From Data to Decision: Why and How Great Data is Key to FSI Innovation and Efficiency

Or the secret value of great data management in an ever-changing, AI-driven world.

HPC | QUANTUM | DATA | AI

Who I Am, What I Do

Alex Woodie

- Managing Editor, Datanami (soon to be BigDATAwire)
- Contributing Editor at HPCwire and AIwire
- Senior Editor at IT Jungle (IBM i and AS/400)

What is “Data Management”?

Carl Olofson, IDC analyst

Barcelona hotel, site of paella

Rules of Data Management

Why not just use stone tablets?

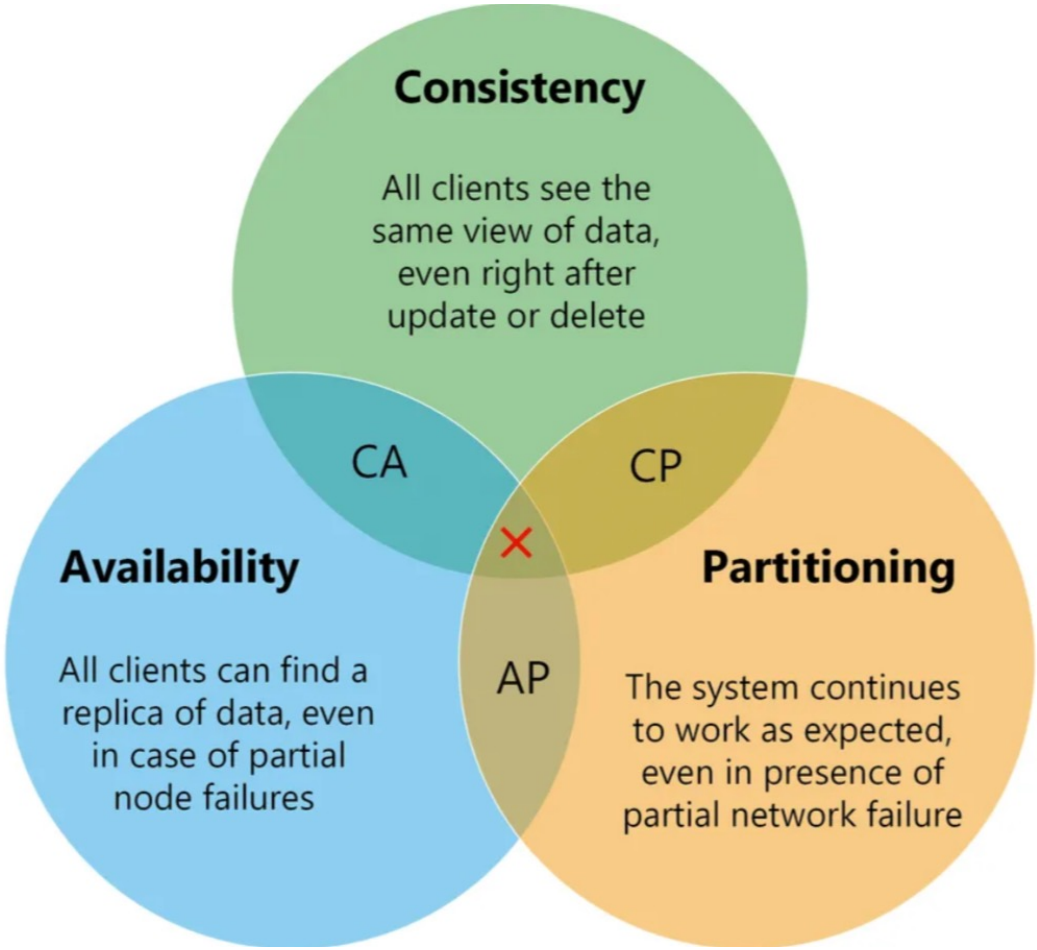
- Because the data is constantly changing
- And because the use cases are constantly changing, too.

But There Are Always Tradeoffs

CAP Theorem says there is a tradeof among:

- Data consistency
- Data availability
- Data partitioning

- You manage data in a way that does everything for everyone



Finding Data Value

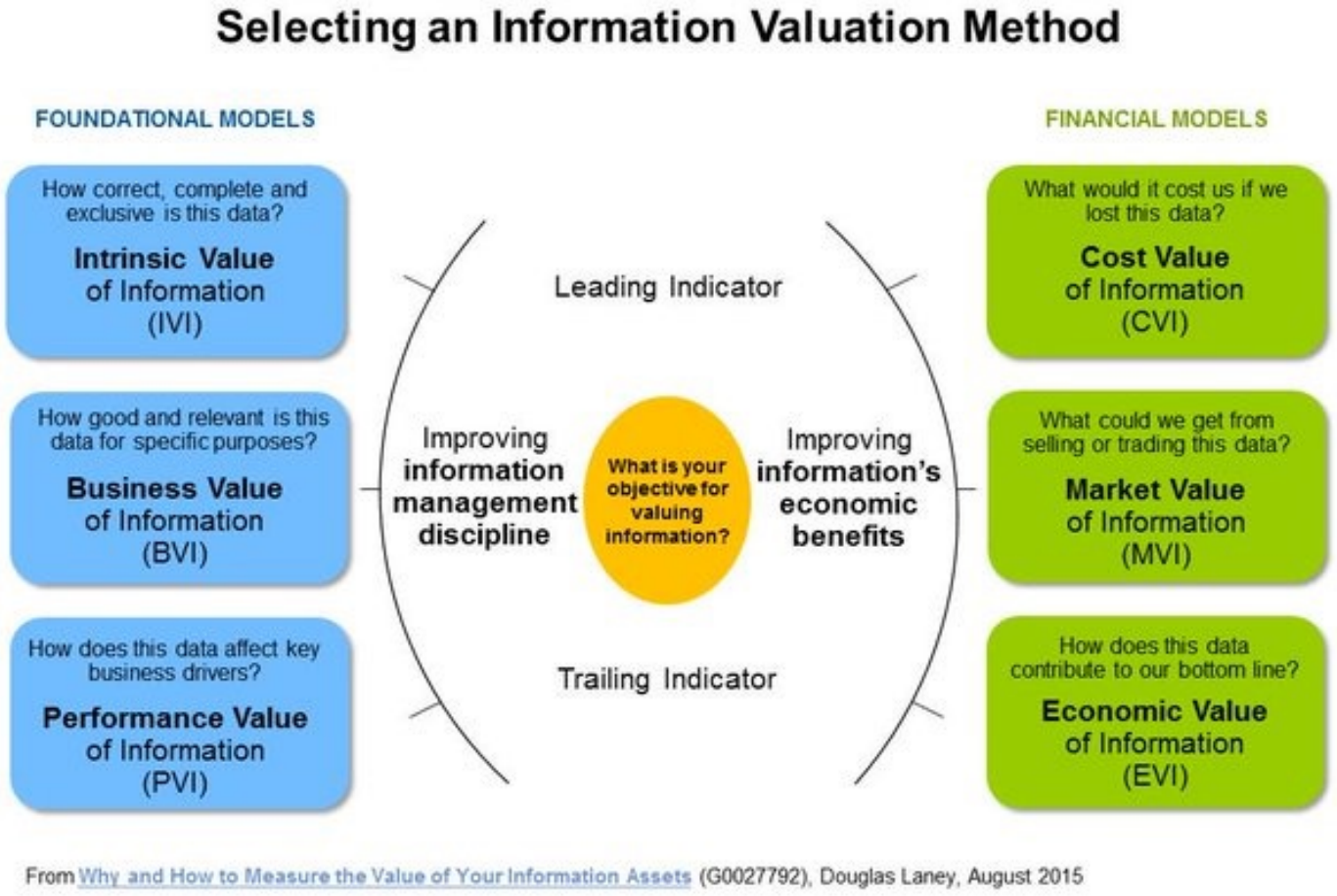
HPC + AI WALL STREET

- 9/11 showed a core lack of data valuation principles

- Doug Laney from META Group spearheaded new approach

Data Valuation (cont.)

- Multiple methods for calculating the value of data
- Laney's Infonomics uses foundational and financial models



Data Management Challenges in Financial Services

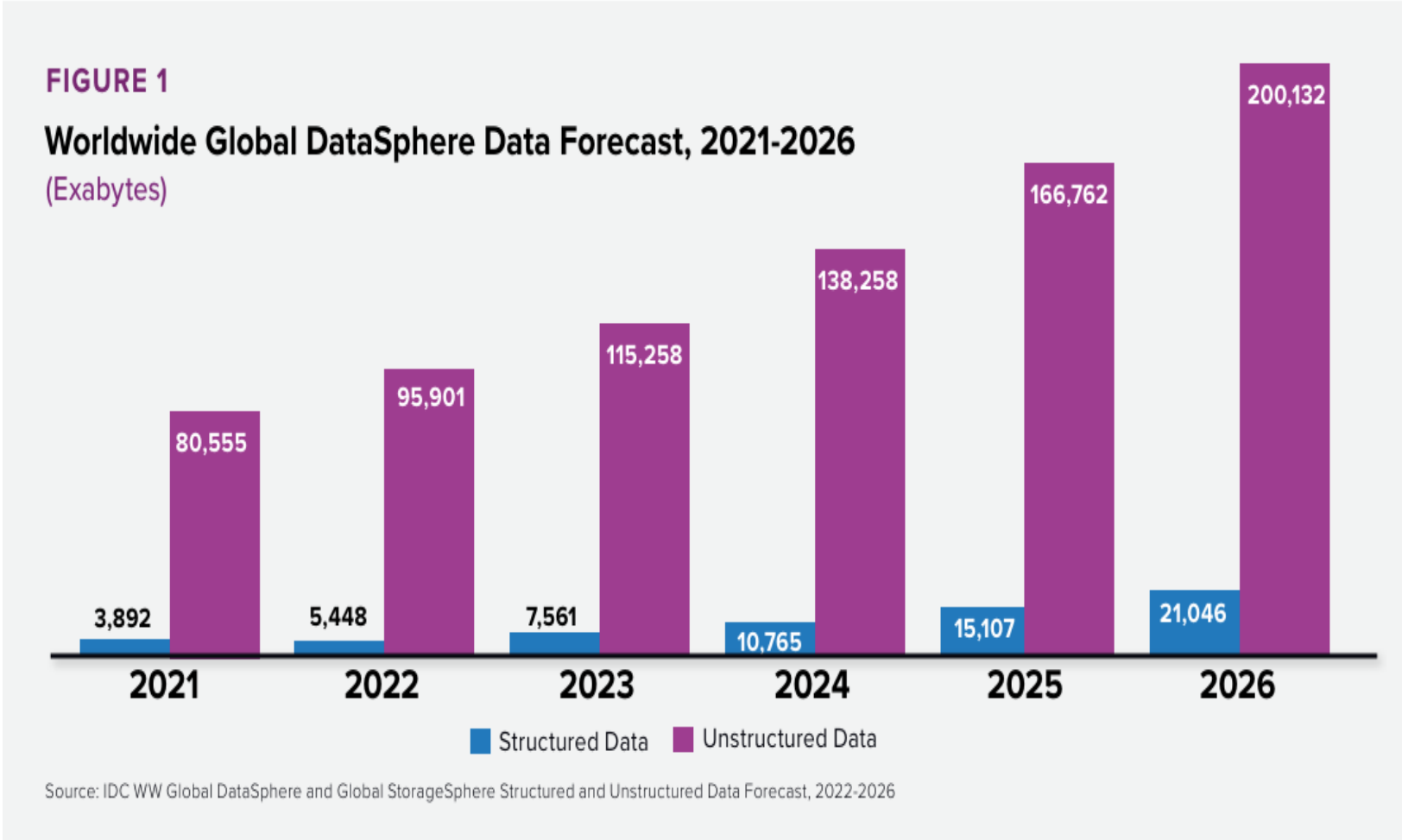
HPC + AI WALL STREET

- Regulations at state, national, and international levels increase cost and complexity
- Data sovereignty requirements
- M&A creates data silos/integration challenges
- Legacy data systems create friction
- Data privacy and security risks amplified

Growth of Unstructured Data

IDC Global DataSphere study predicted 175 zettabytes of data created by 2025

- ~80% unstructured
- ~20% structured



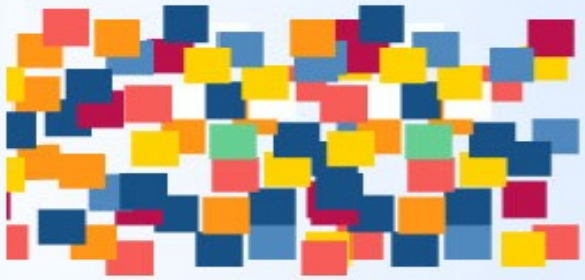
What Is Unstructured Data?

- Text (email, PDFs, etc.)
- Images
- Audio
- Video
- Social media data
- IoT data
- Sensor data

Structured vs Unstructured Data

HPC + AI WALL STREET

Structured Data	Unstructured Data
• Consistent data types and formats	• Varied data types and formats
• Easily searchable and analyzed	• Difficult to search and analyze
• Well-defined schema and relationships	• No predefined schema or relationships
• Stored in databases or spreadsheets	• Stored in documents, emails, social media, etc.
• Examples: Sales transactions, financial statements	• Examples: Text documents, images, videos



Semi-Structured Data

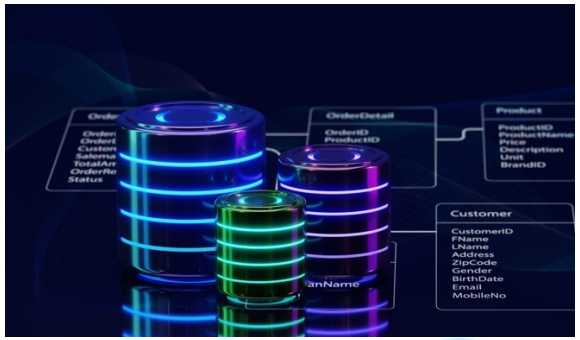
- Semi-structured data sits between structured and semi-structured data
- Examples:
  - JSON
  - Web logs
  - Sometimes email (if it has metadata)

{JSON}
JavaScript Object Notation

```
document
{
    "sellerid": 123456,
    "type": "car",
    "style": "sedan",
    "year": 2013,
    "trim": "performance",
    "model": "s"
}
```

Storage for Structured, Unstructured, and Semi-Structured Data

- Relational databases for structured data



- NoSQL databases for semi-structured data

- File systems and object stores for unstructured data

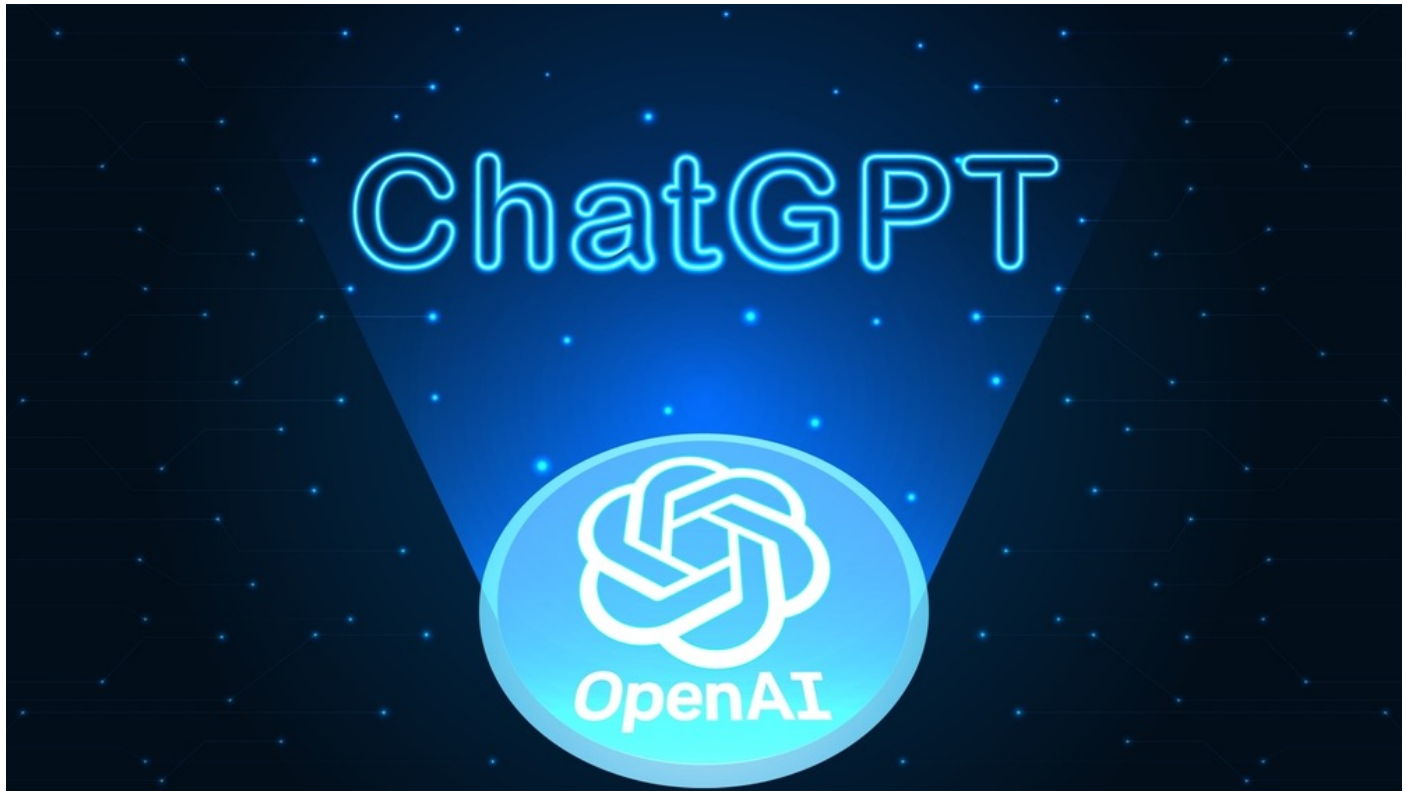
Why Unstructured Data is Important

Unstructured data — words and images — is the fuel for neural networks and deep learning.

- Recurrent neural nets
- Convolutional neural nets
- Transformers (i.e. large language models)

GenAI Relies on Unstructured Data

- Generative AI via LLMs and other foundation models run on unstructured data
- But storing unstructured data is not easy
- Big problemo!

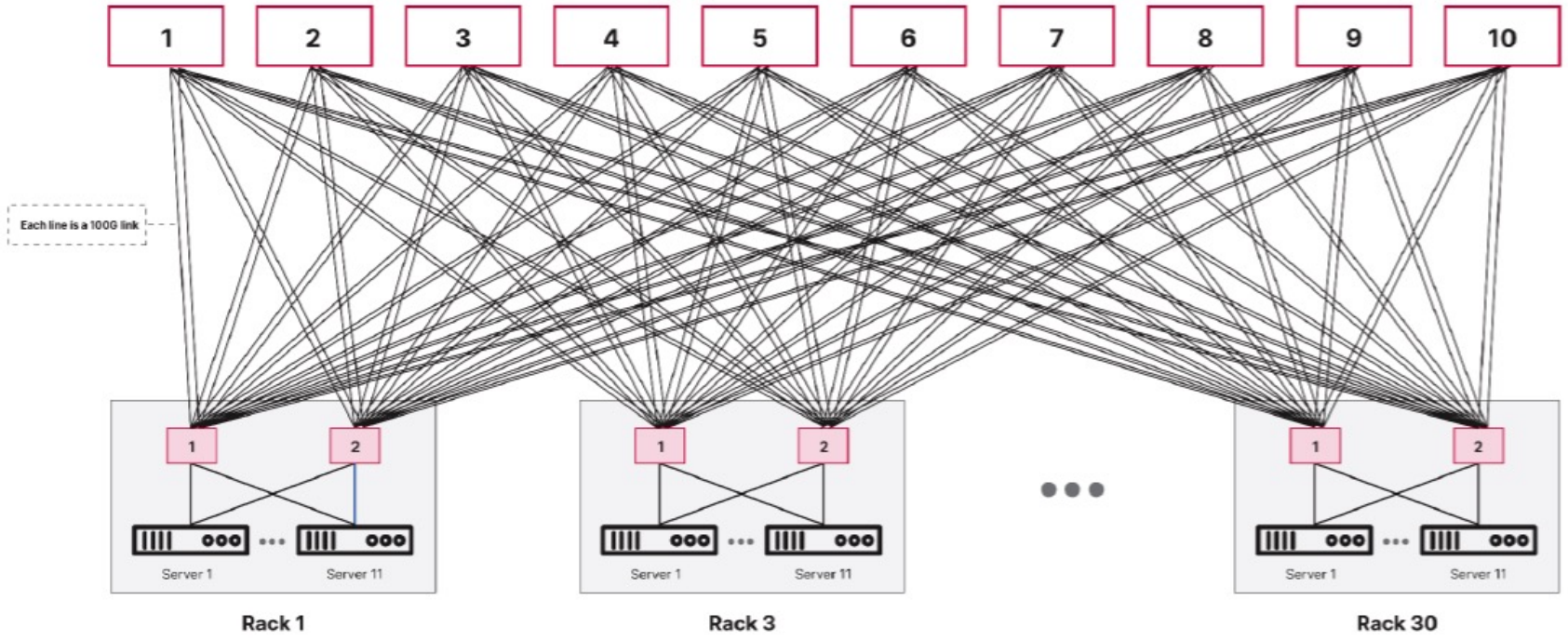


Enter the Exabyte Era

HPC + AI WALL STREET

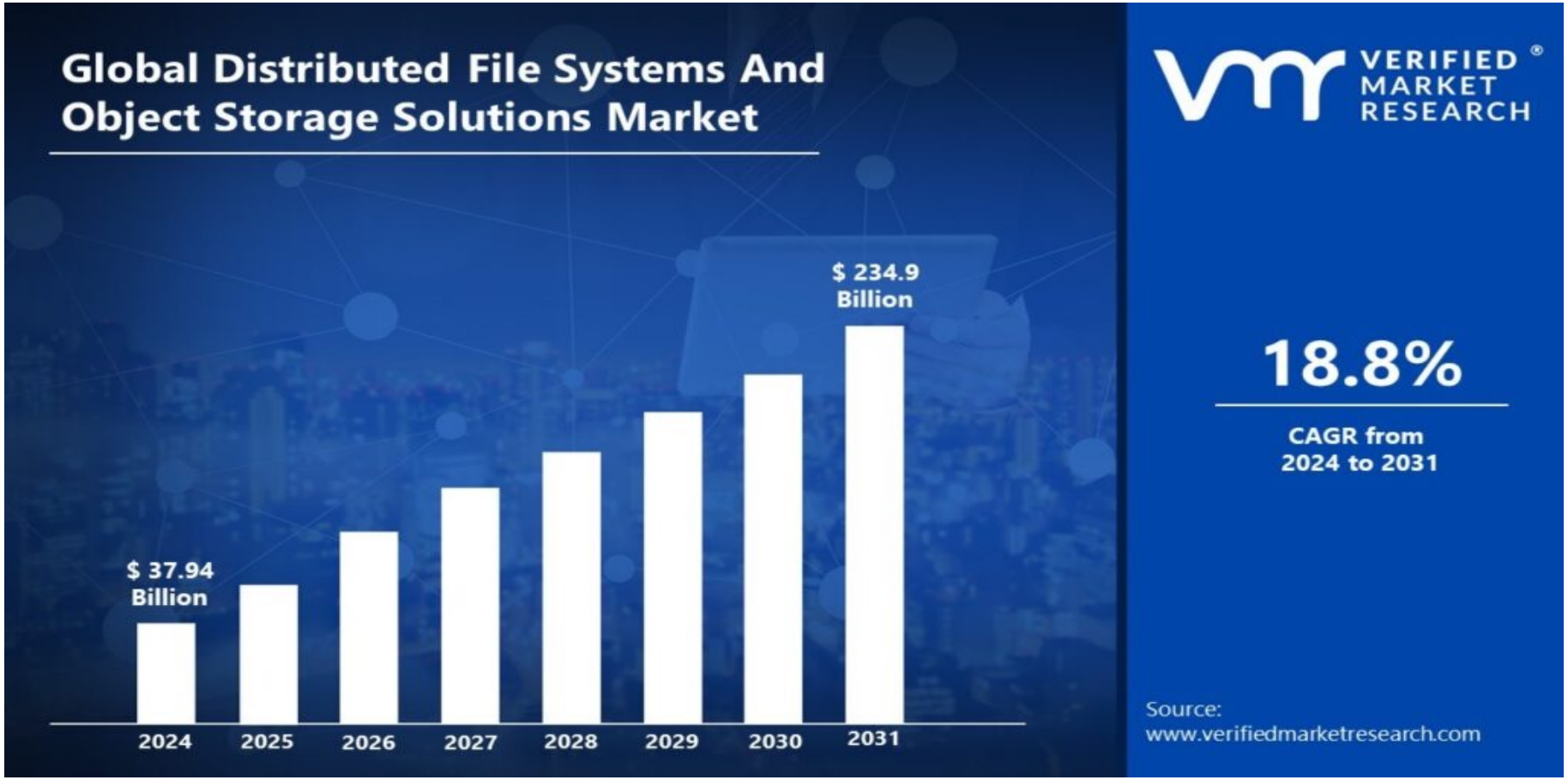
MinIO Datapods

- 24 x 30TB NVMe drives per server
- 11 servers per rack
- 30 racks per cluster



"100 to 200 petabytes is the new single-digit petabytes."
– *MinIO CEO AB Periasamy*

Growth of File Systems and Object Stores



Healthy Market for File Systems

Proprietary

Open

Object Stores Galore (As Long As It Looks Like S3)

Amazon S3 is the defacto standard for object stores.

Google Cloud Storage

Azure Blob Storage

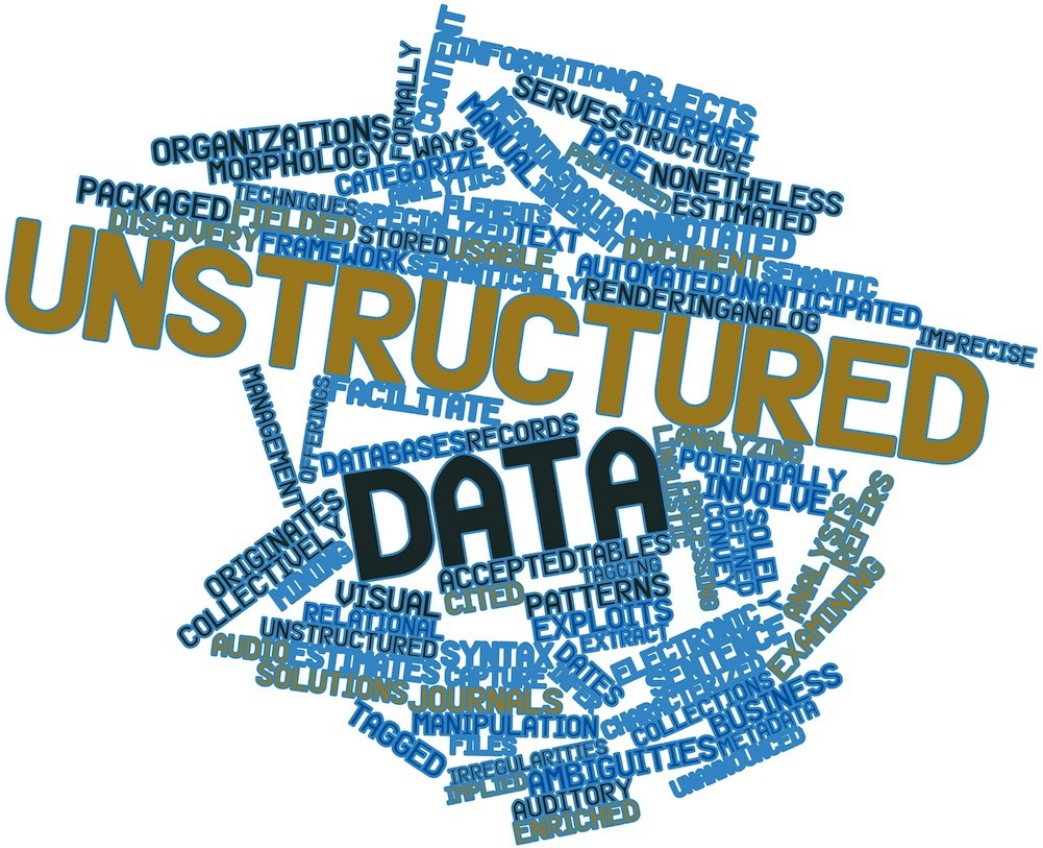
IBM Cloud Object Storage

SCALITY

Unstructured Data Management in its Infancy

HPC + AI WALL STREET

Companies spend 30% of their IT budgets on storing and backing up unstructured data (Komprise)



"They don't know what they have, most of what they have is crap, and they don't even have access to it."
— *Jacob Farmer, founder of Starfish Storage, on unstructured data*

The Data Is (Still) Not Alright

HPC + AI WALL STREET

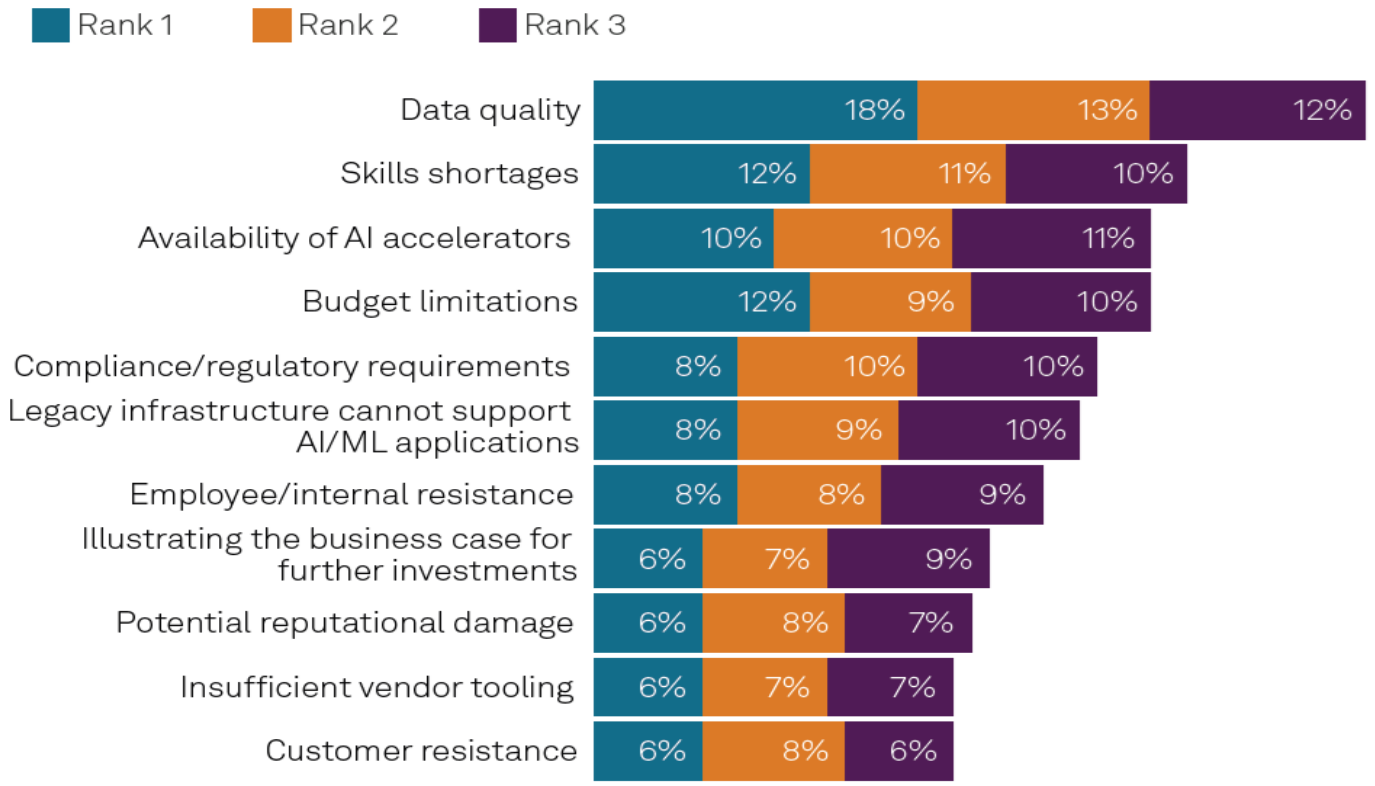
- GenAI exposing bad data
- GIGO still a thing
- Some companies caught unprepared
- Good data management and governance being rewarded

Data Quality is Not Up To Par

Weka survey finds:

- "Legacy data platforms" partly to blame for poor data quality
- Companies that invest in data management have better AI outcomes

Figure 5: Top three impediments to organizations moving an AI/ML application from pilot to production environments



Q. What are the primary challenges or impediments to moving an AI/ML application from proof-of-concept/piloting stages to production environments?
Base: All respondents (n=1,519).
Source: S&P Global Market Intelligence 451 Research Global Trends in AI custom survey, 2024.

Data Governance is Lacking

HPC + AI WALL STREET

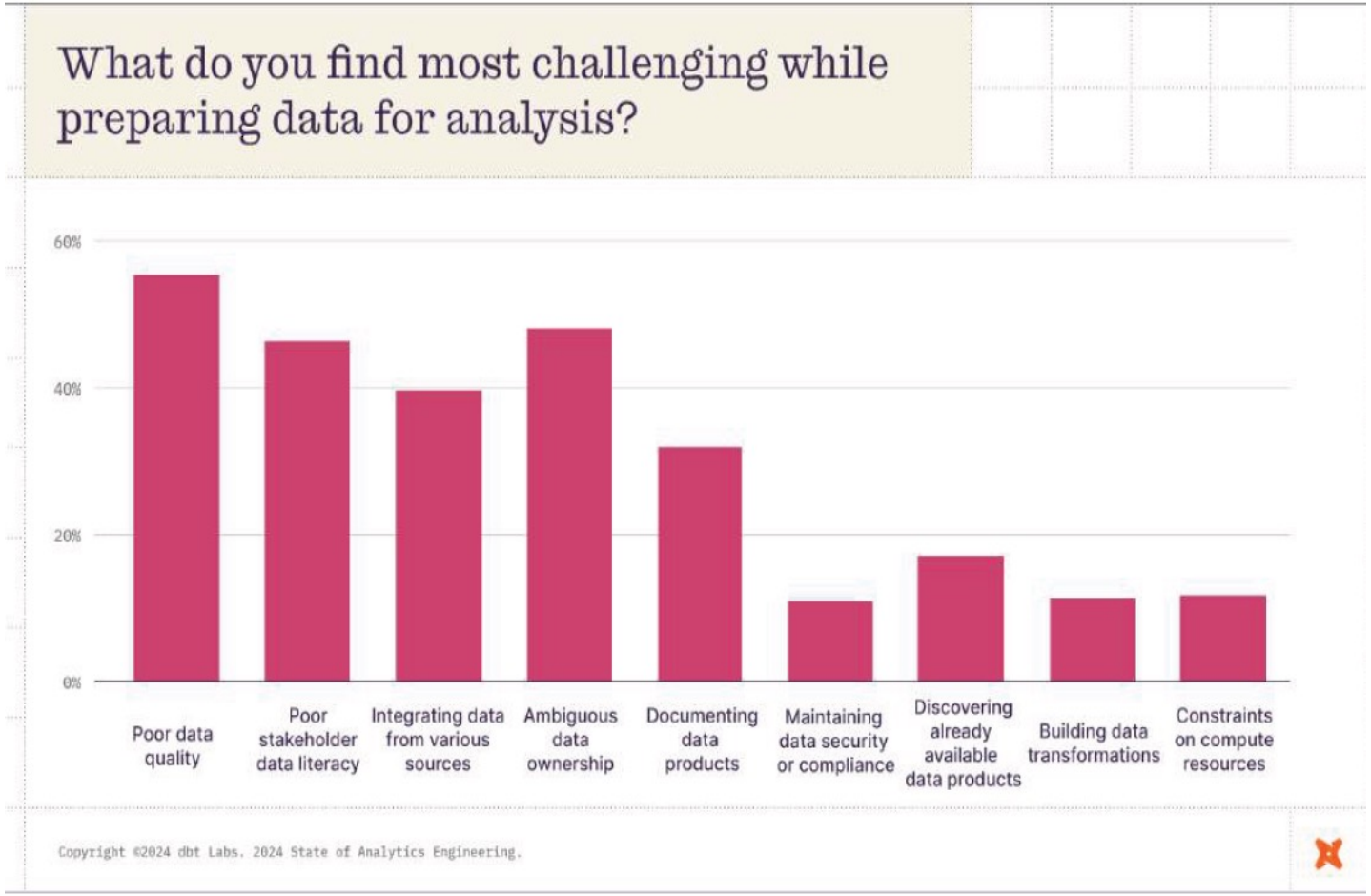
What data challenges inhibit your organization's progress in relation to AI initiatives?

62%
say governance of data used for AI is a challenge

- **62%** Governance of data used for AI
- **54%** Availability of data attributes to increase relevance of AI outcomes
- **52%** Quality of data to use for training or inference
- **50%** Data privacy and security
- **41%** Lack of access to sufficient data to prevent bias

Lack of Data Quality (cont.)

- Dbt Labs' "State of Analytics Engineering 2024" survey
- 57% of data professionals cited data quality as a top challenge, up from 41% a year ago
- 60% said they were increasing investment in data quality solutions



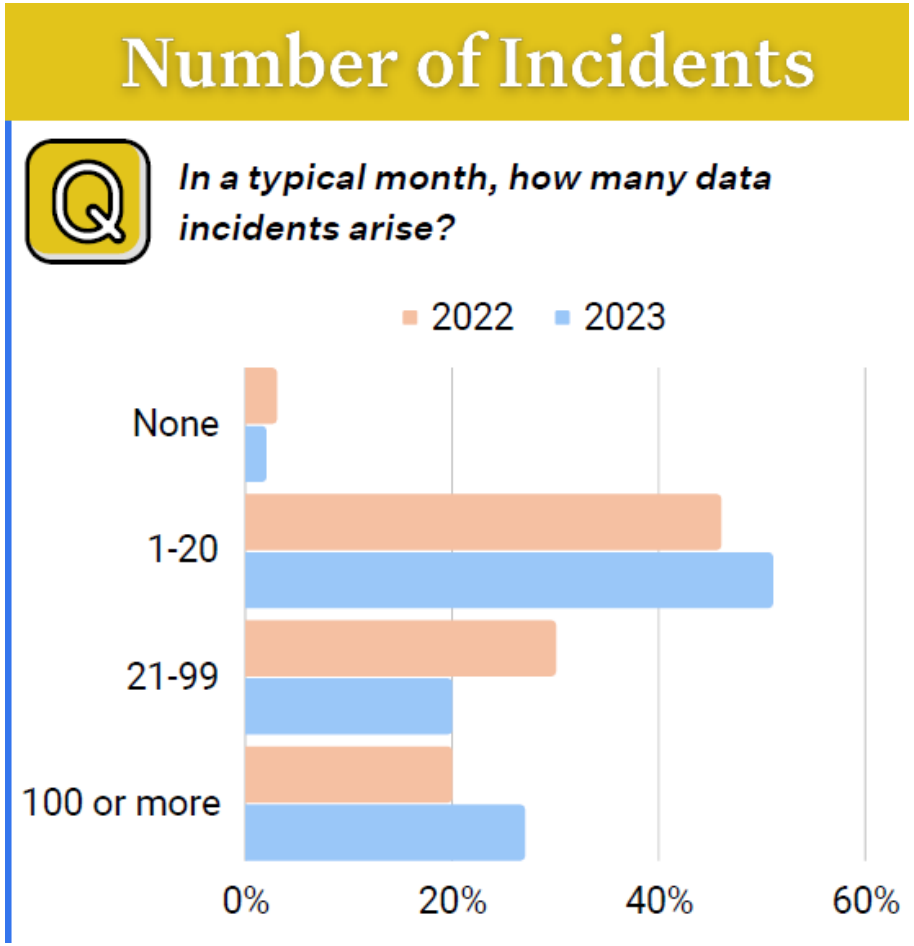
Data Incidents on the Rise

A Monte Carlo study finds:

- Average number of data incidents per month increased from 59 per organization in 2022 to 67 in 2023.

"Basically, people are having more issues, spending more time on them, and generally getting into situations where their stakeholders and business are impacted before they can actually respond and fix things."

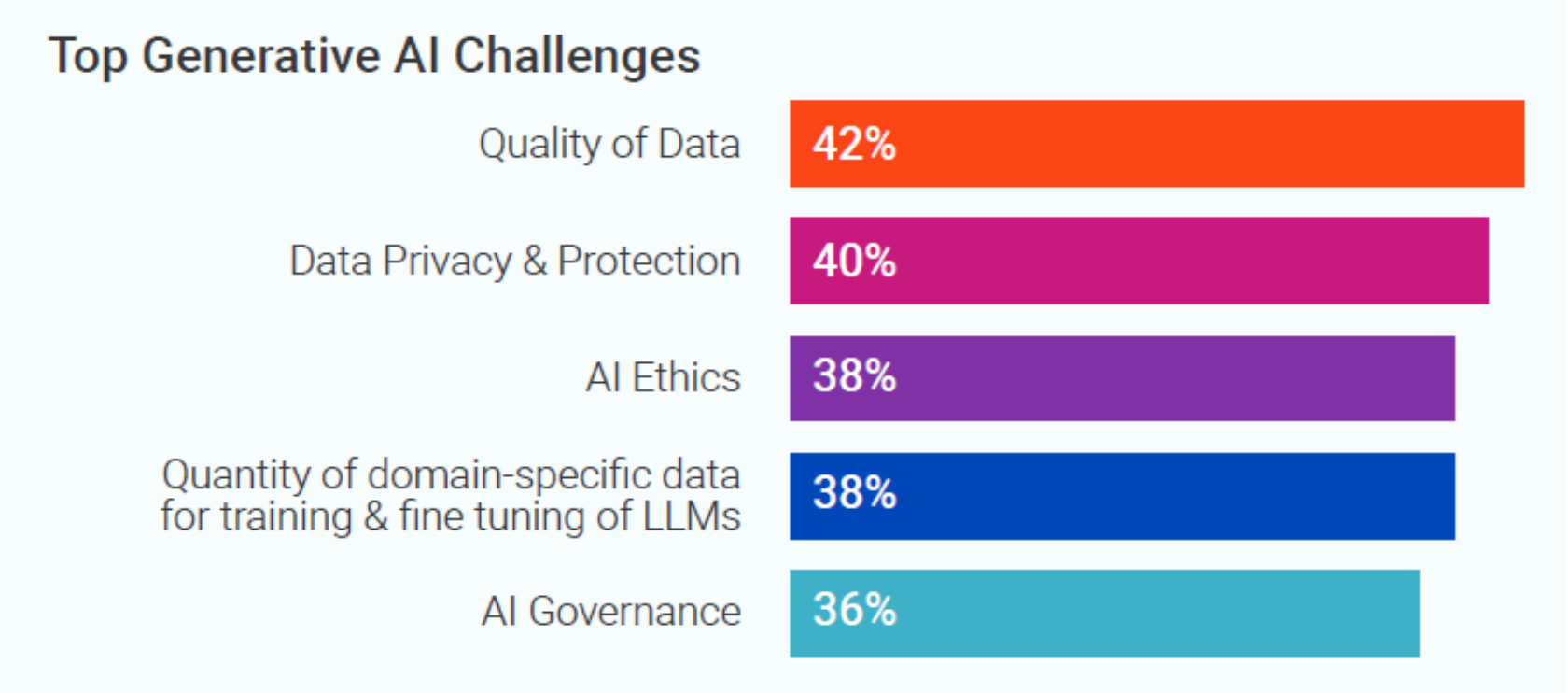
– *Lior Gavish, Monte Carlo CTO and co-founder*



Data Quality vs. GenAI

Informatica's CDO Insights 2024

"…Highly integrated data management capabilities are the key to unlock the vast potential of GenAI."



Data Intelligence as Opportunity

Data intelligence spans:

- Data discovery, governance, lineage, provenance, quality, security, and privacy

Data culture is critical too.

Data Catalogs Keep Popping Up

DataHub

Amundsen

Quest

data.world

collibra

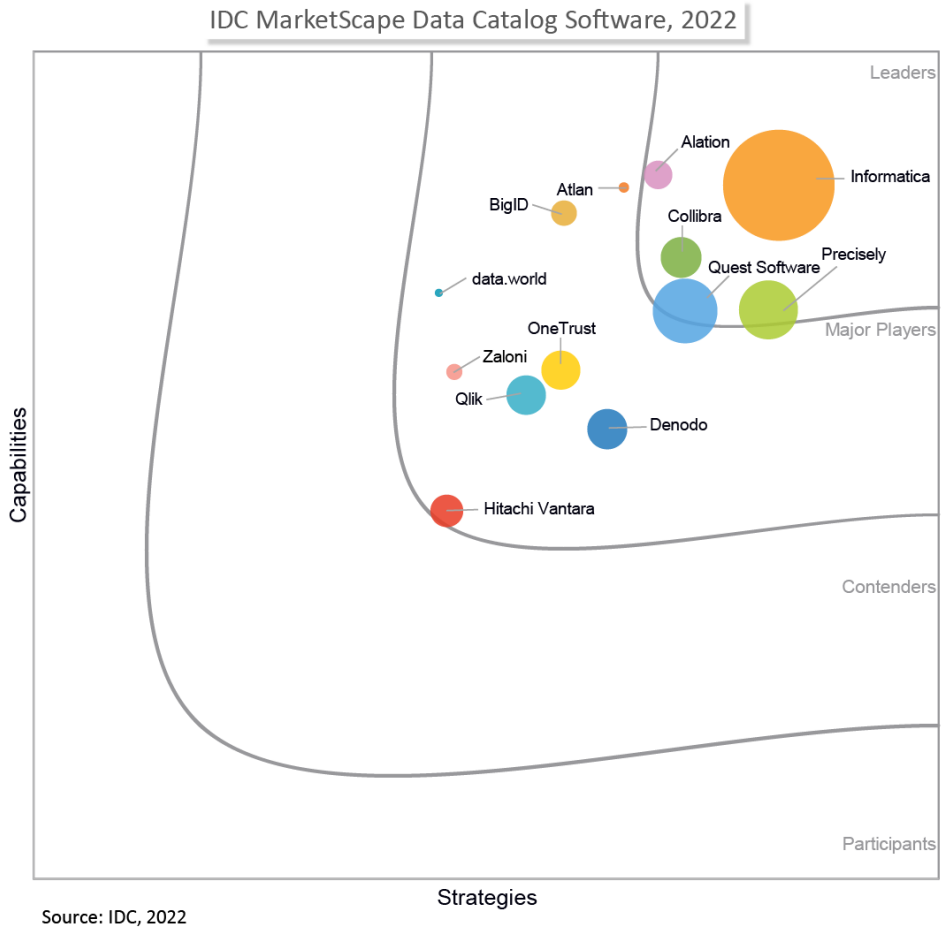
ataccama

Informatica

precisely

atlan

Alation



Source: IDC, 2022

Data Governance

BigID

collibra™

Alation

PRIVACERA

IMMUTA™

Informatica™

Data Transformation, ETL, ELT

MATILLION

IBM

dbt

Fivetran

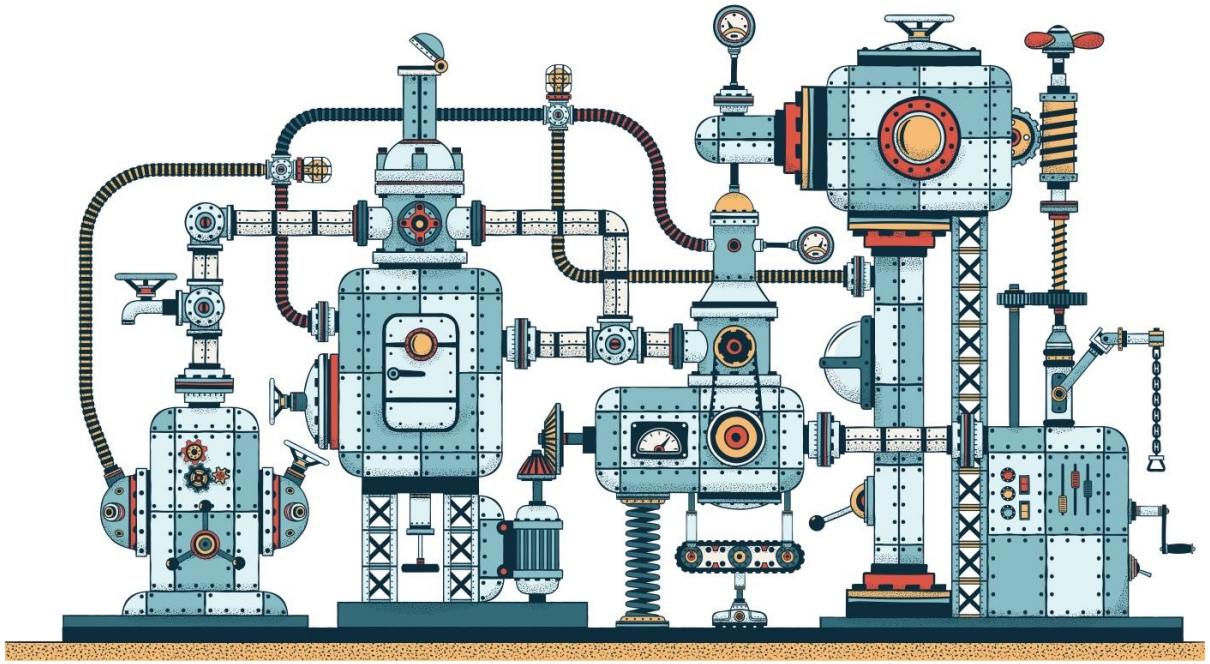
meltano

Informatica

Apache Airflow

talend

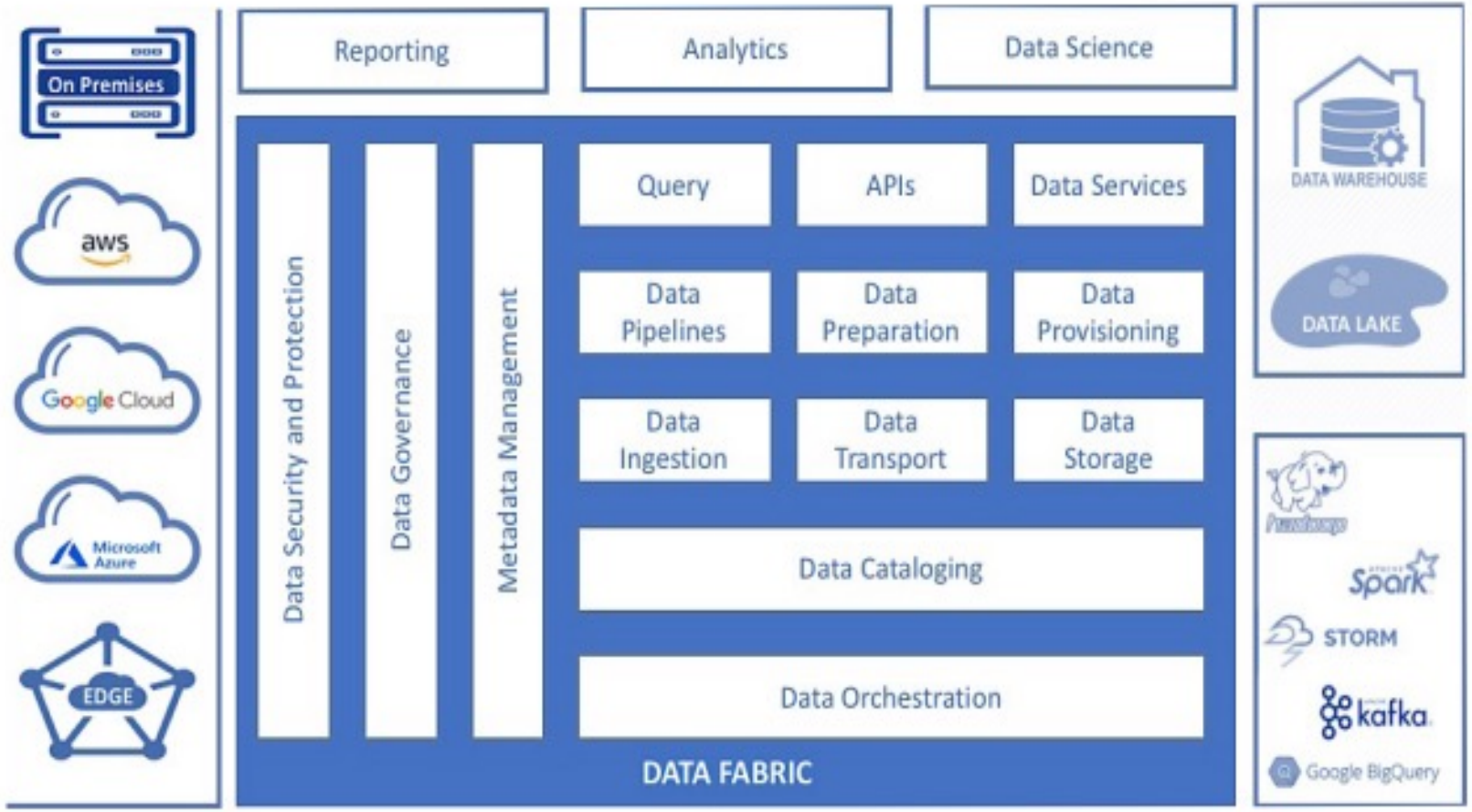
SINGER



Data Fabric: Centralized Data Management

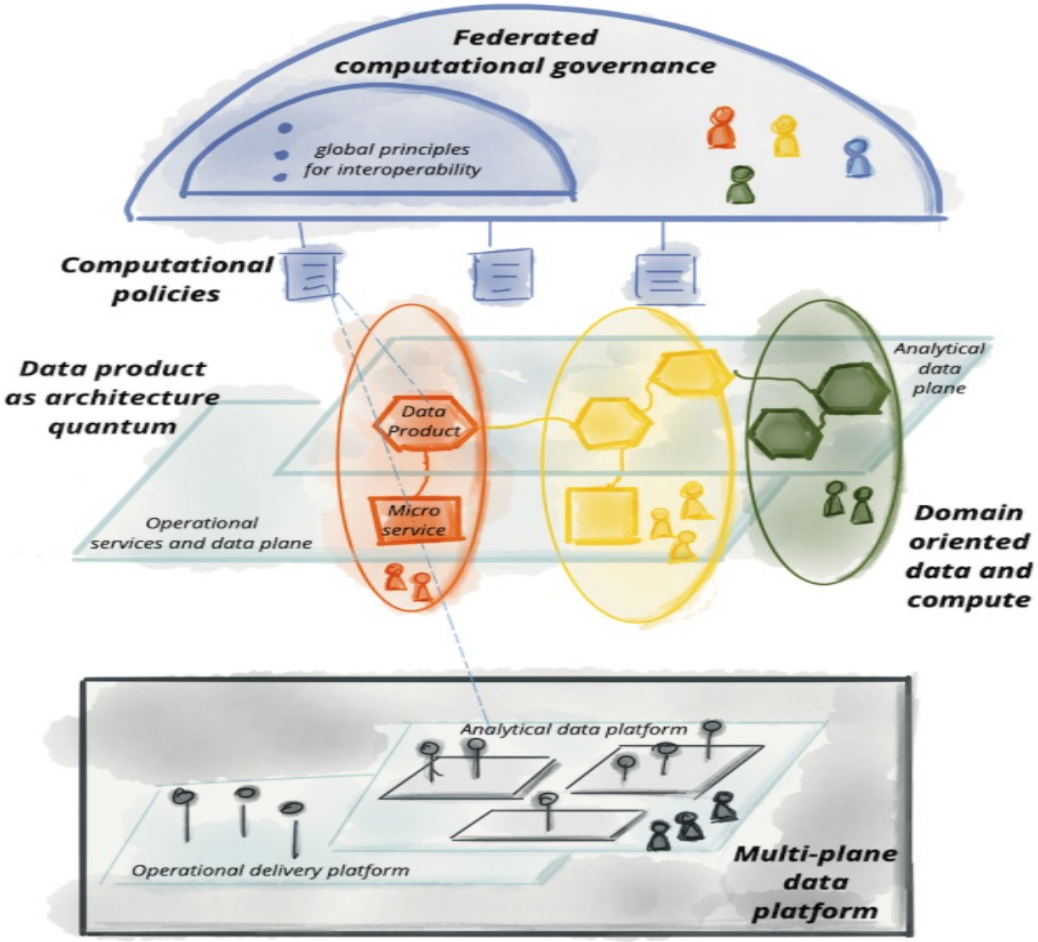
Provides metadata-driven centralization of:

- Data integration
- Data catalogs
- Data governance
- Data prep
- Data security

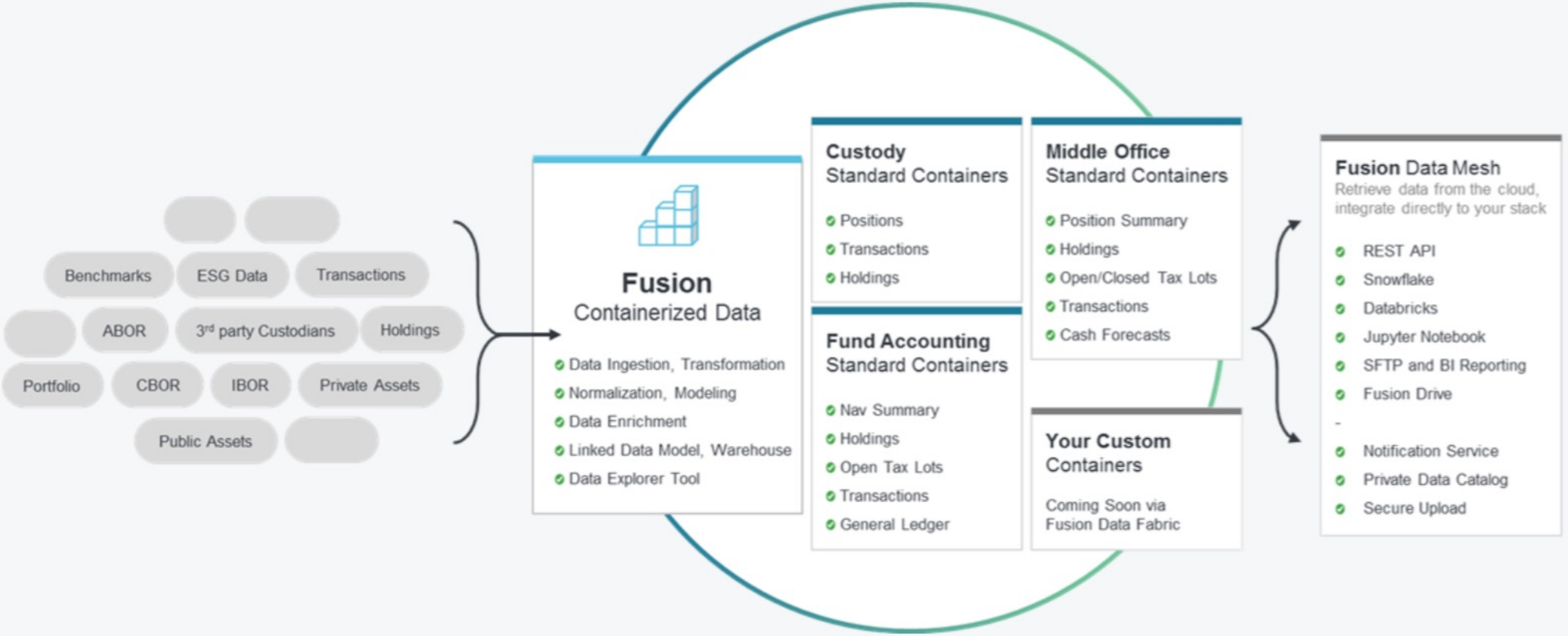


Data Mesh: A More Organic Approach

- Data mesh concept created by Zhamak Deghani
- Data left where it is
- Distributed data teams
- Treat data as a product
- Centralized governance



JPMorgan Containerized Services



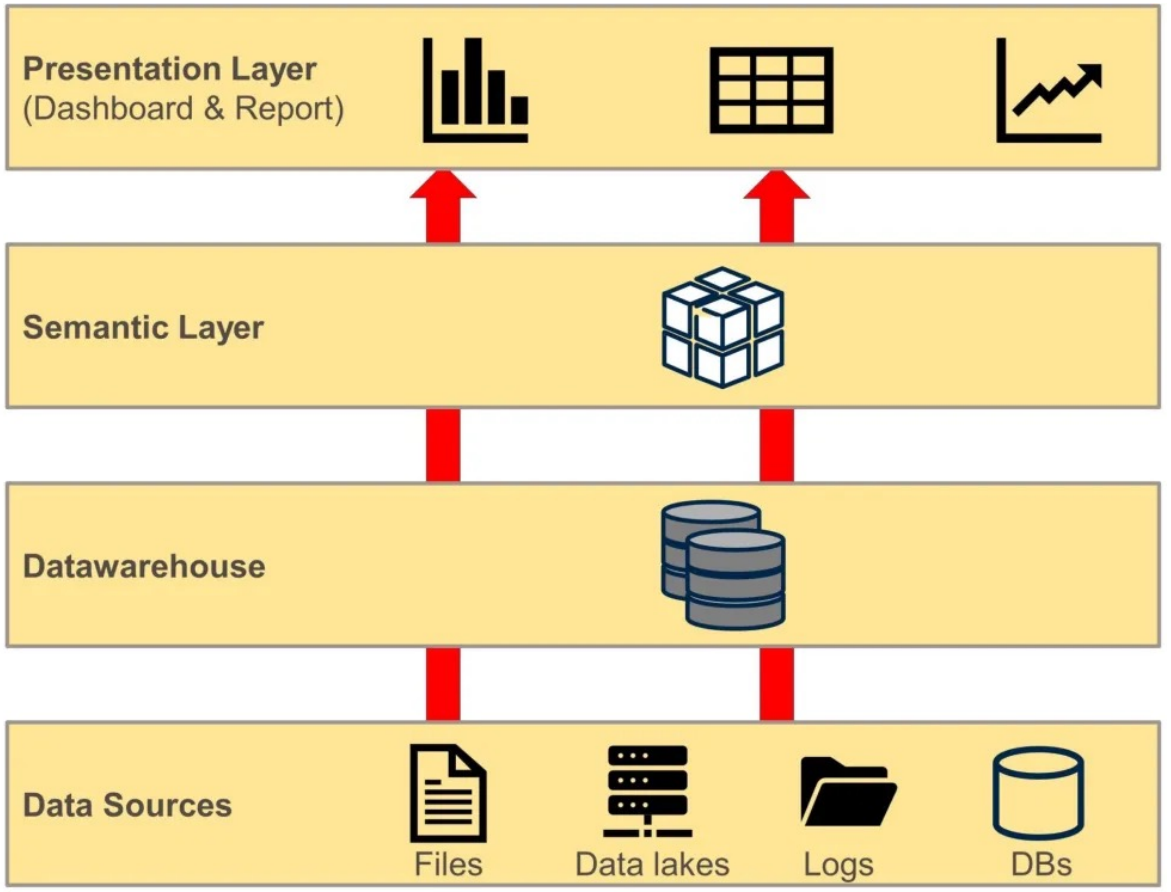
Rise of Independent Semantic Layers

- Maps your data model to business metrics and terms
- Ensures you get correct answers with SQL queries
- Now with GenAI for NLQ

dbt Labs

cube

ATSCALE



Rise of Lakehouse and Open Table Formats

- Lakehouse blends data lakes (based on HDFS or S3-based object storage) and data warehouses (based on relational database tech)
- Maintain scalability and flexibility of data lake
- Maintain data accuracy and transactionality of warehouse

Open Table Formats

Table formats deliver:

- ACID transactions
- Support for multiple query engines
- Time-travel functions
- Granular access control

Iceberg Wins — Polaris Emerges

Ali Ghodsi and Ryan Blue

Apache Polaris – metadata catalog

Convergence on Apache Iceberg

Databricks paid $1B to $2B for Tabular, the company behind Apache Iceberg, unifying the open table format community.

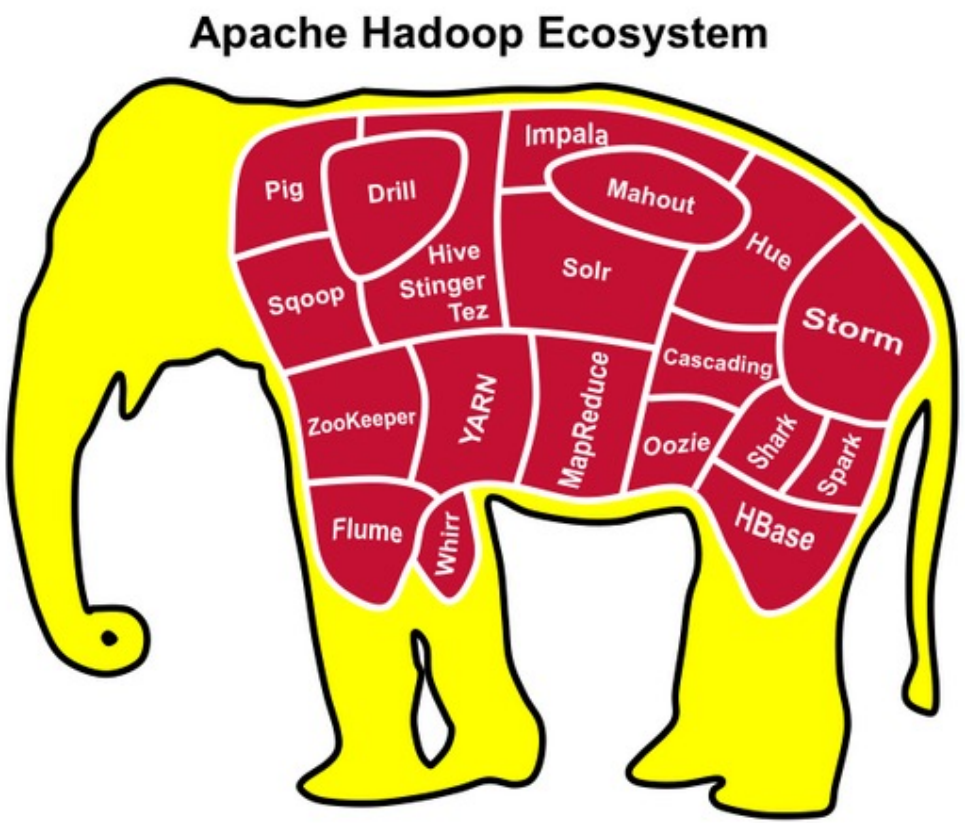
"If you're in the Iceberg community, this is **Go Time** in terms of entering the next era"

— *Read Maloney, CMO, Dremio*

Hadoop Redux?

Iceberg delivered the Hadoop Dream

Companies are free to run Spark, Presto, Trino, Dremio, Flink, and other Compute engines on data stored in Iceberg or Delta Lake.



Conclusion

Questions or comments?

Email at alex@datanami.com