databricks

Big Data Management

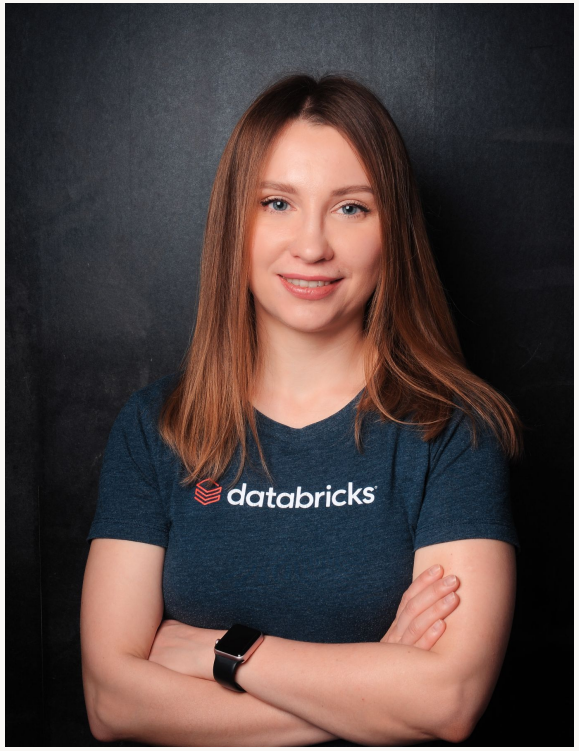
Simplifying Data, Analytics & AI

Tania Sennikova Sr. SA Databricks

Bio Tania

Sr. Solution Architect, Field Engineering

- 2005-2010 Master Degree with major in cryptography
- 2010-2014 Data Analyst in tech startups (Moscow)
- 2014-2016 MSc Web Science University of Koblenz and Landau
- 2017-2021 Data Scientist Consultant (Publicis Group)
- Since 2021 Solutions Architect in Databricks



Agenda

1. Introduction to Databricks
    a. Current state of the industry
    b. Customer Stories
2. Quiz

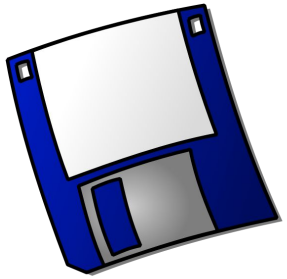
Introduction to Databricks

The history of data management and analytics

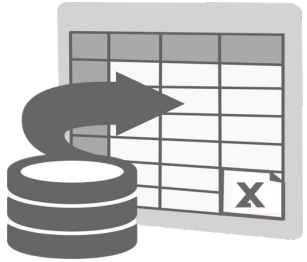
Let's Rewind...

...to the 1990s

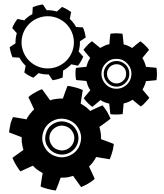
Small amounts of data were created at a relatively **sluggish pace.**



Data was mostly documents and in the **form of rows and columns.**



Storing and processing this data was not much trouble as a **single storage and processing engine would do the job.**



The old way of storing and managing data

Teradata and IBM built this thing called the database. It was literally a box. You threw all of your data in there.

teradata.

ORACLE

IBM

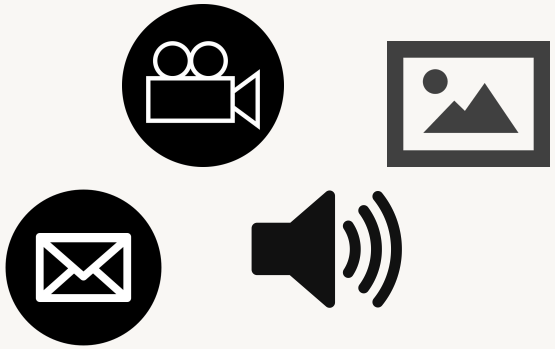
...Until it filled up. **Then you had to call your rep and have no other choice but to buy more.**

As the years passed, the internet changed everything.

The Early 2000s

Data was generated in a **multitude of formats at a volume never before seen.**

Structured and semi structured data were created in the form of **audio, video, images and emails, just to name a few.**



All of this data became collectively known as **big data.**

Enter Hadoop

The Late 2000s

Multiple storage units and processors were required to handle Big Data.

This concept was incorporated in the **framework of Hadoop**.

Hadoop could store and process vast amounts of **any data efficiently using a cluster of commodity hardware.**

The beginnings of Apache Spark at UC Berkeley

AMPLab funded

by tech companies:

Google facebook amazon

Saw impactful projects:

...leveraging massive amounts of data

...doing high impact ML/AI






The Genesis of

APACHE Spark™

We'll provide a large set of anonymous movie rating data.

Can you improve our recommendation system prediction accuracy by 10%?

Spark storms onto the scene

2009–2013

In 2009, Spark started as a project at the UC Berkeley AMPLab

In 2010, Spark was open sourced under a BSD license. **In 2013, Spark became an Apache top level project.**

Databricks is founded in 2013, built off of their initial success of Spark.


Cloud
Big Data + ML
Open Source
Three Bets on Three Trends

Our Vision

Help **data teams** solve the
world's toughest problems

Our Mission

Help **make big data simple and democratize big data solutions,** such as **machine learning** and **AI**.

Help **businesses transform to data-driven organizations** and easily access the **value gained from data**.

databricks

The data and AI company

Gartner Magic Quadrant — LEADER — 2023 Cloud Database Management Systems

Gartner Magic Quadrant — LEADER — 2024 Data Science & Machine Learning

FORRESTER® WAVE LEADER 2023 — Cloud Data Pipelines

FORRESTER® WAVE LEADER 2024 — Data Lakehouses

FORRESTER® WAVE LEADER 2024 — AI Foundation Models For Language

IDC ANALYZE THE FUTURE LEADER — Analytic Stream Processing

10,000+ global customers

$2.4B+ in annual revenue

14B+ in investment

Inventor of the **lakehouse** and pioneer of **generative AI**

Creator of:

DELTA LAKE

Unity Catalog

APACHE Spark™

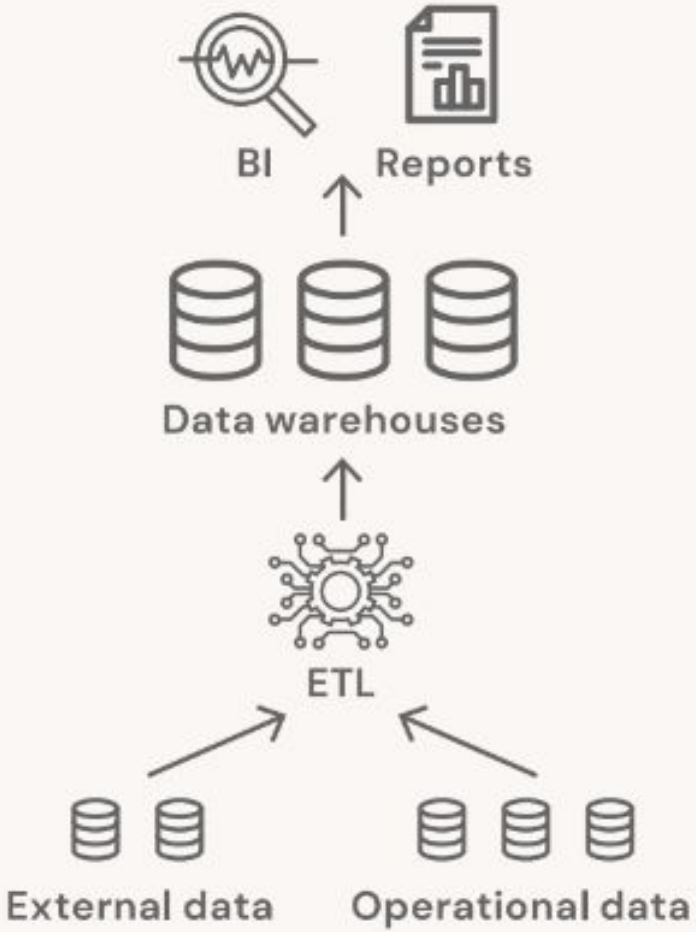
mlflow

DBRX

What is a Lakehouse?

Businesses need more than relational databases

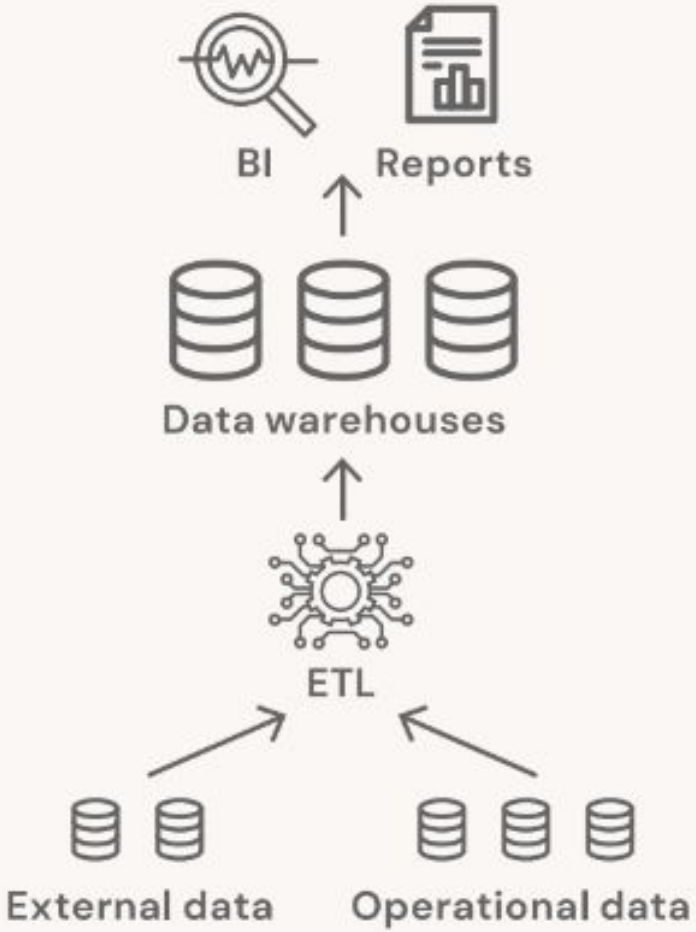
Data warehouse



Pros:

- Business intelligence (BI)
- Analytics
- Structured & clean data
- Predefined schemas

Data warehouse



Cons:

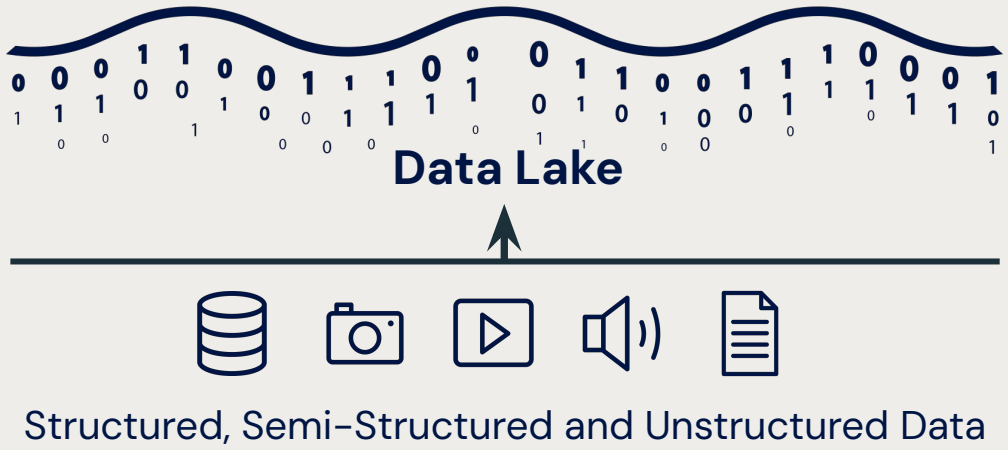
- No support for semi or unstructured data
- Inflexible schemas
- Struggled with volume and velocity upticks
- Long processing time

2000s
Big Data explosion

Data Lakes

Pros:

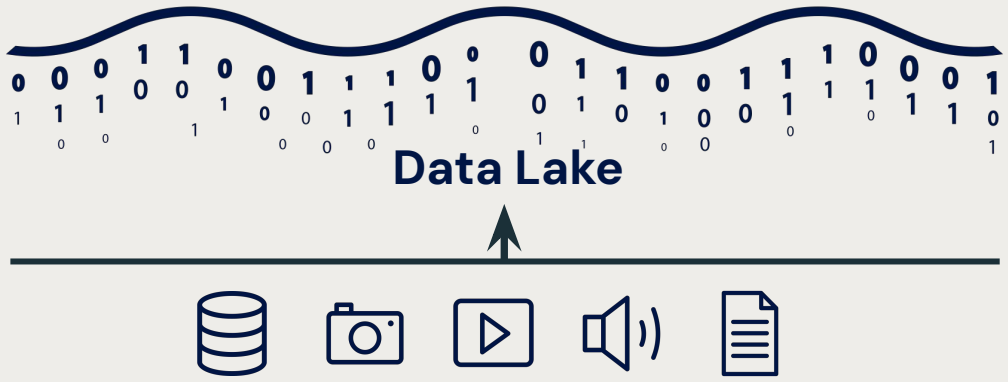
- Flexible data storage
- Streaming support
- Cost efficient in the cloud
- Support for AI and Machine Learning



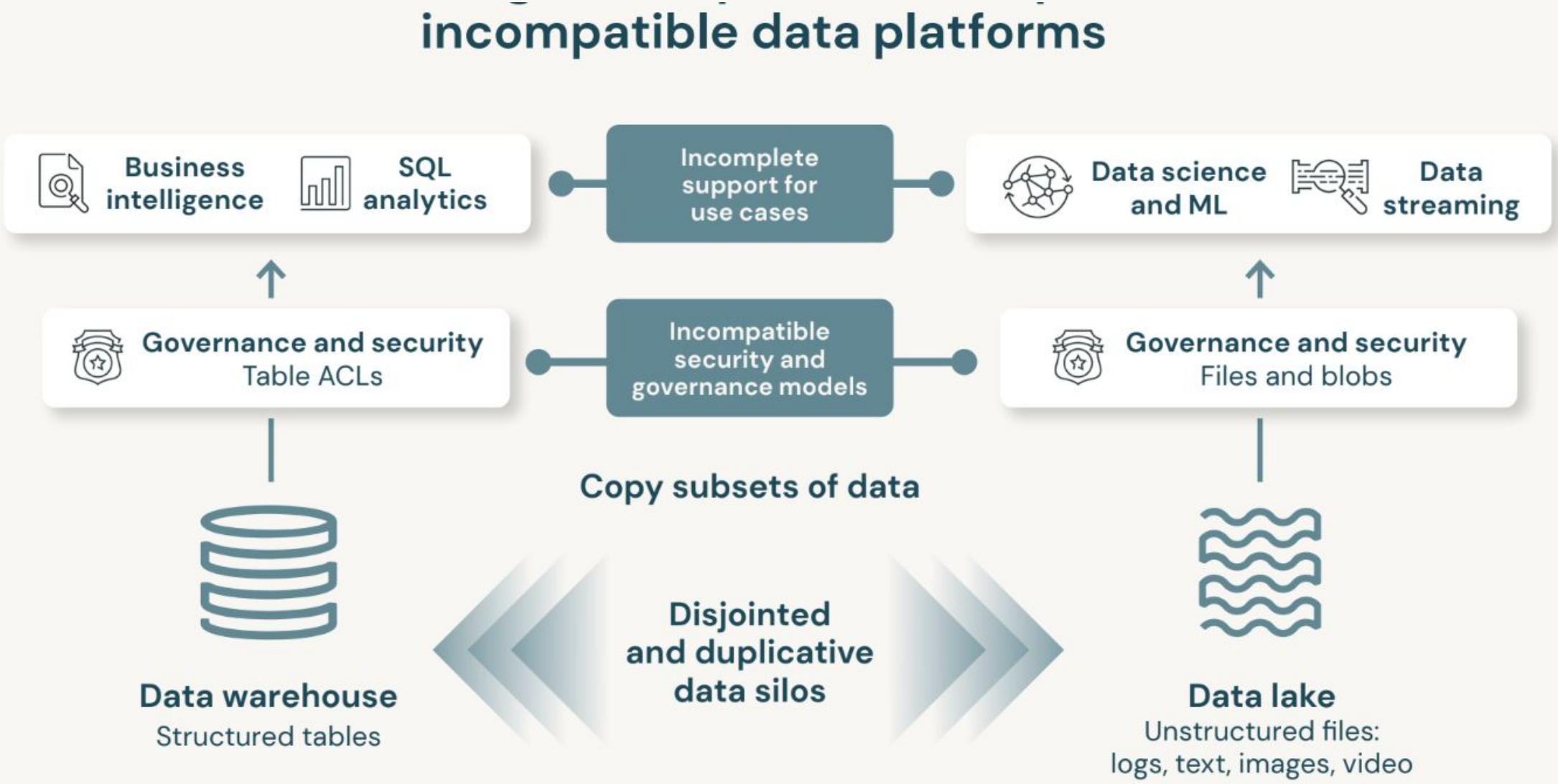
Data Lakes

Cons:

- No transactional support
- Poor data reliability
- Slow analysis performance
- Data governance concerns
- Data warehouses still needed



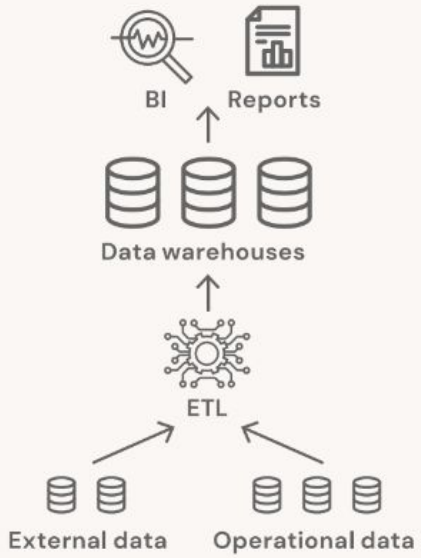
Business required two disparate, incompatible data platforms



Data lakehouse

One platform to unify all your data, analytics and AI workloads

Data warehouse



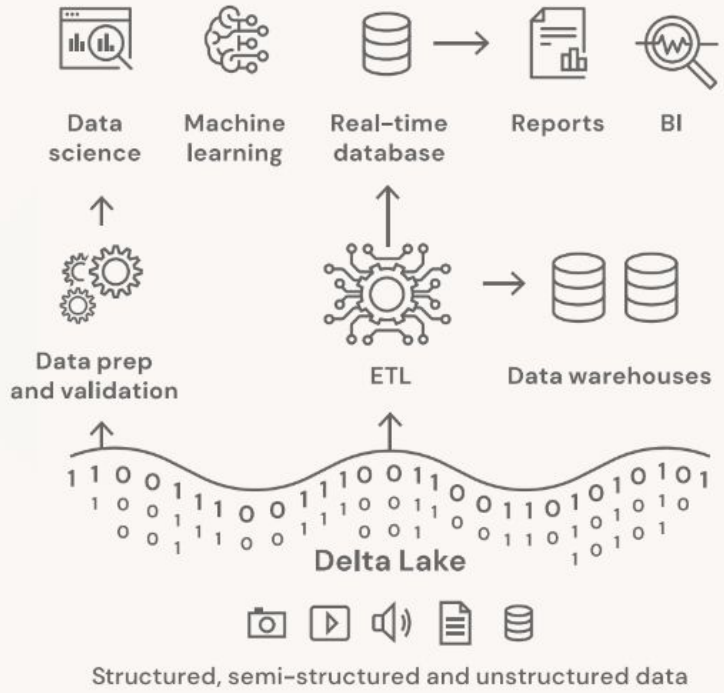
Lakehouse Platform

All machine learning, SQL, BI, and streaming use cases

One security and governance approach for all data assets on all clouds

An open and reliable data platform to efficiently handle all data types

Delta Lake



Key features of a data lakehouse:

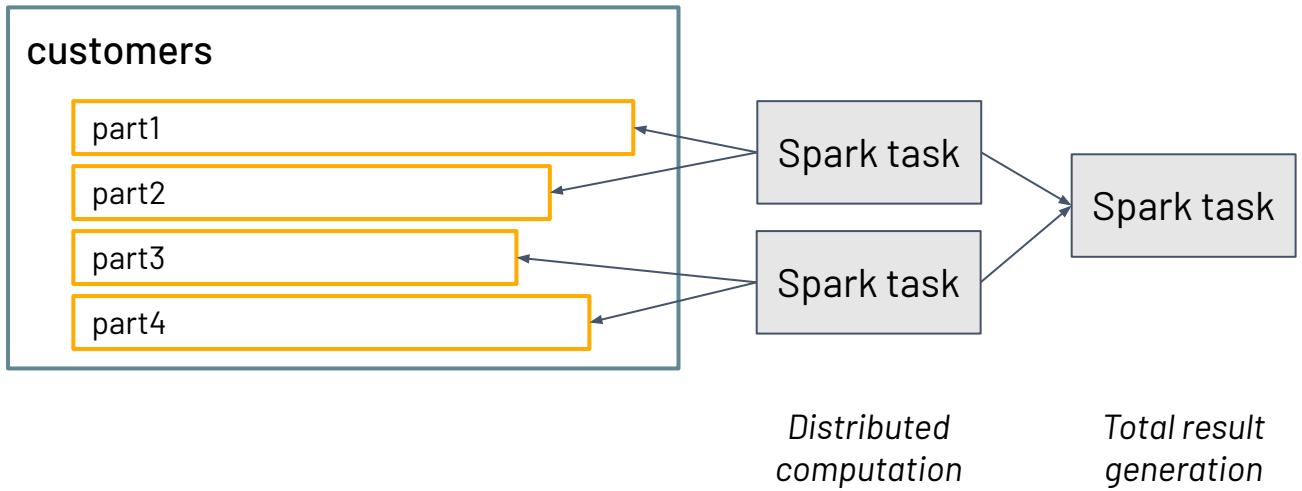
- Transaction support
- Schema enforcement and governance
- Data governance
- BI Support
- Decoupled storage from compute
- Open storage formats
- Support for diverse data types
- Support for diverse workloads
- End-to-end streaming

DELTA LAKE

What does Parquet look like?

```
%fs ls /tmp/bernhard/loan_by_state_delta
```

path
dbfs:/tmp/bernhard/loan_by_state_delta/part-00000-cd80fd12-457b-4058-a904-3628c6e90a57-c000.snappy.parquet
dbfs:/tmp/bernhard/loan_by_state_delta/part-00004-69b2735a-18d1-474e-8318-ecd13ca410a7-c000.snappy.parquet
dbfs:/tmp/bernhard/loan_by_state_delta/part-00009-05545d85-f051-4a37-852a-18298fb4f3cd-c000.snappy.parquet
dbfs:/tmp/bernhard/loan_by_state_delta/part-00010-3fa88ba6-61cc-45cd-8c0d-e0949d1bbcce-c000.snappy.parquet



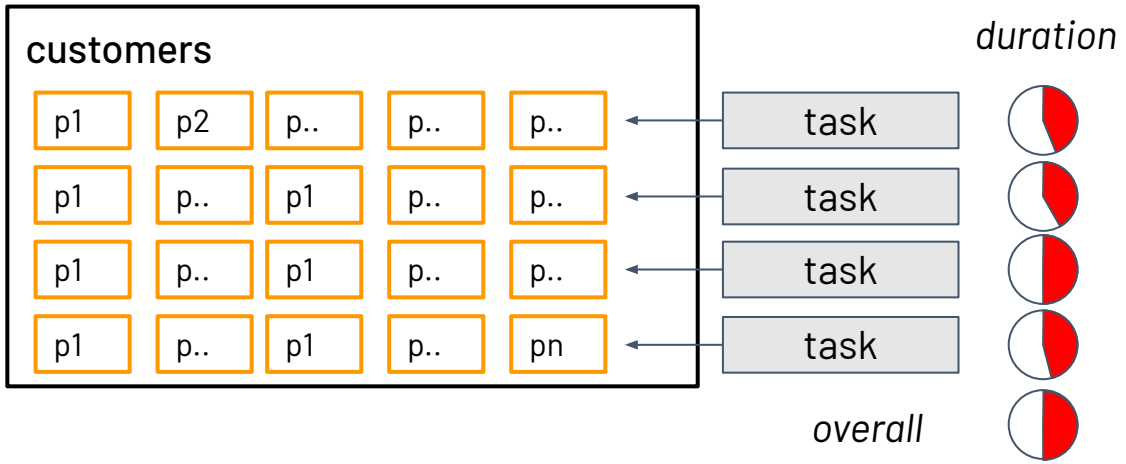
Parquet data

- is a folder
- contains several part files
- designed for being read by distributed tasks

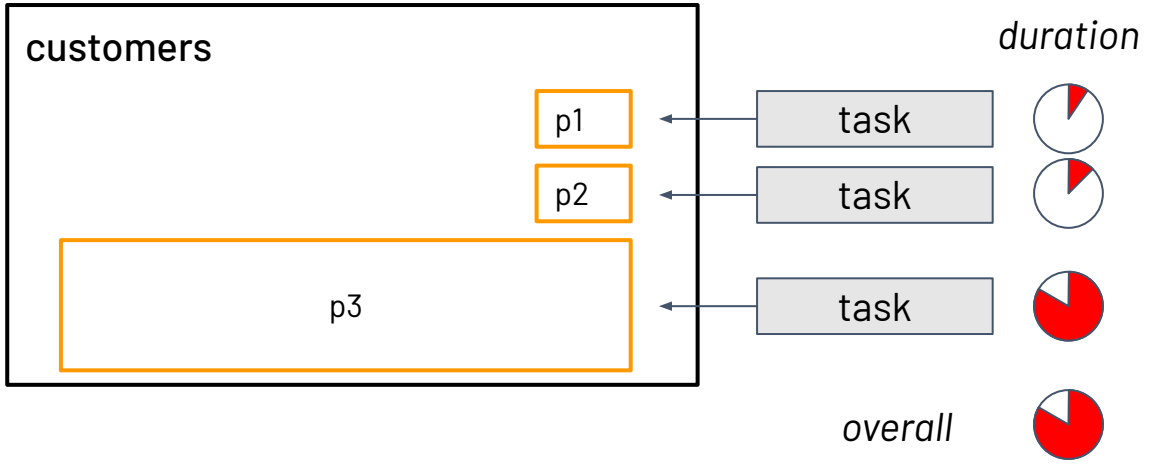
databricks

What can go wrong with Parquet?

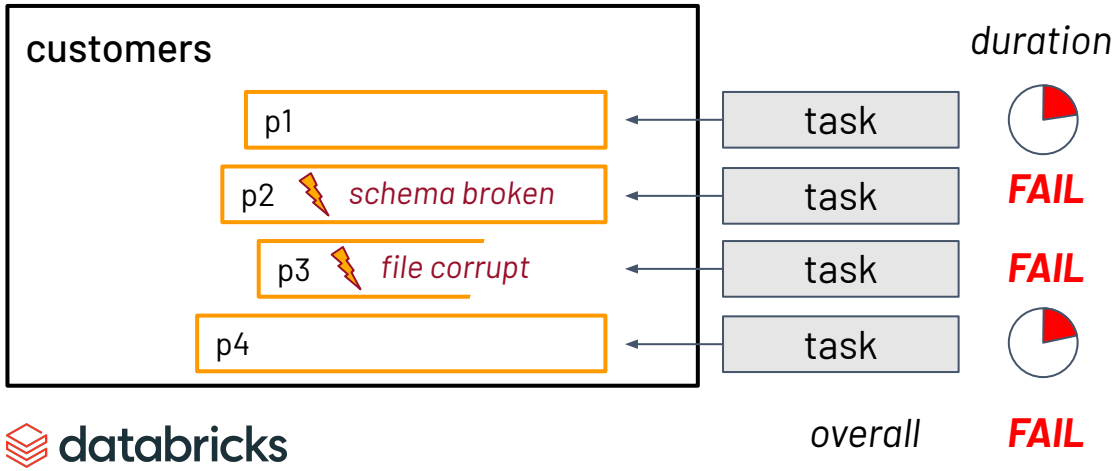
Small file problem



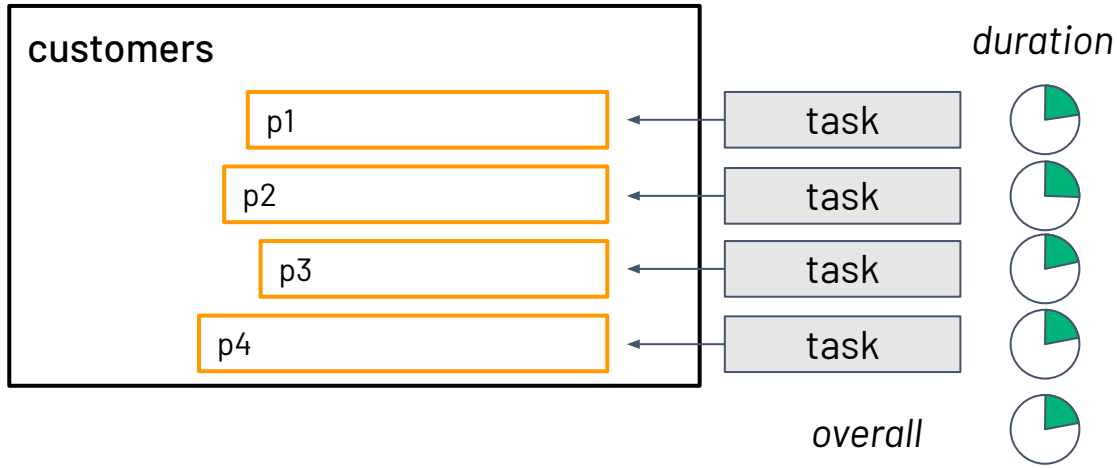
Non-linear file sizes



Corrupt data

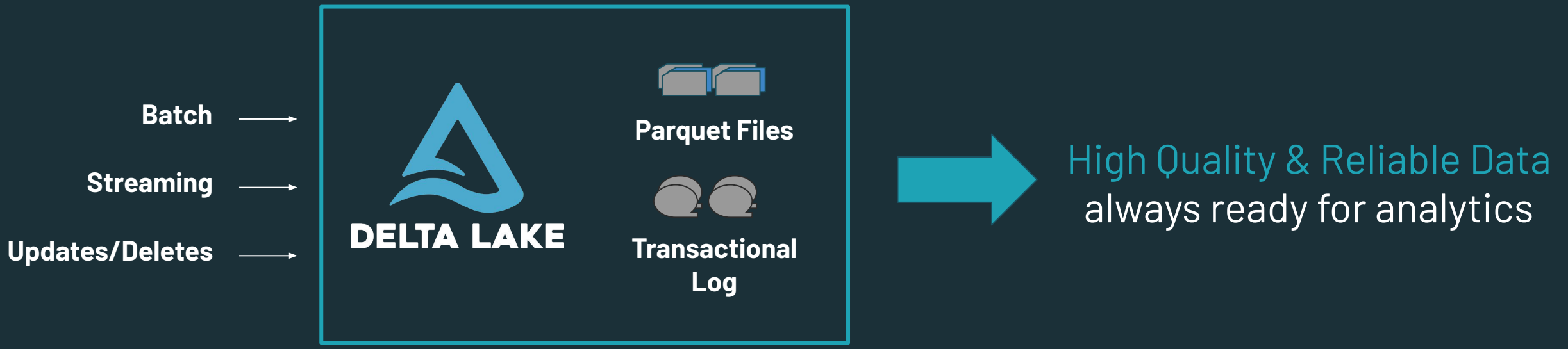


Goal



databricks

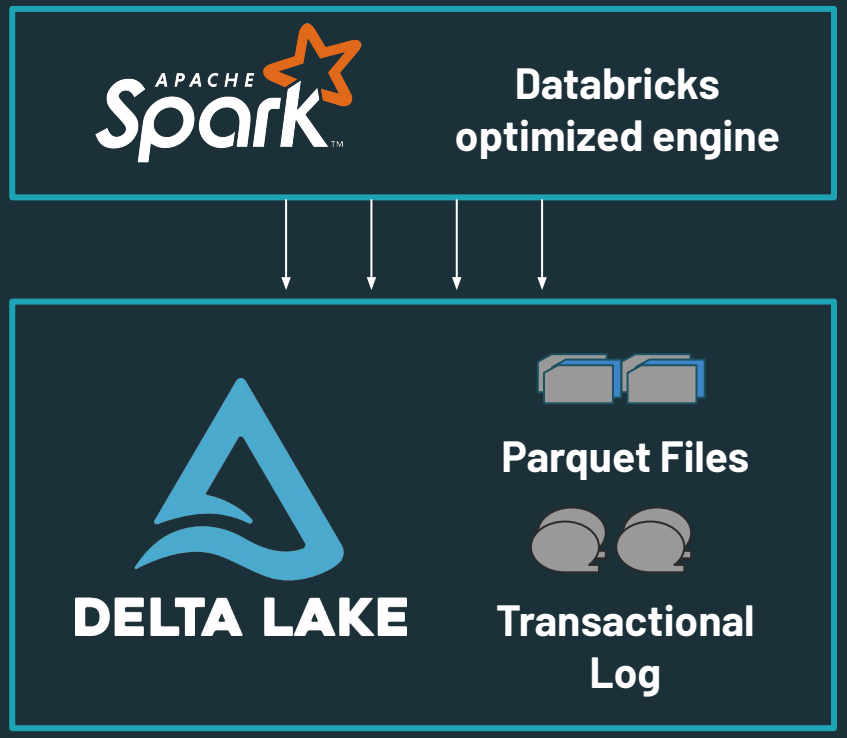
Delta Lake ensures data reliability



Key Features

- ACID Transactions
- Schema Enforcement
- Unified Batch & Streaming
- Time Travel/Data Snapshots

Delta Lake optimizes performance



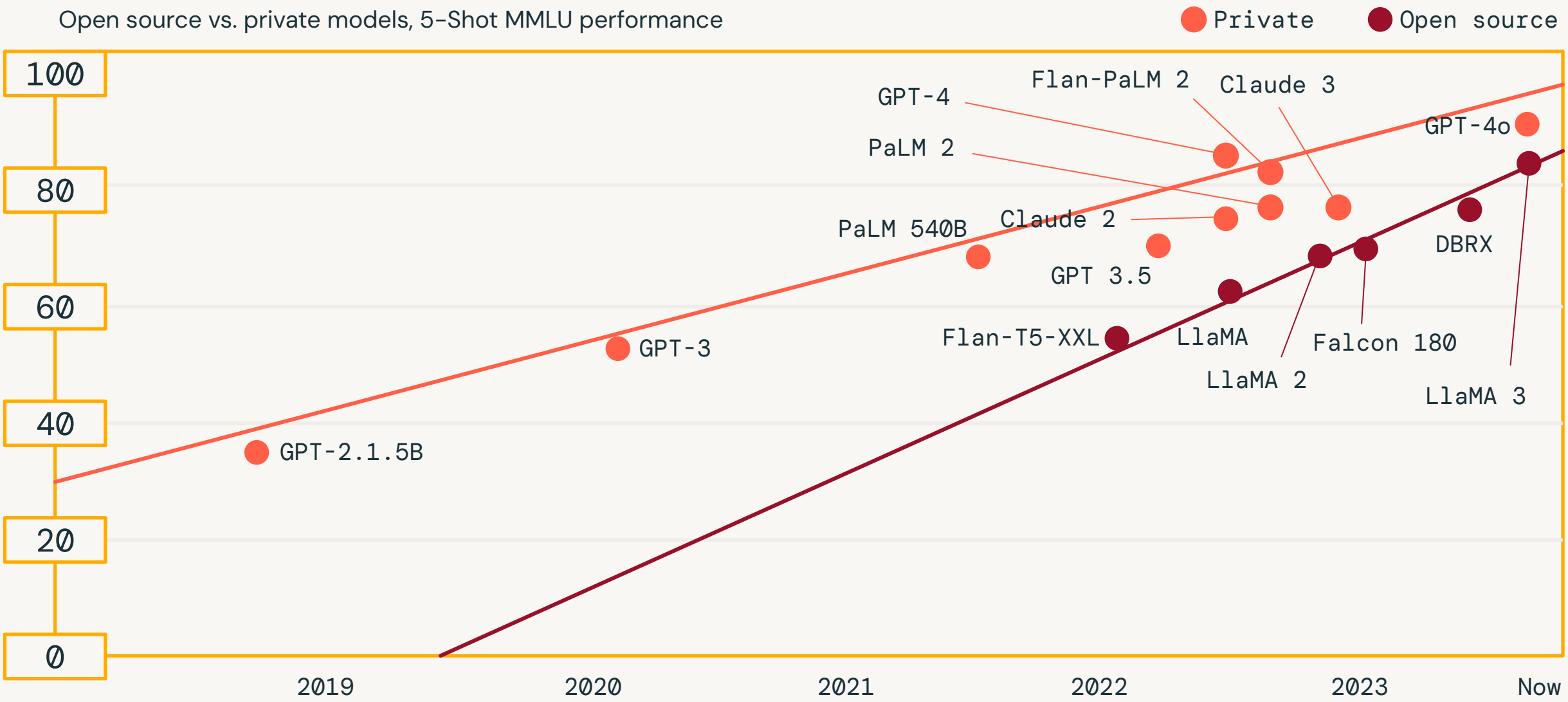
Highly Performant
queries at scale

Key Features

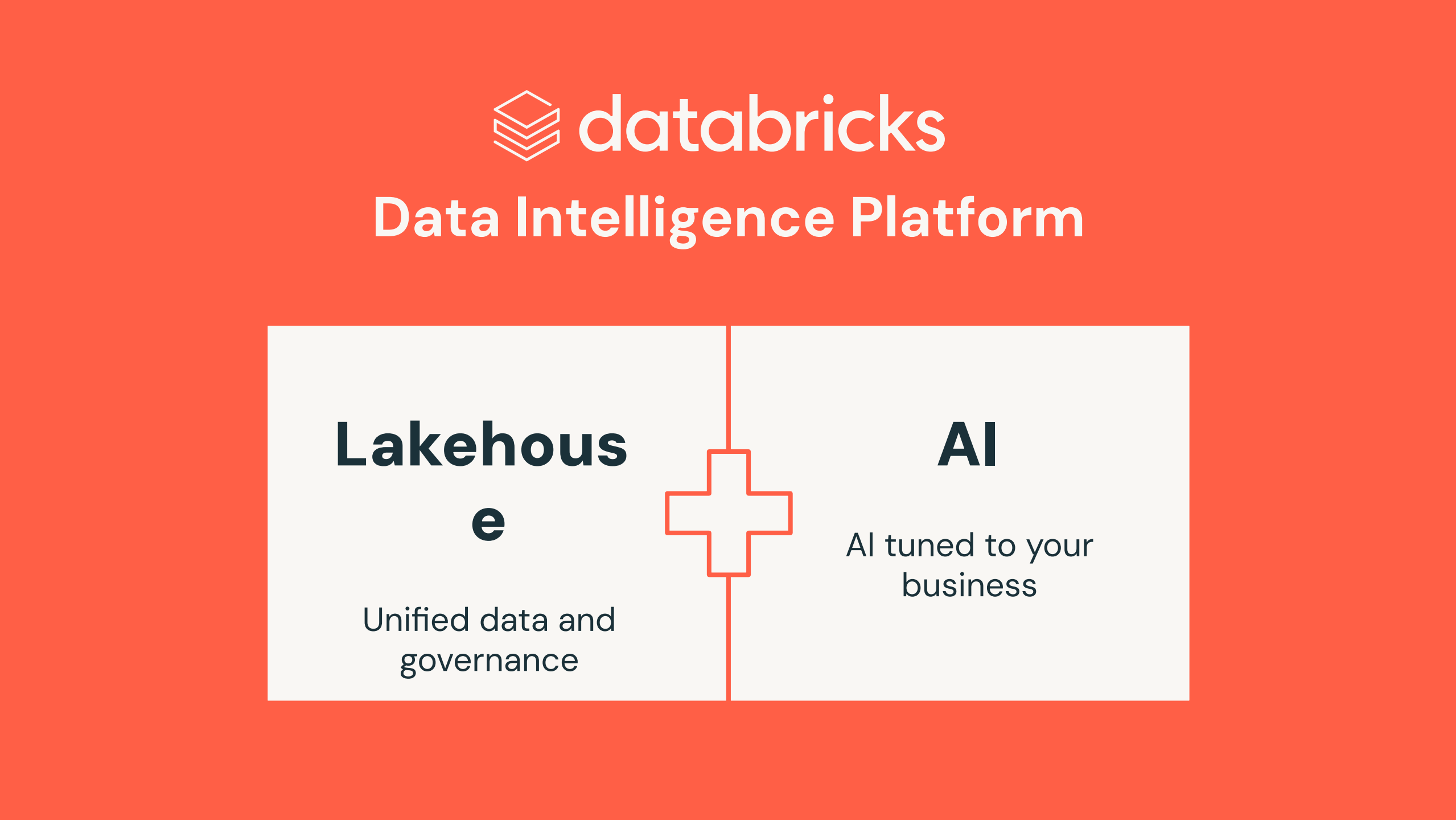
- Indexing
- Compaction
- Data skipping
- Caching

But there is a new kid on the block...

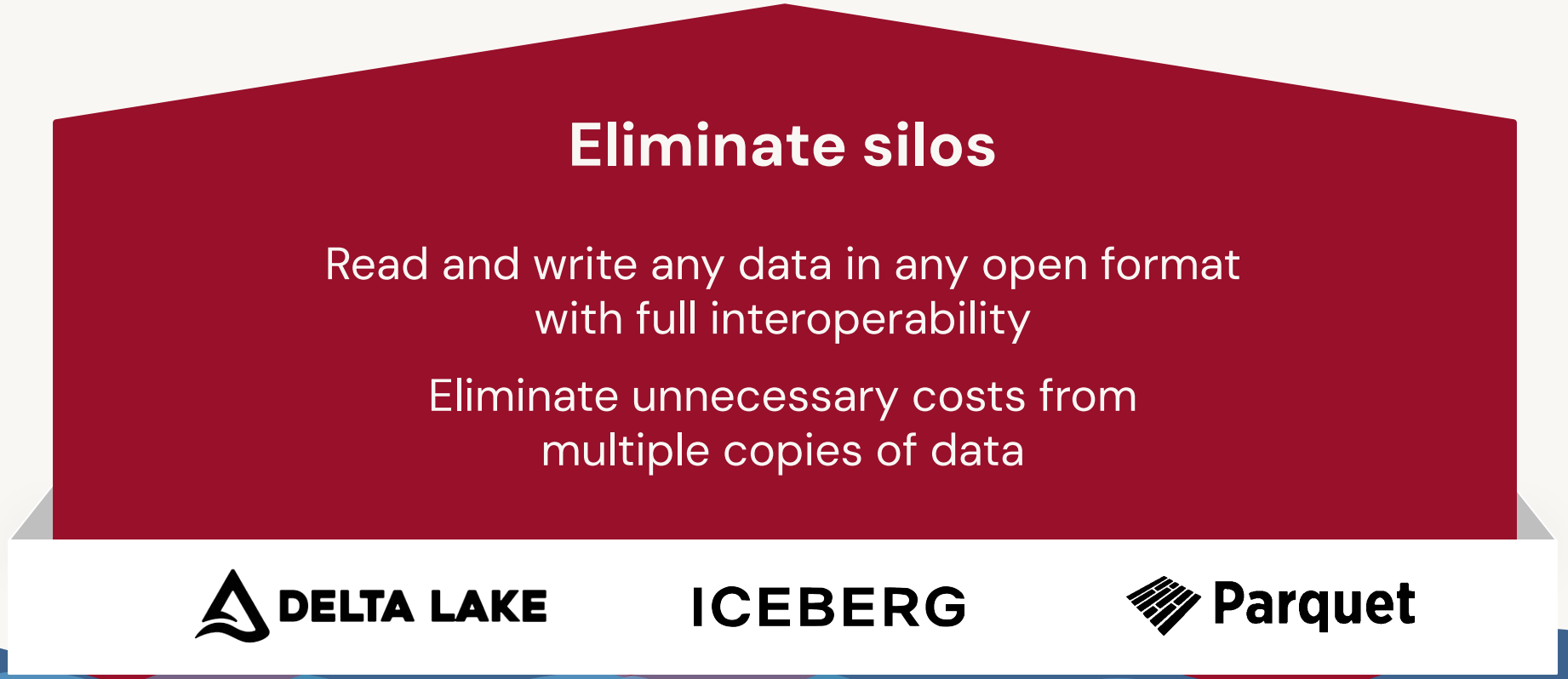
LLMs maxing out on general intelligence tests



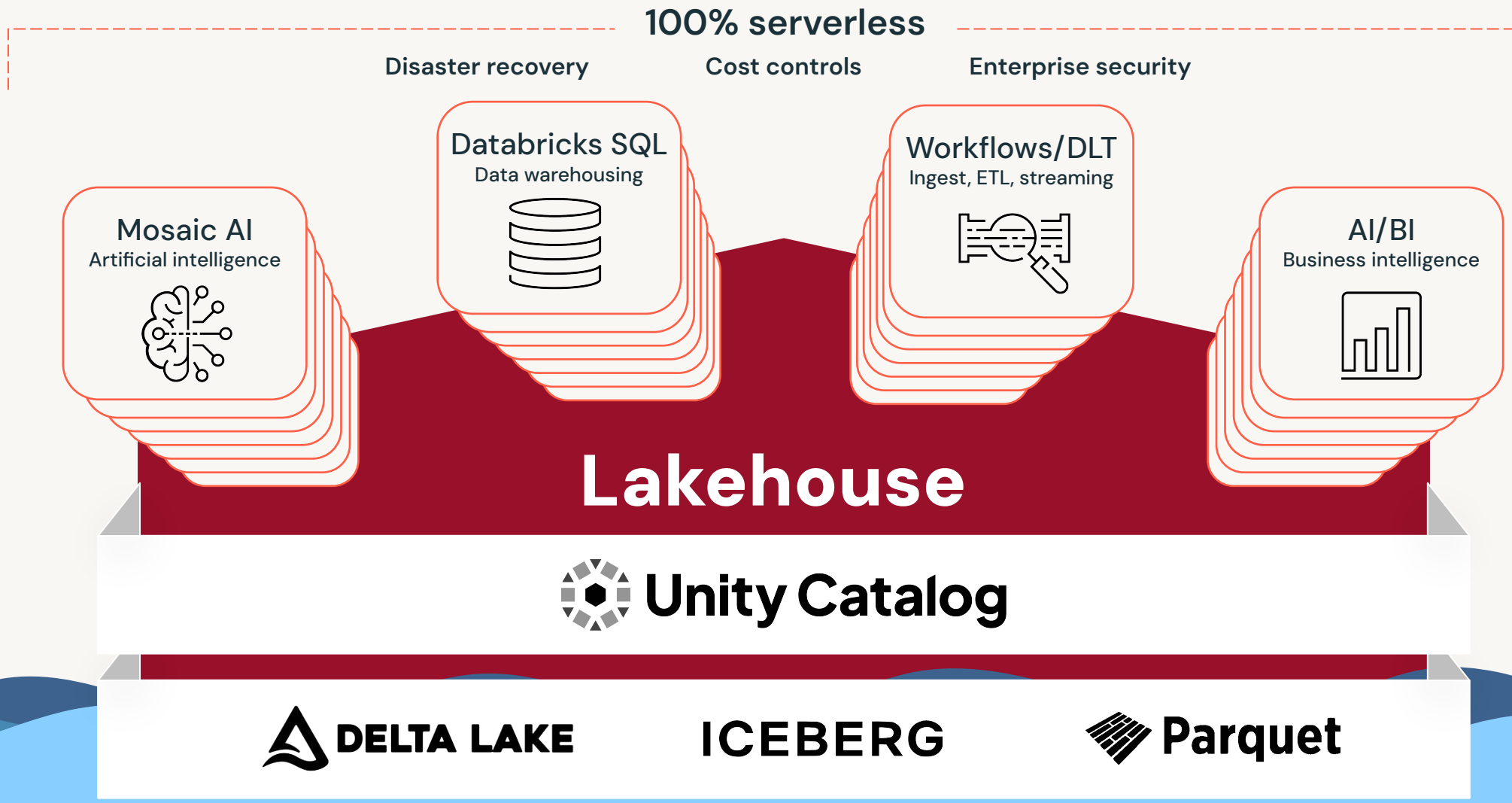
Data Intelligence Platform


databricks
Data Intelligence Platform
Lakehouse
Unified data and governance
AI
AI tuned to your business

Own your data



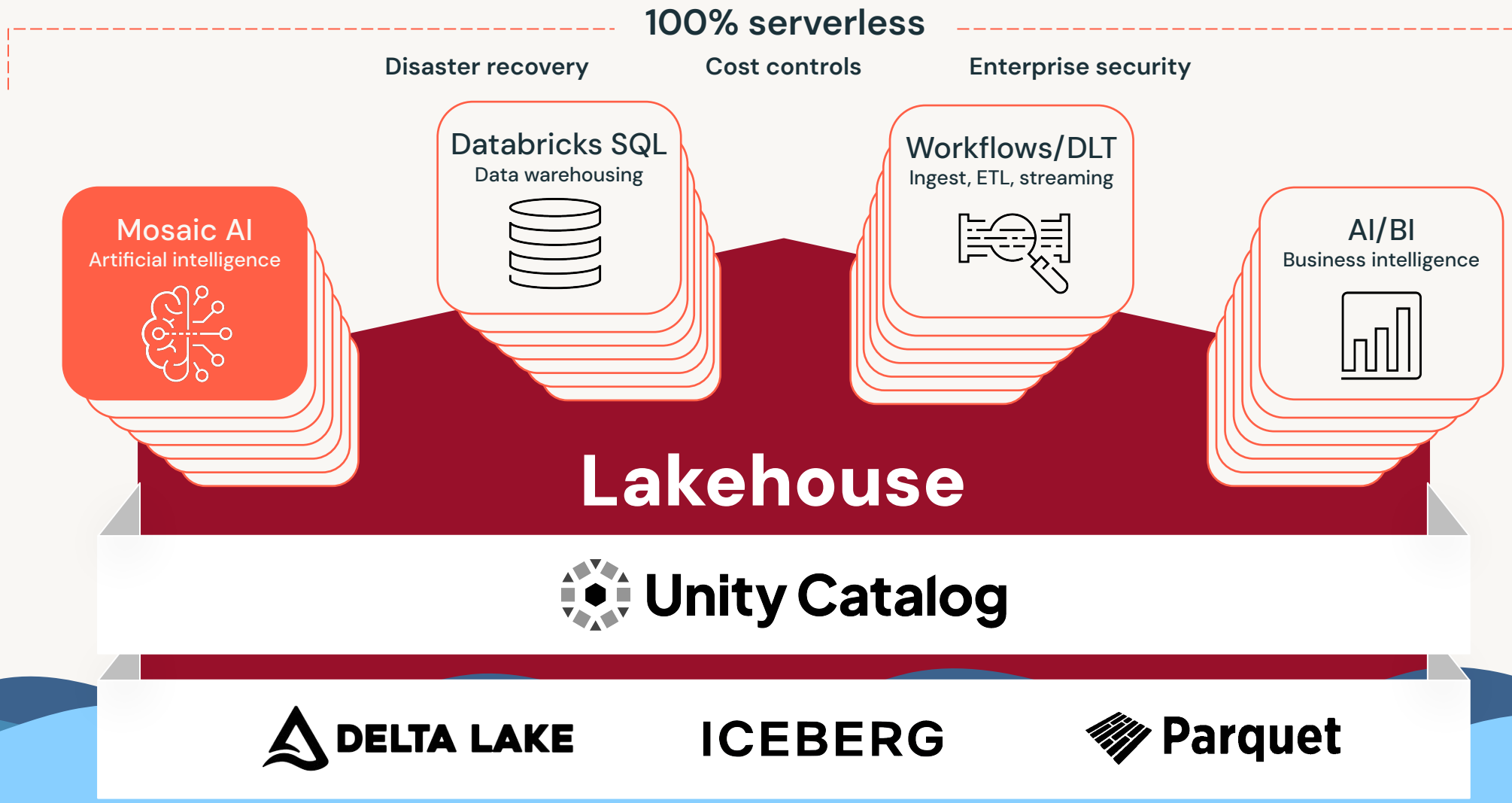
Databricks Data Intelligence Platform



Data intelligence is...

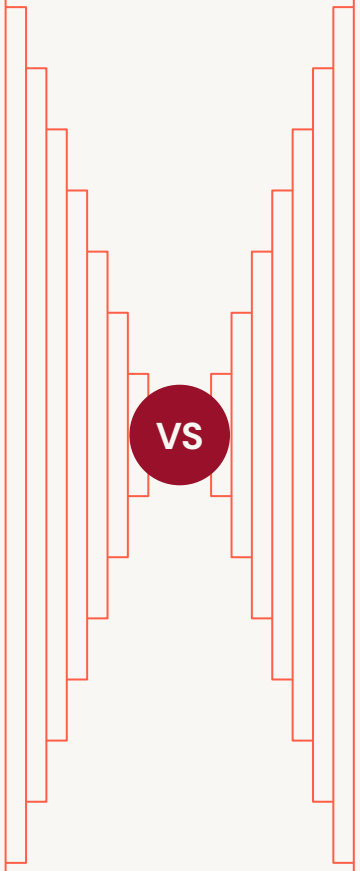
Democratized data + **Democratized AI**

Databricks Data Intelligence Platform



General intelligence

Consumer models trained on a broad dataset **disconnected** from your business data



Data intelligence

AI **connected** to your customer data and able to solve domain-specific problems

Factset is a Fortune 500 Financial Services company

FQL (Factset Query Language) is a proprietary data retrieval language used to access FactSet data

```
FF_BASIC_DERIVED(
  FF_PRD_EPS(ANN_R, #ALL#, "0CY") AS
"EPS_Current_Year", FF_PRD_EPS(TTM_R,
#ALL#, "0CY") AS "EPS_TTM" )
WHERE FF_SECURITY_TYPE("#ALL#") IN ("EQ")
AND FF_COUNTRY("#ALL#") IN ("US")
```

FQL

Problem

Goal

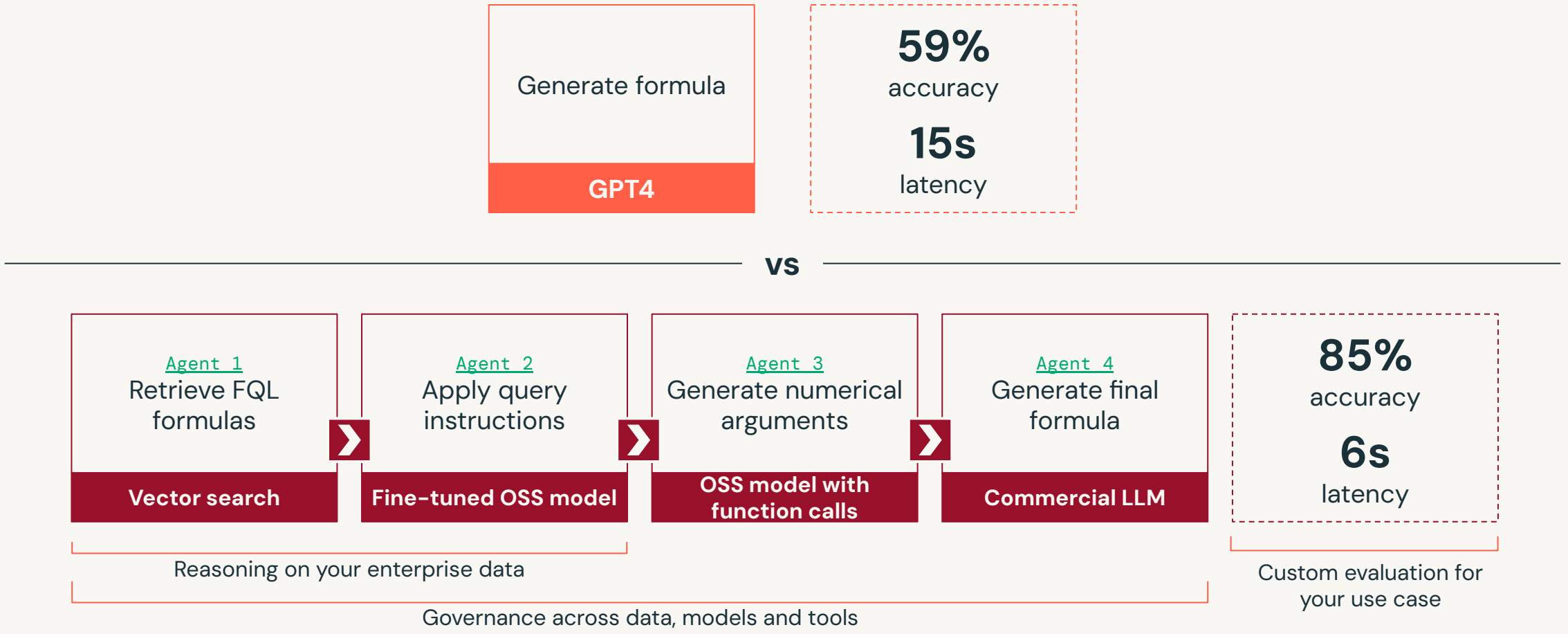
Factset wanted GenAI to take English queries and render the FQL for them

"Give me the current year and trailing EPS for all US-listed equities"

English equivalent

General intelligence does not work for enterprise use cases

Agents with data intelligence delivered production quality

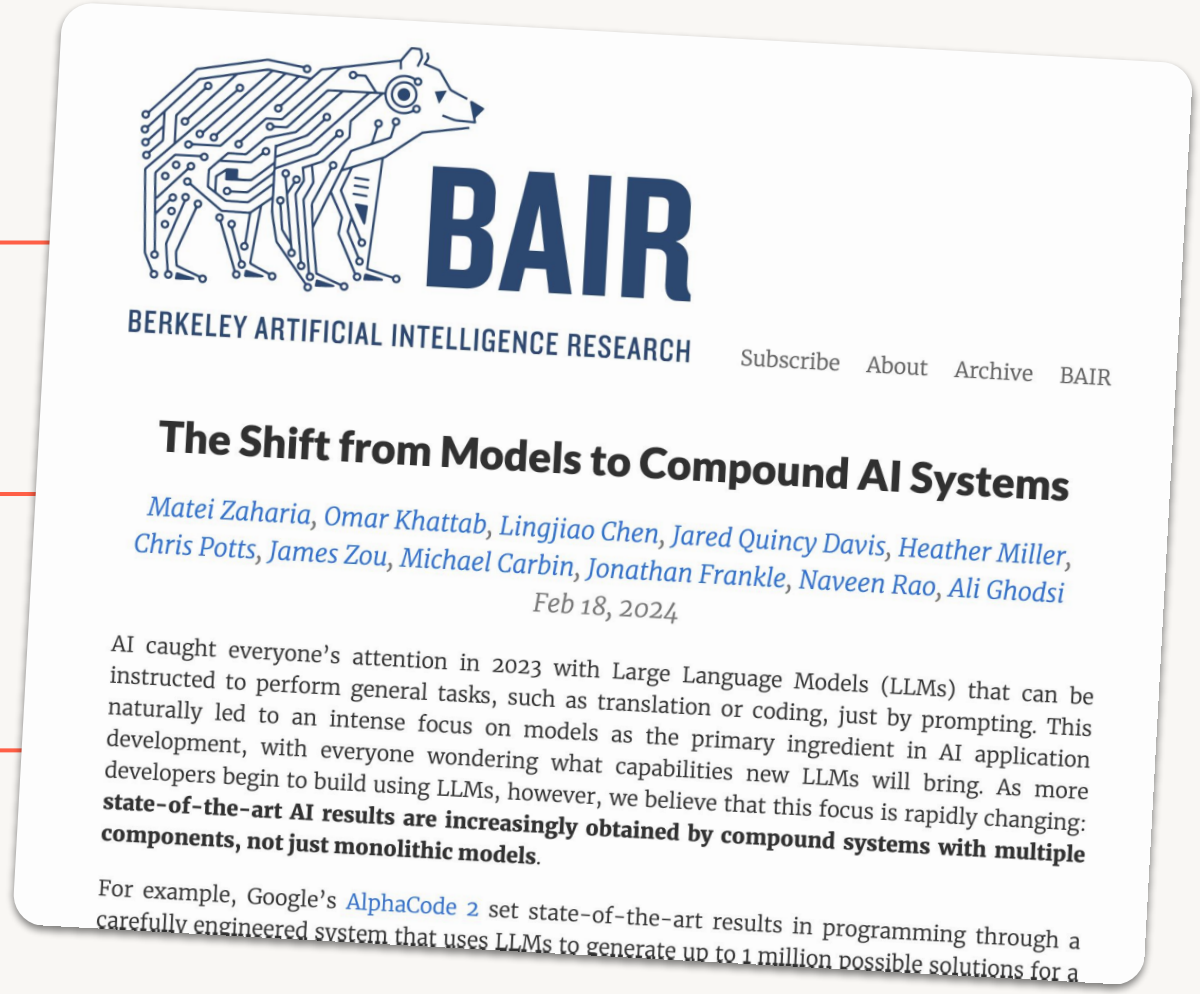


Data intelligence with an agent system

Reasoning on your enterprise data

Custom evaluation for your use case

Governance on data, models and tools



BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Subscribe About Archive BAIR

The Shift from Models to Compound AI Systems

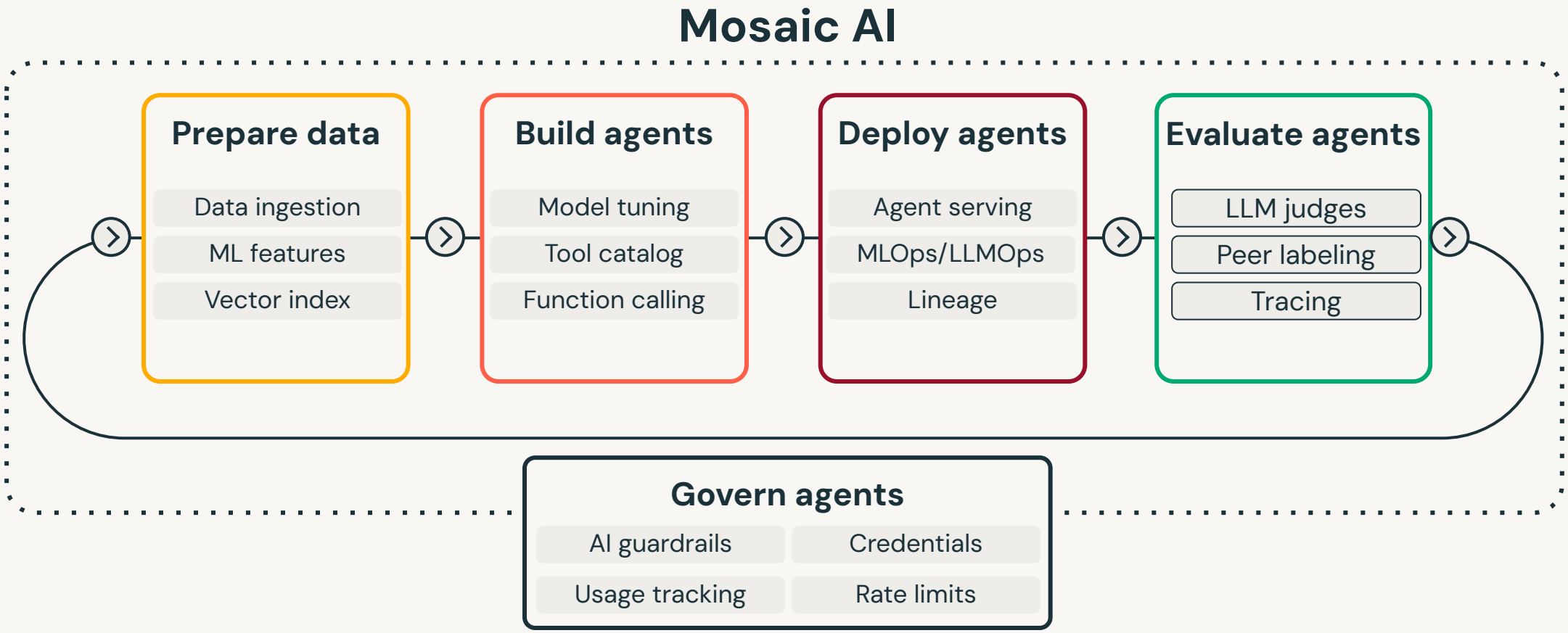
Matei Zaharia, Omar Khattab, Lingjiao Chen, Jared Quincy Davis, Heather Miller, Chris Potts, James Zou, Michael Carbin, Jonathan Frankle, Naveen Rao, Ali Ghodsi

Feb 18, 2024

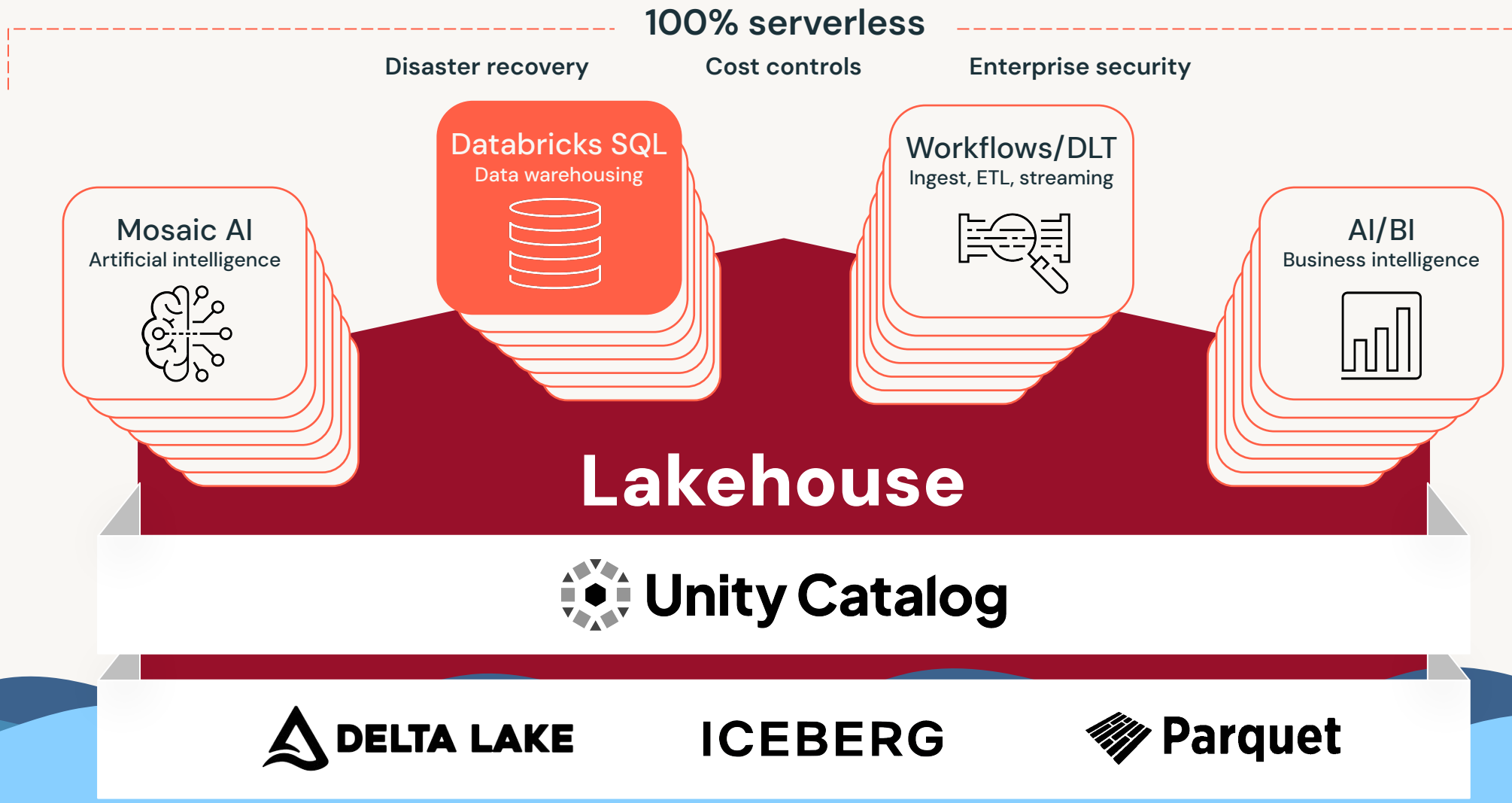
AI caught everyone's attention in 2023 with Large Language Models (LLMs) that can be instructed to perform general tasks, such as translation or coding, just by prompting. This naturally led to an intense focus on models as the primary ingredient in AI application development, with everyone wondering what capabilities new LLMs will bring. As more developers begin to build using LLMs, however, we believe that this focus is rapidly changing: **state-of-the-art AI results are increasingly obtained by compound systems with multiple components, not just monolithic models**.

For example, Google's AlphaCode 2 set state-of-the-art results in programming through a carefully engineered system that uses LLMs to generate up to 1 million possible solutions for a

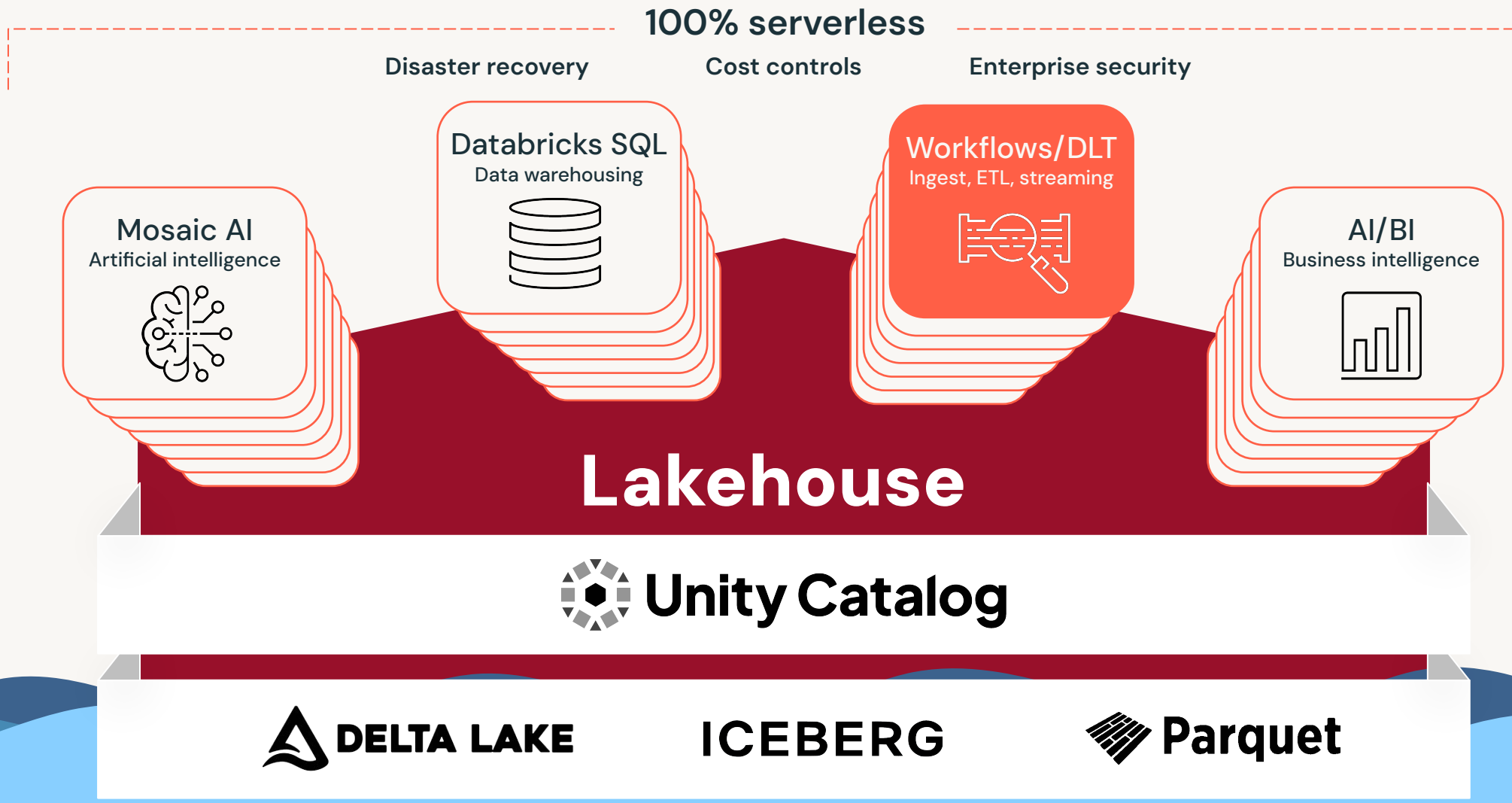
Build agent systems with Databricks



Databricks Data Intelligence Platform



Databricks Data Intelligence Platform



LakeFlow Connect

Native ingestion connectors for SaaS applications and databases

Simple and low maintenance

No code, quicker time to value, democratized data

Lakehouse-first design

Secure and healthy pipelines that live where you do your work

End-to-end efficiency

Lower costs, better performance, better scalability

CONNECTORS

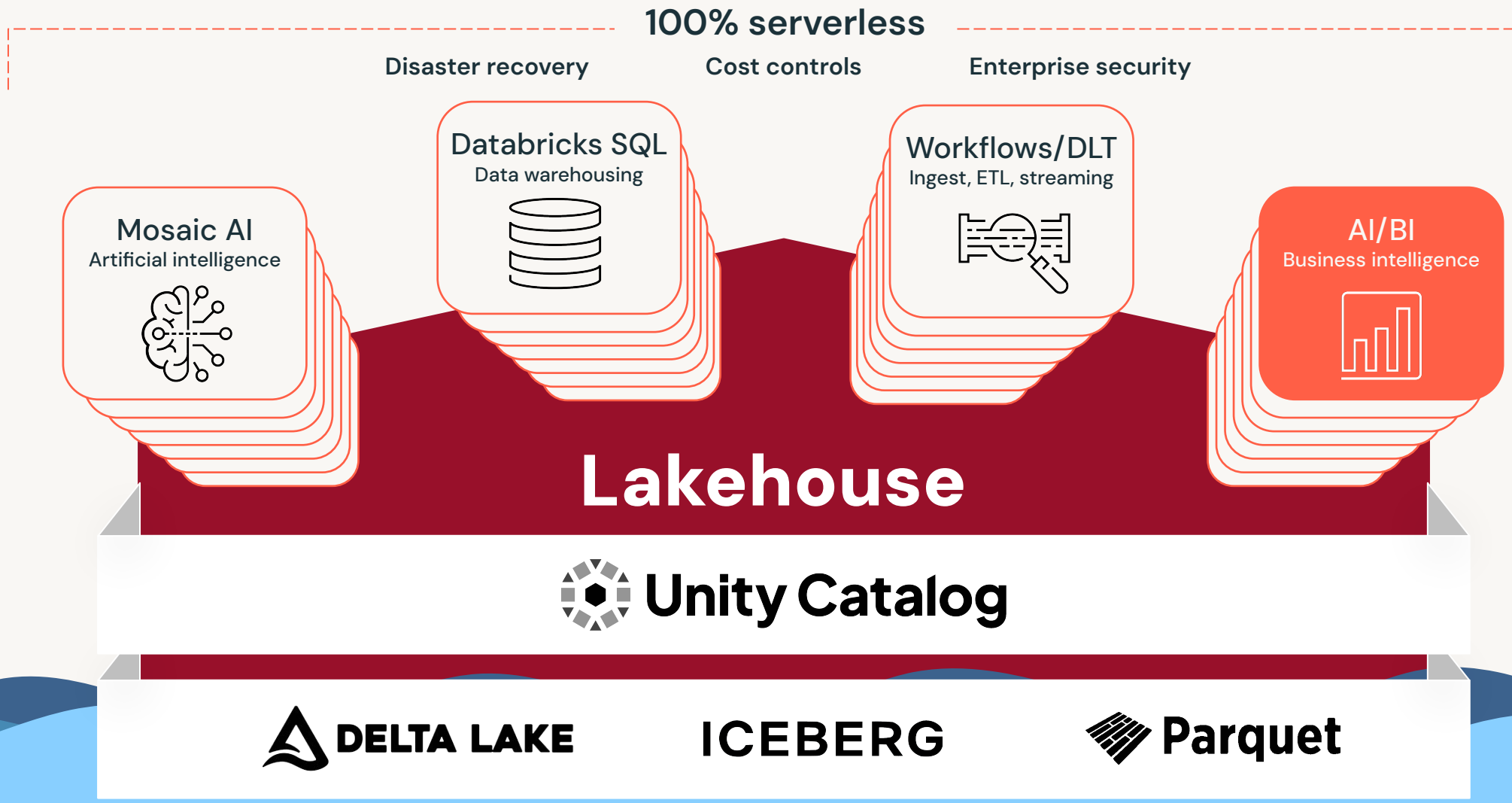
Applications

salesforce, servicenow, SharePoint, Google Analytics, Dynamics 365, ORACLE NETSUITE, workday, Google Ads, Meta

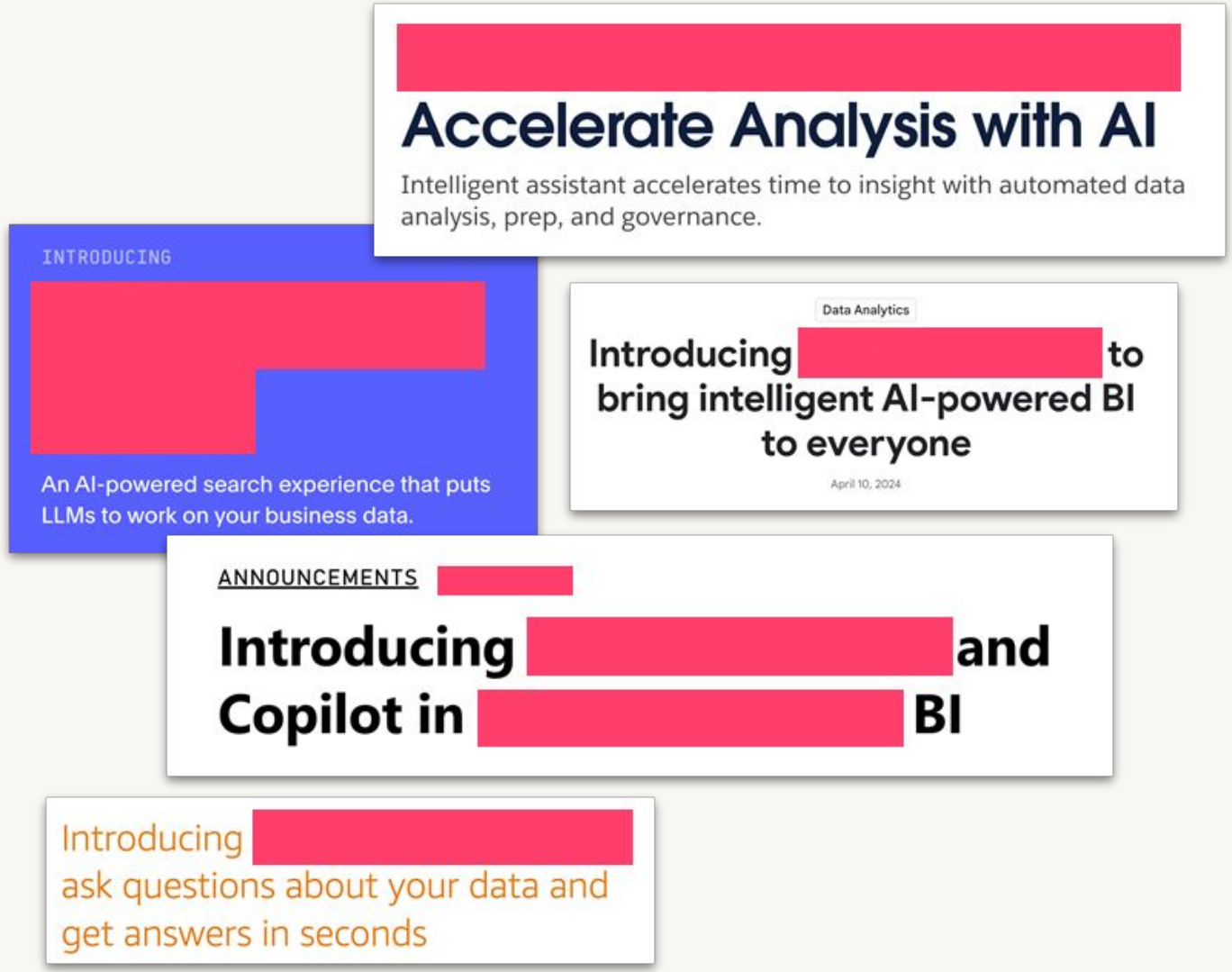
Databases

Microsoft SQL Server, IBM DB2, ORACLE DATABASE, MySQL, amazon DynamoDB, PostgreSQL, mongoDB

Databricks Data Intelligence Platform



Everyone is looking to Gen AI to enable self-service analysis



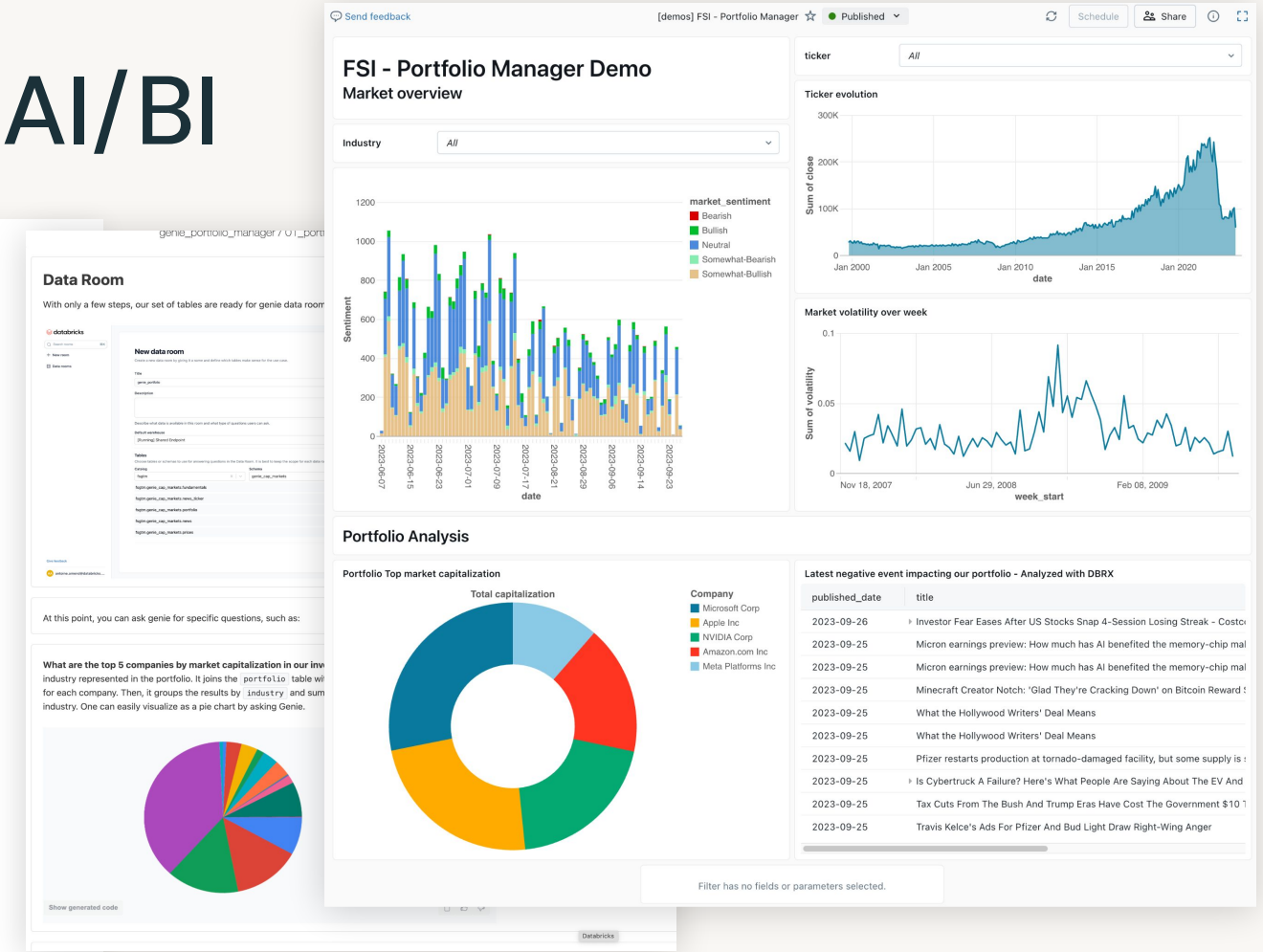
DEMO

FSI - Portfolio Assistant AI/BI

This leverages 5 datasets from the marketplace, news and daily stock prices

Genie Questions:

- How diversified is my portfolio by market cap?
- Visualize as a pie chart
- What are the top 5 companies by market capitalization in our investment portfolio?
- How has the stock price of Apple changed over time?
- Was there a stock split for AAPL in the 2014 2015 timeframe?
- Show me the market volatility for Technology companies during the financial crisis by week
- What is the market sentiment for companies in the retail industry?
- Visualize as a bar chart
- What was the top 10 most recent negative event for banking companies in my portfolio
- How has the market sentiment changed by week for my portfolio?

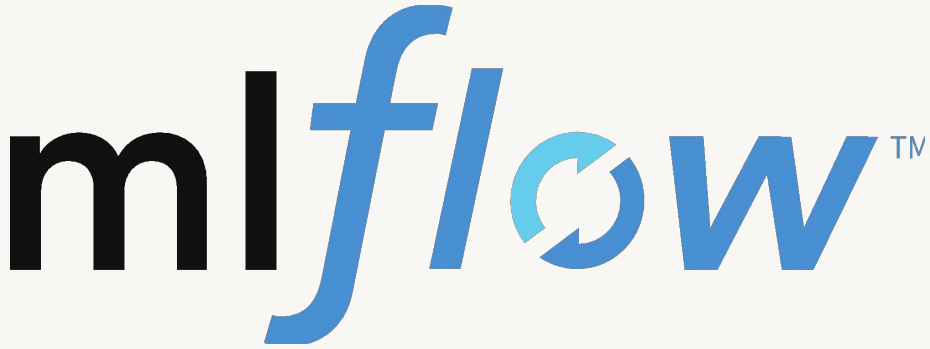



```
%pip install dbdemos
import dbdemos
dbdemos.install('aibi-sales-pipeline-review')
```

Genie Space

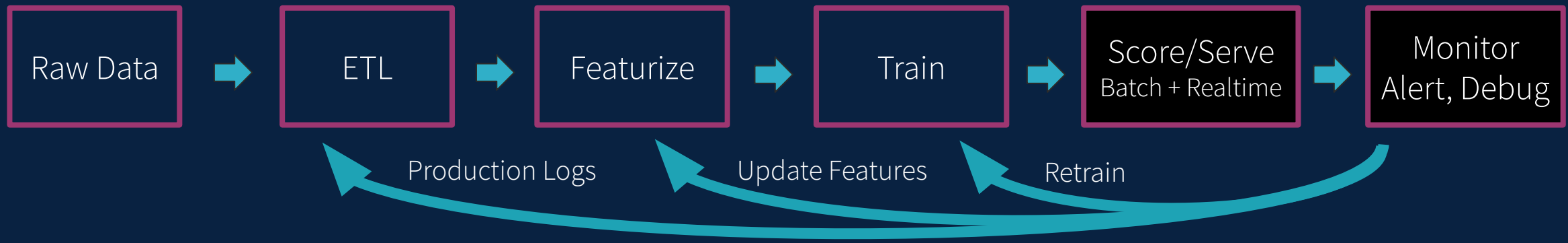
Dashboard

Catalog access



ML Lifecycle and **Challenges**

mlflow
An open source platform for the machine learning lifecycle



Tuning | Deploy | Model Mgmt

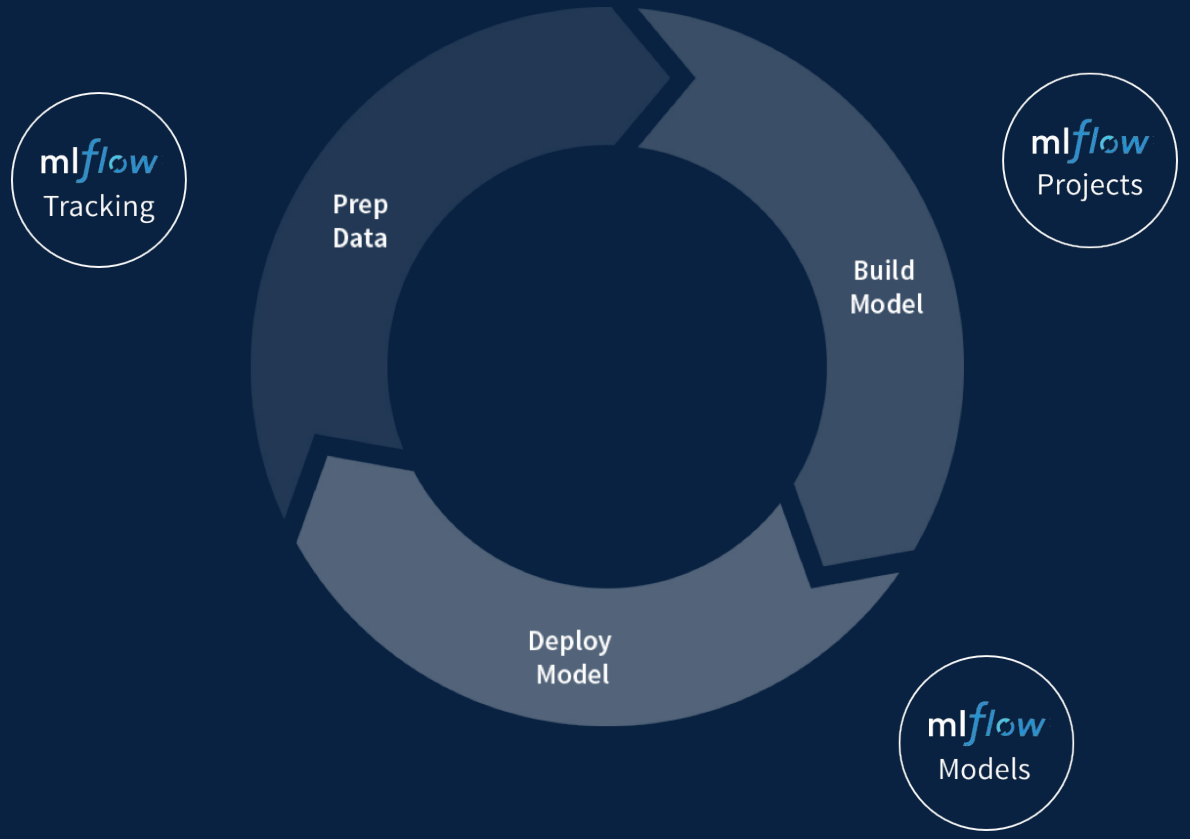
Collaboration | Scale | Governance

Feature Repository	Experiment Tracking	AutoML, Hyper-p. search	Remote Cloud Execution	Project Mgmt (scale teams)	Model Exchange	A/B Testing	CI/CD/Jenkins push to prod	Orchestration (Airflow, Jobs)	Lifecycle mgmt.	Data Drift	Model Drift

mlflow

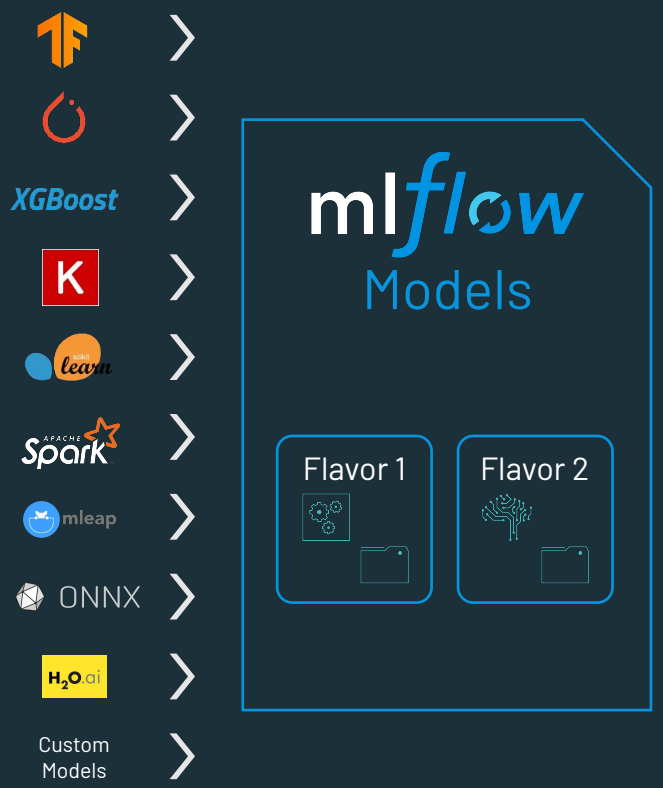
Introducing MLflow

Unveiled in June 2018, MLflow is the only open source framework designed to manage the complete Machine Learning Lifecycle.



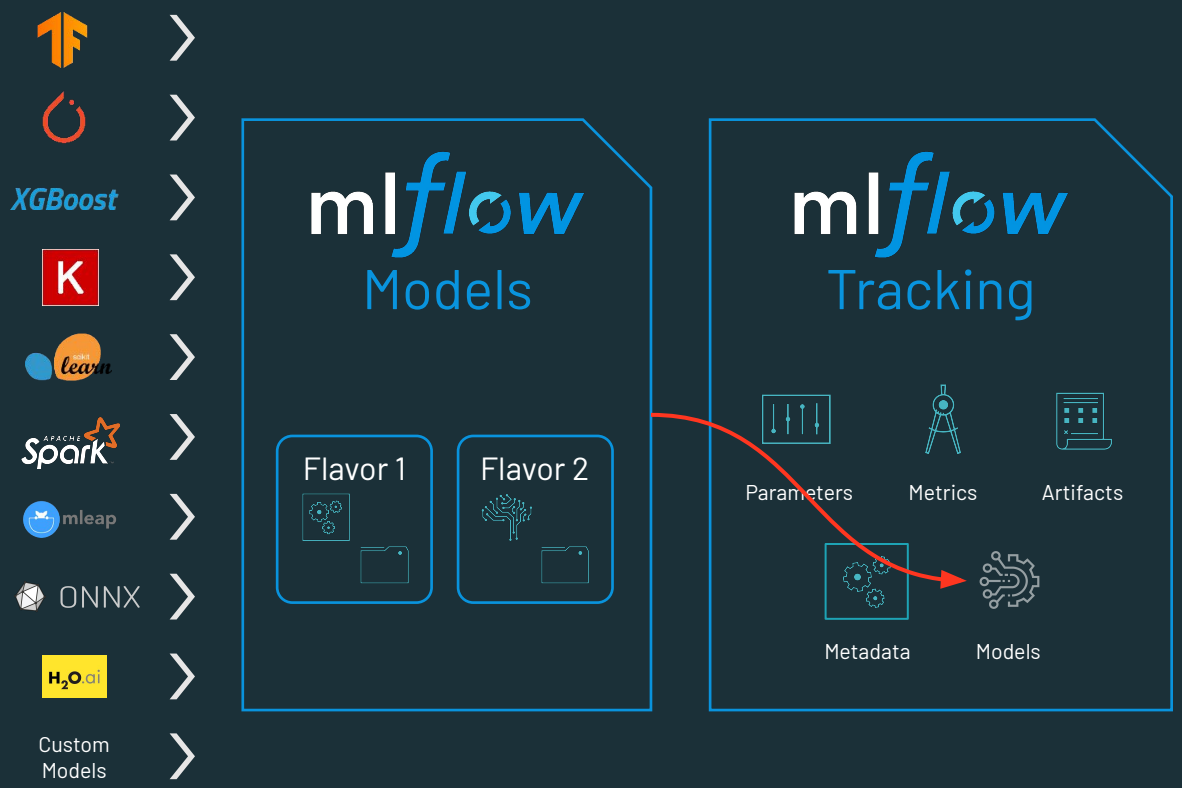
mlflow

mlflow Model Lifecycle



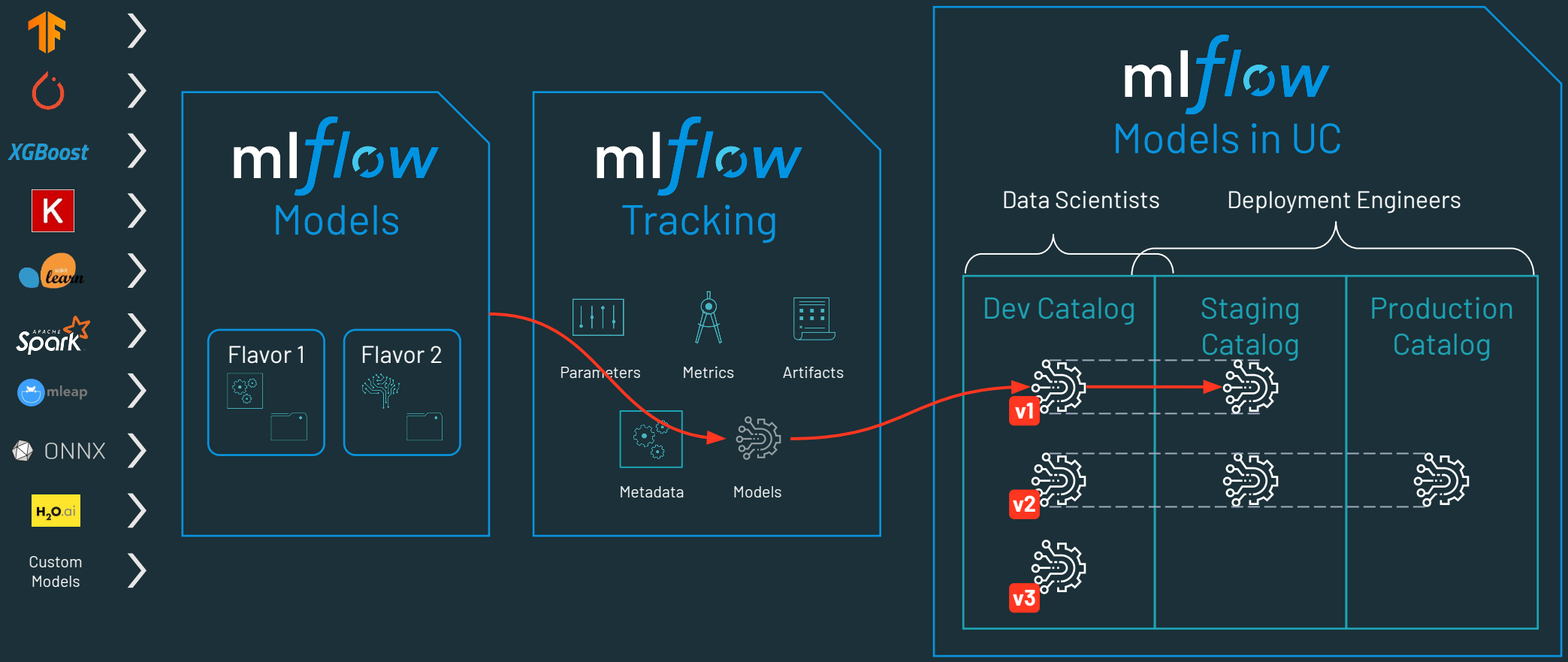
databricks

mlflow Model Lifecycle



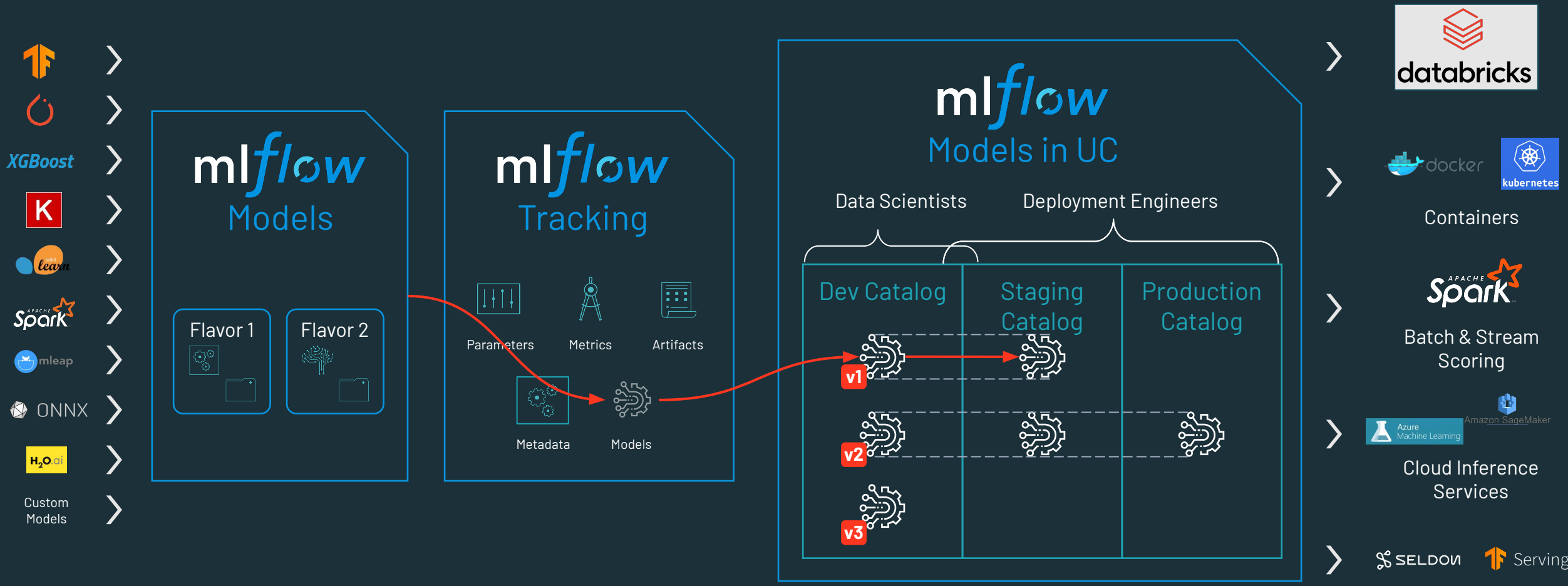
databricks

mlflow Model Lifecycle

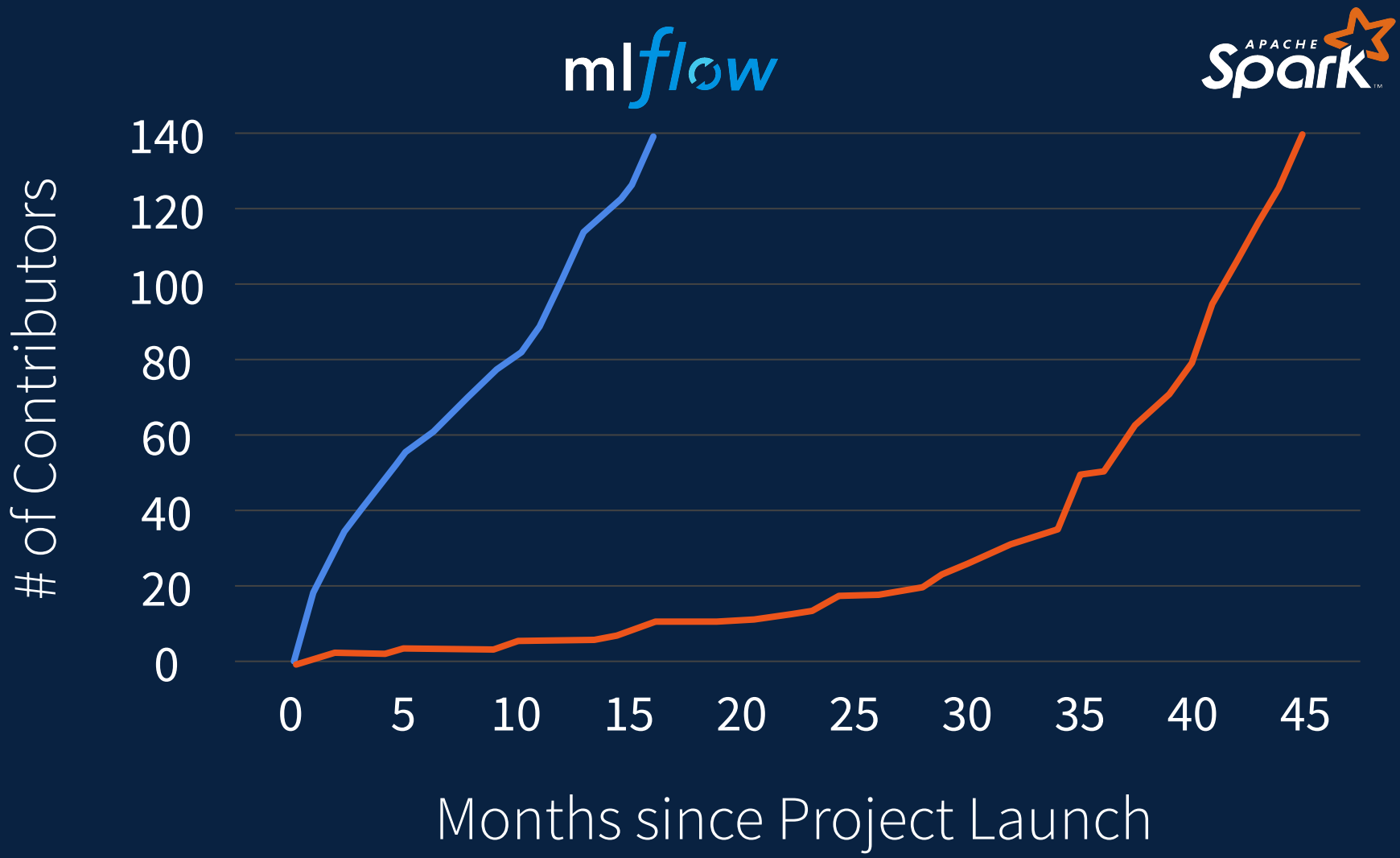


databricks

mlflow Model Lifecycle



databricks


mlflow
APACHE Spark
# of Contributors
140
120
100
80
60
40
20
0
0
5
10
15
20
25
30
35
40
45
Months since Project Launch

Use Case Video

RIVIAN

Challenge

With over 11,000 EAVs on the road, generating terabytes of IoT data daily, Rivian struggled to scale their legacy cloud tooling and spent significant resources on maintenance — slowing their ability to be truly data driven.

Solution

Databricks Lakehouse standardizes all vehicle data — such as core telemetry data, connected car systems and perceptions systems — so they drive holistic insight into everything that happens with vehicles after they leave assembly.

Impact

$37M+

from improved vehicle reliability, improved battery and charging system efficiency and faster delivery of autonomous driving features


zipline

zipline

Challenge

Every Zipline flight generates a gigabyte of data with potential life-or-death consequences. But accessing and federating the data for both internal and external decision making was easier said than done before Databricks, as they didn't have an efficient way of harnessing and sharing the data across the organization and their supply chain partners.

Solution

With Databricks, Zipline's data team and their colleagues throughout the company are able to access all of the information they need to accurately measure success, find the metrics that relate to customer experiences or logistics, and improve on them exponentially as more data is ingested and machine learning models are refined.

Impact

44%

Faster time-to-market for predictive maintenance solutions

50%

Less time to gather insights to monitor traffic congestion

10

Months to develop 30 analytics solutions, compared to 1.5 years

databricks
DATA+AI SUMMIT

WORLD'S LARGEST
DATA, ANALYTICS
AND AI CONFERENCE

IN-PERSON
JUNE 9-12
SAN FRANCISCO

REGISTRATION OPEN

20K+
ATTENDEES
ON SITE

700+
BREAKOUT
SESSIONS

20+
TRAINING
SESSIONS

81%
Larger ASP
for accounts who attend*

$383M
Pipeline Influenced in 2024

950+
Exec Forum Attendees in 2024

14%
Higher Win Rate
for accounts who attend

Virtual Event // June 11 - 12 // Free live stream of keynotes + select sessions