MCKNIGHT
CONSULTING GROUP

Remembering Data Quality when the Data Spigot is Turned Way Up

Presented by: William McKnight

President, McKnight Consulting Group

3 X Inc 5000

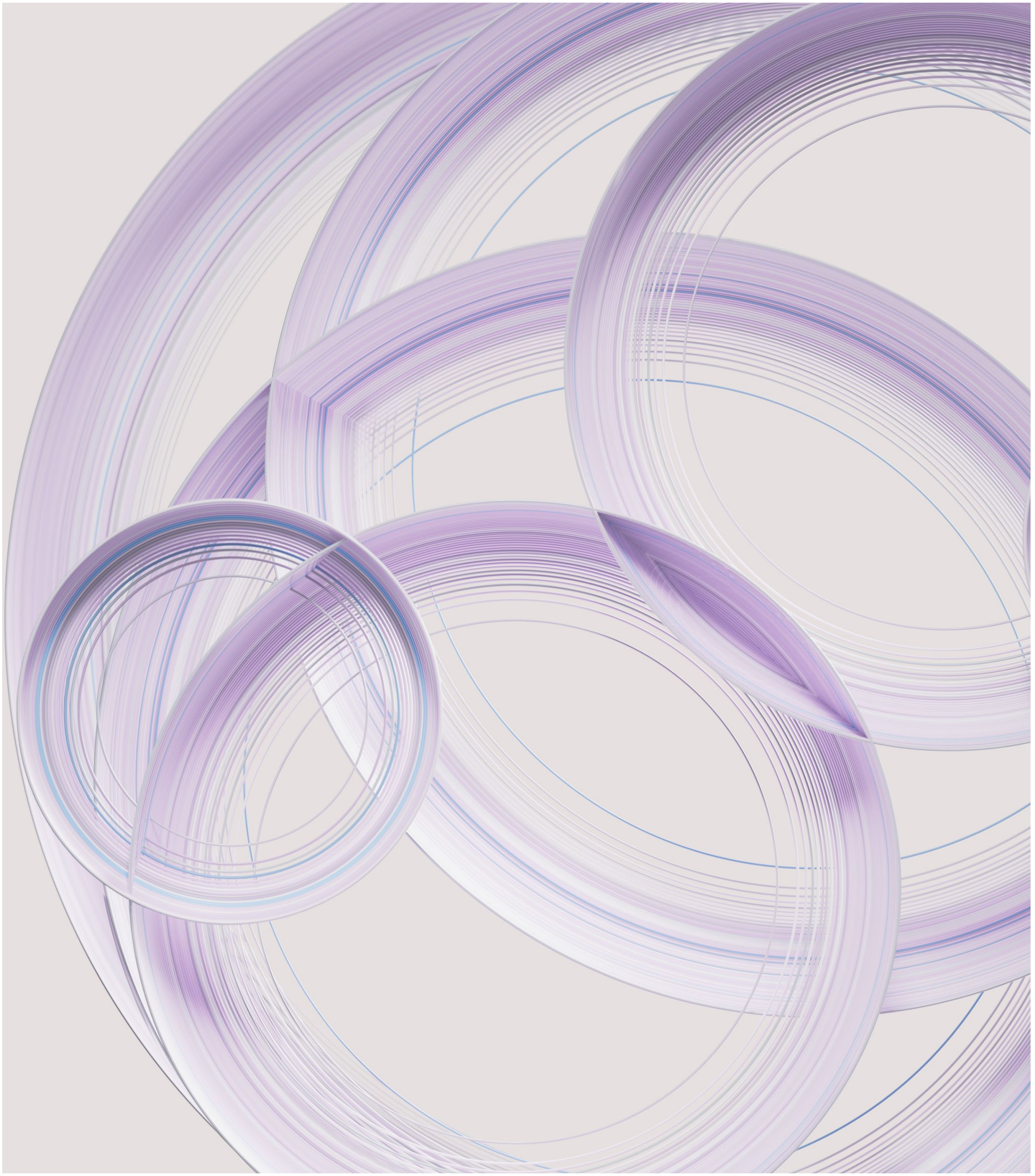
/in/wmcknight

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

The Next Generation of Data & AI



Global GDP is estimated to increase due to Generative AI by...

$10T

Which is why most companies are in piloting mode...

+45%

But the majority are not in production...

90%

Source: 1. 2024 J.P. Morgan Global Research study
Source: 2. October 2023 Gartner Inc. poll of 1400+ leaders

Enterprise Generative AI Data Problem

There is an unprecedented change in your data strategy →

Fragmented Data Stack

2X

more software vendors serving Fortune 2000 client data stacks.[1]

Increased Productivity Pressure due to Generative AI

60%

of occupations could be partially automated by Generative AI. [2]

Lack of Enterprise Data Readiness for Generative AI

6%

of enterprises have a Generative AI application in production.[3]

Source: 1. Wall Street Journal
Source: 2. Goldman Sachs
Source: 3. Harvard Business Review

Preparing your data for AI

Design principles of the next-generation data fabric →

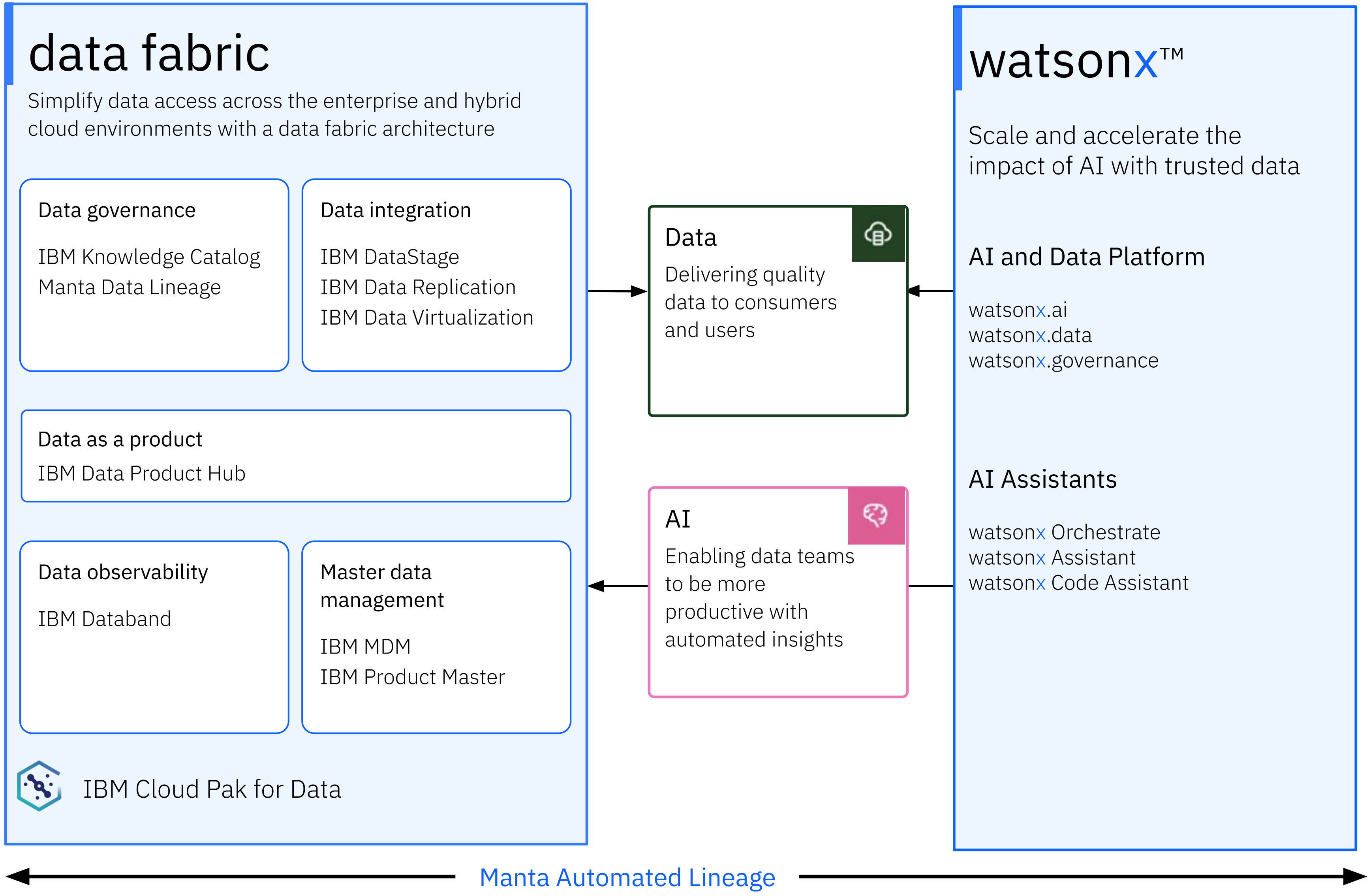
1. Data Integration across a range of integration styles
2. Data Intelligence for curating and delivering trusted data
3. Hybrid-Ready any location, any style, any integration
4. Generative AI for Data to achieve unprecedented productivity
5. Data for Generative AI across structured and unstructured data

IBM Data Fabric + watsonx

- Quality and integrity
- Trust
- Fit for purpose model
- Openness

Preparing Your Data

Deliver AI Outcomes



data fabric

Simplify data access across the enterprise and hybrid cloud environments with a data fabric architecture

Data governance
IBM Knowledge Catalog
Manta Data Lineage

Data integration
IBM DataStage
IBM Data Replication
IBM Data Virtualization

Data as a product
IBM Data Product Hub

Data observability
IBM Databand

Master data management
IBM MDM
IBM Product Master

IBM Cloud Pak for Data

Data
Delivering quality data to consumers and users

AI
Enabling data teams to be more productive with automated insights

watsonx™

Scale and accelerate the impact of AI with trusted data

AI and Data Platform
watsonx.ai
watsonx.data
watsonx.governance

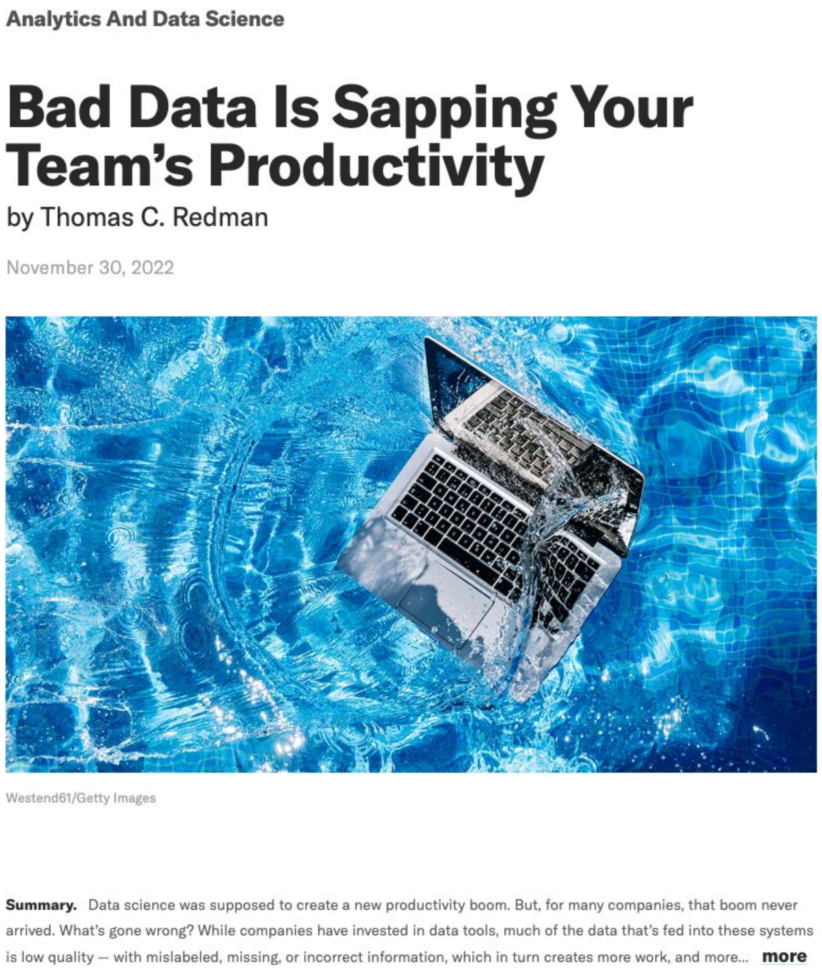
AI Assistants
watsonx Orchestrate
watsonx Assistant
watsonx Code Assistant

Manta Automated Lineage

Gain deeper visibility into the data and its journey from source to end-use for regulatory compliance and AI use cases with Manta, an IBM company ibm.com/artificial-intelligence

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



Analytics And Data Science

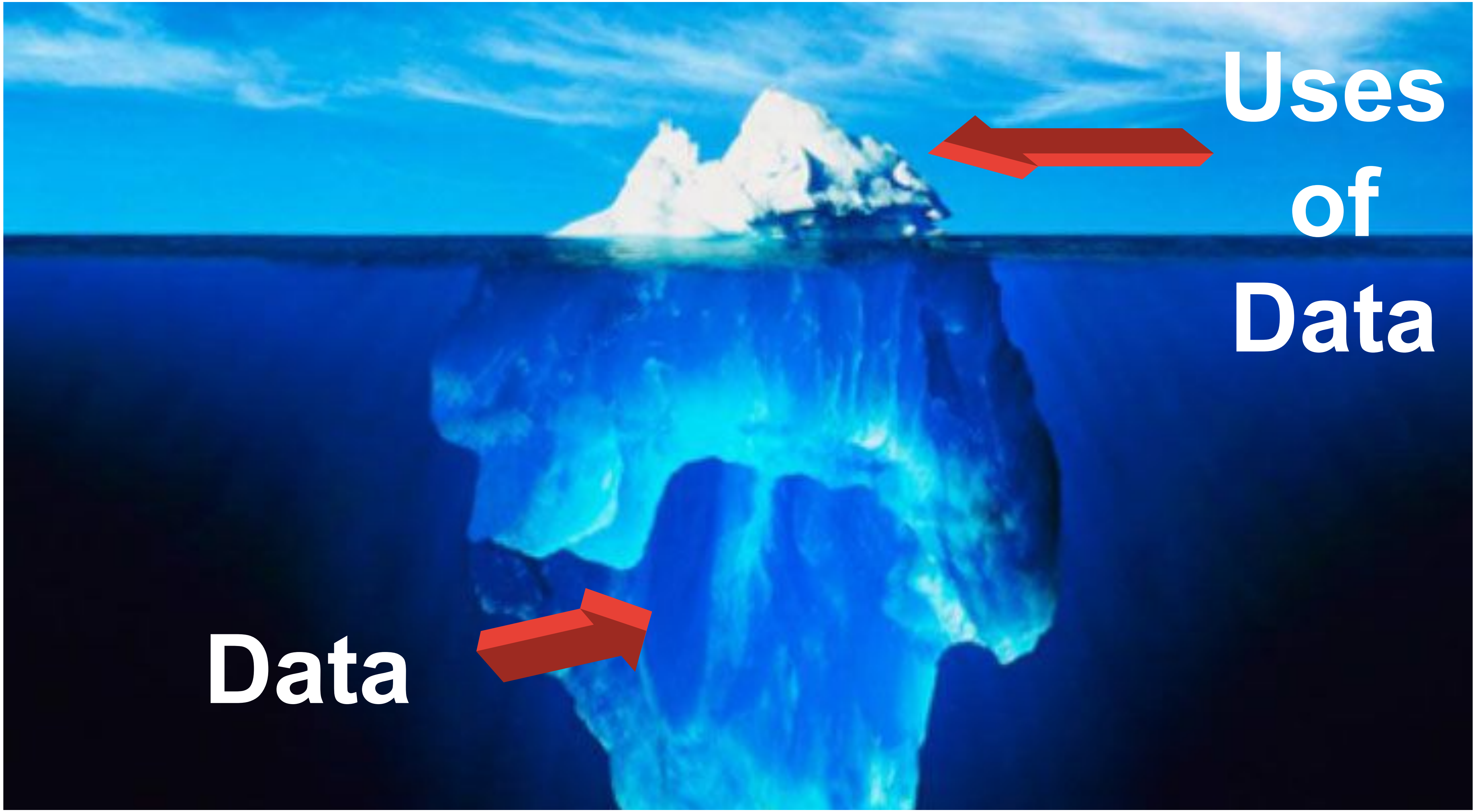
Bad Data Is Sapping Your Team's Productivity

by Thomas C. Redman

November 30, 2022

Westend61/Getty Images

Summary. Data science was supposed to create a new productivity boom. But, for many companies, that boom never arrived. What's gone wrong? While companies have invested in data tools, much of the data that's fed into these systems is low quality — with mislabeled, missing, or incorrect information, which in turn creates more work, and more... **more**



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 They Do

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

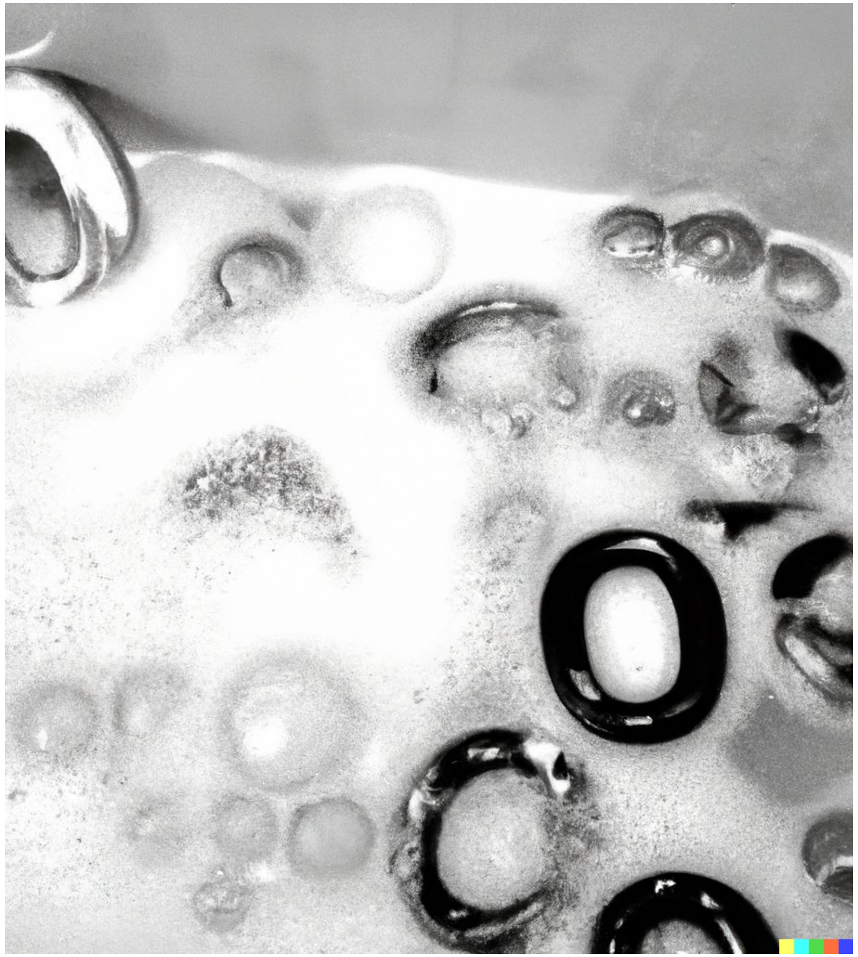
Investments in Data Quality

- 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)

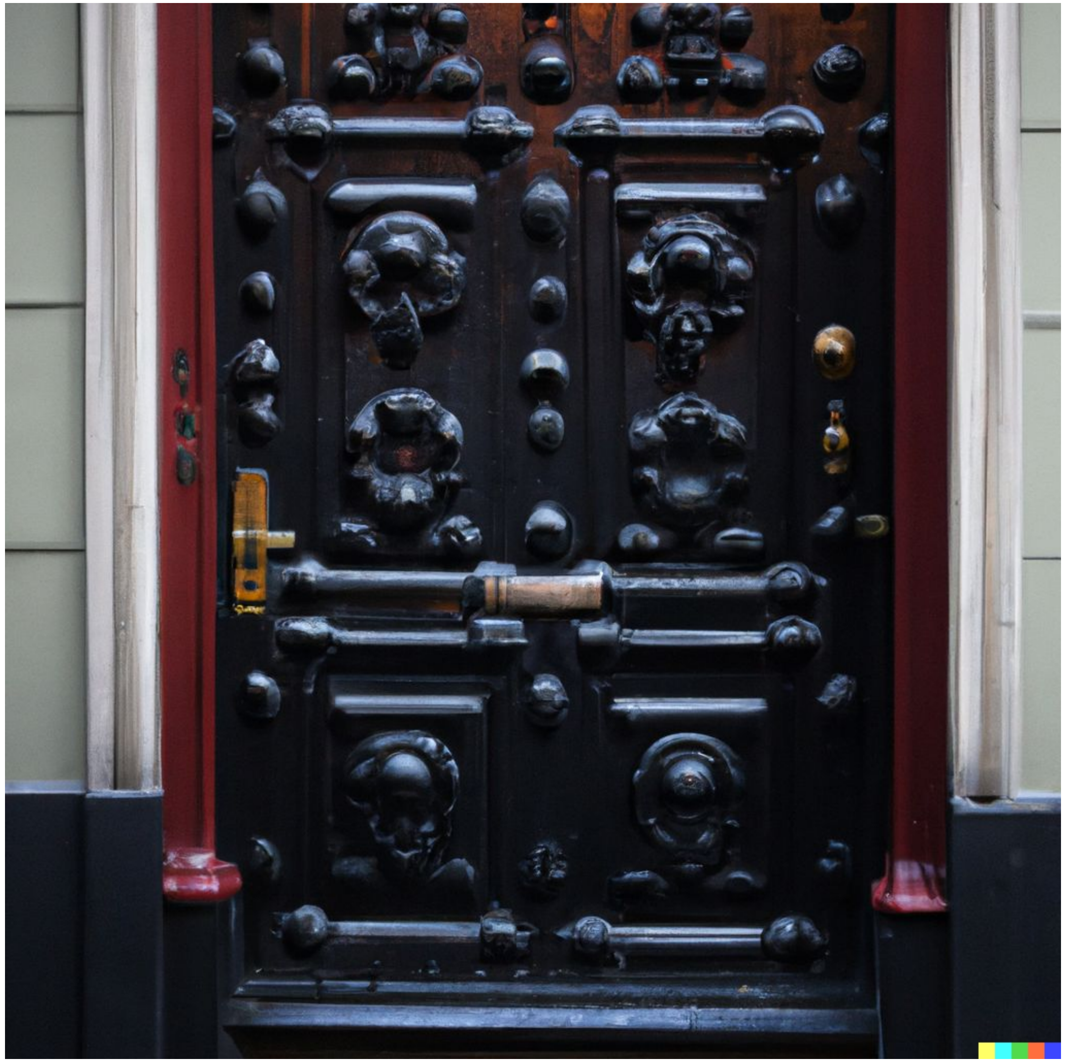
- Improved Decision Making
- Increased Efficiency
- Reduced Risk
- Improved Customer Satisfaction
- Increased Profitability
- Enhanced Security
- Improved Compliance

A Good Data Governance Program Keeps Business Interest In Data Quality

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

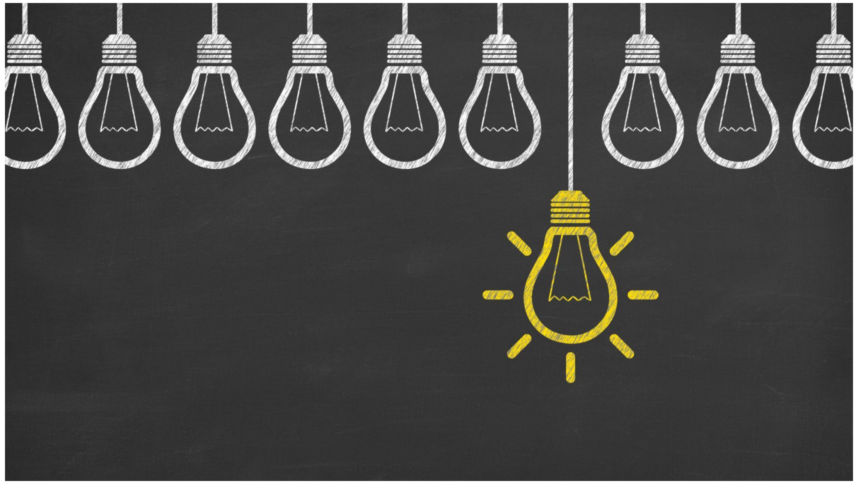
Five Actions to Perform for Data Quality

- 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
- Vector Database
- Feature Store

Data Quality Expectations



1. Appropriate
2. Suitable
3. Adequate
4. Proper
5. Satisfactory
6. Acceptable
7. Relevant
8. Serviceable
9. Functional
10. Usable
11. Fit-for-purpose

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.

Improve by 1 point, pays for Project

- Reduce fake claims, inflated claim amounts, multiple claims for the same incident
- Reduce Misrepresentation of policy coverage, policyholder impersonation, claims for ineligible vehicles
- Fraud
- Returns
- Customer Retention
- Supply Chain Efficiency
- Predictive Model Accuracy
- Marketing Effectiveness
- Claims Processing
- Inventory Management
- Revenue Cycle Management
- Operational Efficiency
- Risk Management

Streaming Data Data Quality

Big Data Collection Systems

- Exactly once versus At Least Once
- At Least Once
  - Guarantees Order of Delivery
    - Apache Kafka/Amazon MSK
    - Kinesis Data Streams
  - Does not Guarantee Order of Delivery
    - SQS in Standard Mode
    - Kinesis Data Firehose
- Exactly Once with Guaranteed Order
  - SQS in FIFO Mode
  - Dynamo DB Streams

Streaming Data Presents Unique Data Quality Challenges

- **Data Quality Challenges in Streaming Data**
  - **High velocity**: Data is generated at an incredible pace, making it difficult to ensure quality.
  - **High volume**: The sheer amount of data generated can be overwhelming, making it hard to detect quality issues.
  - **Variety**: Streaming data comes in various formats, structures, and sources, adding complexity to quality checks.
- **The Need for Real-time Data Quality**
  - **Real-time data validation**: Validate data as it's generated to detect quality issues immediately.
  - **Real-time data cleansing**: Cleanse data in real-time to prevent quality issues from propagating downstream.
  - **Real-time data monitoring**

Examples of Data Quality Challenges with Streaming Data

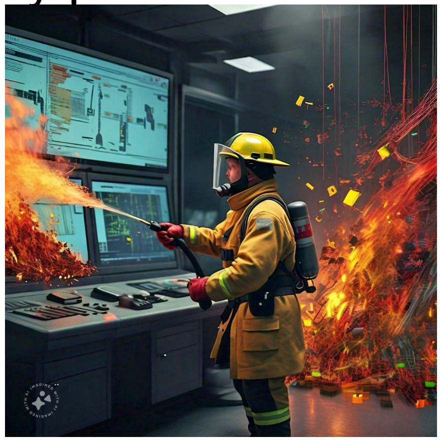
- Inaccurate Predictive Models: Inaccurate data leads to incorrect predictions and wasted resources on unnecessary retention efforts.
- Inefficient Resource Allocation: Incomplete and inconsistent data causes inefficient resource allocation and worsens traffic congestion.
- Missed Business Opportunities: Noisy and duplicate data leads to missed business opportunities and inaccurate market analysis.
- Regulatory Non-Compliance: Erroneous data causes false negatives and regulatory non-compliance, resulting in fines and reputational damage.
- Ineffective Personalization: Incomplete and inconsistent data leads to inaccurate recommendations and a poor user experience.
- Supply Chain Disruptions: Erroneous data causes inaccurate insights and supply chain disruptions, resulting in delayed production and costly penalties.
- Security Threats: Noisy data with false positives leads to unnecessary alerts and wasted resources, allowing security threats to go undetected.

Streaming Data Data Quality

- Streaming data generates a large amount of data in real time, making it extra difficult to assess and maintain data quality.
- Streaming data can come from diverse sources and have different formats, structures, and semantics, posing challenges in data harmonization and standardization.
- Data quality issues need to be identified and resolved quickly in streaming scenarios to ensure the accuracy and reliability of real-time analytics and decision-making.

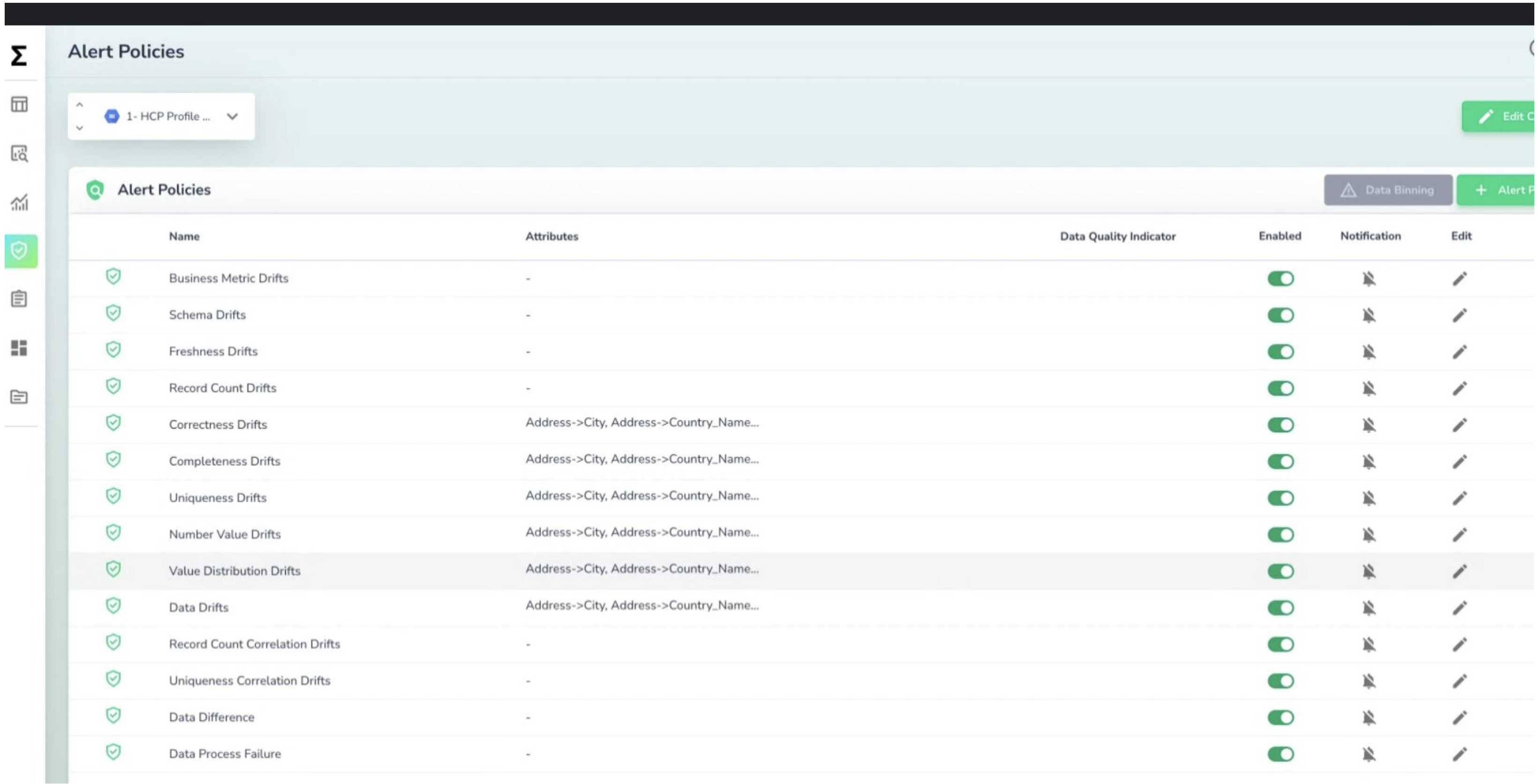
So…

- 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.



The Data Quality Needs of Streaming Data Leads to Data Observability

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



	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

- 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.

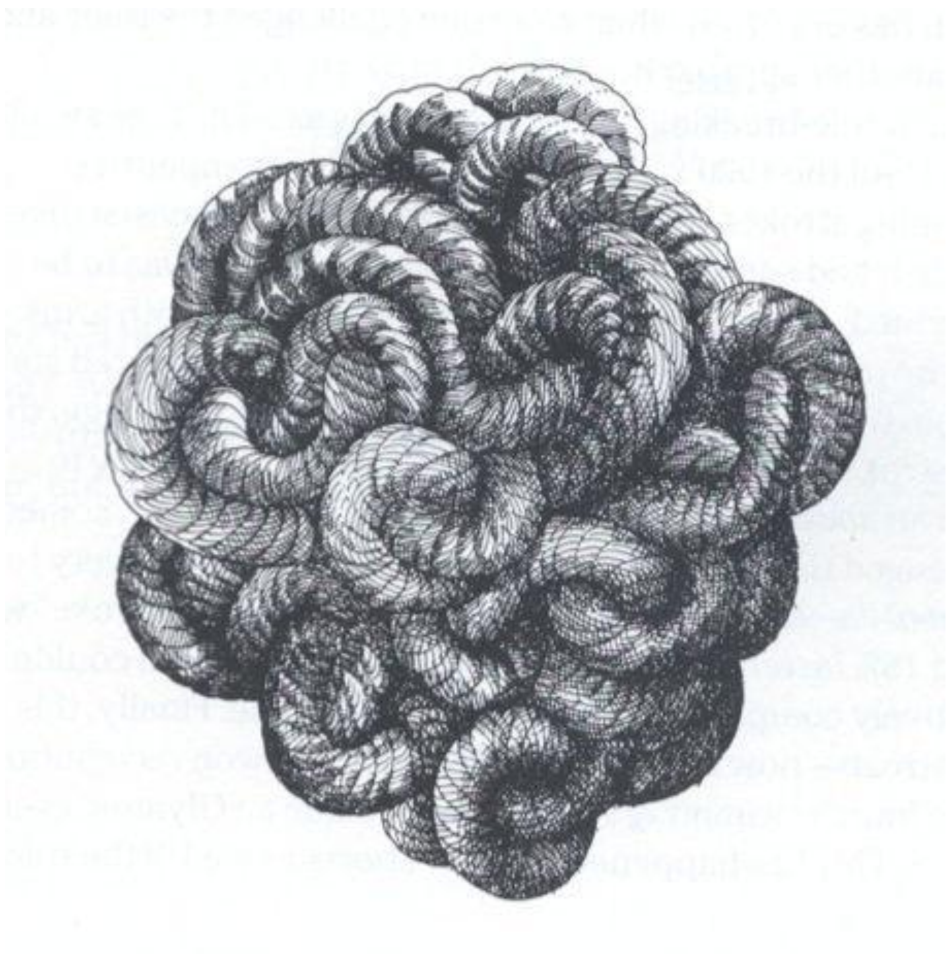
We Divide the Data Observability Vendors Into Two Camps

- 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.

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.



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**

Summary

- Data quality can and should have a value proposition
- Data quality is never an accident
- Consider data quality when considering applications
- Establish the Value Proposition for Data Quality in Project Improvement
- 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 streaming 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

MCKNIGHT
CONSULTING GROUP

Remembering Data Quality when the Data Spicket is Turned Way Up

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