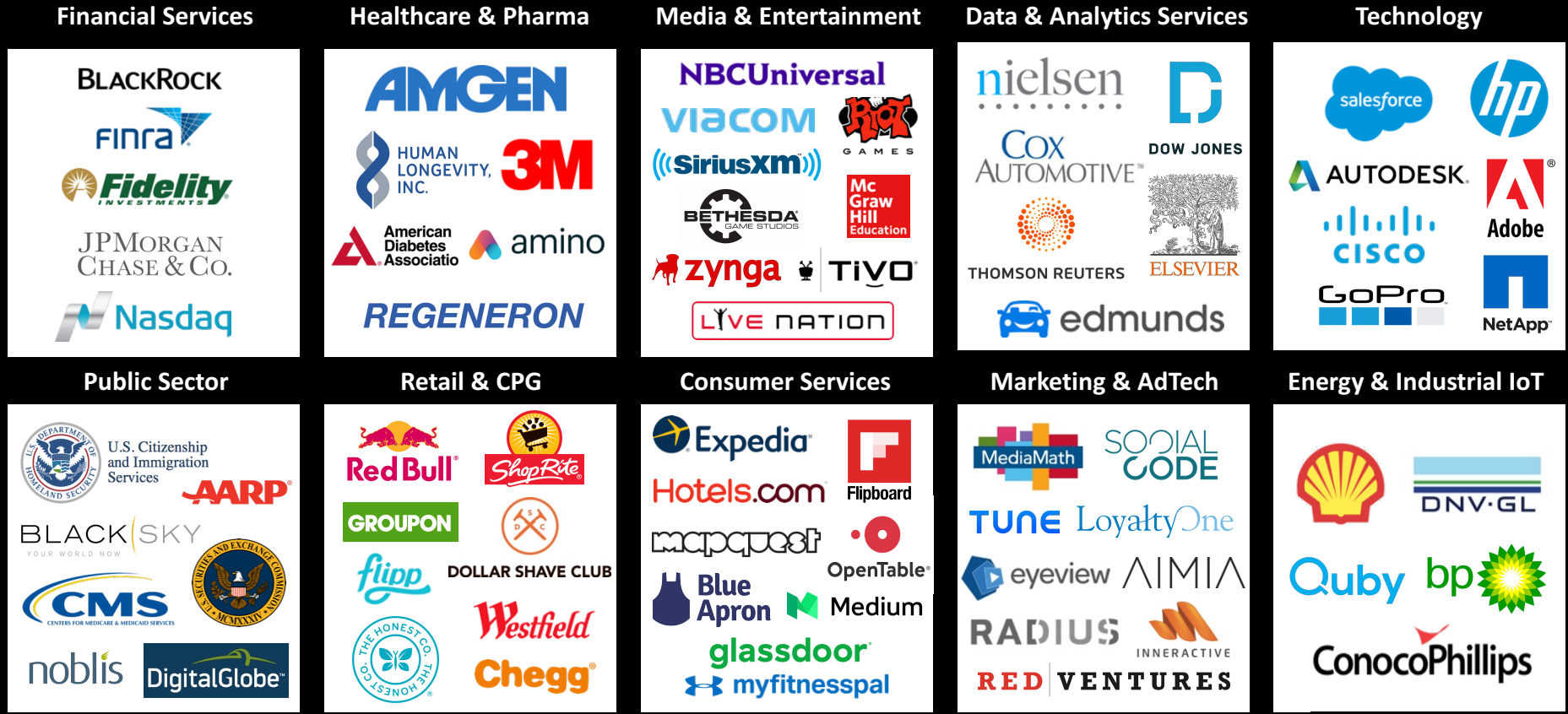
Unifying Data Warehousing with Data Lakes

Ali Ghodsi, Co-Founder & CEO

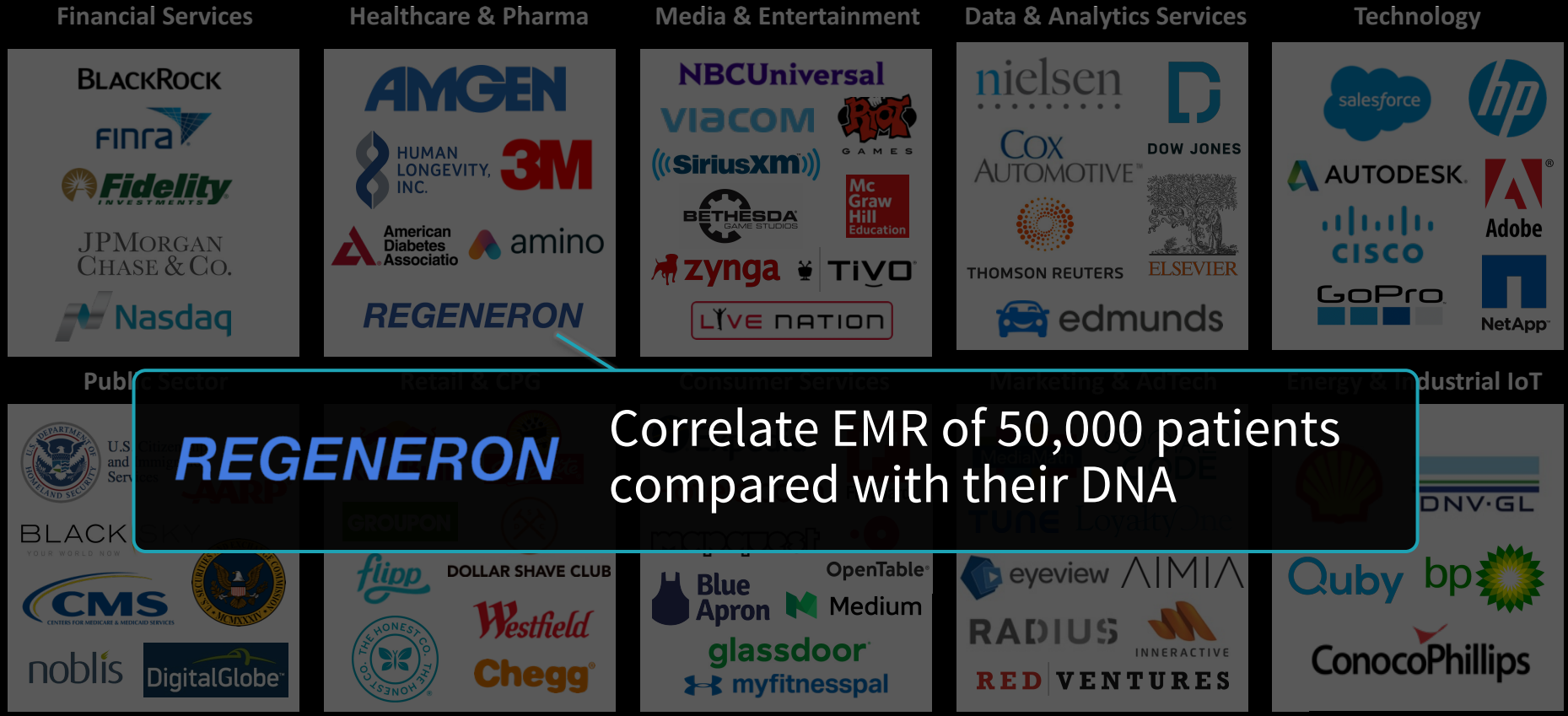
Oct 25, 2017

Many enterprises are undergoing a data transformation

Databricks Customers Across Industries



Health care AI cloud dataset use case



Enterprise AI use case

Financial Services

Healthcare & Pharma

Media & Entertainment

Data & Analytics Services

Technology

Public Sector

Retail & CPG

Consumer Services

Marketing & Advertising

Energy & Industrial IoT

Real-time AI use-case

Financial Services

Healthcare & Pharma

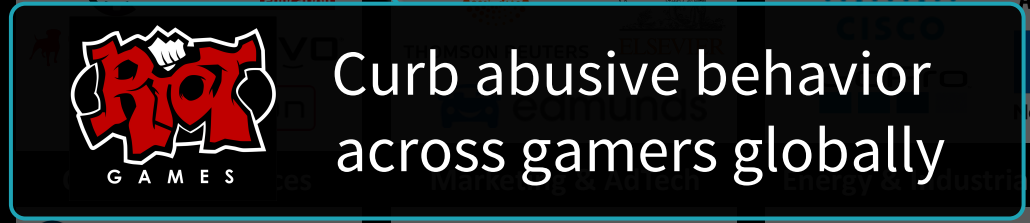
Media & Entertainment

Data & Analytics Services

Technology

Public Sector

Retail & CPG



Big Data was the Missing Link for AI

BIG DATA

Customer Data

Emails/Web pages

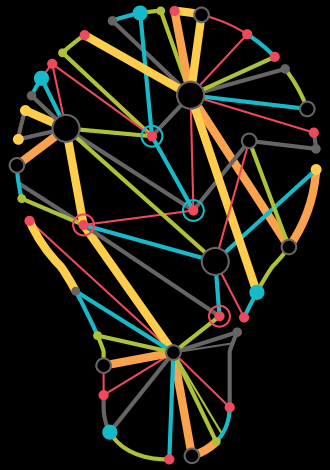
Click Streams

Sensor data (IoT)

Video/Speech

…

GREAT RESULTS



Most companies are Struggling with Big Data

Hardest part of AI isn't AI

"Hidden Technical Debt in Machine Learning Systems", Google NIPS 2015

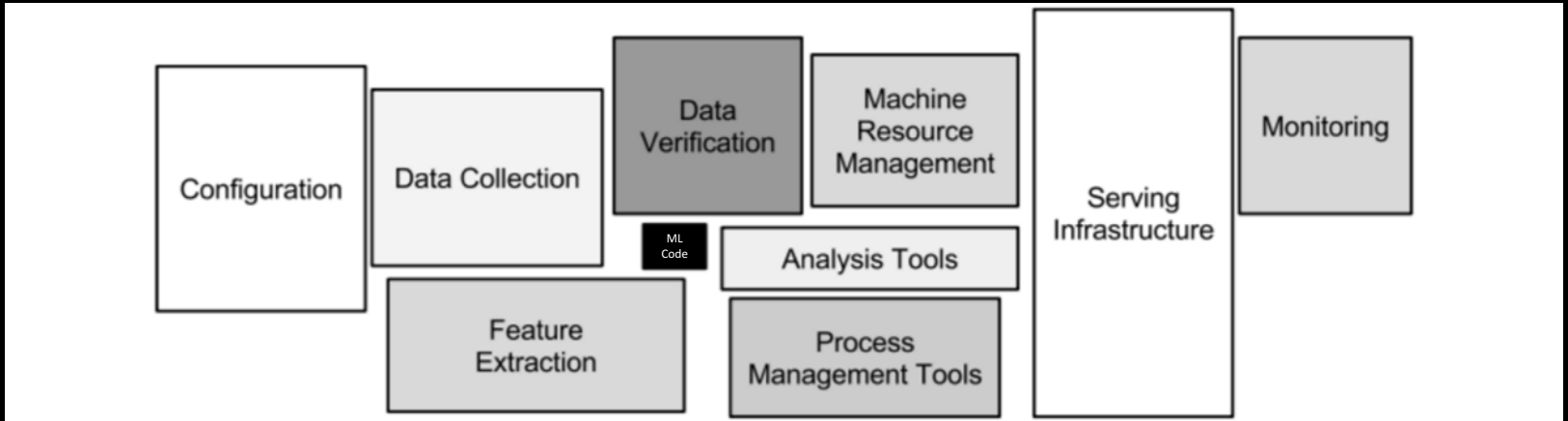


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

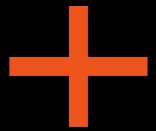
The hardest part of AI is Big Data

Building Predictive Applications is really Hard!

databricks

Unified Analytics Platform

UNIFIED
PROCESSING
ENGINE



UNIFIED
EXPERIENCE
ACROSS TEAMS

The Evolution of Big Data

The Era of the Data Warehouse

Data Warehouse (DW)

ETL important data to central DW and get Business Intelligence (BI)

THE GOOD

- Pristine Data
- Fast Queries
- Transactional

THE BAD

- Expensive to Scale, not Elastic
- Requires ETL, Stale Data, No Real-Time
- No Predictions, No ML
- Closed formats (lock in)

Not Future Proof – Missing Predictions, Real-time, Scale

The Era of the Data Lake

Hadoop Data Lake

ETL all data to central scalable open lake for all use cases

THE GOOD

- Massive scale
- Inexpensive Storage
- Open Formats (Parquet, ORC)
- Promise of ML & Real Time Streaming

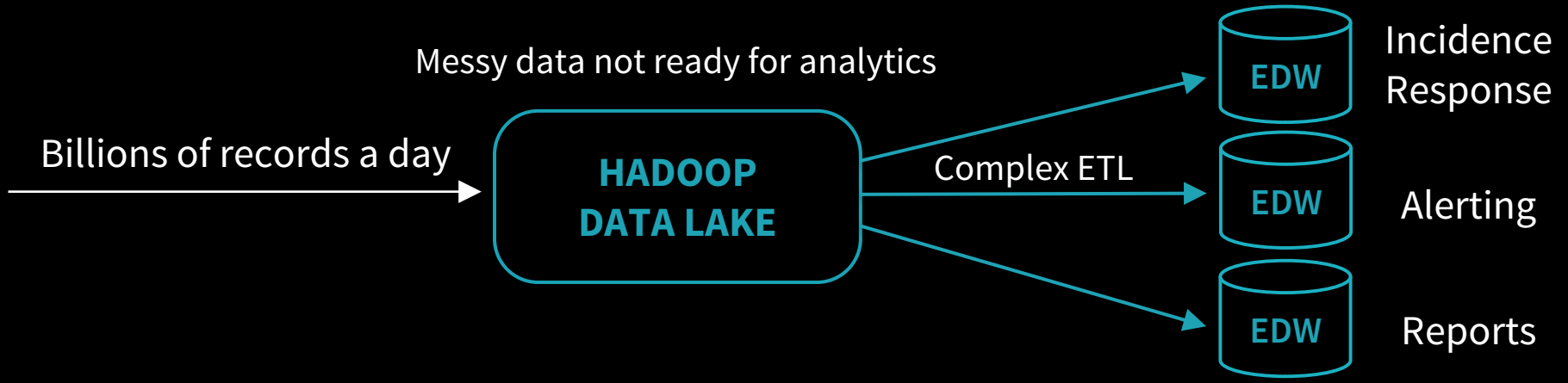
THE BAD

- Inconsistent Data
- Unreliable for Analytics
- Lack of Schema
- Poor Performance

Become a cheap messy data store with poor performance

The Current State of Data Platforms

Info Sec at a Fortune 100 Company



ENTERPRISE DATA WAREHOUSE

- Only 2 weeks of data
- Very expensive to scale
- Proprietary Formats
- No Predictions (ML)

DISADVANTAGES OF ARCHITECTURE

- Poor agility in responding to new threats
- Scale Limitations, no historical data
- 6 Months and twenty people to build

The Next Generation Data Platform

Announcing Databricks Delta

First **UNIFIED** data management system that delivers:

databricks


The
SCALE
of data lake
The
RELIABILITY &
PERFORMANCE
of data warehouse
The
LOW-LATENCY
of streaming

Databricks Delta

THE GOOD OF DATA LAKES

- Massive scale on Amazon S3
- Open Formats (Parquet, ORC)
- Predictions (ML) & Real Time Streaming

THE GOOD OF DATA WAREHOUSES

- Pristine Data
- Transactional Reliability
- Fast Queries (10-100x)

Enables Predictions, Real-time and Ad Hoc Analytics at Massive Scale

Databricks Delta Under the Hood

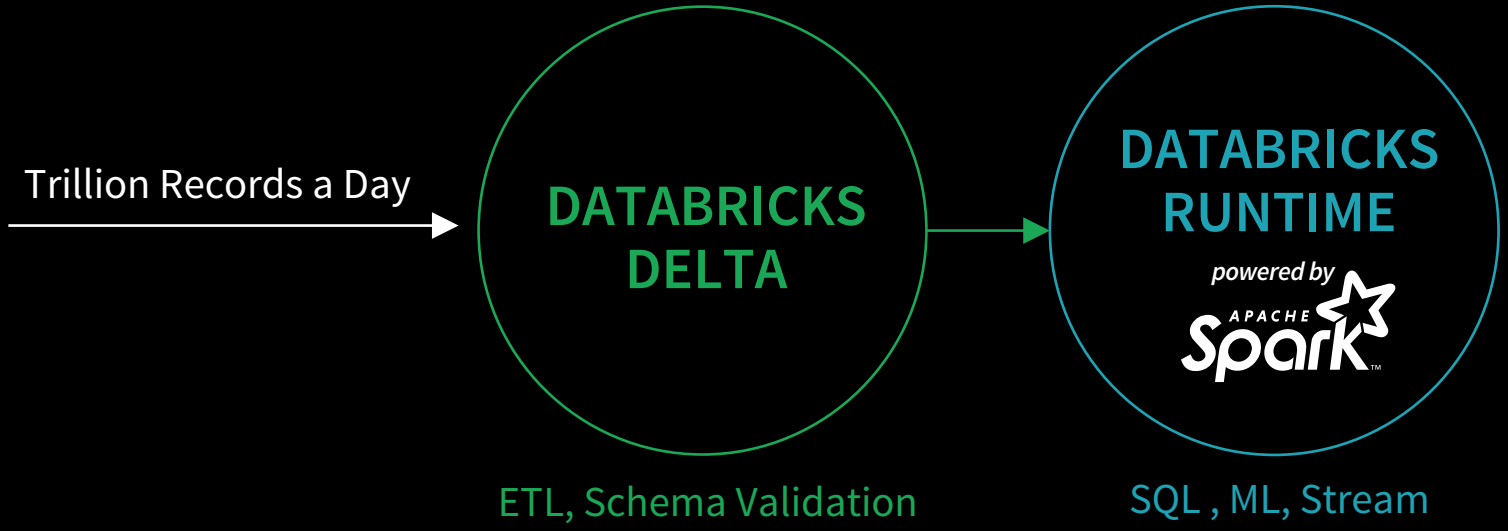
MASSIVE SCALE
- Decouple Compute & Storage

RELIABILITY
- ACID Transactions & Data Validation

PERFORMANCE
- Data Indexing & Caching (10-100x)

LOW-LATENCY
- Real-Time Streaming Ingest

Info Sec with Databricks Delta



ADVANTAGES

- AI capable data warehouse at the scale of a data lake
- Interactive analysis on 2 years of data
- 2 Weeks to build with a 5 person data platform team

databricks®

Unified Analytics Platform

UNIFIED
EXPERIENCE
ACROSS TEAMS

Notebooks, Dashboards, Reports

databricks®

Unified Analytics Platform

UNIFIED
DATA MANAGEMENT
Reliable Transactions, Performance

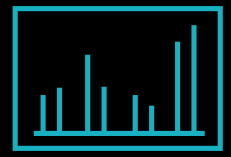
+

UNIFIED
EXPERIENCE ACROSS TEAMS
Notebooks, Dashboards, Reports

Demo by Michael Armbrust

Evolution of a Cutting-Edge Data Pipeline

Events

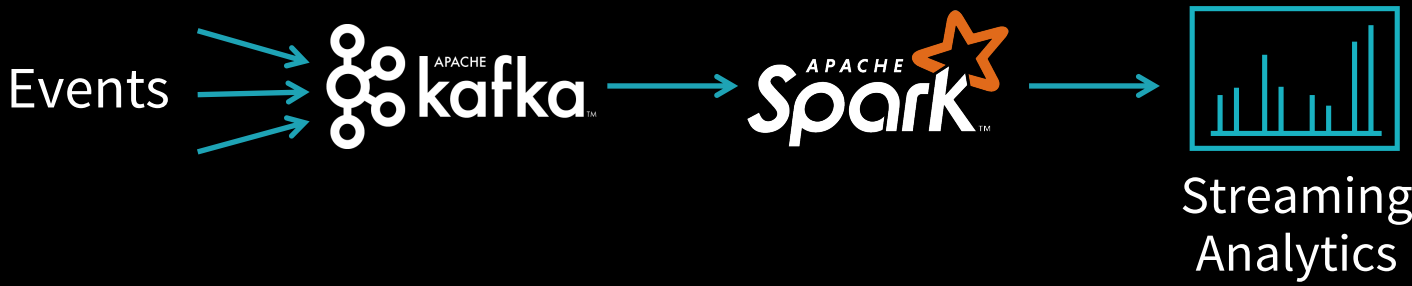


Streaming Analytics

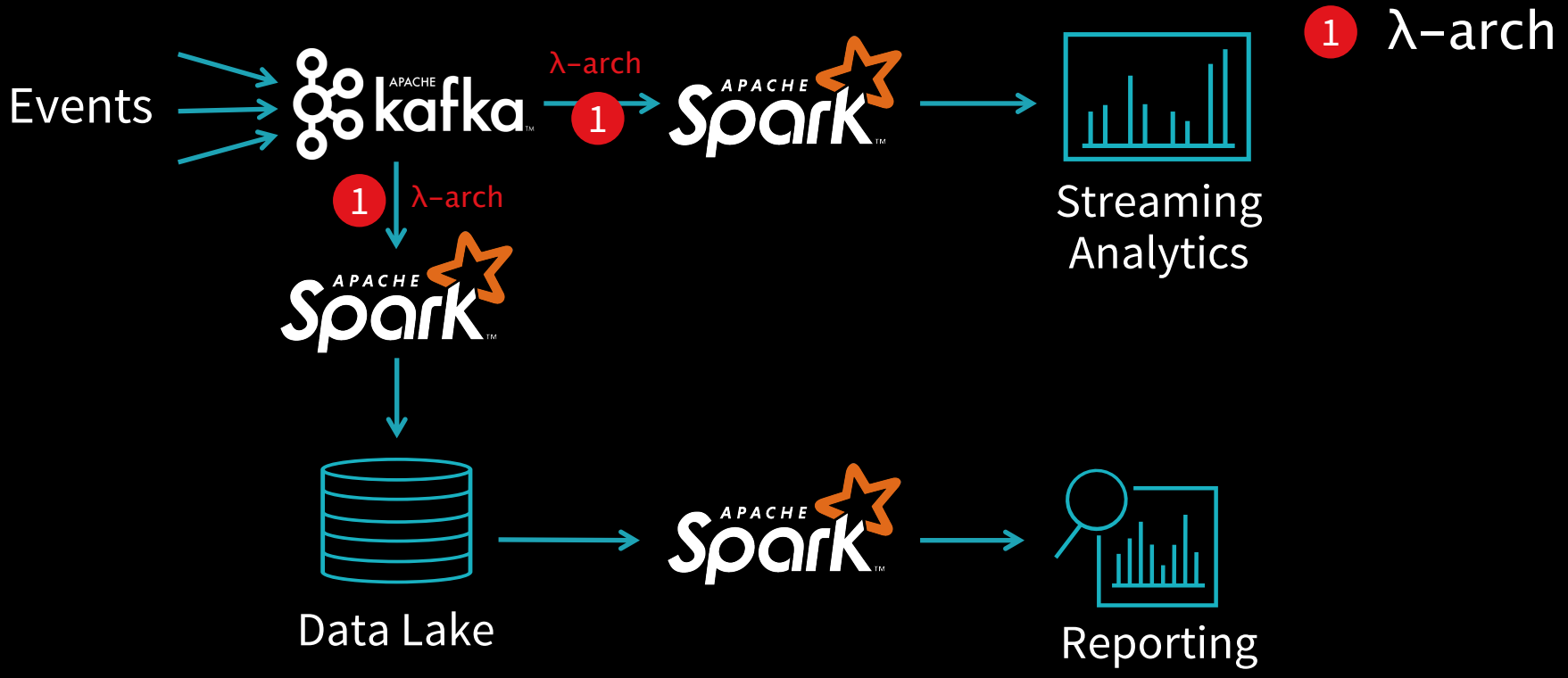
Data Lake

Reporting

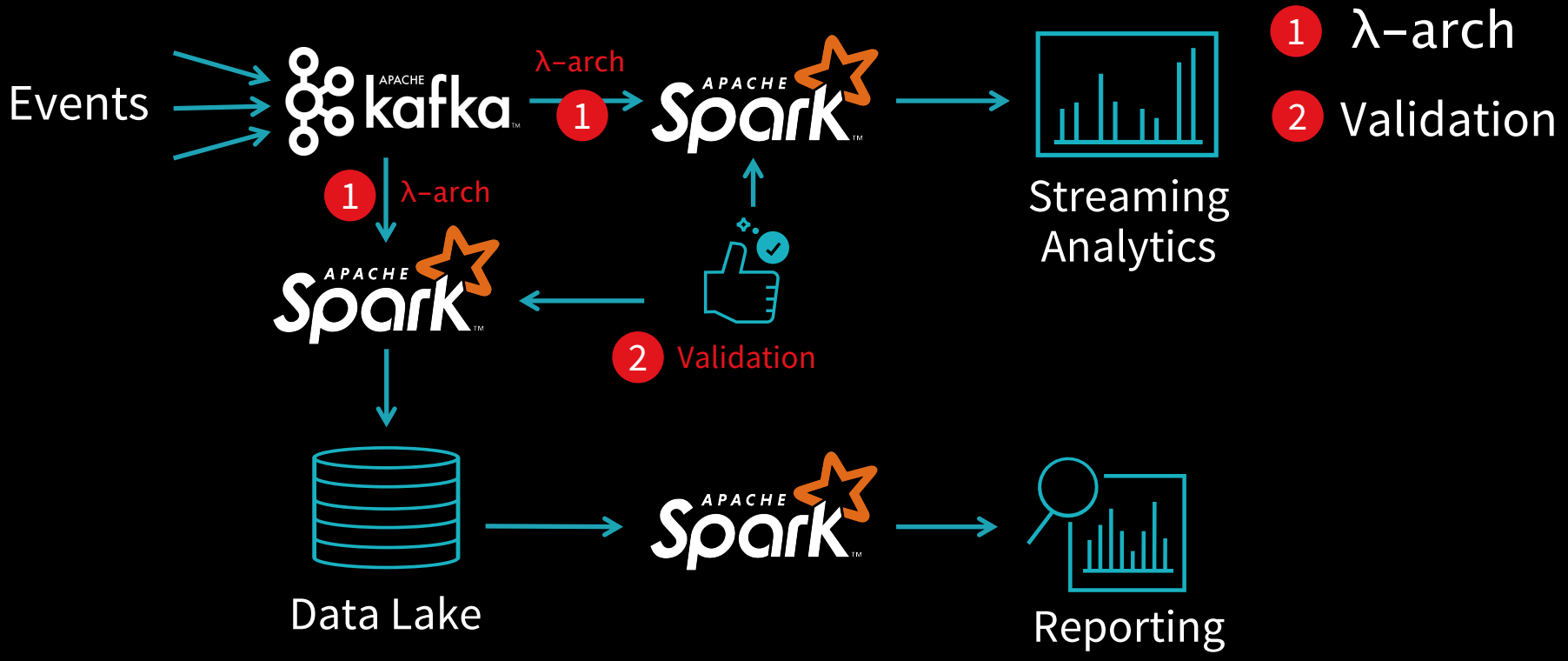
Evolution of a Cutting-Edge Data Pipeline



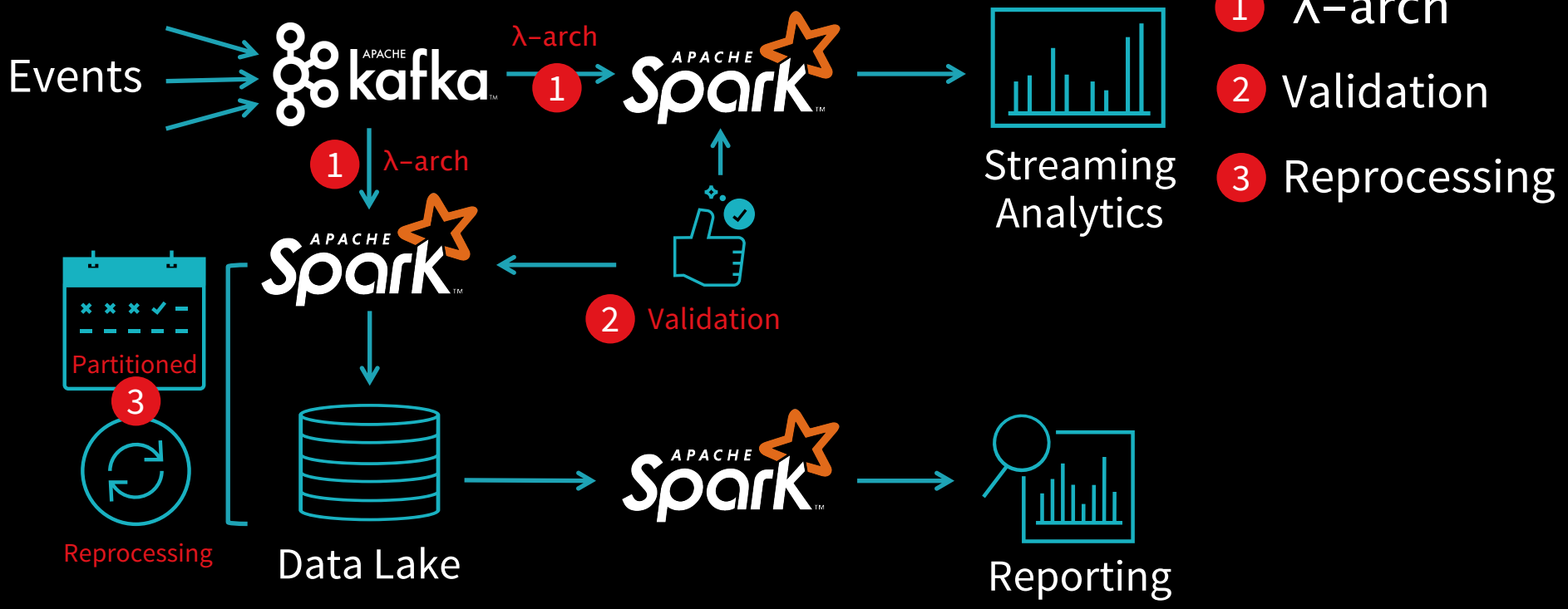
Challenge #1: Historical Queries?



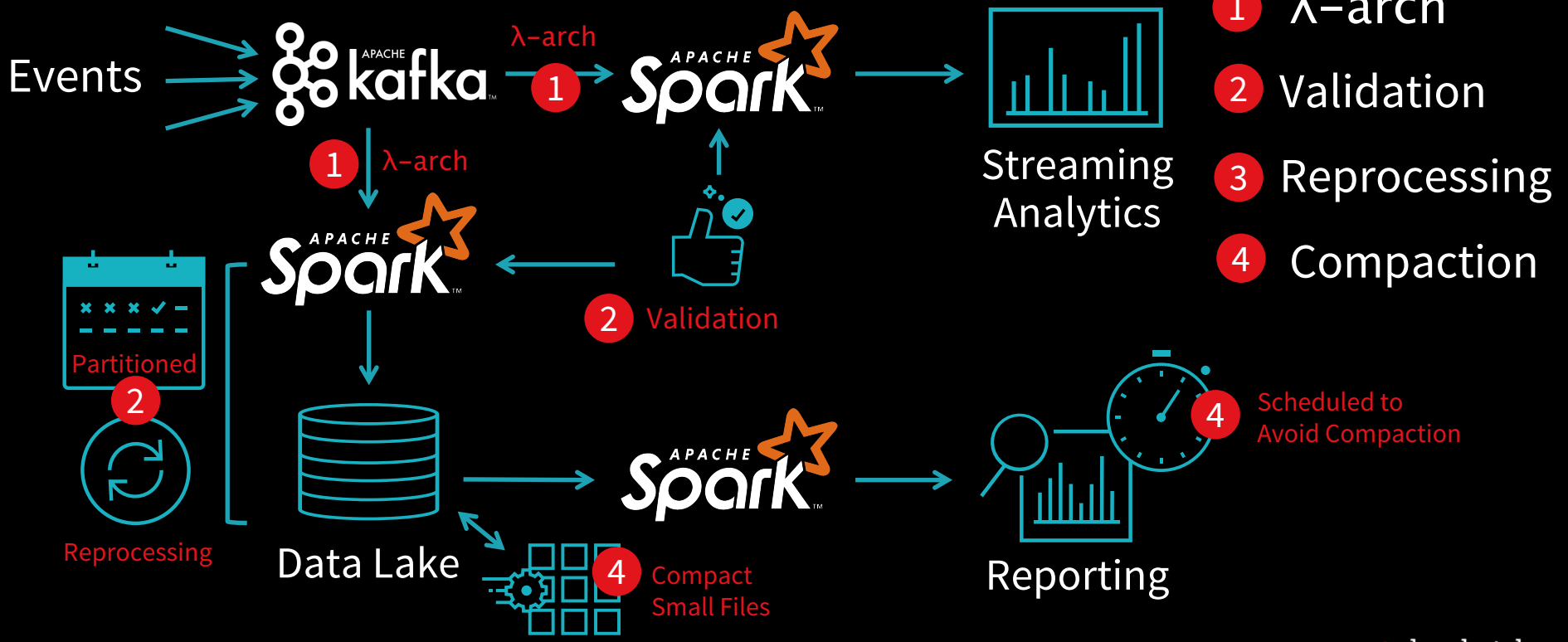
Challenge #2: Messy Data?



Challenge #3: Mistakes and Failures?

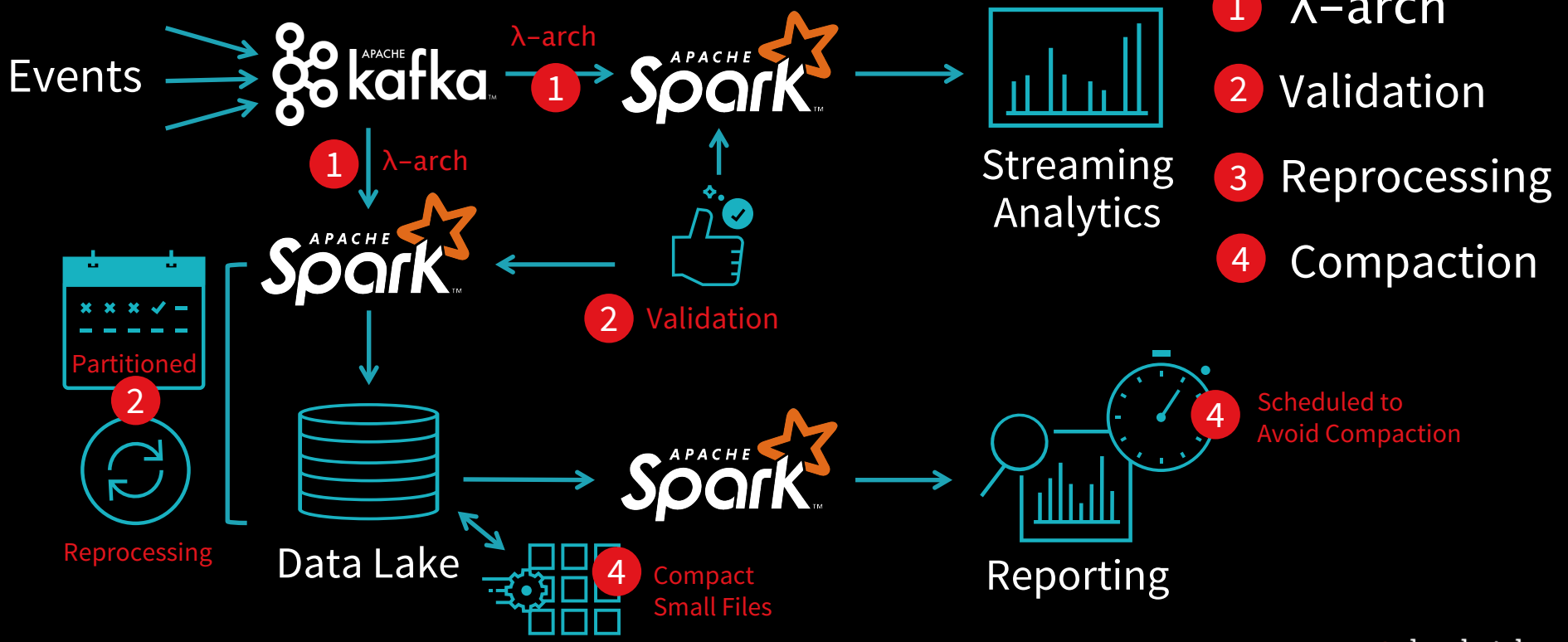


Challenge #4: Query Performance?



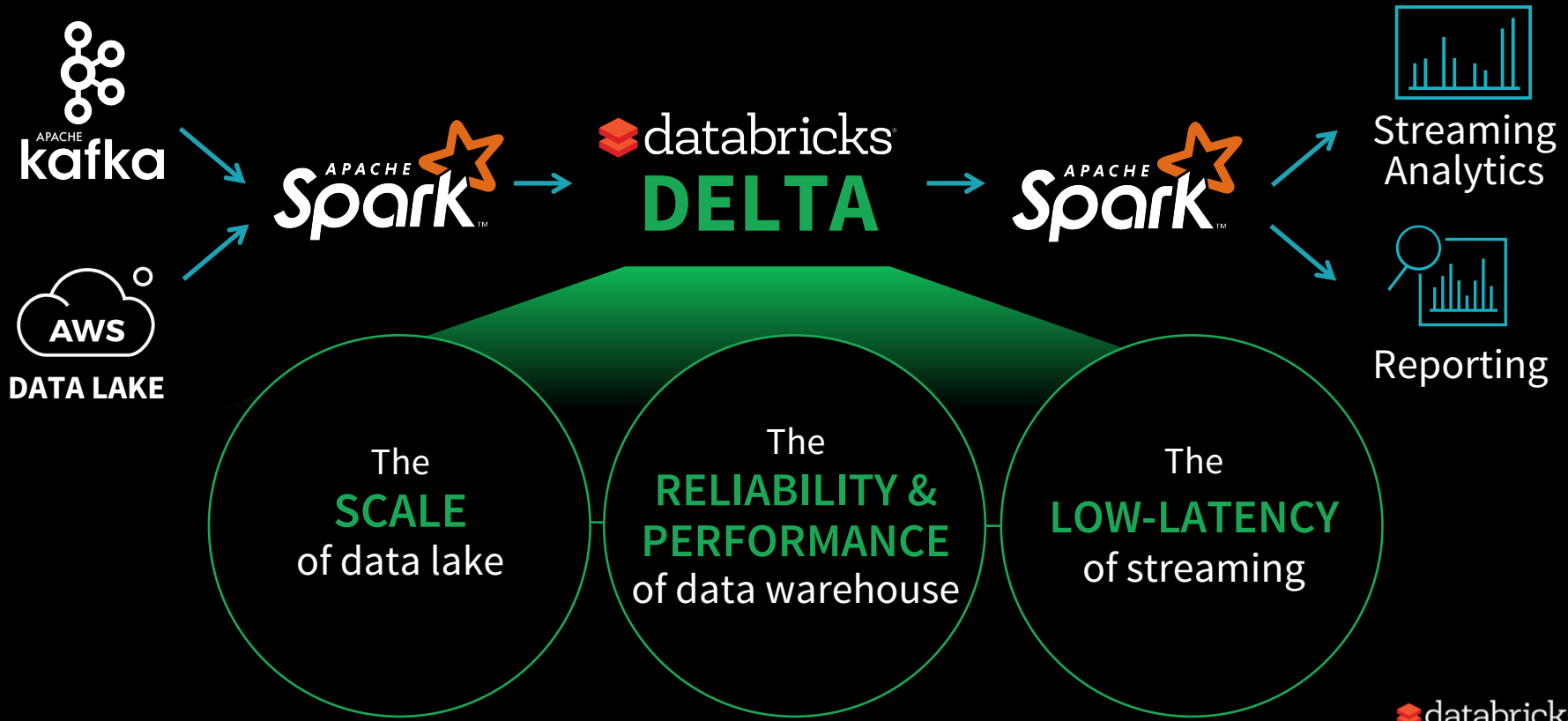
Let’s try it instead with **DELTA**

The Canonical Data Pipeline



databricks

The Delta Architecture



Sign up for the Private Beta
visit **databricks.**