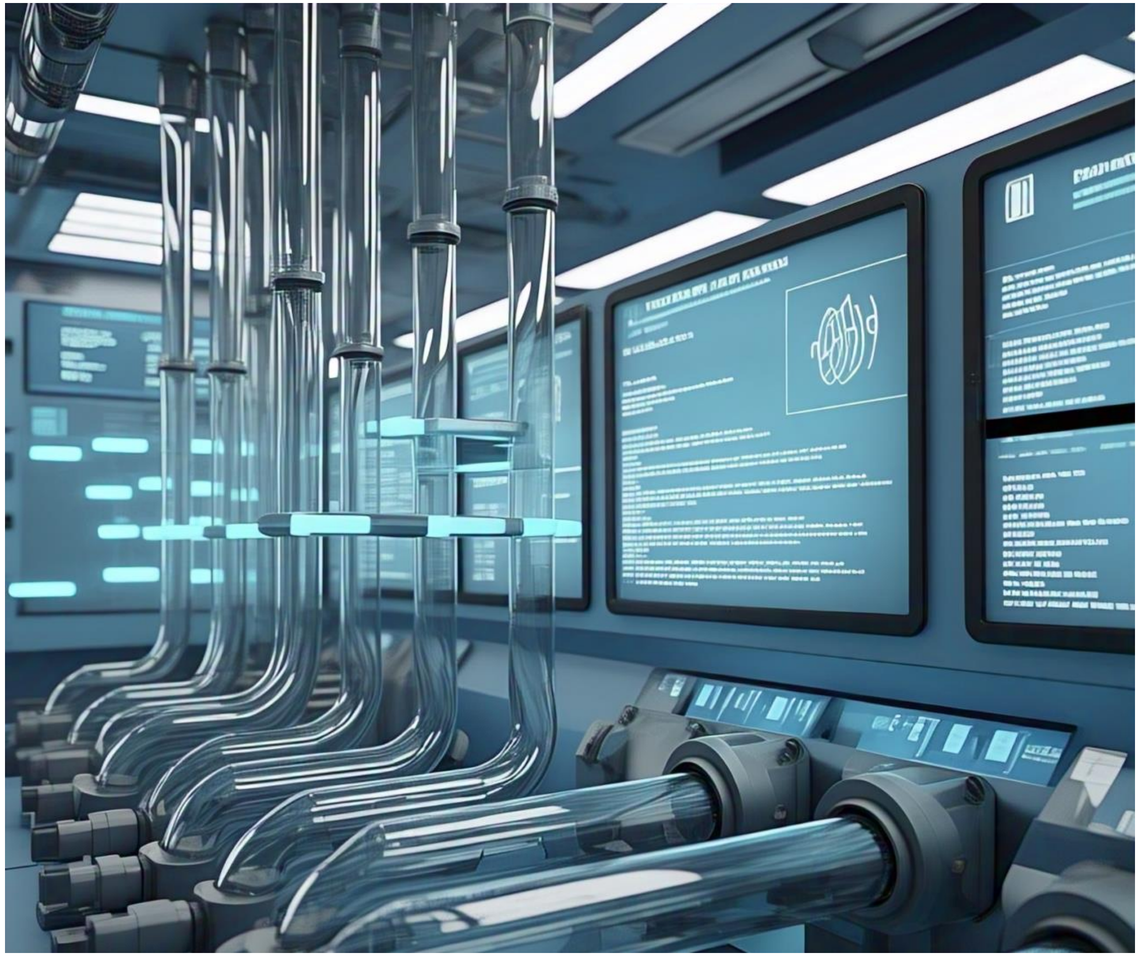
- "Correct" data is a widespread need
- Yet, data quality lacks consistent definition
  - » You can't improve what you can't measure
  - » Tangible benefits accrue from improved efficacy of the applications using the data

Most People Don’t Care About It Until it’s a Problem

- Usually not considered critical path
- You must be an advocate
- Cite improved chance of success
- Slippery slope

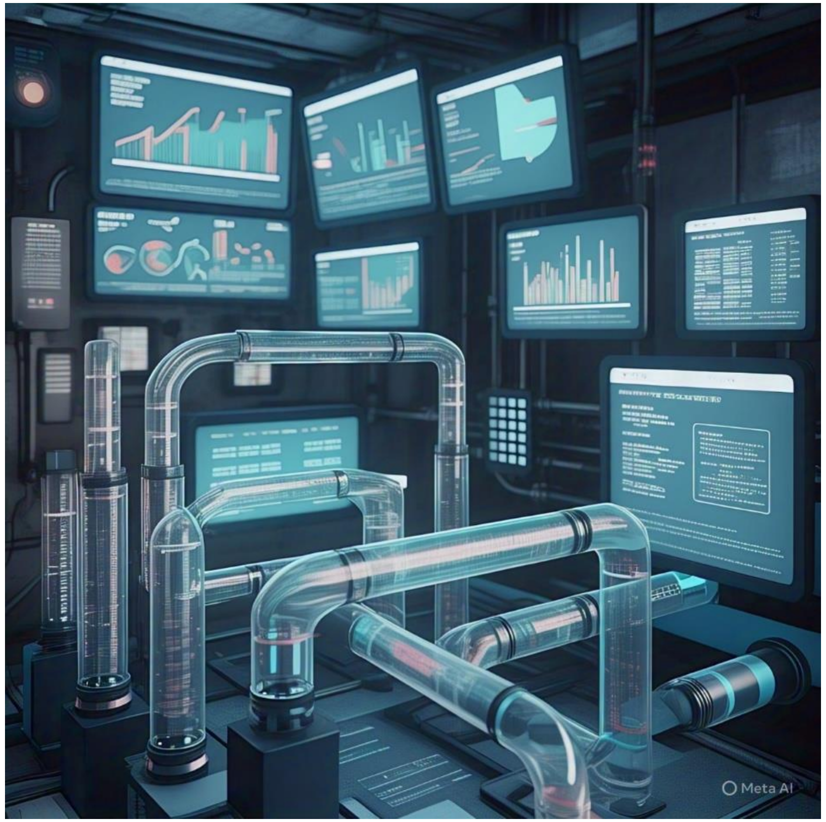
Investments in Data Quality

- Investments Yield “Cleaner” Data
- Business objectives cannot be met without quality data in support
  - Data Quality Returns are in the improved efficacy of projects targeting business objectives
- Data Quality should be an integral part of most projects

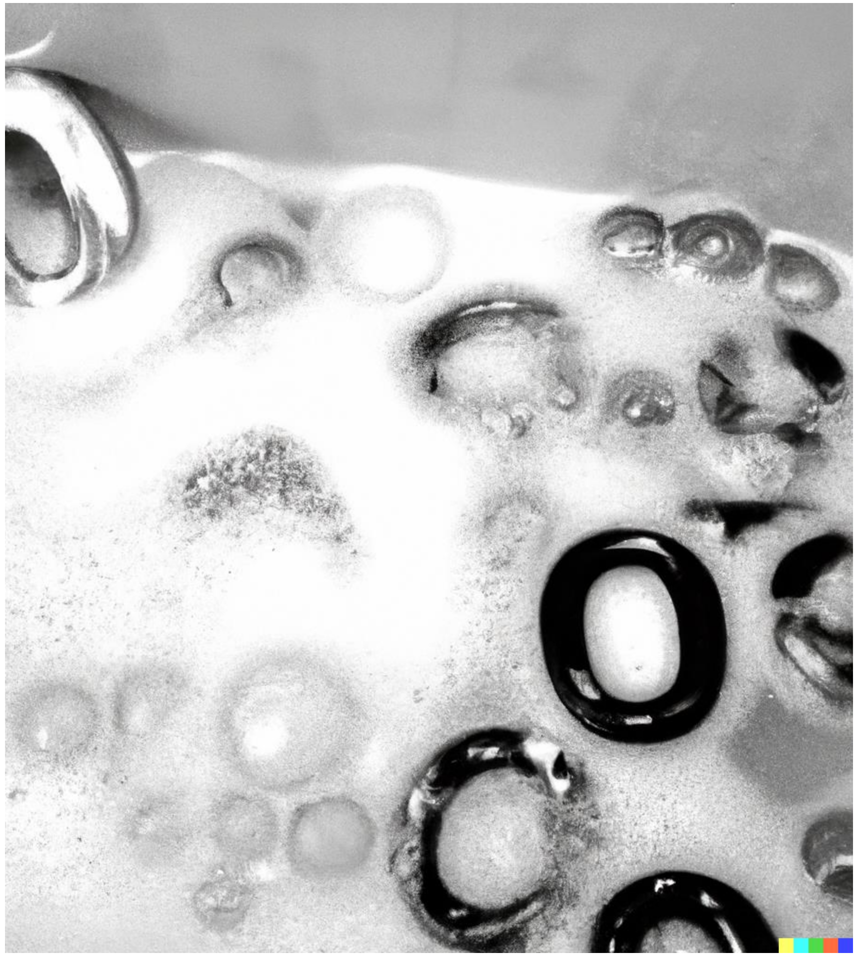


Cost to the Enterprise of Poor Data Quality

- One-off DQ repeated remediations
- Poor/Failed Enterprise Initiatives
- Misguided roadmaps
- Compliance cost
- $x per data record attributed to:
  - Failed outreach
  - Losing customers
  - Storage space and effort with duplicate records
  - Incorrect marketing segmentation and personalization
- Cost expansion
  - On average corporate data grows at 40% per year



The Benefit Of Clean Data Is Not Enough



- ROI
- Strategic Benefit
- Lower TCO

Data Quality Should Have a Value Proposition To Project(s)

- Improved Decision Making
- Increased Efficiency
- Reduced Risk
- Improved Customer Satisfaction
- Increased Profitability
- Enhanced Security
- Improved Compliance

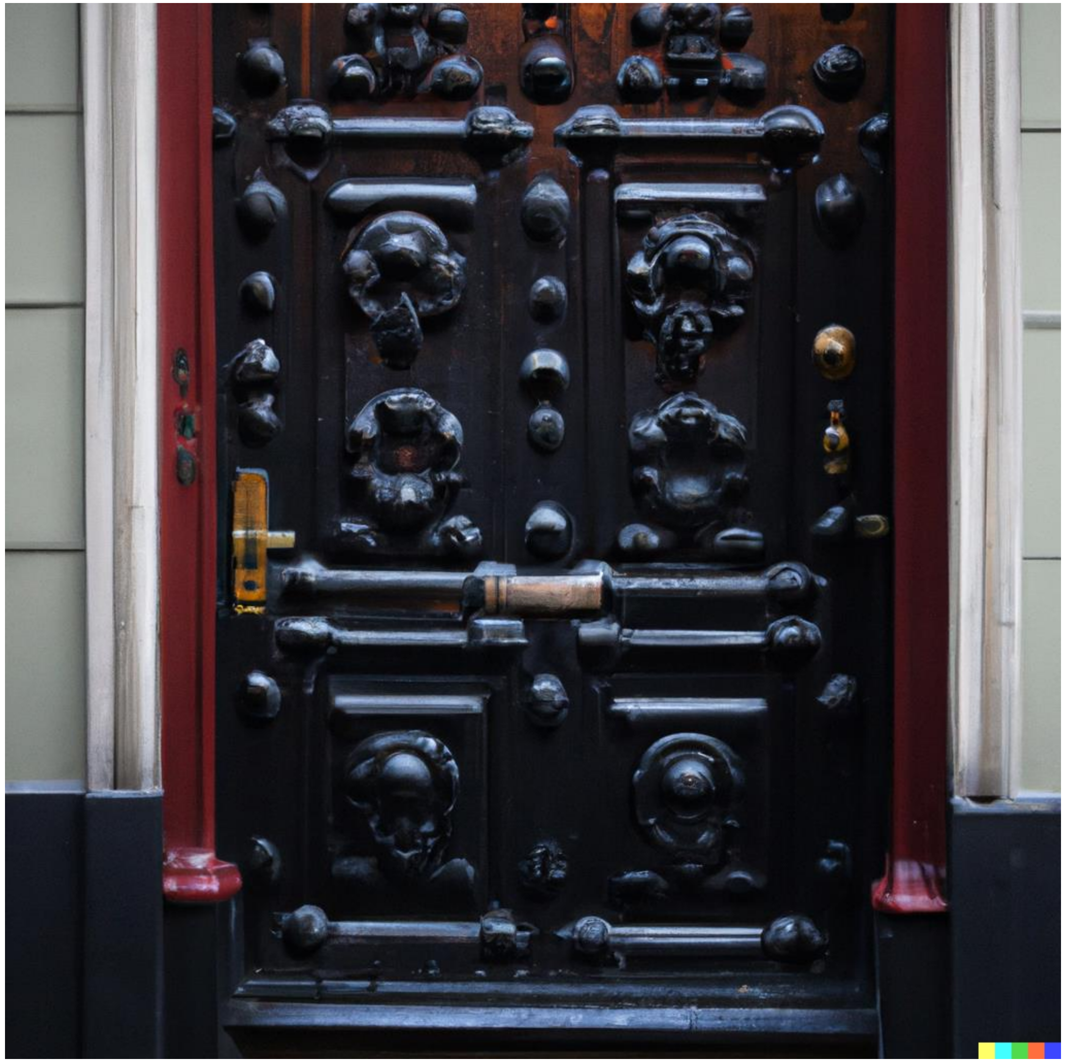
A Good Data Governance Program Keeps Business Interest In Data Quality

- Data Governance helps organizations understand what data they have, where it's stored, and how it's used.
- It ensures data privacy, security, and compliance with regulations.
- It tracks data access and usage, ensuring data is used for intended purposes.
- It provides insights into data meaning, lineage, and impact.

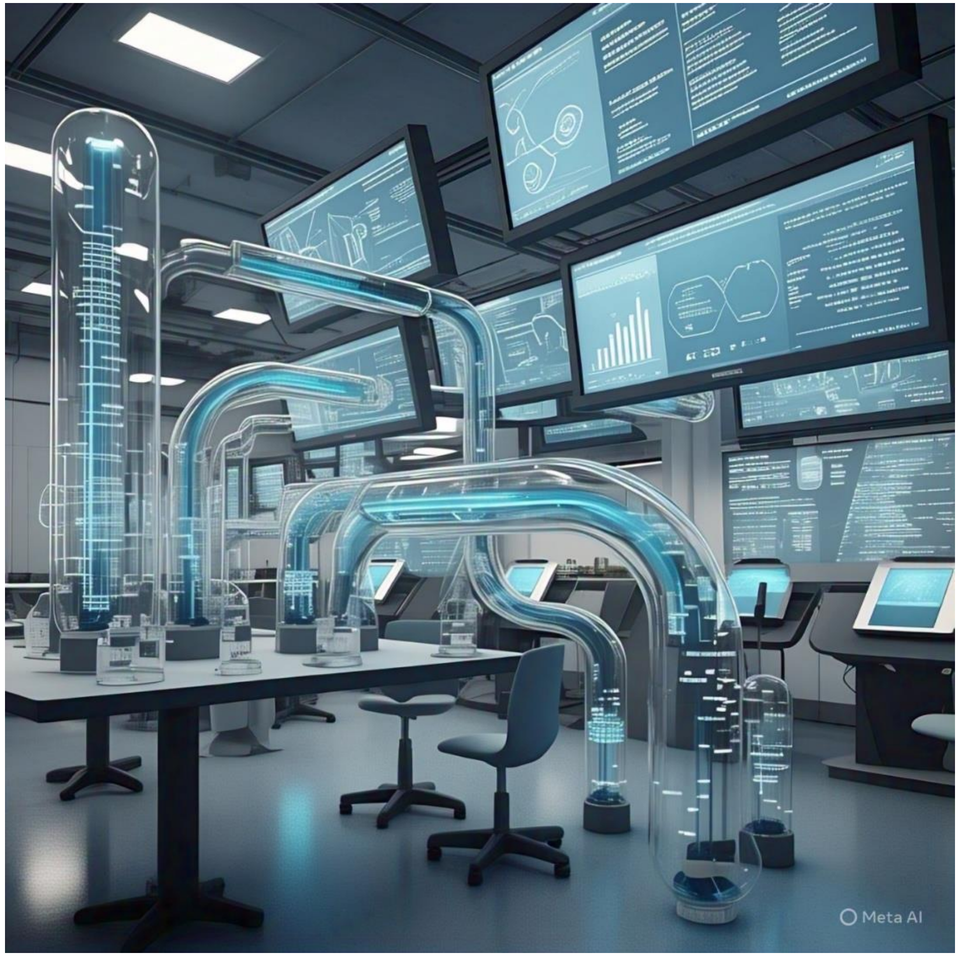
Only a Methodological Approach Will Work

- Repeatable Process
- Progressive Improvement
- New Data
- Requirements Change

The Causes Of Poor Data Quality Keep Coming In The Front Door



Data Quality Improvement Program



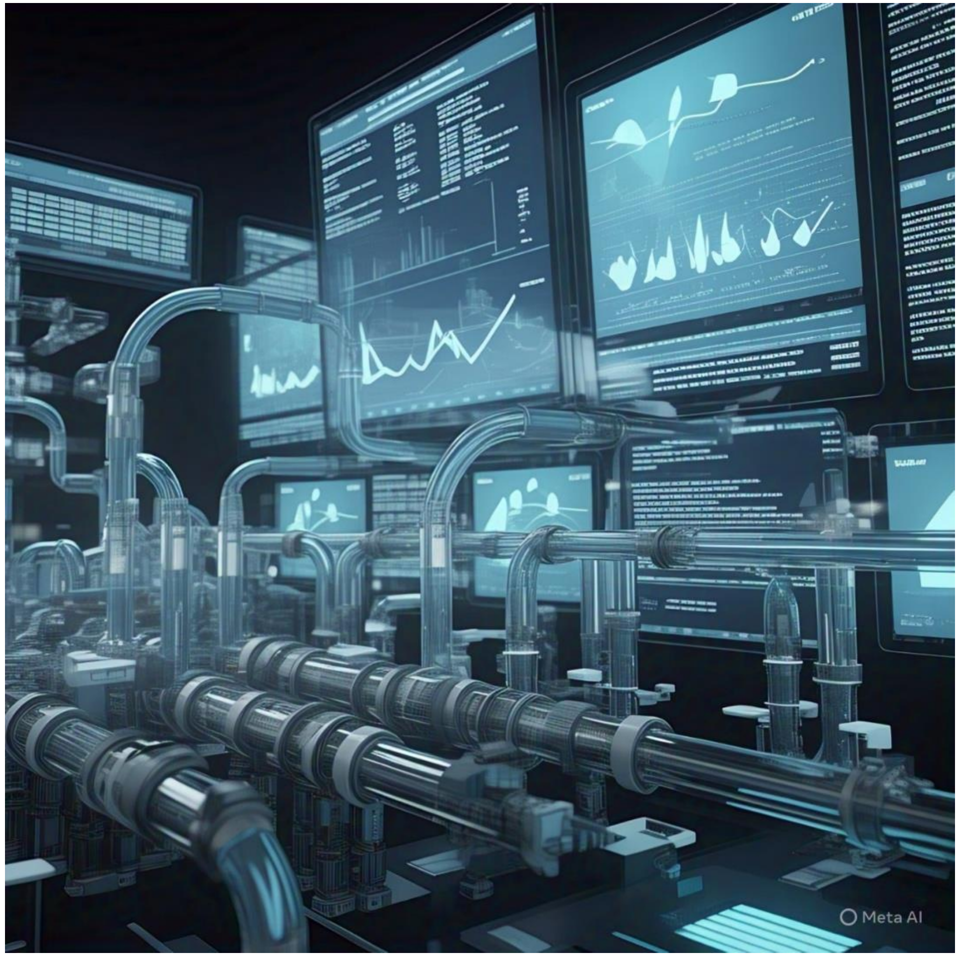
- Define the quality expectations
- Profile data
- Measure data quality improvement options
- Select the best option
- Improve the quality of data and improve the business

Data Quality Rule Categories

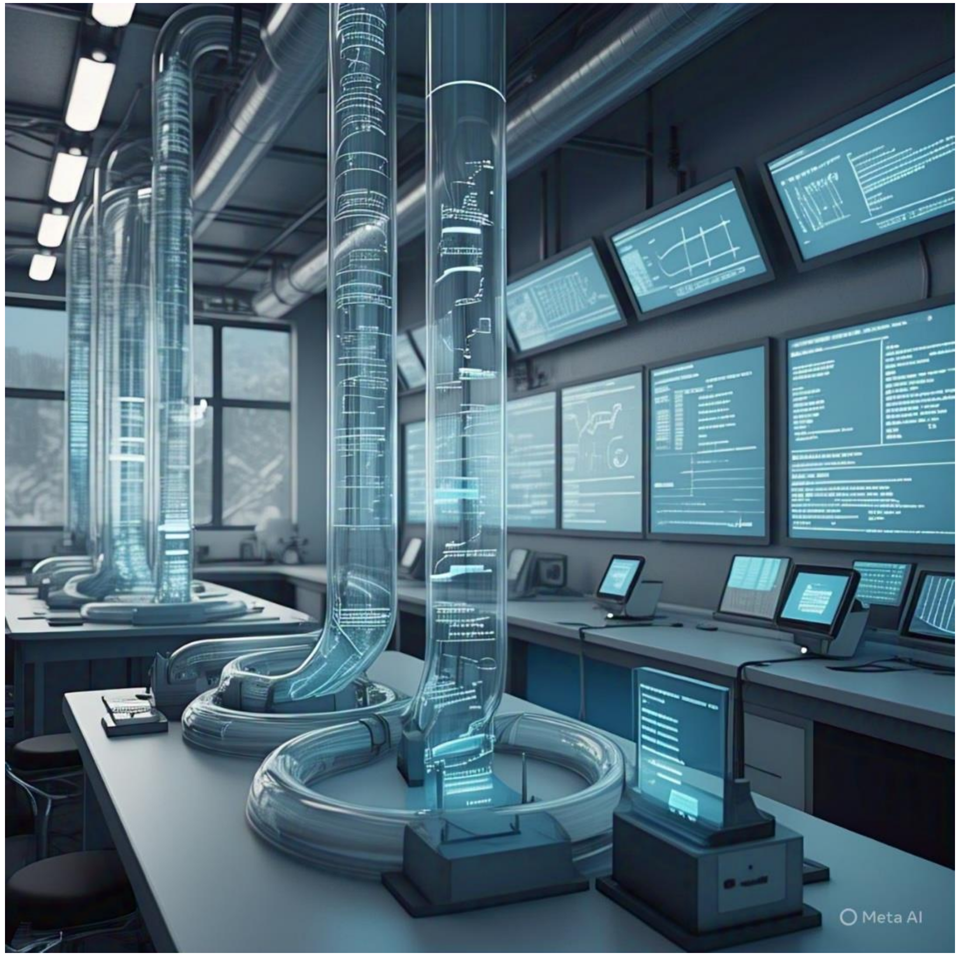
- Align business processes with data-driven insights
- Data-Driven Decision Making
- Referential Integrity
- Uniqueness
- Cardinality
- Subtype/Supertype constructs
- Value reasonableness
- Consistency
- Formatting
- Data derivation
- Completeness
- Correctness

Five Actions to Perform for Data Quality

- Screen Data Entry
- Add Cross-Checking
- Quarantine Data (for decisions)
- Report on Quality Violations (still may require additional action to fix DQ)
- Change or Repair Incorrect Data to Conform to DQ



Put Quality Data in a Leveragable Platform



- Data Warehouse
- Data Lake
- Master Data Management
- Data Hub

Every Project Needs a Focus On Data Quality

- Clean data is the key to unlocking the power of many business processes, including:
  - Information-based in-store and contact center cross- and up-selling (NEEDS clean customer and product data)
  - Credit card fraud detection (NEEDS clean customer and transaction data)
  - Supply chain efficiencies and just-in-time production capabilities (NEEDS clean product and location data)
  - Predictive churn management (NEEDS clean customer and transaction data)
- Having clean customer, product, transaction, and location data is essential for these projects to be successful.

Streaming Data Data Quality

- Continuously monitor streaming data to identify anomalies, outliers, and deviations from expected patterns, indicating potential data quality issues.
- Implement real-time data validation rules and cleansing techniques to detect and correct errors, inconsistencies, and missing values on the fly.
- Supplement streaming data with additional information from external sources to enhance its completeness and context, enabling more accurate analysis and decision-making.
- Track the origin, transformations, and usage of streaming data to ensure transparency, traceability, and auditability of data quality processes.

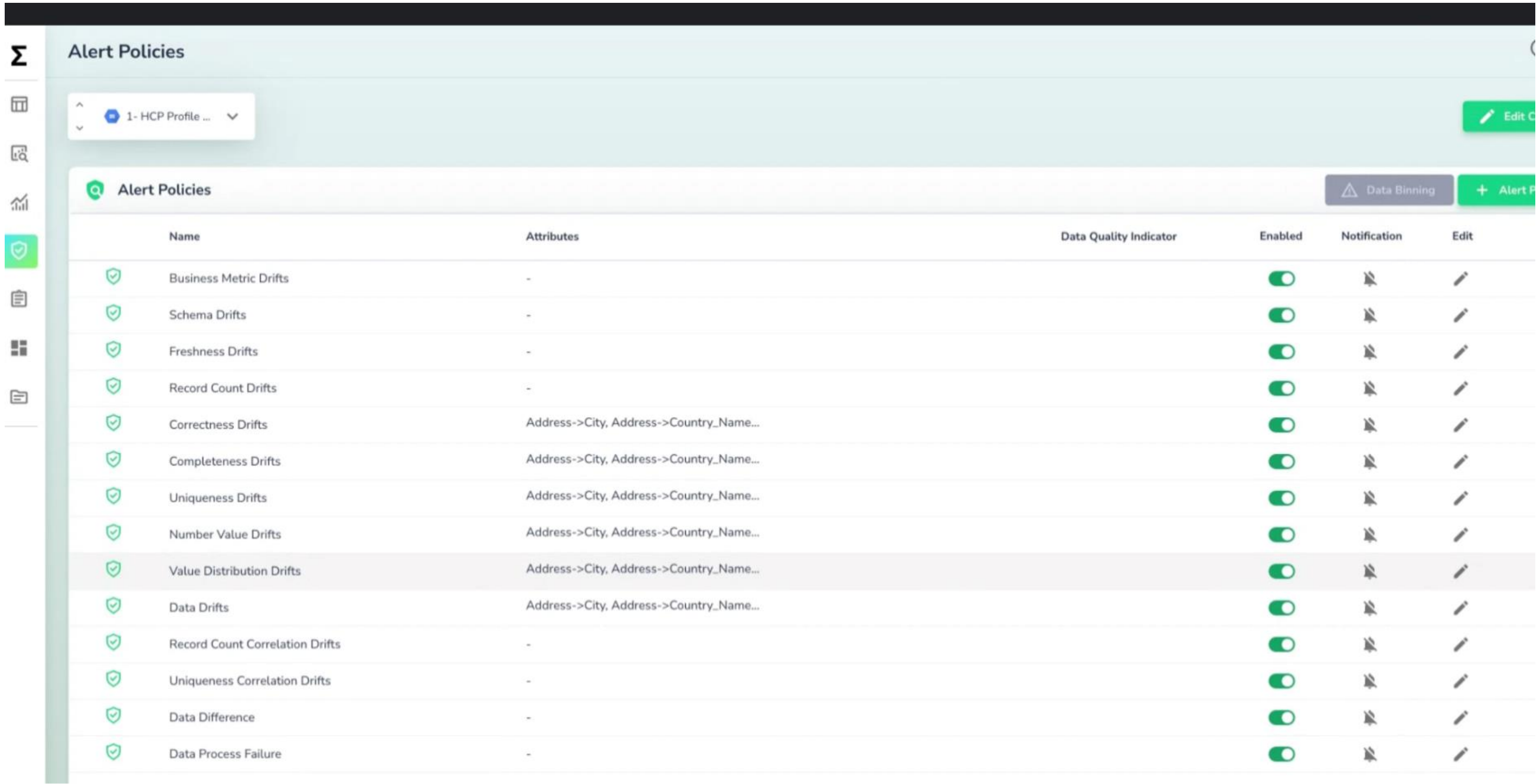
Data Observability

Data Observability tools differ from legacy Data Quality tools

- **Proactive vs. Reactive**: Data observability tools are proactive, detecting issues in real-time, whereas legacy tools often rely on reactive approaches, identifying issues after they've occurred.
- **Automated Monitoring**: Data observability tools continuously monitor data pipelines, automatically detecting anomalies and changes, whereas legacy tools may require manual checks.
- **Advanced Analytics**: Data observability tools often leverage machine learning and statistical analysis to identify complex patterns and anomalies, whereas legacy tools may rely on simpler rules-based approaches.
- **Real-time Alerts**: Data observability tools provide real-time alerts and notifications, enabling swift action to prevent data issues from impacting AI models or business decisions.
- **Holistic Visibility**: Data observability tools provide end-to-end visibility into data pipelines, enabling a more comprehensive understanding of data quality and integrity.

Increasing complexity, reliance on data-driven decision-making, and the need for real-time monitoring lead us to data observability

- Data Observability is a cloud-native way to look at data quality.



Alert Policies

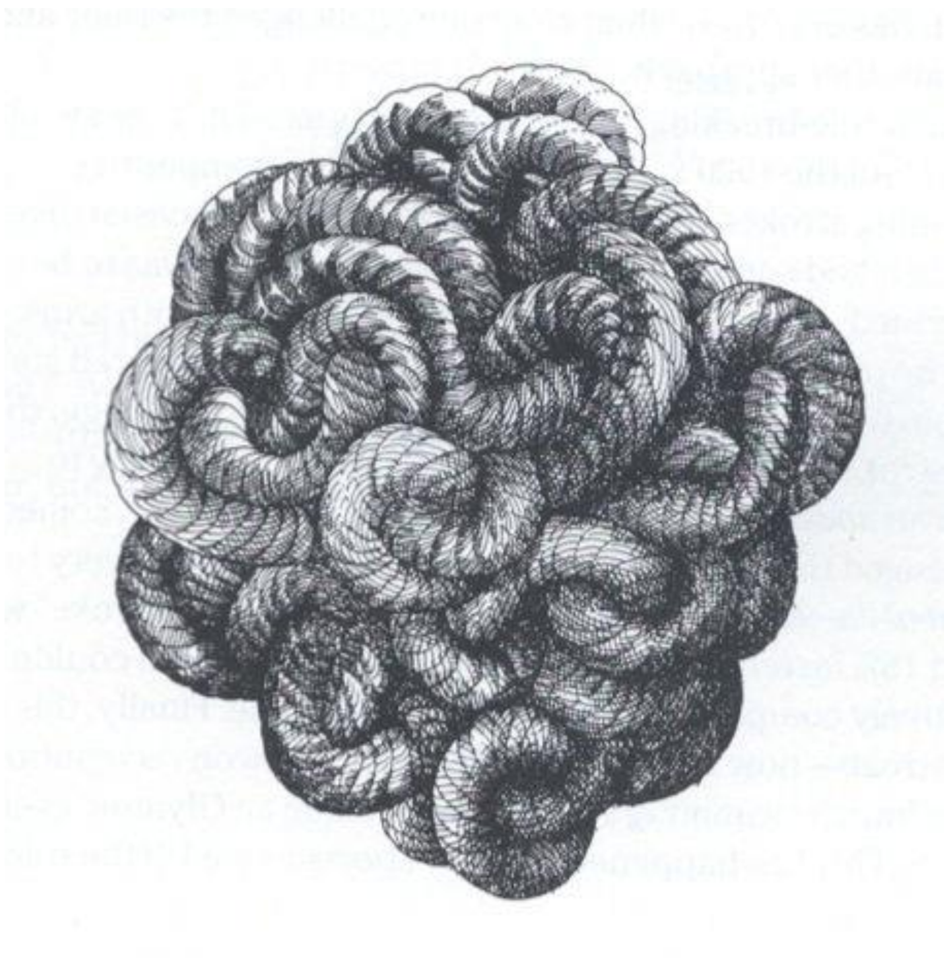
Name	Attributes	Data Quality Indicator	Enabled	Notification	Edit
Business Metric Drifts	-		●	🔕	✎
Schema Drifts	-		●	🔕	✎
Freshness Drifts	-		●	🔕	✎
Record Count Drifts	-		●	🔕	✎
Correctness Drifts	Address->City, Address->Country_Name...		●	🔕	✎
Completeness Drifts	Address->City, Address->Country_Name...		●	🔕	✎
Uniqueness Drifts	Address->City, Address->Country_Name...		●	🔕	✎
Number Value Drifts	Address->City, Address->Country_Name...		●	🔕	✎
Value Distribution Drifts	Address->City, Address->Country_Name...		●	🔕	✎
Data Drifts	Address->City, Address->Country_Name...		●	🔕	✎
Record Count Correlation Drifts	-		●	🔕	✎
Uniqueness Correlation Drifts	-		●	🔕	✎
Data Difference	-		●	🔕	✎
Data Process Failure	-		●	🔕	✎

Data Observability

- Data observability is a rapidly **growing** field that provides a **comprehensive** view of an organization's data's well-being, both during its movement and storage.
- It is essential for building a data ecosystem, **identifying and fixing issues before** they become problems for applications, analytics, and user experience.
- Data observability is suited for today's complex distributed data landscapes, including **edge, on-premises, hybrid, and multi-cloud environments.**
- **Automation and orchestration** are key pillars of data observability, as manual approaches to data quality have proven ineffective.
- **AI**, AIOps, and predictive analytics play a role in data quality and observability, with AI being particularly valuable in pinpointing and suggesting solutions for data health issues.

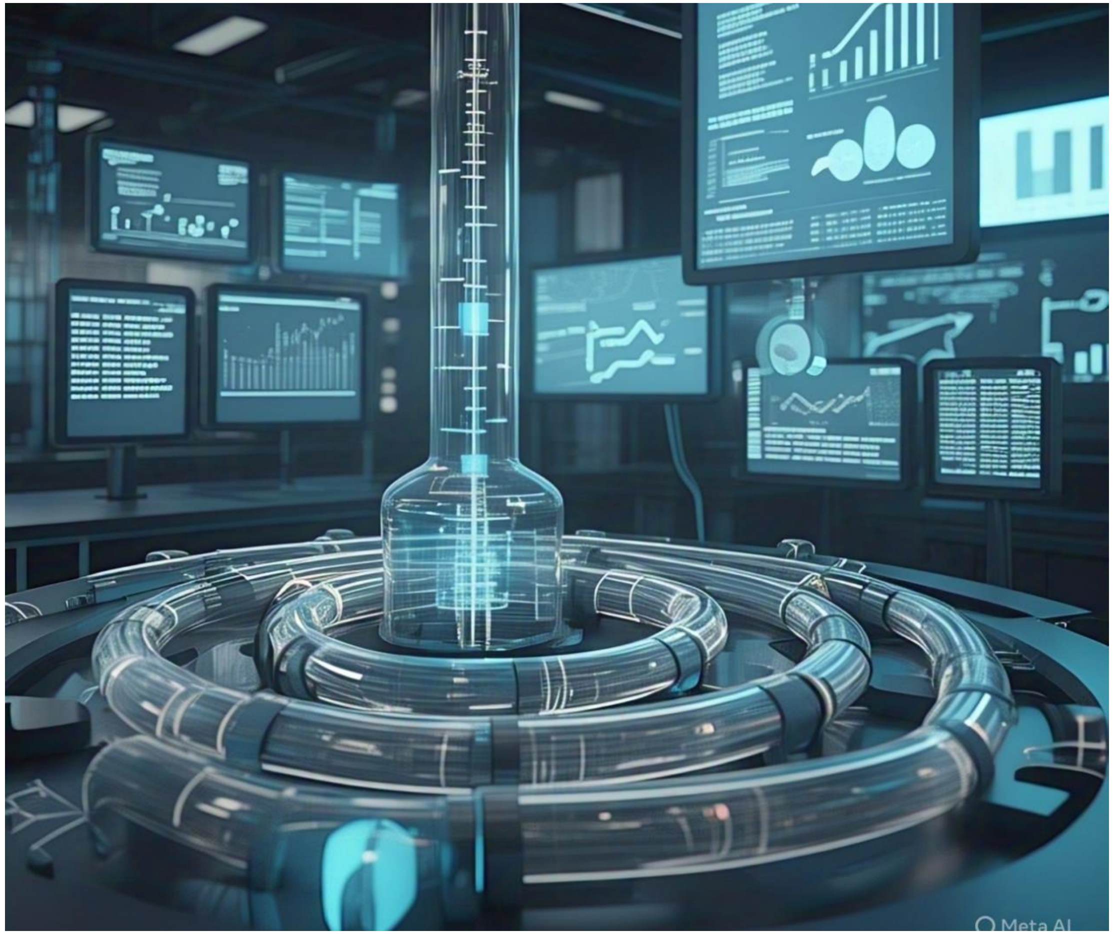
The Gordian Knot of Data Observability

- Data observability tools generate a vast amount of information about data health.
- However, sifting through this data and extracting actionable insights can be overwhelming.
- The "gordian knot" for data and pipeline health observability is the difficulty of filtering out **irrelevant** data noise and identifying the critical signals that require attention.



Data Observability for AI

- **Data observability ensures AI reliability**: By using high-quality data, AI systems can reach accurate conclusions.
- **Proactive data management**: Data observability replaces manual checks and automatically identifies data anomalies and outliers.
- Data observability tools can detect:
  - **Data irregularities**: Sudden spikes or drops in data volume or values.
  - **Data shifts**: Gradual changes in data patterns or distributions over time.
  - **Unusual data points**: Outliers, unexpected changes, or data that deviates from the norm.



Data Observability Benefits for AI

Data Validation: Automated checks ensure data accuracy and consistency.

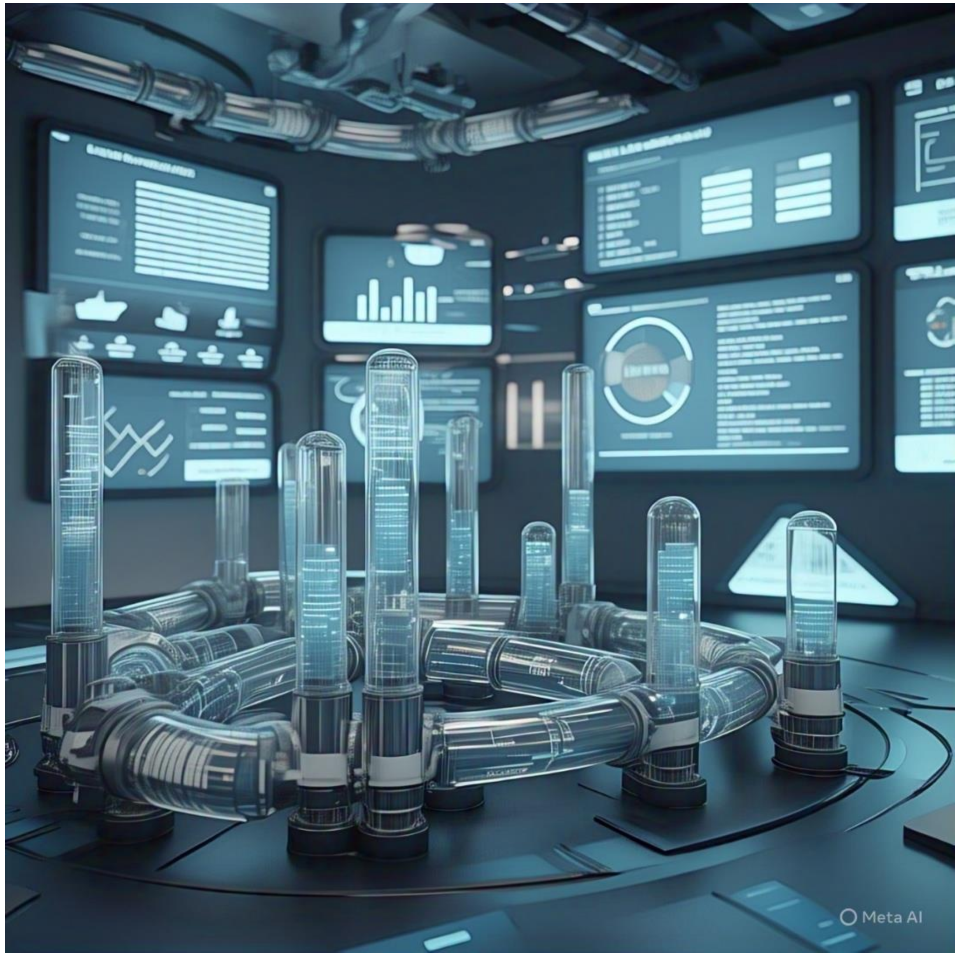
Data Integrity Monitoring: Detection of corruption or loss during data transformations.

Alert System: Notifications for unexpected changes or issues that may impact data quality.

Rule-Based Validation: Verification that business rules are correctly applied to data processing.

Data Observability Use Cases

- Recommendation Systems
- AI-Powered Workflows
- Machine Learning Applications
- Foundational Model Training
- Customer Interactions with Chatbots

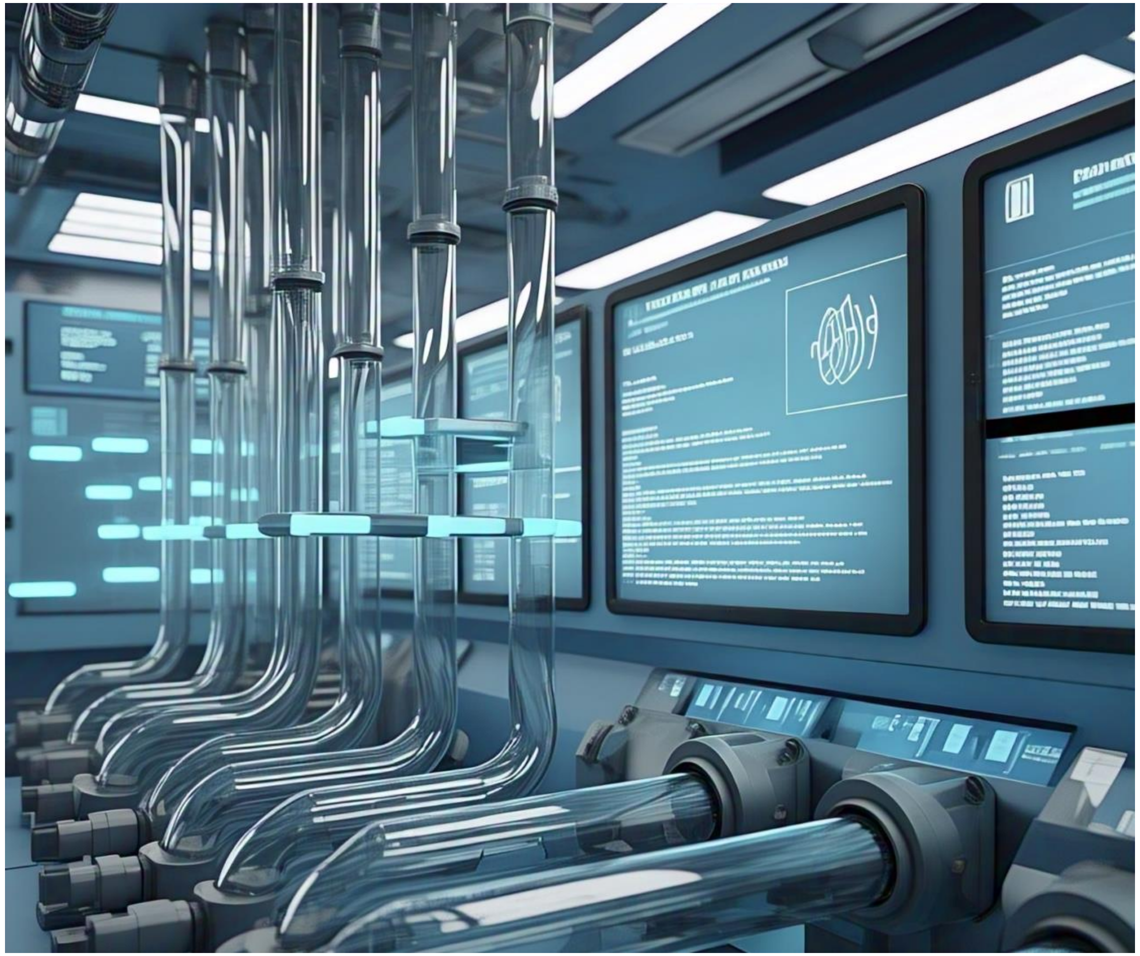


The (Near) Future of Data Observability

- More Multidimensional Anomaly Detection
- Adaptive and Intelligent to Evolving Activity
- Enhanced Efficiency in Identifying Issues
- Correlation
- More Design with Human-in-the-Loop

Multiple Dimension Anomaly Detection

- Moving Beyond One-Dimensional Checks
- Identifying Hidden Anomalies
- Simultaneous Evaluation of Attributes
- Detecting Interdependencies
- Understanding Context
- Spotting Complex Patterns



Critical Capabilities for Data Observability: Data Lineage and Pipelines

- Lineage Visualization
- Impact Analysis
- Pipeline Monitoring

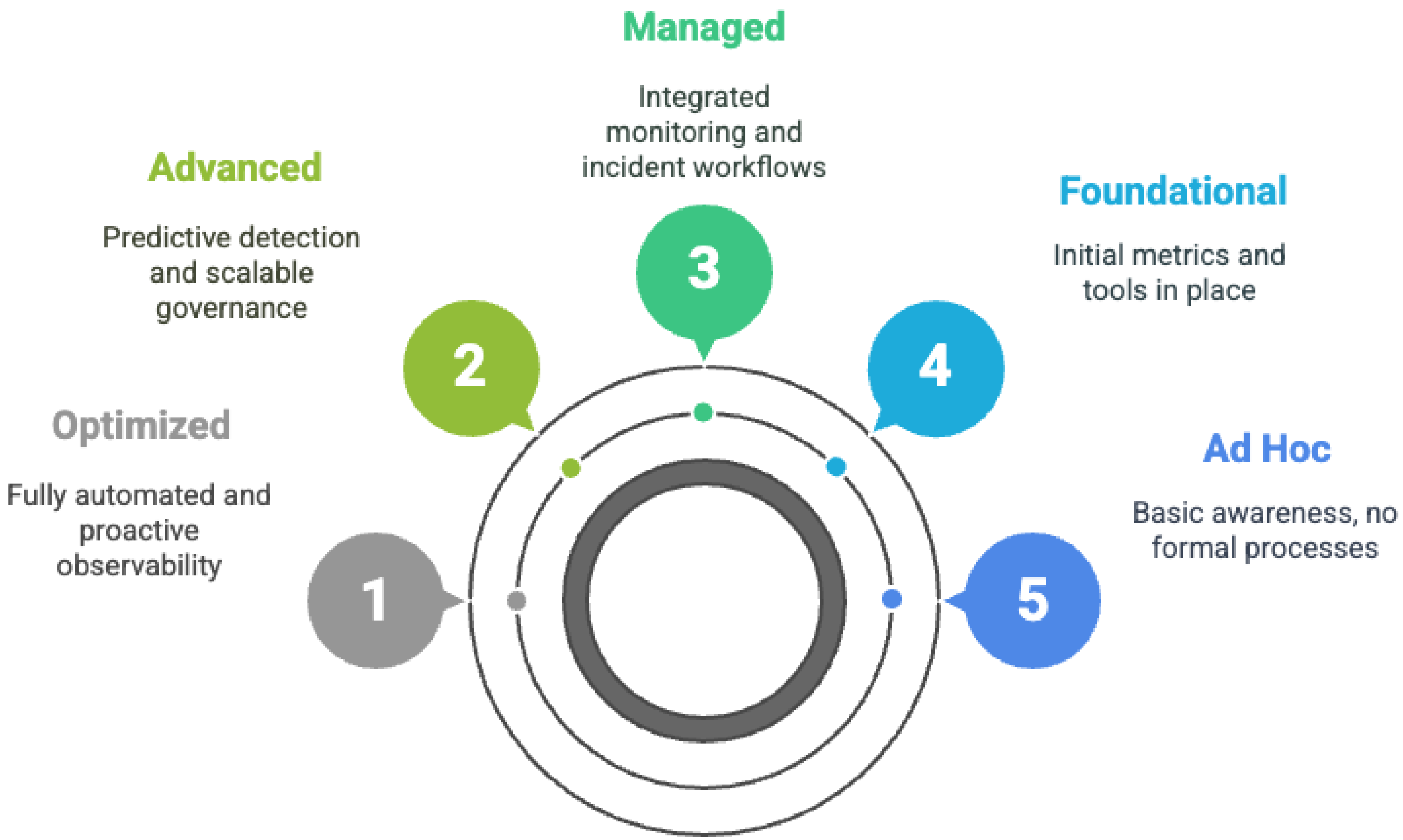
Critical Capabilities for Data Observability: Data Quality and Monitoring

- Alerting & Notifications
- Data Monitoring Dashboards
- Data Validation Rule Completeness

Critical Capabilities for Data Observability: Real-Time Data Processing and Analysis Features

- Real-Time Anomaly Detection
- Automated Metadata Collection
- Source and API Completeness
- Machine Learning Capabilities

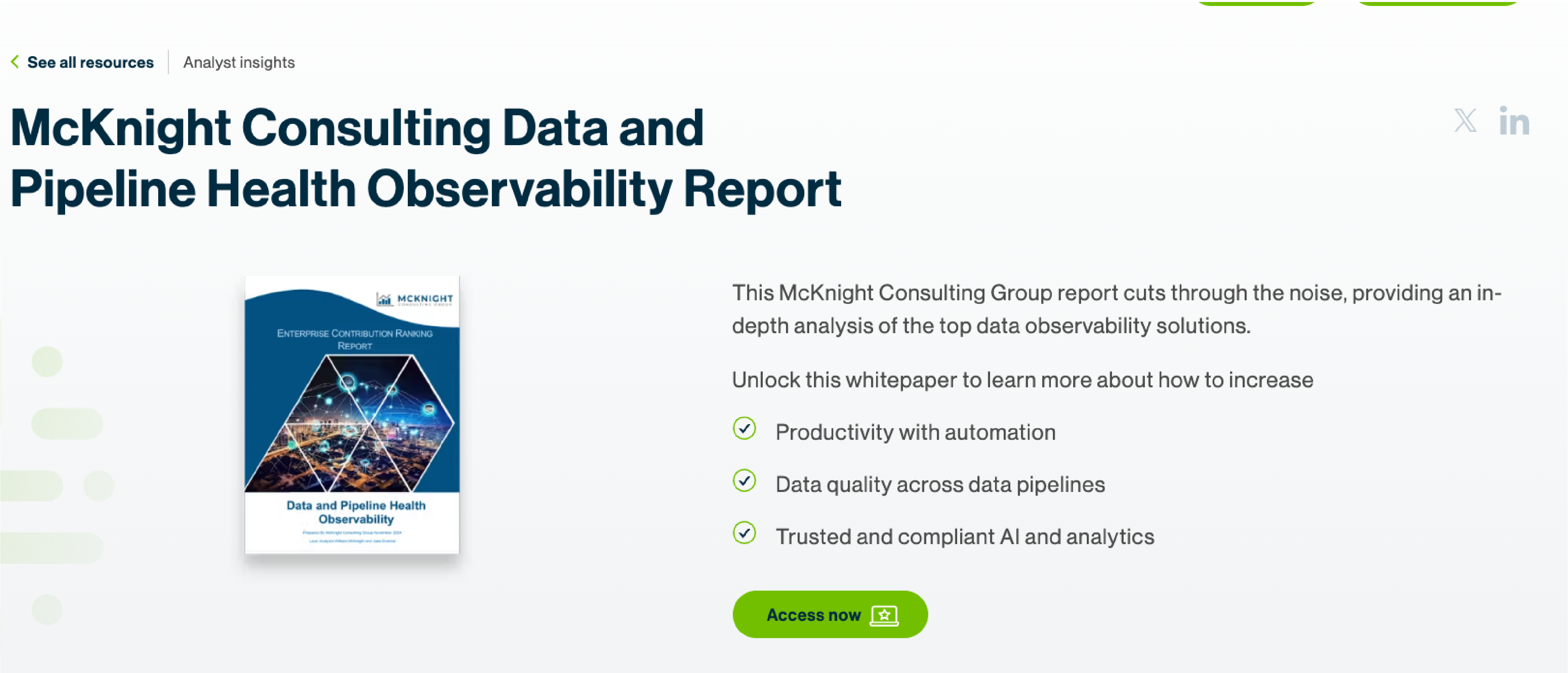
Data Quality/Observability Maturity Model



Managed
Integrated monitoring and incident workflows
Advanced
Predictive detection and scalable governance
Foundational
Initial metrics and tools in place
Optimized
Fully automated and proactive observability
Ad Hoc
Basic awareness, no formal processes
1
2
3
4
5

Made with Napkin

https://shorturl.at/giUIw



‹ See all resources | Analyst insights

McKnight Consulting Data and Pipeline Health Observability Report

MCKNIGHT CONSULTING GROUP

ENTERPRISE CONTRIBUTION RANKING REPORT

Data and Pipeline Health Observability

This McKnight Consulting Group report cuts through the noise, providing an in-depth analysis of the top data observability solutions.

Unlock this whitepaper to learn more about how to increase

- ✓ Productivity with automation
- ✓ Data quality across data pipelines
- ✓ Trusted and compliant AI and analytics

Access now

Summary

- Data Quality is never an accident
- Data Quality is becoming part of Data Observability
- The traditional approach to data quality is no longer sufficient for fast, streaming data
- Data observability is a new approach that is designed specifically for modern data, and it involves monitoring and analyzing data in real-time to detect anomalies, errors, and quality issues
- By adopting data observability for data quality, organizations can improve data quality, reduce downtime, and increase efficiency
- Data observability prevents errors and automates monitoring
- Enterprises should consider the complexity of their project scope and technical environment when selecting a data observability solution, utilizing the quadrant framework (Full Spectrum, Solution Specific, Bespoke, or Framework).Data quality enables reliable decisions

MCKNIGHT
CONSULTING GROUP

The Data Observability Advantage

Unlocking the Secrets to Reliable, High-Quality Big Data

Presented by: William McKnight

President, McKnight Consulting Group

3 x Inc 5000

/in/wmcknight

www.mcknightcg.com
(214) 514-1444

TOP VOICE
★ 2024 ★
thinkers 360
OVERALL

I'VE BEEN FEATURED
IN THE 2024
dataIQ™ 100 USA
THE MOST INFLUENTIAL PEOPLE IN DATA