



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?

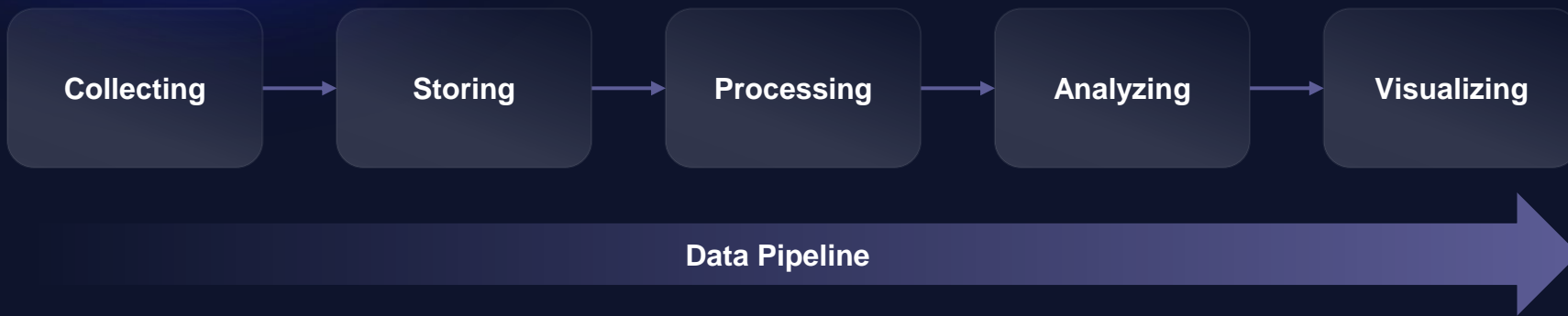


ChatGPT

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.

store, process, analyze, and visualize data efficiently and at scale.

A modern data stack is a integrated ecosystem of tools and technologies designed to collect,



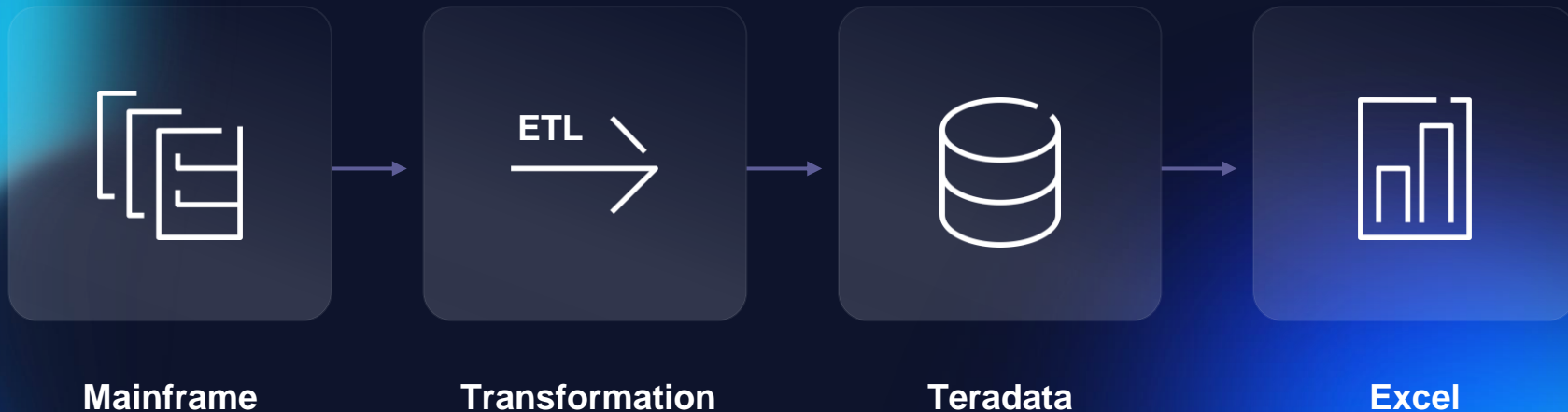
The Modern Data Stack

A small rant



The Modern Data Stack

Looks familiar...







“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

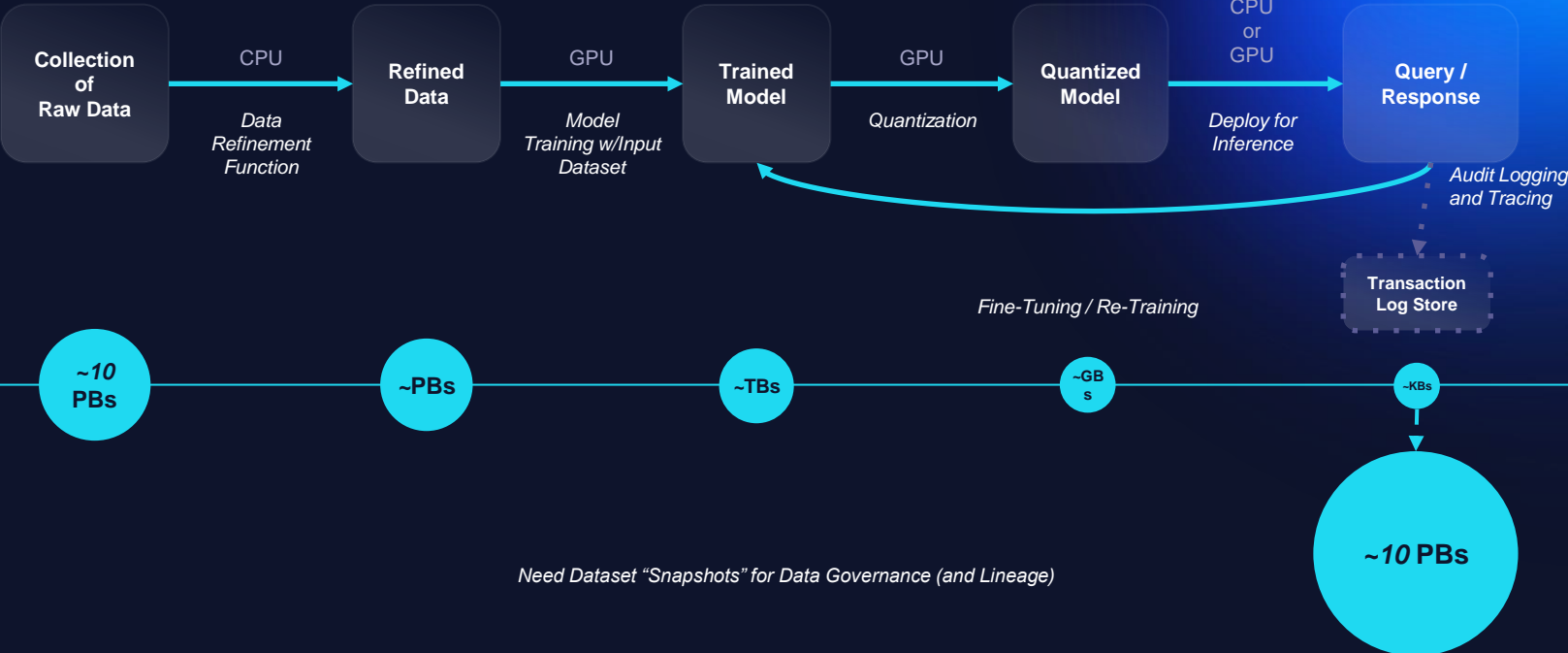
A modern AI data stack typically comprises advanced machine learning frameworks, scalable data processing platforms, robust data infrastructure, and comprehensive analytics tools integrated into a cohesive ecosystem for collecting, storing, processing, analyzing, and leveraging data to train and deploy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

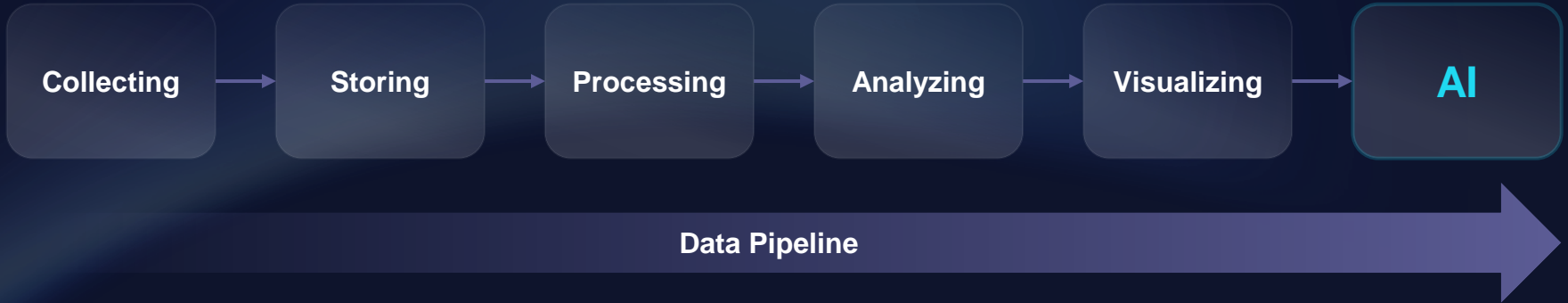


Data Pipeline

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?

Structured data ~ 5%



Unstructured Data ~ 95%

	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

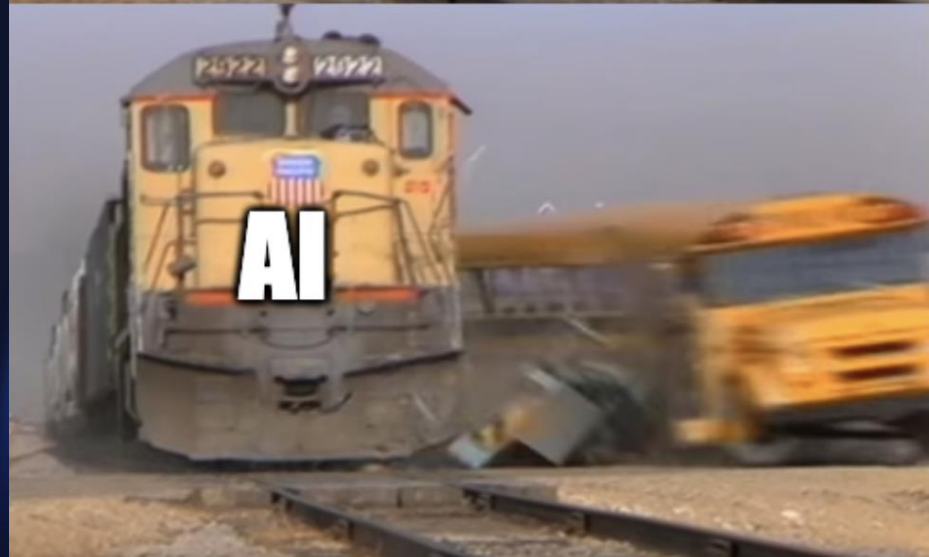
Tools optimized for BI

Performance requirements

Scalability

Future proofing

Cost



Things are getting really complicated

Data Maturity

Small data volume, complexity

Small data organization

Centralized data ownership

Data as a by-product

Centralized governance

Siloed data infrastructure

Large data volume, complexity

Large data organization

Domain-driven data ownership

Data as a product

Federated governance model

IT-driven self-service infrastructure



Data Maturity

Data Maturity

**Modern
Data Stack**

**Data Mesh
Data Fabric**



Data Maturity

Data Management and AI



Well-curated and well-documented datasets and data models



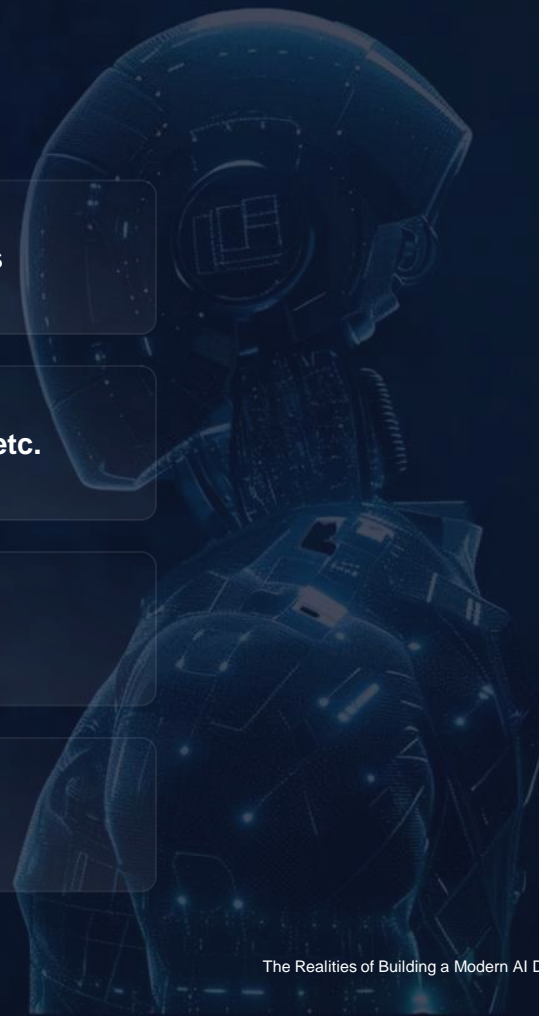
Governance including lineage, versioning, PII management, etc.



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

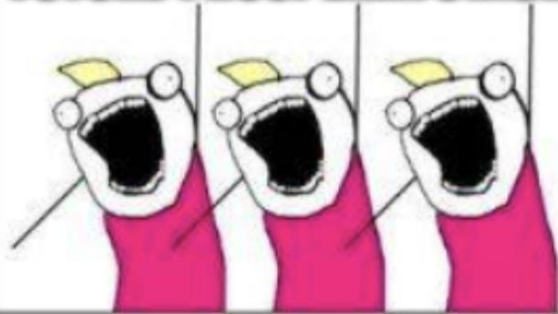
Simplicity

- 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

The top section of the infographic is divided into six main categories, each containing a grid of company logos:

- INFRASTRUCTURE:** Includes STORAGE (AWS, Microsoft, Oracle), MPP DBs (Teradata, Vertica, Exasol), DATA LAKES / LAKEHOUSES (Dremio, Databricks), DATA WAREHOUSES (Snowflake, Amazon Redshift, Microsoft Azure Synapse), STREAMING / IN MEMORY (Kafka, Flink, Apache Druid), NOSQL DATABASES (Cassandra, MongoDB, Redis), REAL TIME DATABASES (InfluxDB, TimescaleDB), GRAPH DBs (Neo4j, Amazon Neptune), GPU DATABASES (Kinetica), DATABASE ABSTRACTION (CockroachDB, Yugabyte), VECTOR DATABASES (Pinecone, Weaviate), and VOLT (Amazon Athena, Snowflake).
- ANALYTICS:** Includes BI PLATFORMS (Looker, Tableau, Power BI), VISUALIZATION (Tableau, Power BI, Qlik), DATA SCIENCE NOTEBOOKS (Databricks, JupyterLab), ENTERPRISE ML PLATFORMS (Databricks, SAS, Tibco), DATA ANALYST PLATFORMS (Alteryx, Alteryx), and CUSTOMER DATA PLATFORMS (Salesforce, Oracle).
- MACHINE LEARNING & ARTIFICIAL INTELLIGENCE:** Includes DATA SCIENCE PLATFORMS (Databricks, SAS, Tibco), ENTERPRISE ML PLATFORMS (Databricks, SAS, Tibco), DATA GENERATION & LABELING (Scale AI, Hive), MLOPS (Weights & Biases, DVC), NLP (OpenAI, Google), HORIZONTAL AI / AGI (Anthropic, OpenAI), AI HARDWARE (NVIDIA, Intel), GPU CLOUD (Paperport, Lambda), and CLOSED SOURCE MODELS (OpenAI, Anthropic).
- APPLICATIONS - ENTERPRISE:** Includes SALES (Salesforce), MARKETING (HubSpot, Marketo), CUSTOMER EXPERIENCE (Salesforce, Adobe), HUMAN CAPITAL (Workday), AUTOMATION & OPERATIONS (UiPath, Automation Anywhere), and DECISION & OPTIMIZATION (Palantir, Alteryx).
- APPLICATIONS - HORIZONTAL:** Includes CODE & DOCUMENTATION (GitHub), TEXT (OpenAI), AUDIO & VOICE (OpenAI), IMAGE (OpenAI), VIDEO EDITING (Runway), ANIMATION & 3D (Runway), and SEARCH (Elasticsearch, Algolia).
- APPLICATIONS - INDUSTRY:** Includes FINANCE & INSURANCE (Capital One), HEALTHCARE (Tempus), LIFE SCIENCES (Tempus), TRANSPORTATION (Uber), AGRICULTURE (Bionity), INDUSTRIAL & LOGISTICS (Bentley), and GOVT & INTELLIGENCE (Palantir).

This row contains logos for various open-source and specialized tools:

- OPEN SOURCE INFRASTRUCTURE:** Includes frameworks like PyTorch, TensorFlow, and libraries like HuggingFace.
- AI FRAMEWORKS & LIBRARIES:** Includes PyTorch, TensorFlow, and others.
- AI MODELS & ARCHITECTURES:** Includes GPT, BERT, and other AI models.
- SEARCH:** Includes Elasticsearch, Algolia, and others.
- LOGGING & MONITORING:** Includes Prometheus, Grafana, and others.
- VISUALIZATION:** Includes Tableau, Power BI, and others.
- COLLABORATION:** Includes Slack, Microsoft Teams, and others.

This row contains logos for data sources, marketplaces, and consulting firms:

- DATA SOURCES & APIs:** Includes various data providers and API services.
- DATA MARKETPLACES & DISCOVERY:** Includes platforms like DataCamp and others.
- FINANCIAL & MARKET DATA:** Includes Bloomberg, Reuters, and others.
- AIR / SPACE / SEA:** Includes various aerospace and maritime data providers.
- PEOPLE / ENTITIES:** Includes various people and entity data providers.
- LOCATION INTELLIGENCE:** Includes various location-based data providers.
- ESG:** Includes various ESG data providers.
- DATA & AI CONSULTING:** Includes Deloitte, IBM, Accenture, and others.

Version 1.0 - Feb 2023 | © Matt Turck (@matturck), Kevin Zhang (@kevinzhang) & FirstMark (©firstmarkcap) | Blog post: matturck.com/MAD2023 | Interactive version: MAD.firstmarkcap.com | Comments? Email MAD2023@firstmarkcap.com



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!

