The Realities of Building a Modern AI Data Stack

Colleen Tartow, Ph.D.
Field CTO & Head of Strategy
VAST Data

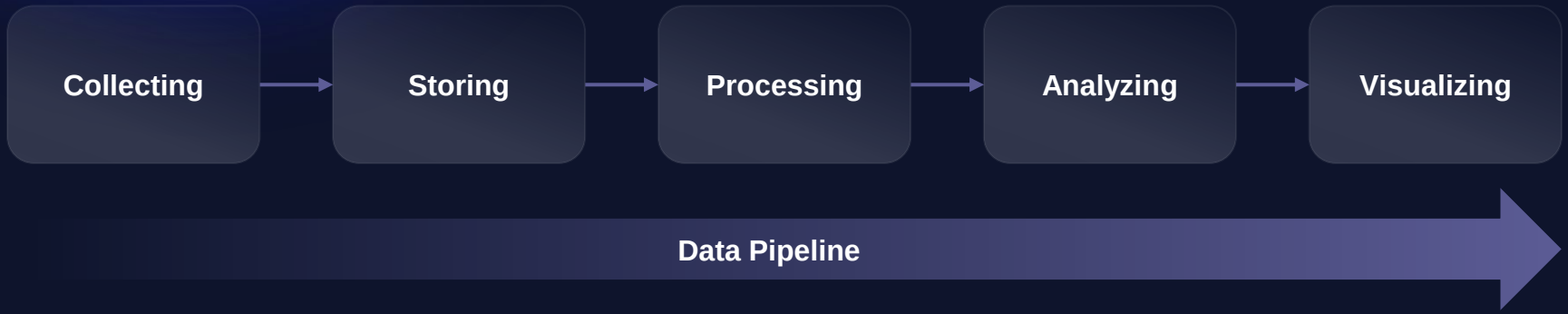
Modern AI Data Stack

You

Hi! Can you please tell me in one sentence what is a Modern Data Stack?

ChatGPT

A Modern Data Stack is a integrated ecosystem of tools and technologies designed to collect, store, process, analyze, and visualize data efficiently and at scale.



The Modern Data Stack

A small rant

Collecting

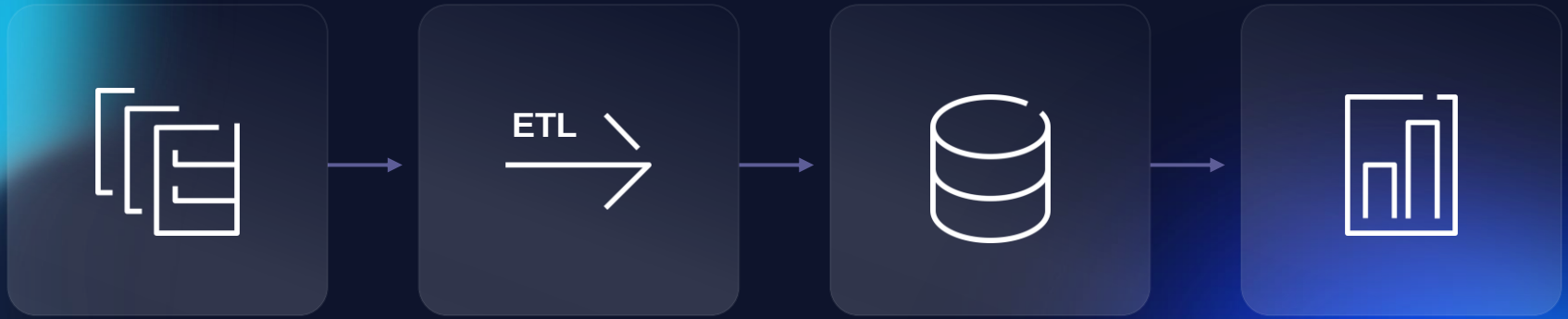
Curating

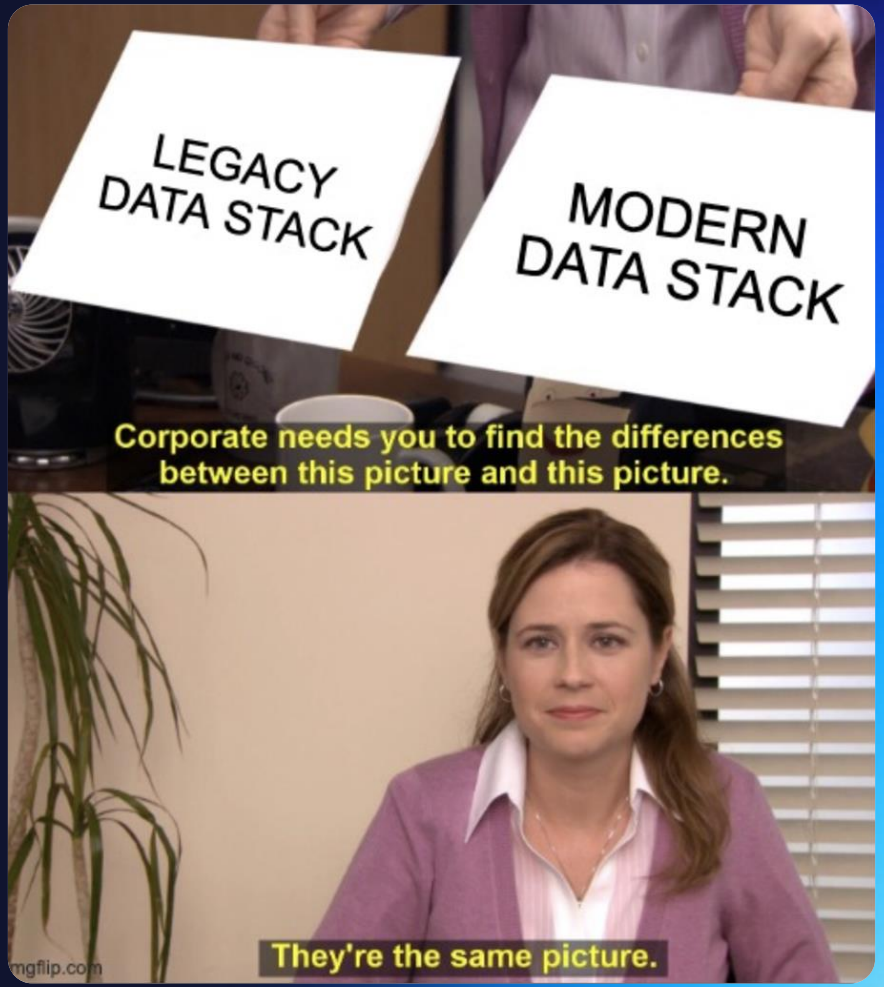
Cloud Data Warehouse

BI Tools

The Modern Data Stack

Looks familiar…




LEGACY DATA STACK
MODERN DATA STACK
Corporate needs you to find the differences between this picture and this picture.
They're the same picture.
imgflip.com

> “The Modern Data Stack is a cloud, SaaS, composable version of the Legacy Data Stack. A truly modern data stack should focus on the most efficient path of data from source to value.”

– Me, 2021

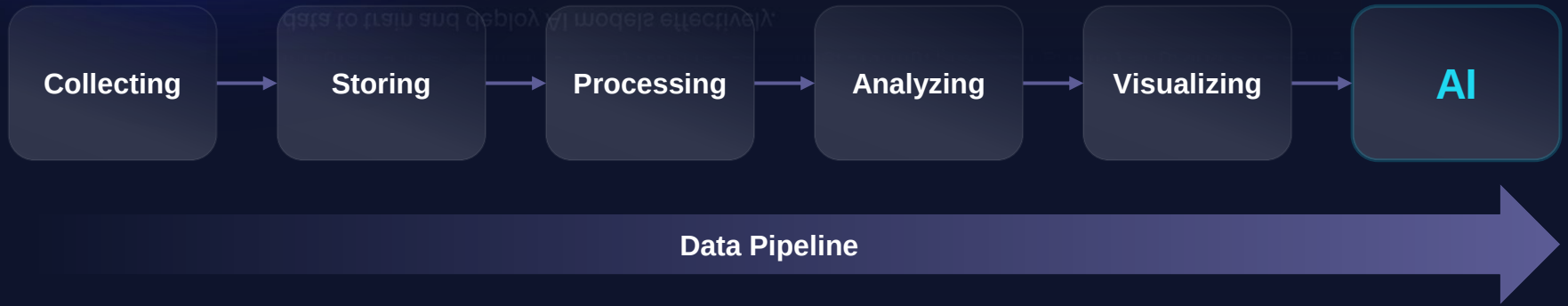
Now is the perfect time to rethink the path to value for data.

You

OK so what does a modern AI data stack look like?

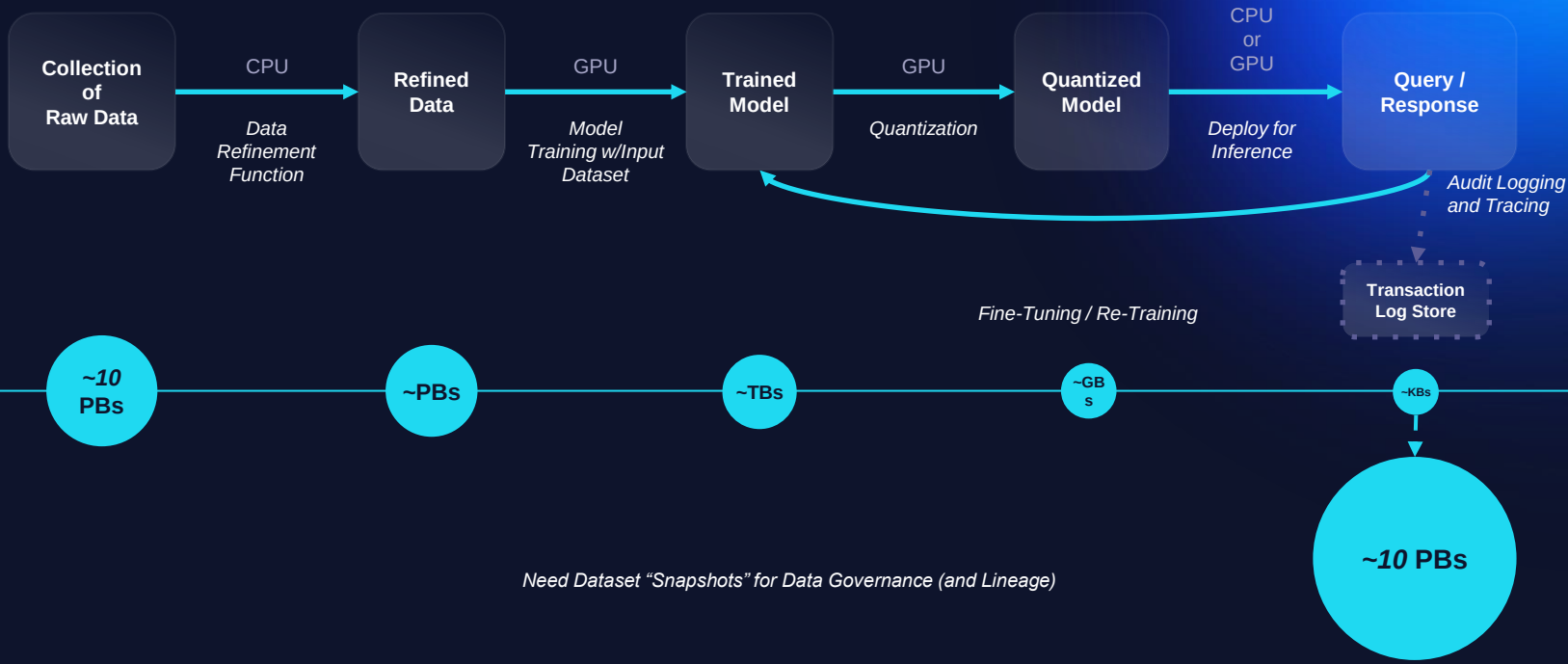
ChatGPT

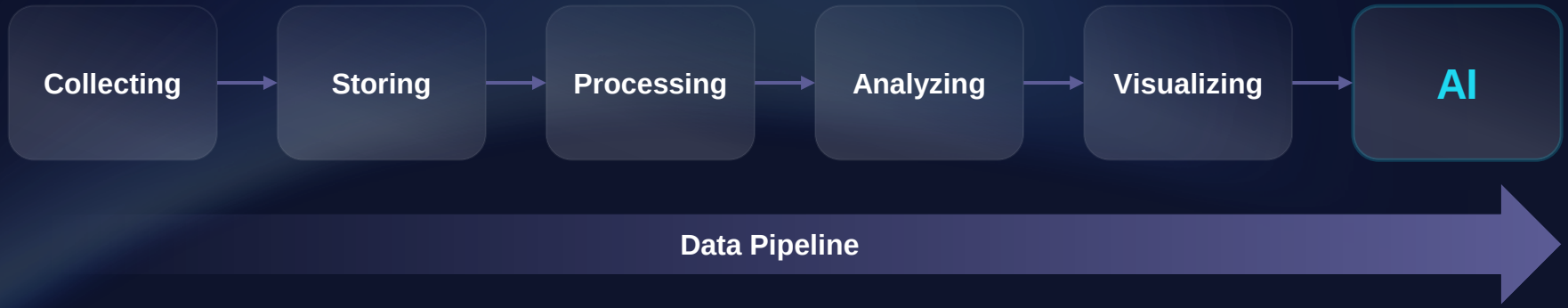
A modern AI data stack typically comprises advanced machine learning frameworks, scalable data processing platforms, robust data infrastructure, and comprehensive analytics tools integrated into a cohesive ecosystem for collecting, storing, processing, analyzing, and leveraging data to train and deploy AI models effectively.



Data Requirements for AI Pipelines End-to-End

Foundational Model Development Requires Copious Amounts of Data Beyond Model Training





Modern Data Stack

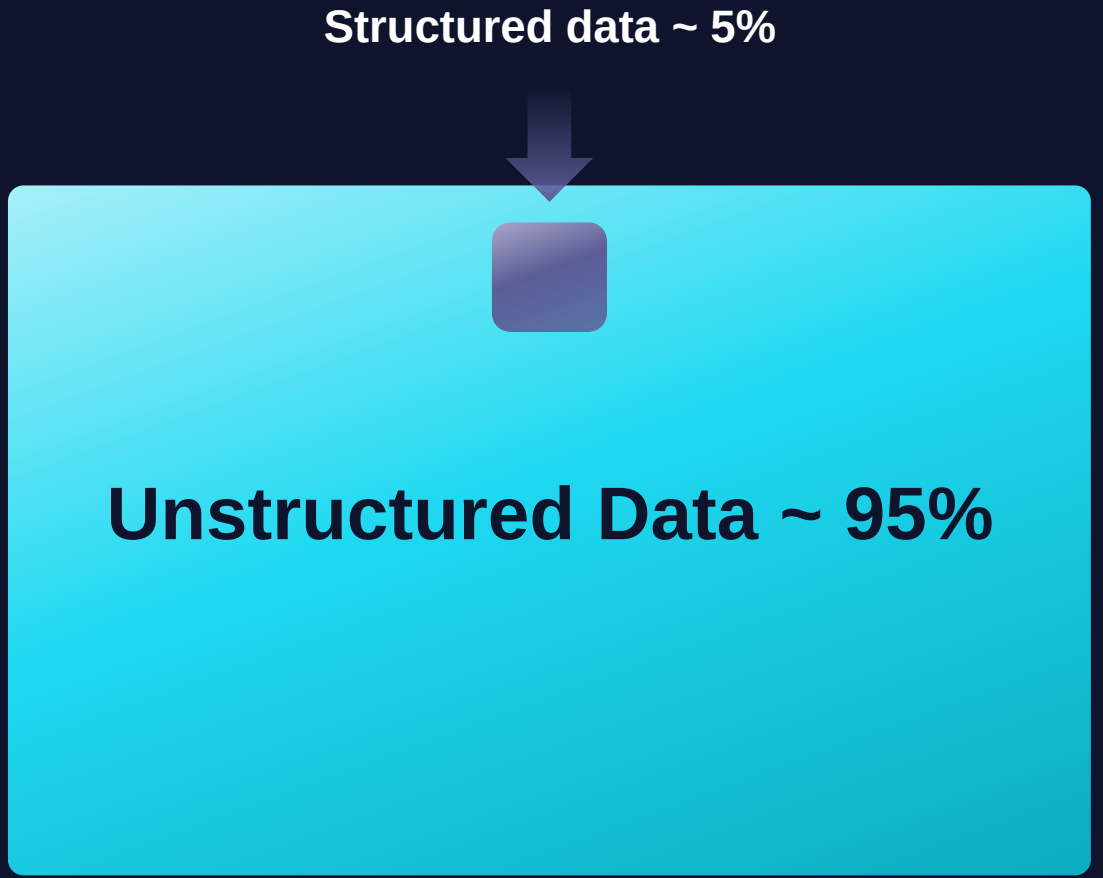
structured,
semi-structured data

Modern AI Data Stack

structured, semi-structured,
unstructured data

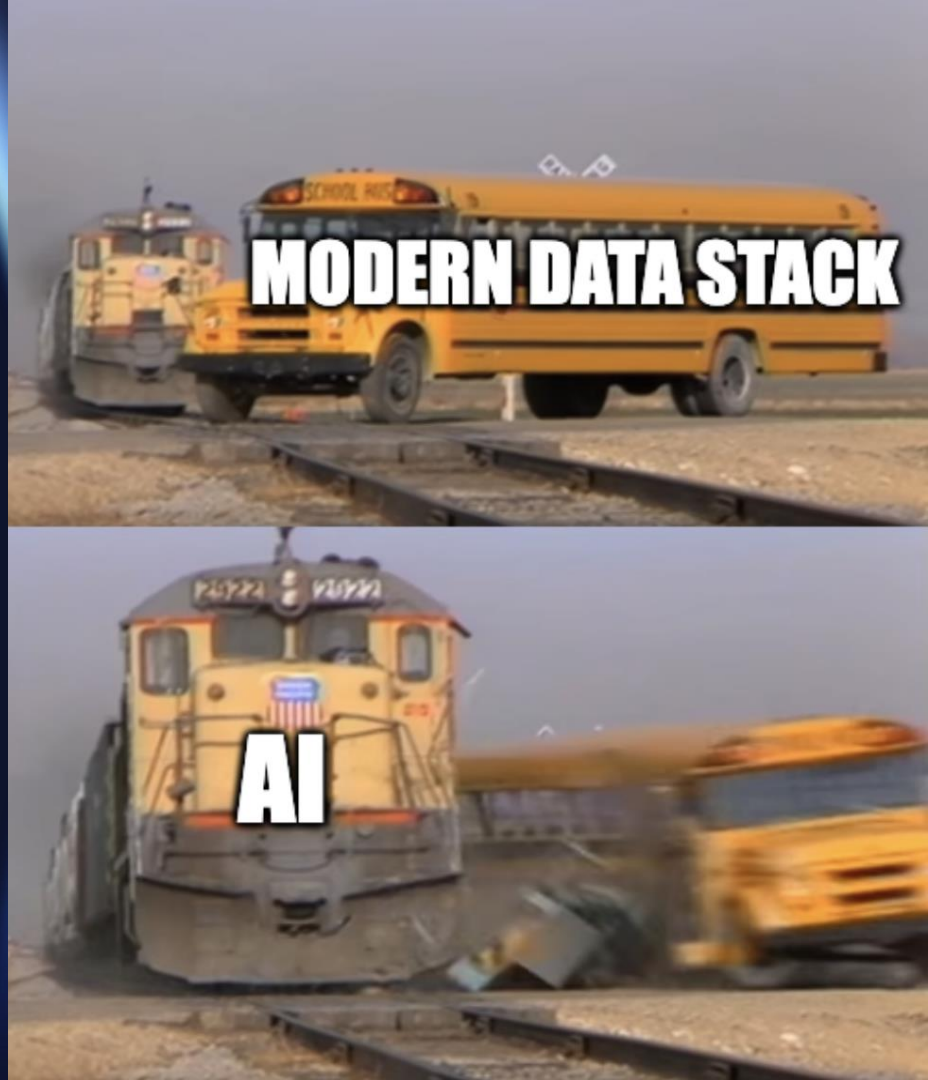
Structured vs Unstructured

Why is this important?



	BI	AI
Data type	Structured, semi-structured	Any
Data volumes	GB, TB	PB, EB
General complexity	Low-medium	High
Real-time processing	Near-real time for operations	Truly real-time ingestion, processing, output
Value	Largely understood	Nascent
Consumption pattern	Human, machine	Machine, human
Governance	Regulated, understood	Regulated*, feared?!
Infrastructure	Cloud, SaaS focus, elastic	Advanced compute and specialized hardware

- Pipeline complexity
- Tools optimized for BI
- Performance requirements
- Scalability
- Future proofing
- Cost



Things are getting really complicated

Data Maturity

Small data volume, complexity	Large data volume, complexity
Small data organization	Large data organization
Centralized data ownership	Domain-driven data ownership
Data as a by-product	Data as a product
Centralized governance	Federated governance model
Siloed data infrastructure	IT-driven self-service infrastructure

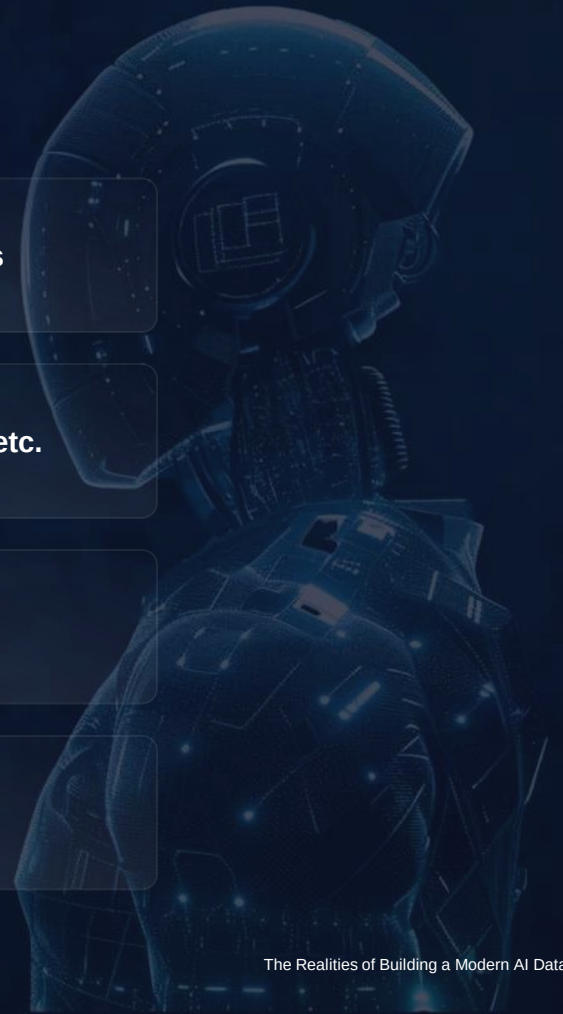
Data Maturity

Modern Data Stack

Data Mesh Data Fabric

Data Management and AI

- Well-curated and well-documented datasets and data models
- Governance including lineage, versioning, PII management, etc.
- Culture of data literacy and domain ownership
- Scalable and well-managed data infrastructure



Data Management

The Modern AI Data Stack

Scalability

Performance

Cost

Accessibility

Simplicity

Scalability

Performance

Cost

Accessibility

Simplicity

- Petabytes, Exabytes
- Scalable database solutions
- Accomodate large models
- Distributed / parallel
- Ease of upgrade and maintenance

Scalability

Performance

Cost

Accessibility

Simplicity

- Real-time ingest, processing, integration, curation, consumption
- Query response time at edge, cloud, core
- CPUs, GPUs, DPUs, etc.
- Checkpointing and recovery

Scalability

Performance

Cost

Accessibility

Simplicity

- Infrastructure costs
- Hybrid architectures
- Compute resource evolution
- Consumption model
- Organizational costs

Scalability

Performance

Cost

Accessibility

Simplicity

- Data security and privacy
- Risk management and regulatory compliance
- High availability / failover, BCDR
- Access control and governance
- Data catalog, discovery

Scalability

Performance

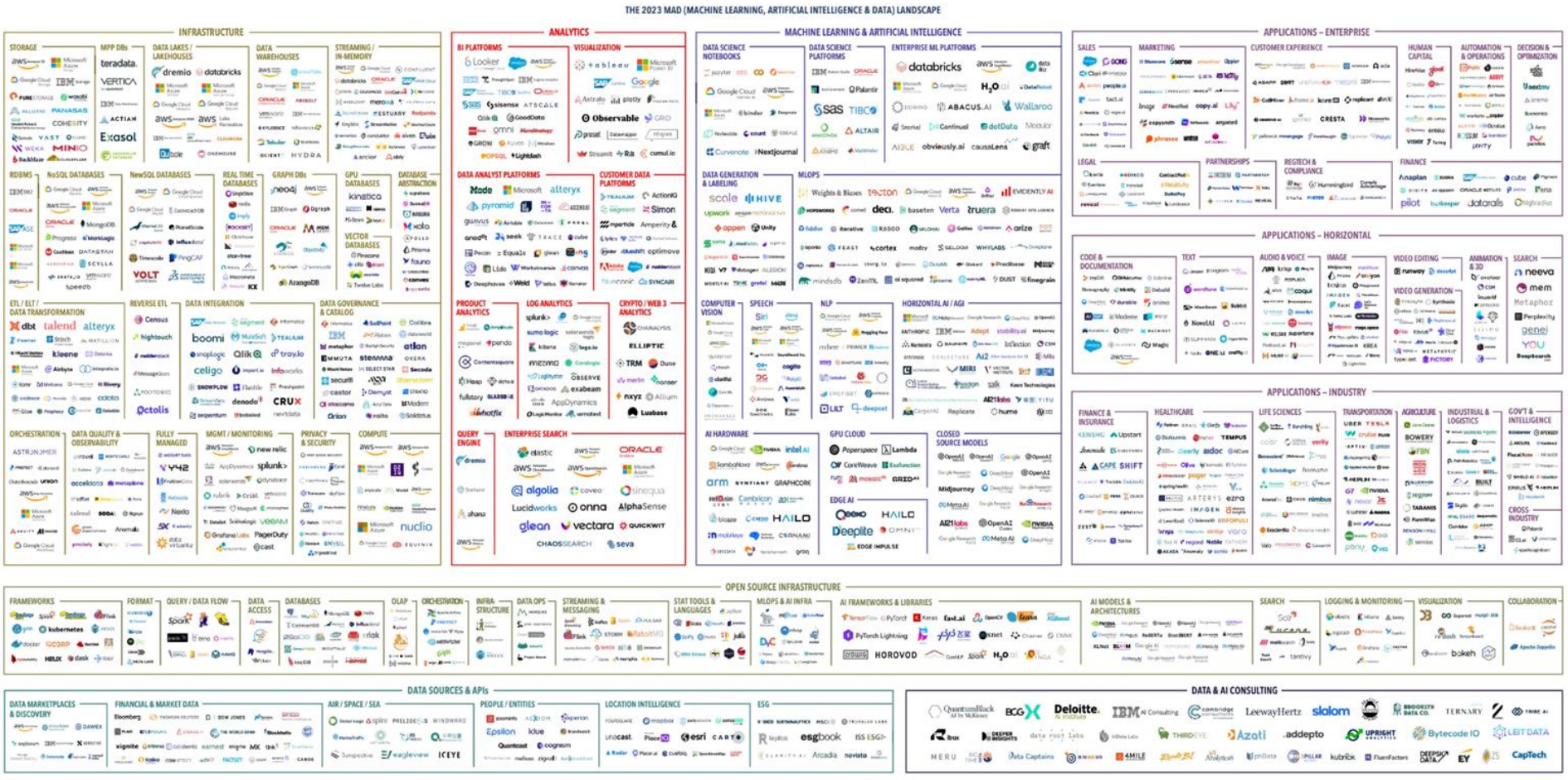
Cost

Accessibility

Simplicity

- Streamlined and well-architected data management
- Data products
- Replication
- Real time vs. batch
- Enables scalability, cost, performance, accessibility


WHAT DO WE WANT?
THE PERFECT FUTURE-PROOF DATA STACK!
WHEN DO WE WANT IT?
YESTERDAY!


THE 2023 MAD (MACHINE LEARNING, ARTIFICIAL INTELLIGENCE & DATA) LANDSCAPE
INFRASTRUCTURE
ANALYTICS
MACHINE LEARNING & ARTIFICIAL INTELLIGENCE
APPLICATIONS – ENTERPRISE
APPLICATIONS – HORIZONTAL
APPLICATIONS – INDUSTRY
OPEN SOURCE INFRASTRUCTURE
DATA SOURCES & APIs
DATA & AI CONSULTING
Version 1.0 - Feb 2023
© Matt Turck (@mattturck), Kevin Zhang (@ykevinzhang) & FirstMark (@firstmarkcap)
Blog post: mattturck.com/MAD2023
Interactive version: MAD.firstmarkcap.com
Comments? Email MAD2023@firstmarkcap.com
FIRSTMARK
EARLY STAGE VENTURE CAPITAL

1. Focus on data's **path to value**
2. Consider **hybrid** architectures
3. Use **data products**
4. Start **small** and keep it simple

Thank you!