How to scale AI & BI on your data architecture

ATSCALE

Today's Speakers

Kirk Borne, Ph.D.

Chief Science Officer
DataPrime, Inc.

Kirk has been an influential globally recognized leader in the data science space for 20 years. His areas of passion and focus include Big Data & Data Science, Artificial Intelligence (AI), and Astrophysics.

Kirk is also the co-creator of the field of Astroinformatics.

Bill Inmon

Father of the Data Warehouse

Bill Inmon – the "father of data warehouse" – has written 60 books published in nine languages. Bill's latest adventure is the building of technology known as textual disambiguation (textual ETL) – technology that reads the raw text in a narrative format and allows the text to be placed in a conventional database so that it can be analyzed by standard analytical technology, thereby creating unique business value for Big Data/unstructured data.

Bill was named by ComputerWorld as one of the ten most influential people in the history of the computer profession.

Soham Bhatt

EDW Modernization Practice Lead, Databricks

Soham Bhatt is a Senior Solutions Architect leading the EDW and ETL modernization practice at Databricks. Before Databricks he worked at Toyota Motors on building their next-generation Big Data Platform.

Prior to that, his background was in building Enterprise Data Warehouses and ETL architectures for Fortune 100 companies with Inmon and Kimball methodologies. In his current role, he loves guiding his customers with best practices as they modernize their EDWs to Data Lakehouses.

Andrew Sohn

VP of Enterprise Data & Data Products Delivery, Inspire Brands

Andrew is the VP of Enterprise Data and Data Products Delivery at Inspired Brands, the second-largest Restaurant Company in the US.

As an Executive-level Data and Analytics leader, Andrew is responsible for delivering tangible business outcomes by leveraging data-related technologies and processes throughout the entire data supply chain – data acquisition, governance, processing, integration, security and compliance, distribution, analytics, and reporting, and external monetization.

Data & Analytics Maturity Model

Capabilities	Level 0 Initial	Level 1 Centralized	Level 2 Proactive	Level 3 Leading
Infrastructure	Data Marts	Data Warehouse	Cloud Data Platform	Data Enrichment
Access	Data Extracts	Data Pipelines	Direct Query	Data Sharing
Modeling	Dataset	Physical Tables	Logical Views	Dimensional
Consumption	SQL	Dashboards	Ad Hoc	Embedded
Insights	Descriptive	Diagnostic	Predictive	Prescriptive

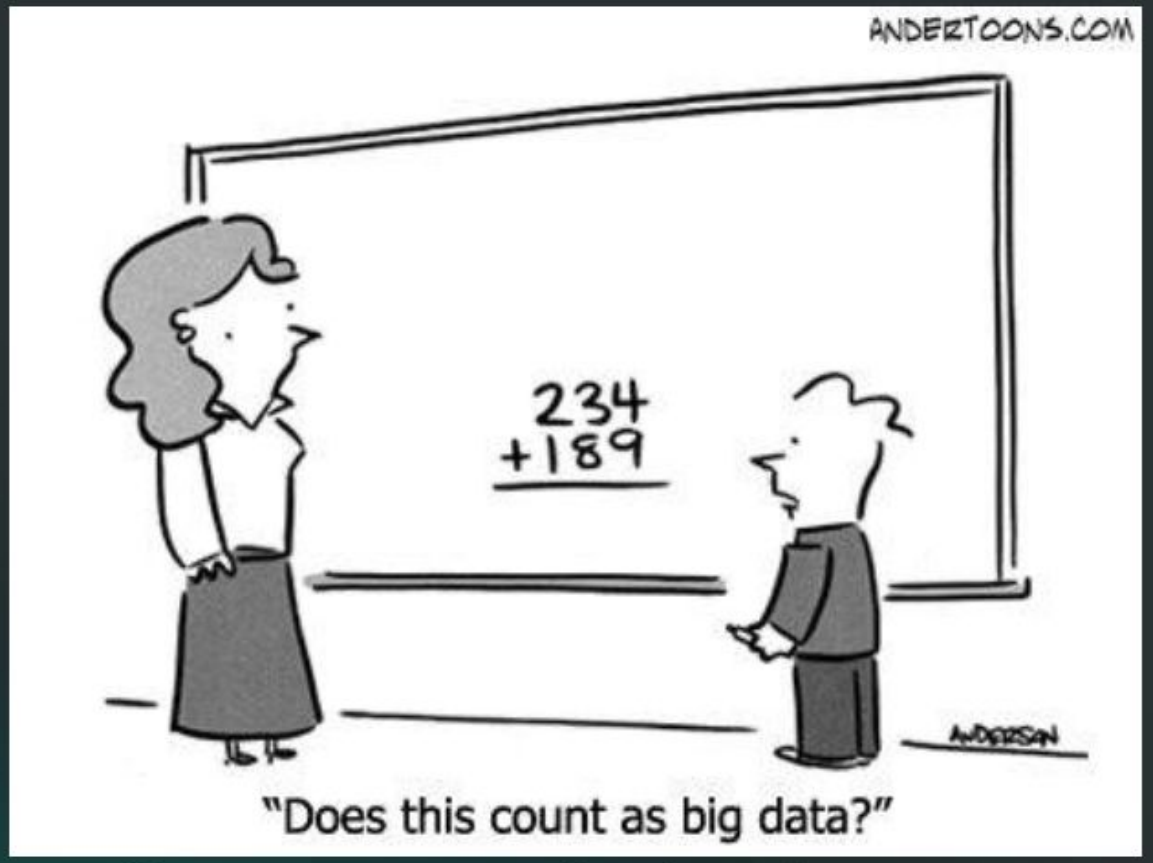
Smarter Decisions

Data Innovation, Empowerment, and Business Value from a Semantically Integrated Data Lake

Presented by Kirk Borne
Chief Science Officer
DataPrime Inc.
@KirkDBorne

What is One of the Biggest Challenges of Big Data, BI, and Data Analytics?

Hint: it is not Volume.



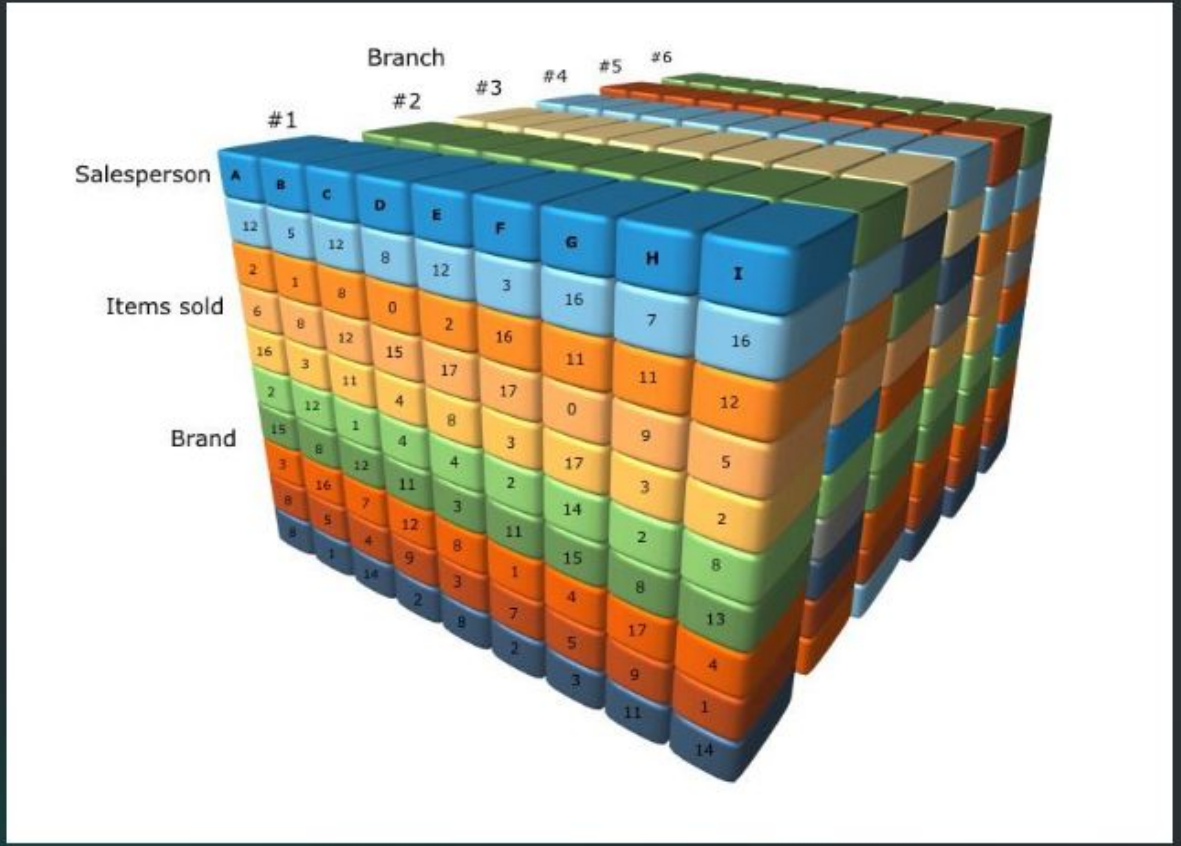
https://cdn.andertoons.com/img/toons/cartoon6517t.png

What is One of the Biggest Challenges of Big Data, BI, and Data Analytics?

Hint: it is not Volume.

Answer: it's the Variety = Data Complexity!

Remember… Variety is the Spice of Discovery!



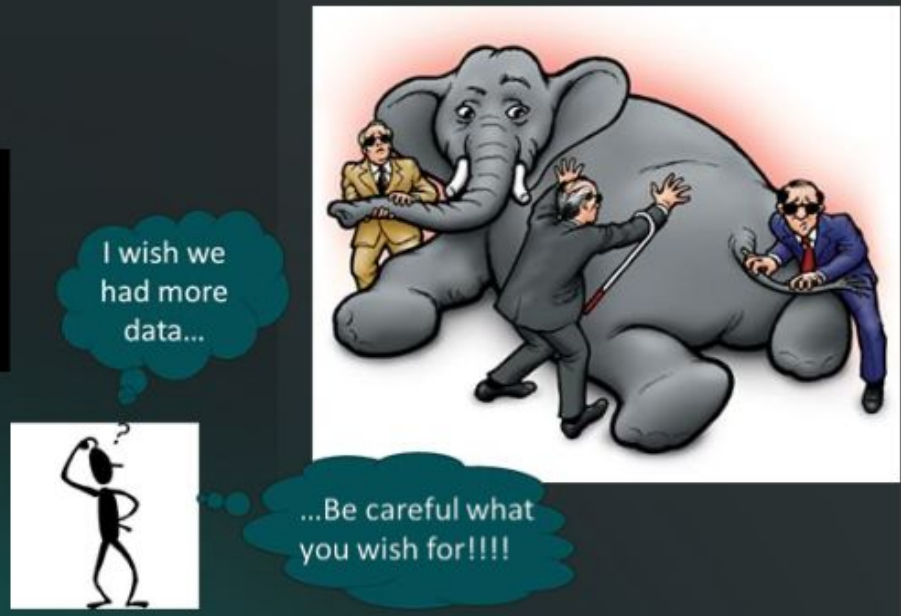
Source: https://medium.com/@nayanika2017/datacubes-do-we-mean-cubes-b6c584fc2cb0

Volume is not a problem. Storage is Manageable. But…

- Analytics and discovery on diverse, distributed data sources is hard!
- Every organization collects many different (complex) sources of data.
- These multiple diverse data sets are often stored in separate silos.
- Silos inhibit data teams from integrating multiple data sets that (when combined) could yield deep, actionable insights to create business value.

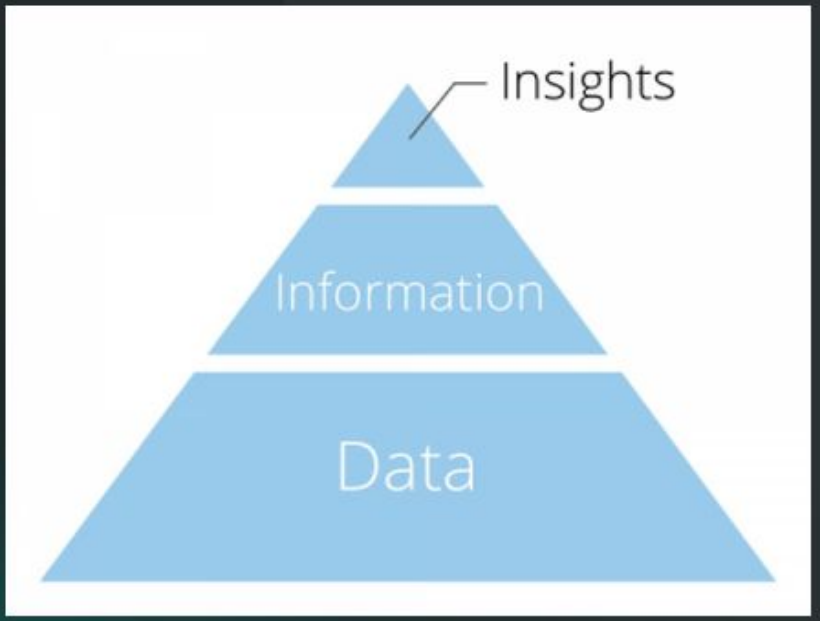
Consider the blind-folded men and the elephant …

"What we have here is a lot of information, but very few insights."



The Power of The Semantic Layer

- With a Semantic-powered Data Lake & Data Warehouse, businesses have the power to integrate those multiple diverse data sources and to change that entire story!
- Access, exploration, and exploitation of rich diverse data sources leads to:
  - ✓ Empowerment of Data Scientists, Data Engineers, Data Analysts, Everyone
  - ✓ Cultural Transformation: Data Literacy, Data Democratization, Experimentation
  - ✓ Insights Discovery
  - ✓ Innovation (new products, processes, services)
  - ✓ Business Value Creation



Data Innovation, Empowerment, and Business Value with the STELLAR Analytics Scorecard for BI and Analytics Mastery

1) **Streaming Data Analytics:** Real-time access to, interaction with, and discovery from data…
   - Detects POI (Person, Pattern, Product, Process, or Point Of Interest)
   - Detects BOI (Behavior Of Interest from any "dynamic actor")

Detect Emerging Classes (Segments), Correlations, Surprises (Anomalies), and Links (Associations) in incoming data.

2) **Team Analytics:** Culture of experimentation, diversification, collaboration, data-sharing, data re-use, and data democratization (Data is Everyone's Job!)

3) **Edge Analytics:** Locality in Time, at the moment of data collection (What else is happening now?)

4) **Location Analytics:** Locality in Geospace, in that context (What else is happening at that place?)

7) **Learning Business System Analytics:** Data-driven knowledge-generation business processes, with continuous feedback, learning, and improvement – embedded in daily business practice (REF)

6) **Agile Analytics:** Analytics By Design, Iterative, Build Proofs of Value, Fail-fast to Learn Fast, MVP (and MLP = minimum lovable product), CI / CD (Continuous Integration / Delivery)

5) **Related-entity Analytics:** Locality in Data Feature Space (What else is like this entity / event?)

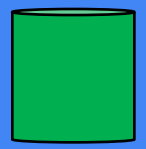
(REF: https://en.wikipedia.org/wiki/Learning_health_systems)

The evolution of data architecture

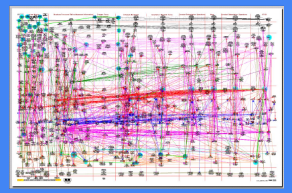
machine generated (data lake)

Text (DW 2.0)

Data Warehouse (integrated data)

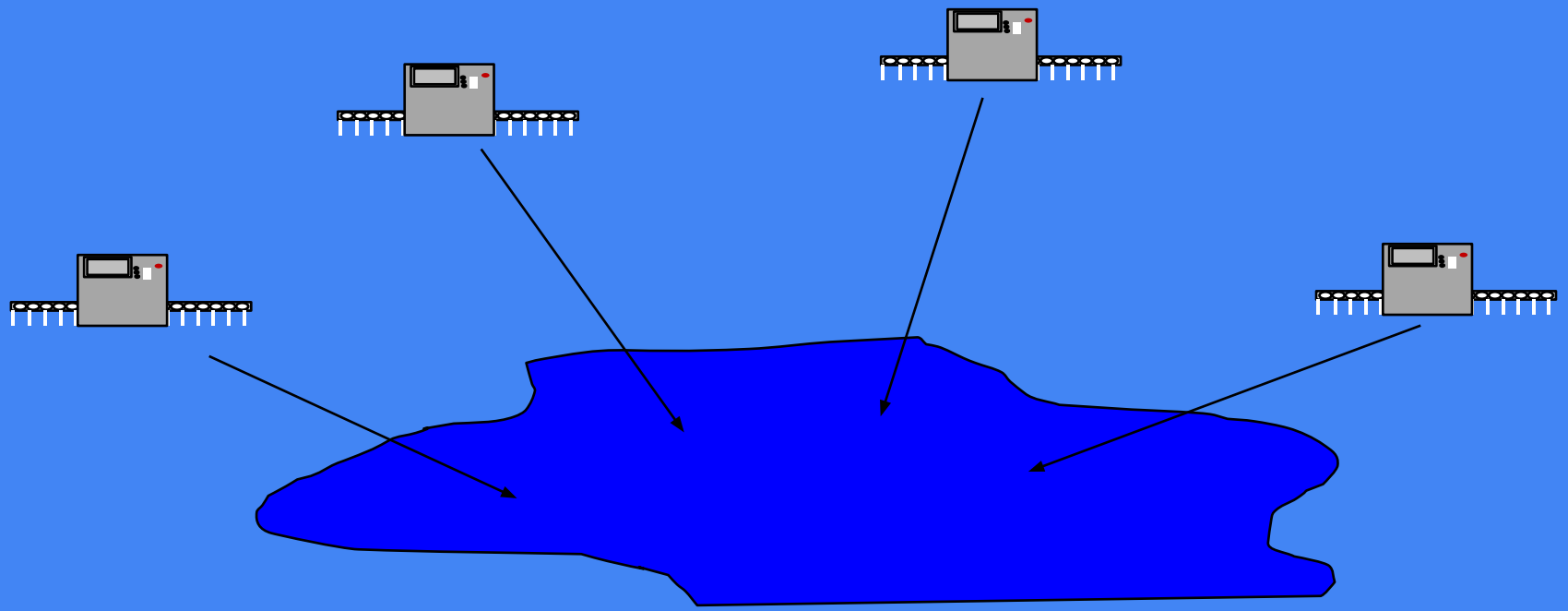


Applications (spider web systems)

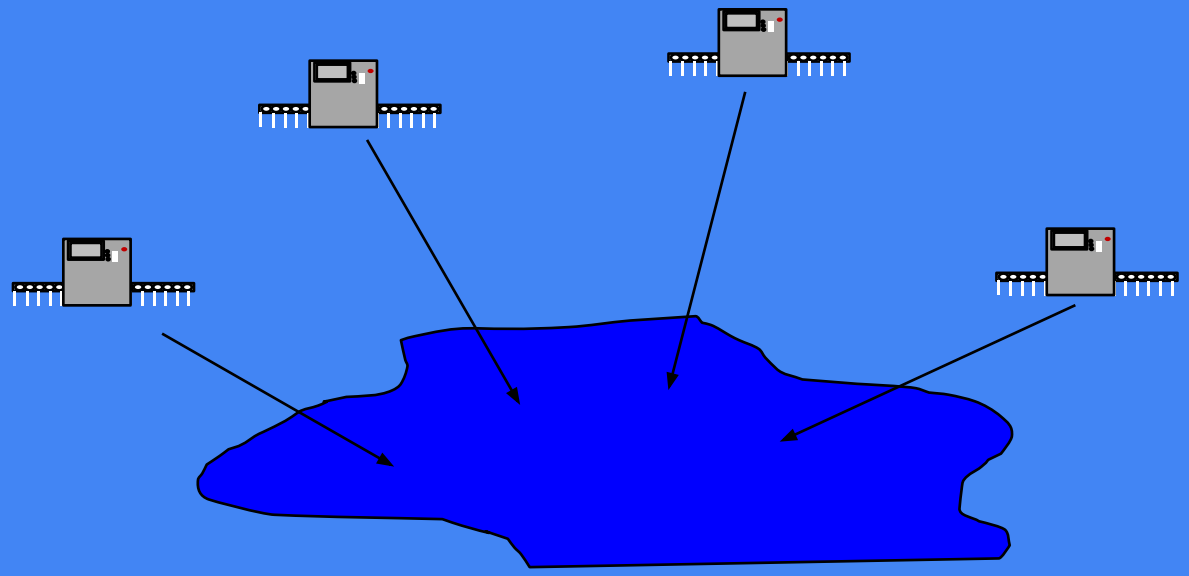


FOREST RIM TECHNOLOGY INC

One day someone decided to dump their machine generated data into a data lake



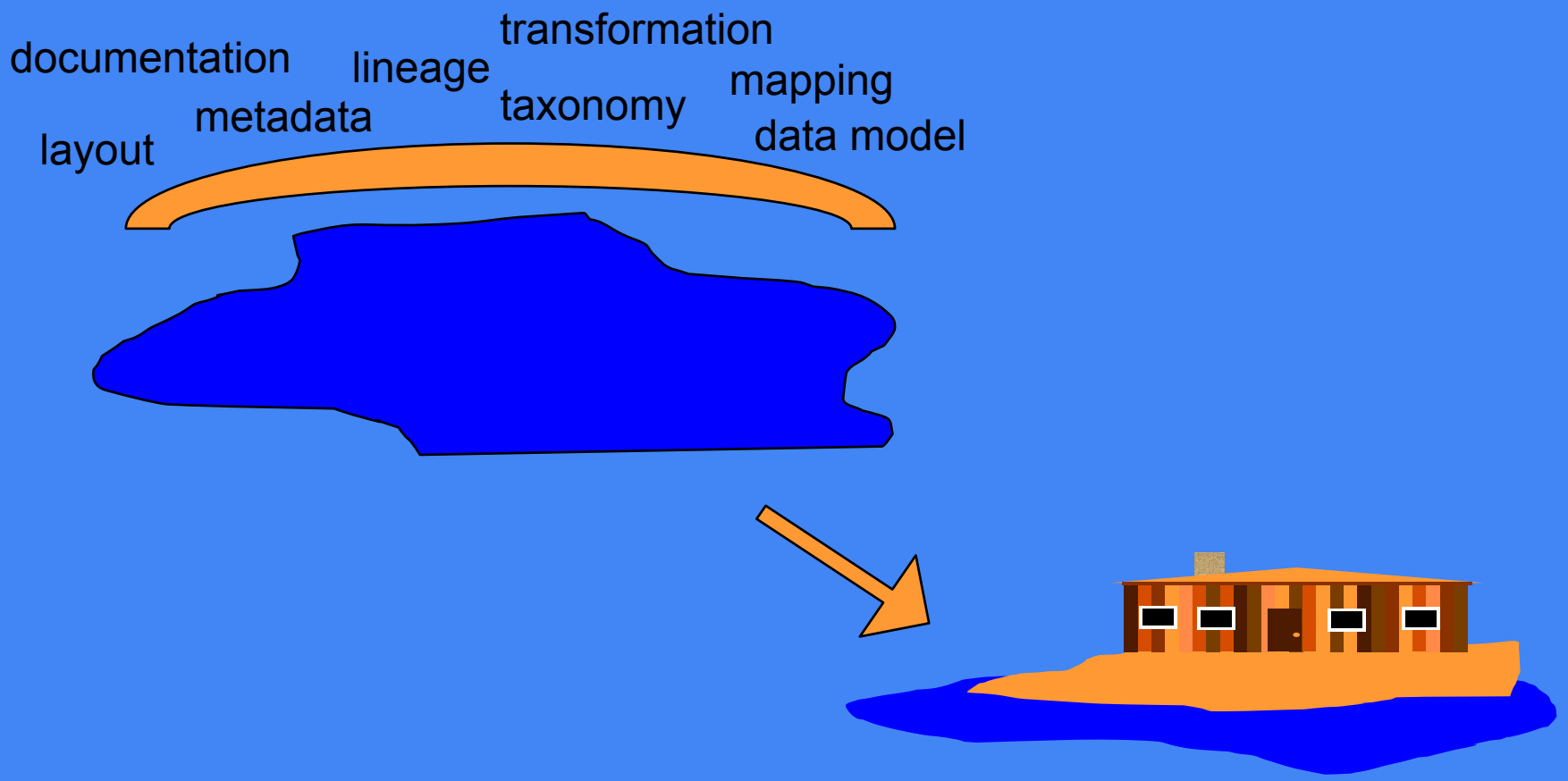
Data lake



Data lake

And in short order the data lake turned into a swamp
No one could find anything

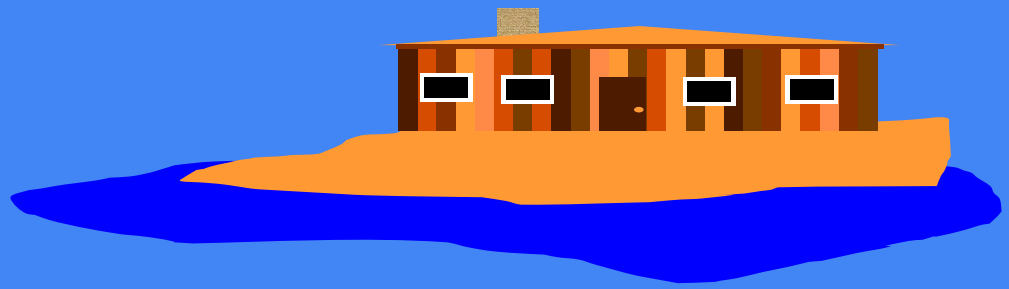
FOREST RIM TECHNOLOGY INC



Then an infrastructure was added to the data lake, turning it into the data lakehouse

And with the infrastructure that was added, the data lakehouse found out that it could start to interface with other kinds of data

Data lakehouse



Data lakehouse

applications	Data warehouse	Dw 2.0	Machine generated

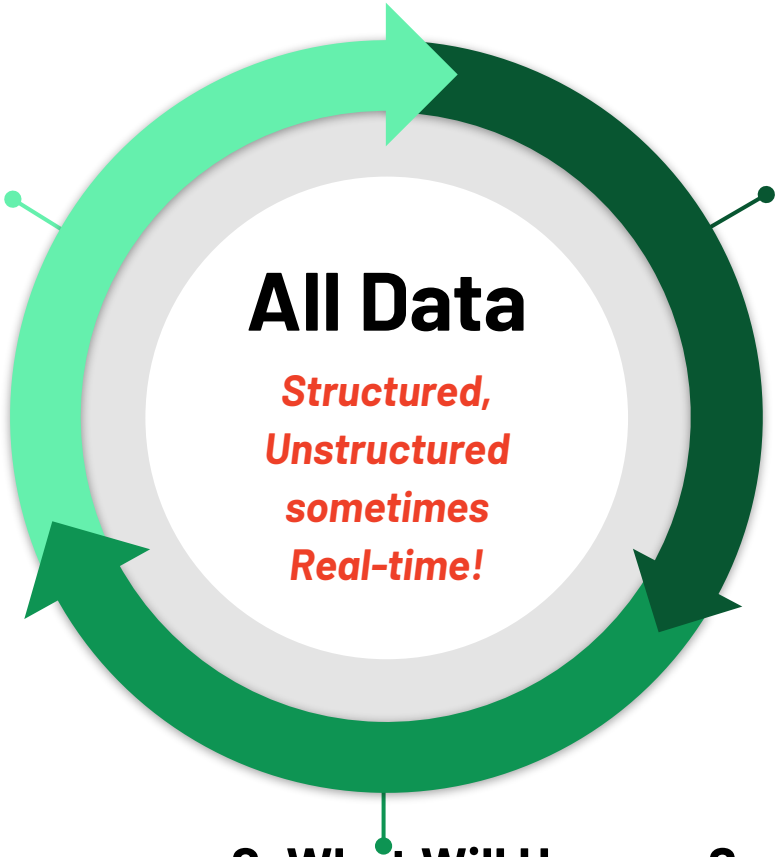
Why BI and AI are converging on a Lakehouse platform?

Business needs Advanced Analytics

BI, AI and Business teams working to get predictive and prescriptive analytics

3. What action can I take

Business, Data Scientists, Data Enggs working together collaboratively



1. What Happened and Why?

Reporting & Dashboarding and Analysts/SMEs empowered with Self-service Analytics

2. What Will Happen ?

Data Science & Machine Learning and AI

Why BI and AI need ALL your data?

BI

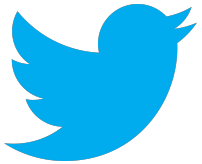
- Siloed Data Marts, Data Warehouses and Data Lakes make Enterprise-wide “Pan-EDW Analytics” impossible.
- You don’t want to base 90% of your decisions on 10% of your structured data! (-Bill Inmon)
- Effective BI requires merging of Structured Data with insights from Unstructured data such as IoT, Text etc.
- Your ML Feature Stores and Model predictions need to be tied to you Universal Semantic layer to expose to all your Reporting teams

AI:

- “Simple AI models with tons of Data always win over complex models with less data”
  From “ Unreasonable Effectiveness of Data” - Google research paper , IEEE, 2009.
- AI algorithms are around from 1970’s - but they only started giving highly accurate predictions when you run them on massive quantities of data.
- ML and AI are iterative processes. Efficient Featuring Engineering needs to run at scale and need access to ALL your data to make the iterations shorter.
- Deep Learning, Anomaly Detection, Time Series Forecasting, Image classification, Recommendation Engines, Sentiment Analysis - all need unstructured data like Images, Videos, text, IoT Data etc.

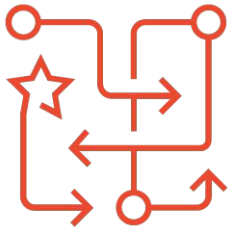
“Next decade will be about Data-centric AI” -Andrew Ng (DeepLearning.ai)

Only 1-5% of enterprises are successful with AI/ML

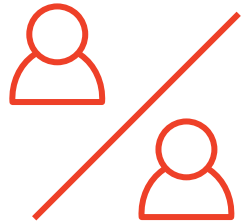


The other 95% struggle

Data is fragmented across many systems

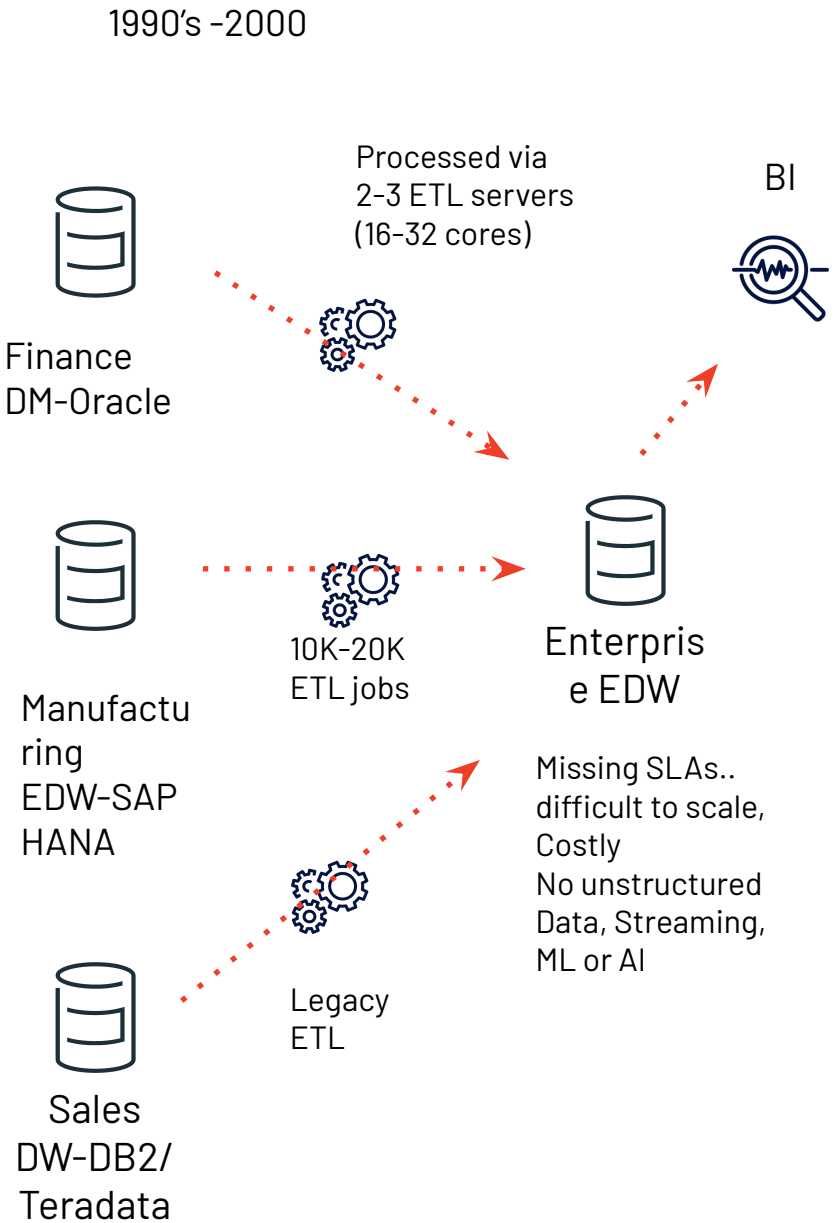


Cost and complexity is a drag on the organization

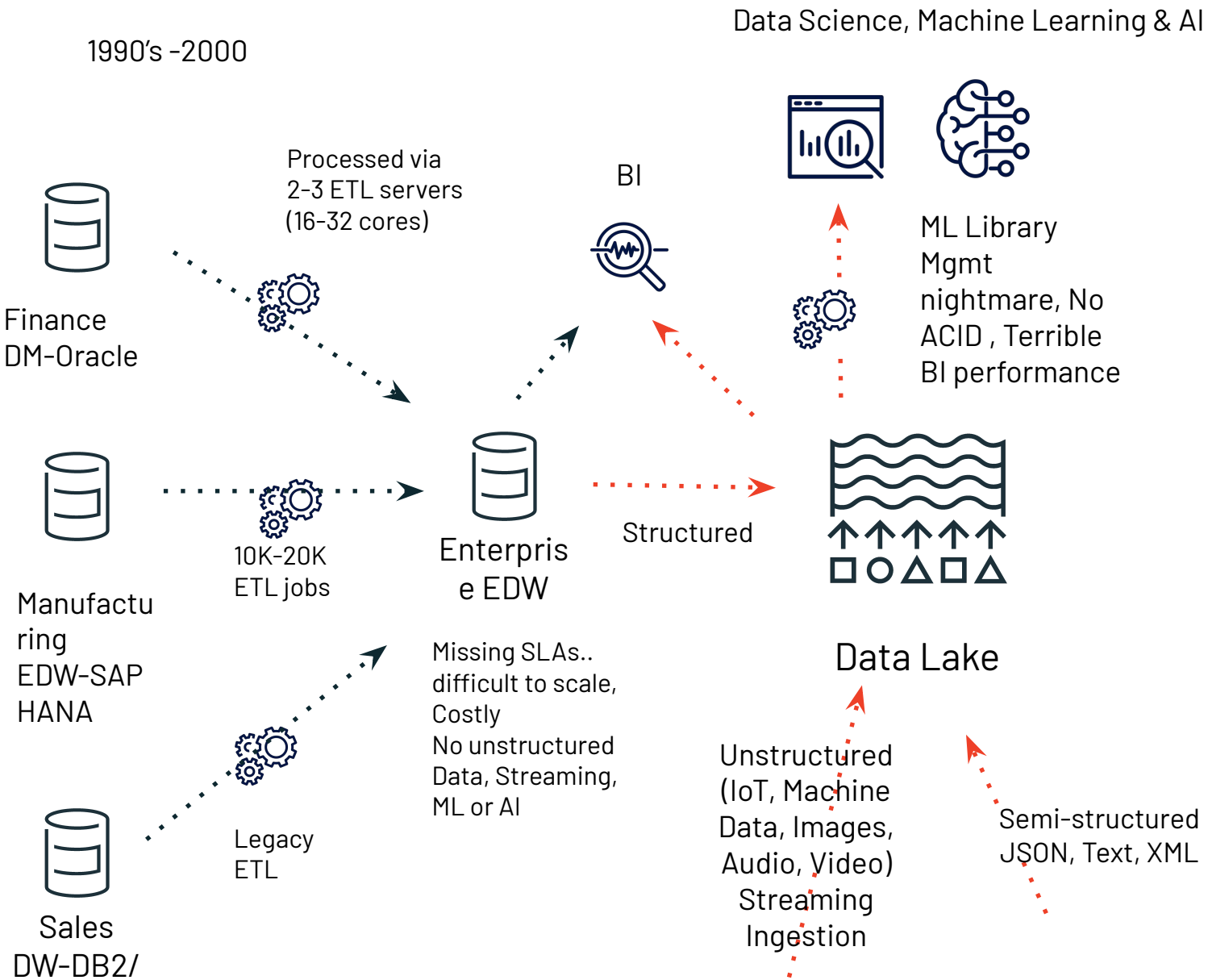


Silos get in the way of data team collaboration

Data platforms are too complicated ..



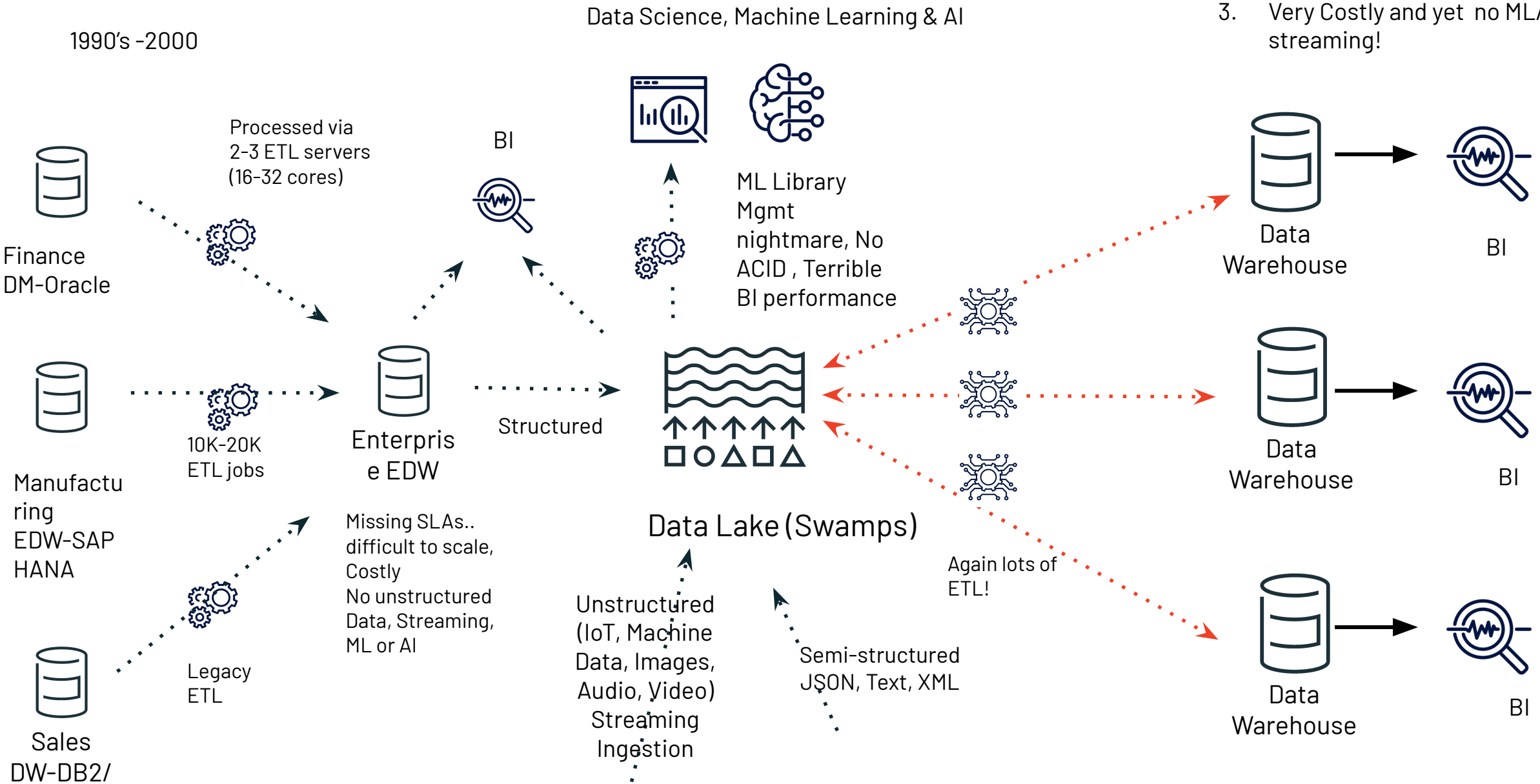
Data platforms are too complicated ..



BI

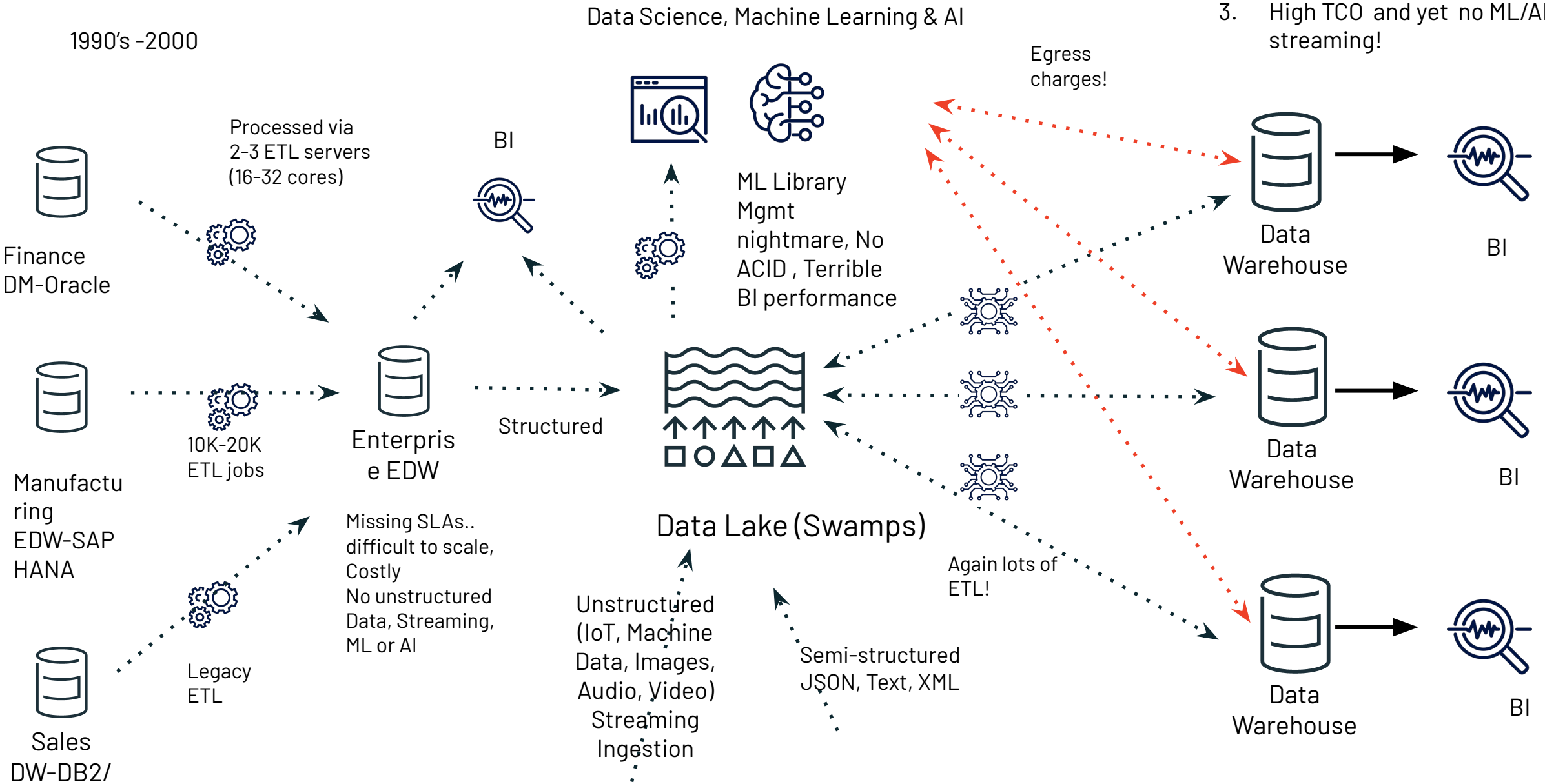
Data platforms are too complicated ..

1. Silos created again - no single sou of Truth for BI
2. Again, proprietary vendor lock-in
3. Very Costly and yet no ML/AI or streaming!



Current State of Data Platforms..

1. Silos created again - no single sou of Truth for BI
2. Again, proprietary vendor lock-in
3. High TCO and yet no ML/AI or streaming!



Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics

Michael Armbrust[1], Ali Ghodsi[1,2], Reynold Xin[1], Matei Zaharia[1,3]
[1]Databricks, [2]UC Berkeley, [3]Stanford University

Abstract

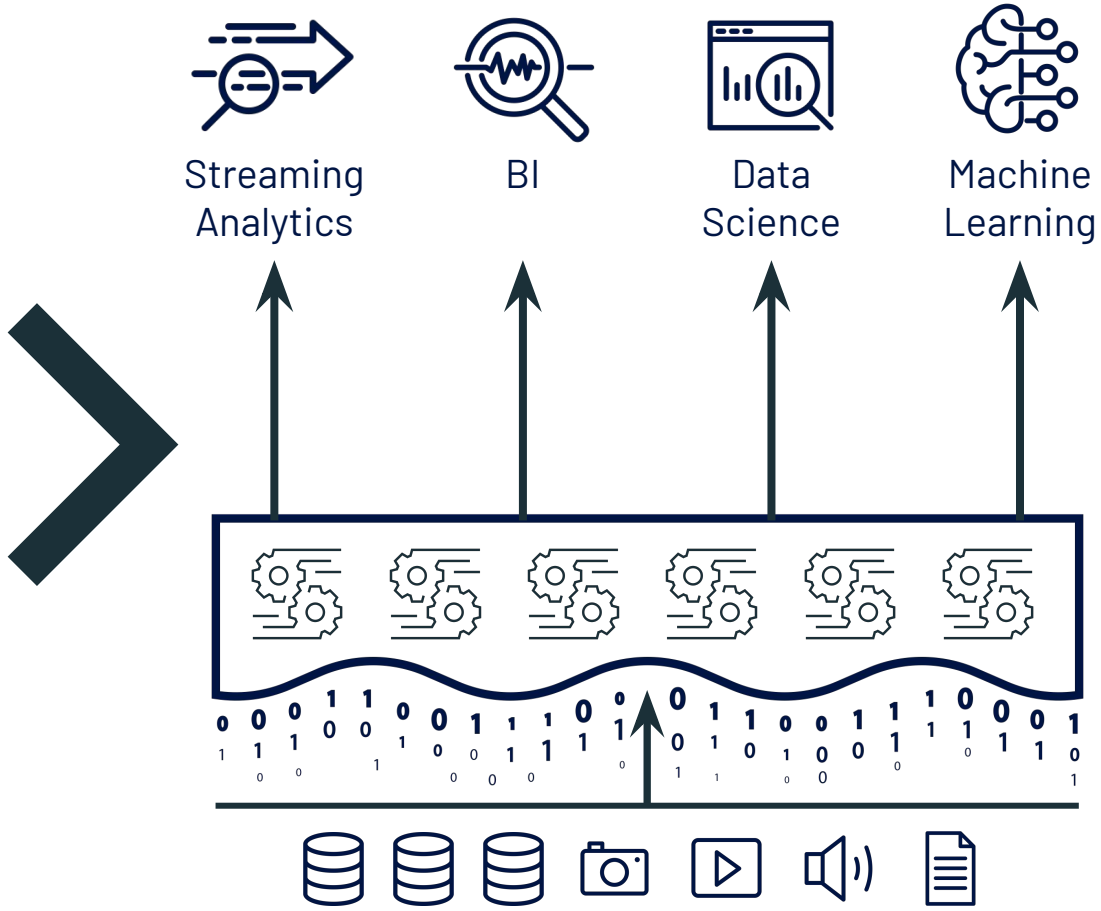
This paper argues that the data warehouse architecture as we know it today will wither in the coming years and be replaced by a new architectural pattern, the Lakehouse, which will (i) be based on open direct-access data formats, such as Apache Parquet, (ii) have first-class support for machine learning and data science, and (iii) offer state-of-the-art performance. Lakehouses can help address several major challenges with data warehouses, including data staleness, reliability, total cost of ownership, data lock-in, and limited use-case support. We discuss how the industry is already moving toward Lakehouses and how this shift may affect work in data management. We also report results from a Lakehouse system using Parquet that is competitive with popular cloud data warehouses on TPC-DS.

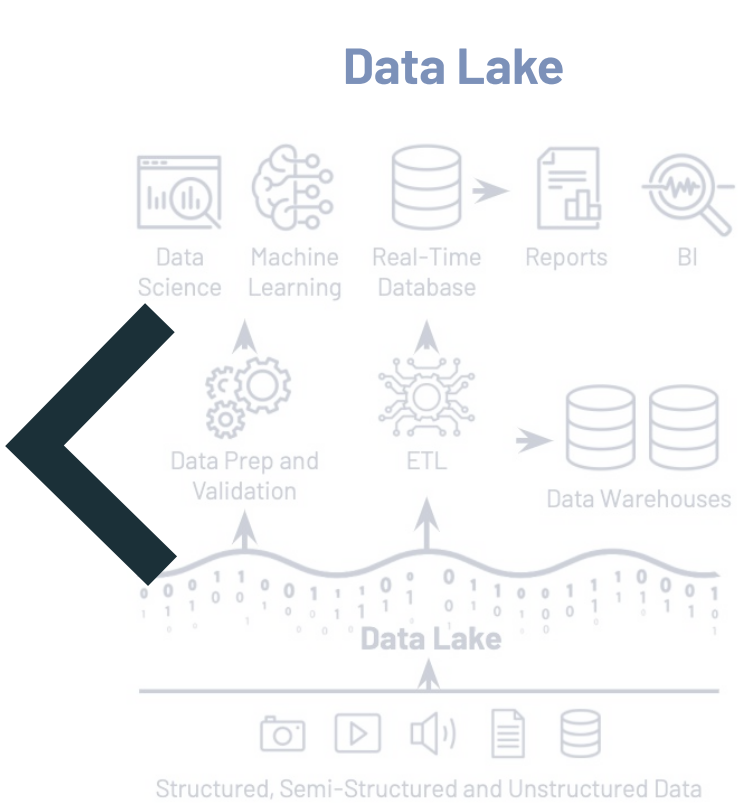
quality and governance downstream. In this architecture, a small subset of data in the lake would later be ETLed to a downstream data warehouse (such as Teradata) for the most important decision support and BI applications. The use of open formats also made data lake data directly accessible to a wide range of other analytics engines, such as machine learning systems [30, 37, 42].

From 2015 onwards, cloud data lakes, such as S3, ADLS and GCS, started replacing HDFS. They have superior durability (often >10 nines), geo-replication, and most importantly, extremely low cost with the possibility of automatic, even cheaper, archival storage, e.g., AWS Glacier. The rest of the architecture is largely the same in the cloud as in the second generation systems, with a downstream data warehouse such as Redshift or Snowflake. This two-tier data lake + warehouse architecture is now dominant in the industry in our experience (used at virtually all Fortune 500 enterprises).

This brings us to the challenges with current data architectures

Lakehouse

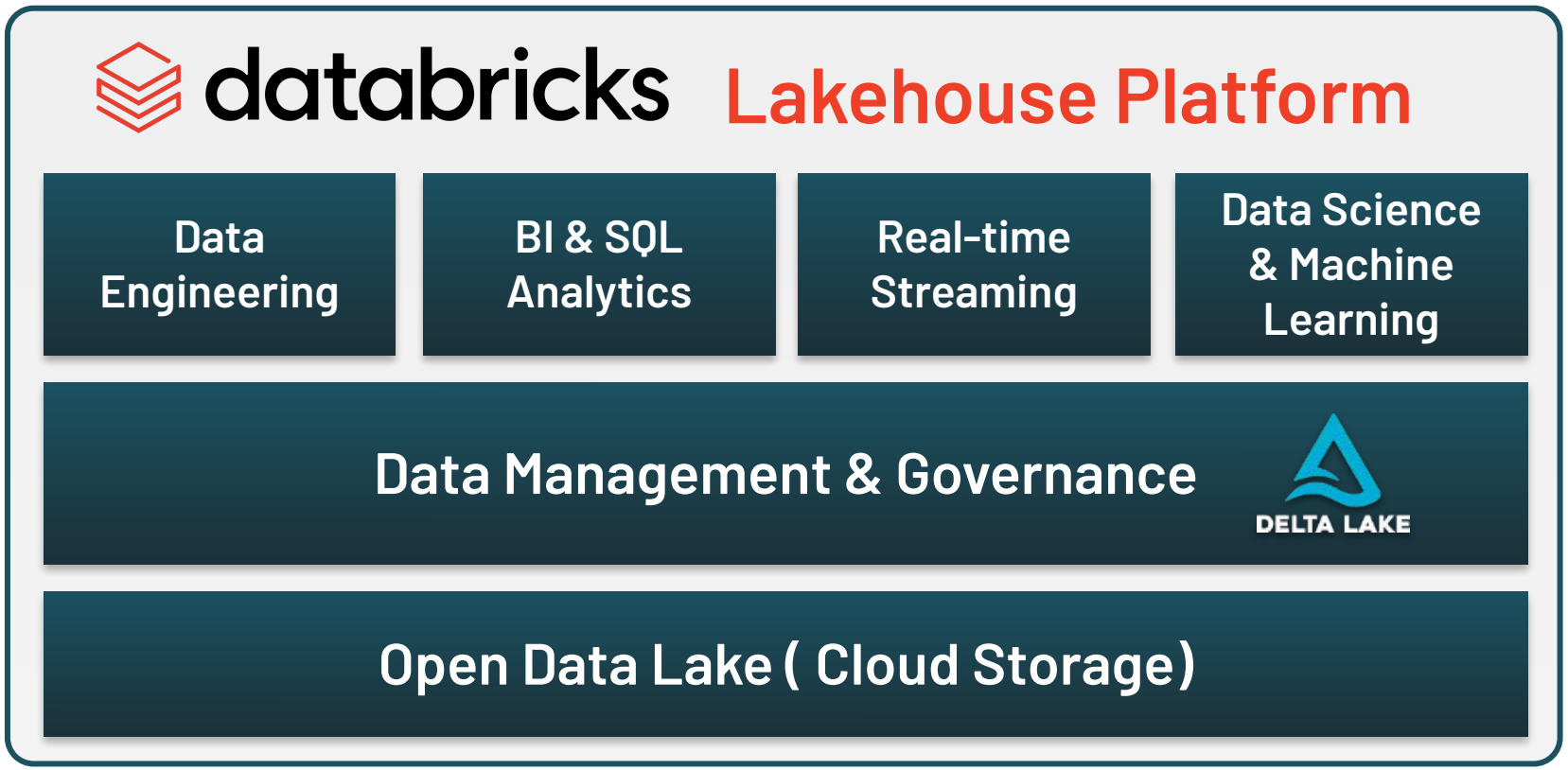




databricks

Lakehouse

Convergence of
your Data, BI and AI Platforms
Open * Collaborative * Simple



aws Microsoft Azure Google Cloud

Key characteristics of a Lakehouse:

1. Single source of Truth - data doesn't move. Compute comes to Data!
2. Direct high-perf BI on ALL your data! (Atscale helps here - universal semantic layer!)
3. First-class ML and AI support (Pytorch, Scikit-Learn, TensorFlow, Keras, Pandas)
4. Unified Fine-grain Security and Governance with Catalog, Lineage etc.
5. Open Data Formats & languages (SQL and Pandas)

Some implementations of a Lakehouse platform for BI and AI

Lakehouse architecture for **unified data warehousing, BI, and ML** on Lakehouse —enabling new use cases not possible before

Large volumes of **streaming IoT data from millions of sensors** difficult to harness for actionable insights and ML

Limitations associated with **legacy Hadoop and Teradata systems;** disparate data difficult to access and unify for analytics ; siloed teams

ML Focused Initiatives:

- Personalized finance
- Risk management
- Fraud detection
- + many more

Disjointed data sources and **legacy data warehousing architecture slowed innovation** and customer-level reporting. Now a single Lakehouse for PBs of data accessed by 3,000+ users across HR, Marketing, Finance, Sales, Support, R&D.

ML use-cases:

- Customer support & service experience
- Marketing personalization
- Anti-abuse & fraud detection

Data Driven Insights at INSPIRE BRANDS

Inspire Brands is the second largest restaurant company in the U.S.

$26B+
in Global System Sales

32,000+
Restaurants

60+
Countries

650,000+
Company & Franchise Team Members

3,200+
Franchisees

$7B
in Collective Supply Chain Purchasing Power

$1.1B
in Annual Ad Fund Spending

4.3B
Guest Interactions Annually

200M
Unique Guests Annually

25M+
Loyalty Members

Why are we building our Unified Data Platform

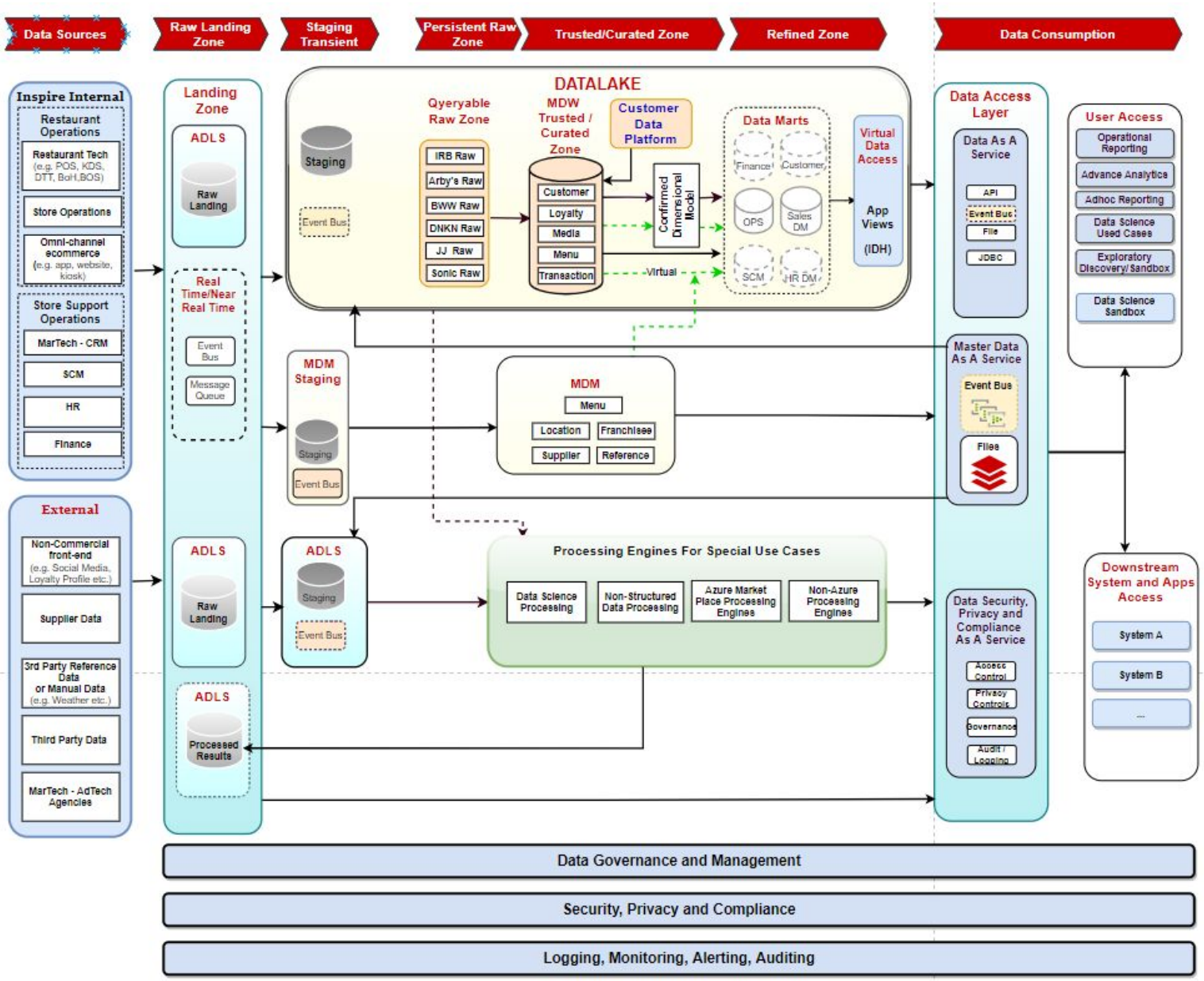
The project was launched to create, embed and operationalize data driven processes into and across our brands.

This is one of many Shared Services we are building to provide our Brand partners with best in class and cost-effective capabilities.

Our brands will benefit through

- Better decision making due to higher quality and faster analytics
- Increased top line sales via better performance all of our marketing channels
- Increased efficiency, guest service levels, and future sales through improved operations (e.g., restaurant labor model, supply chain)

UDP Overview



Guiding Principles

- **Ingest data into the Platform quickly and with minimum friction**
- **Conformation to include all current and future Brands**
- **Enable commercialization of Business Intelligence and Actionable Advanced Analytics Capabilities**
- **Embrace Business Self Service**
- **Adherence to Data Privacy, Security & Compliance**
- **Embed Data Governance & Data Literacy**

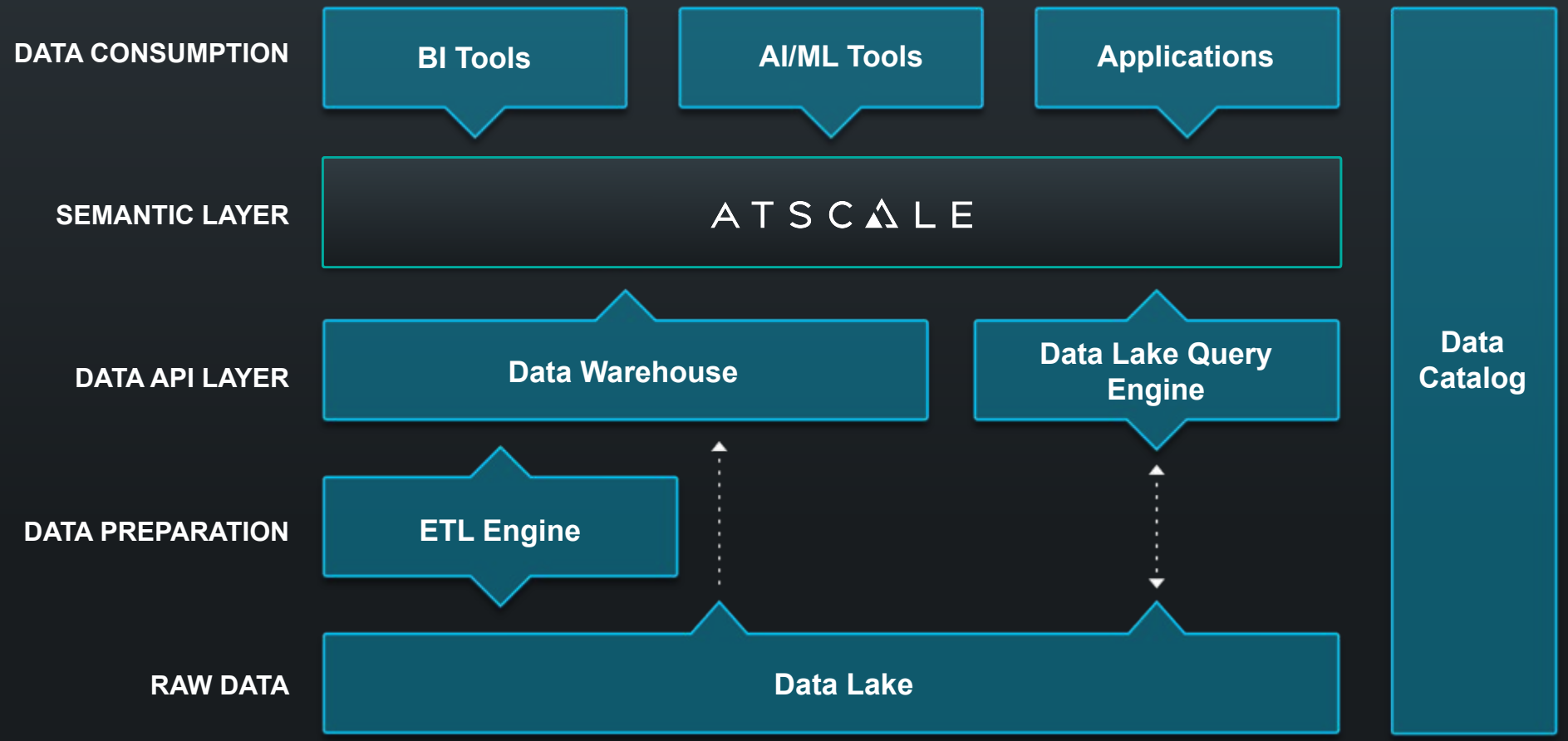
THANK YOU

Contact Andrew Sohn for further information

ASohn@inspirebrands.com

AndrewSohn

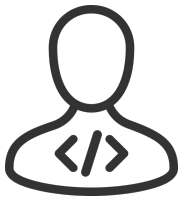
AtScale: *Where we fit.*



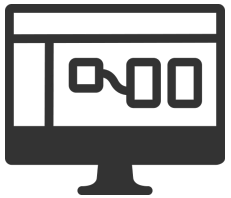
Adding the “I” back to BI

Human Signals

AtScale Modeler



+



Existing Data Context & Relationships

Data Consumers

+

Query Patterns & Frequency

+

Machine Learning

AutoML Engine

DataRobot

Amazon SageMaker

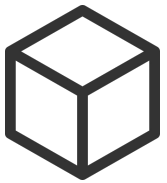
H2O.ai

ML Generated Features

=

Augmented Intelligence

Enriched AtScale Model



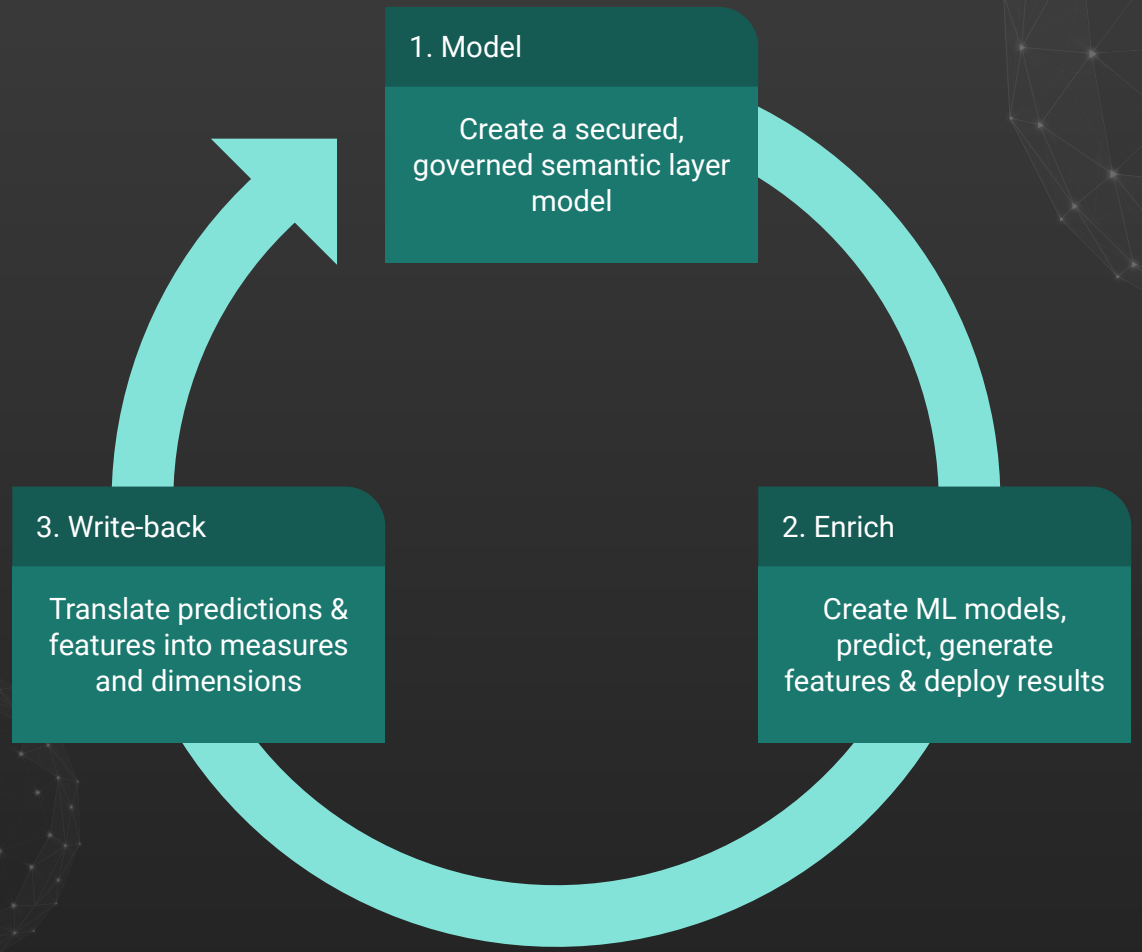
+

Augmented Model

AtScale AI-Link™ Process Flow

BI Teams

- Define common semantics used by the business
- Define dimensionality (e.g. time, geography, product)
- Simplify metrics (i.e. time series analytics)



Data Science Teams

- Develop domain specific features
- Build predictive models based on features
- Score models and understand model drift

www.atscale.com
400 S El Camino Real, Ste 800, San Mateo, CA 94402