Unifying Data Science, Business Analytics and Data Engineering with Azure Databricks

Scott Black

Solutions Architect

databricks

A unified data analytics platform for accelerating innovation across data engineering, data science, and analytics

- Global company with over 5,000 customers and 450+ partners
- Original creators of popular data and machine learning open source projects

Azure Databricks

Combines the best of Databricks and Azure

- First party service
- Enterprise-ready: secure & compliant
- Native integration with Azure services
- Partner ecosystem ready

Azure Databricks

Amazing growth in the first 2 years

Thousands

Azure Databricks customers globally

Millions

Server-hours spinning up every day

2 Exabytes

Data processed per month

30+

Regions available worldwide

Azure Databricks

DELTA LAKE

Open-source storage layer that brings ACID transactions to Apache Spark™ and big data workloads.

75%

of the data processed in Azure Databricks is Delta

Azure Databricks

mlflow™

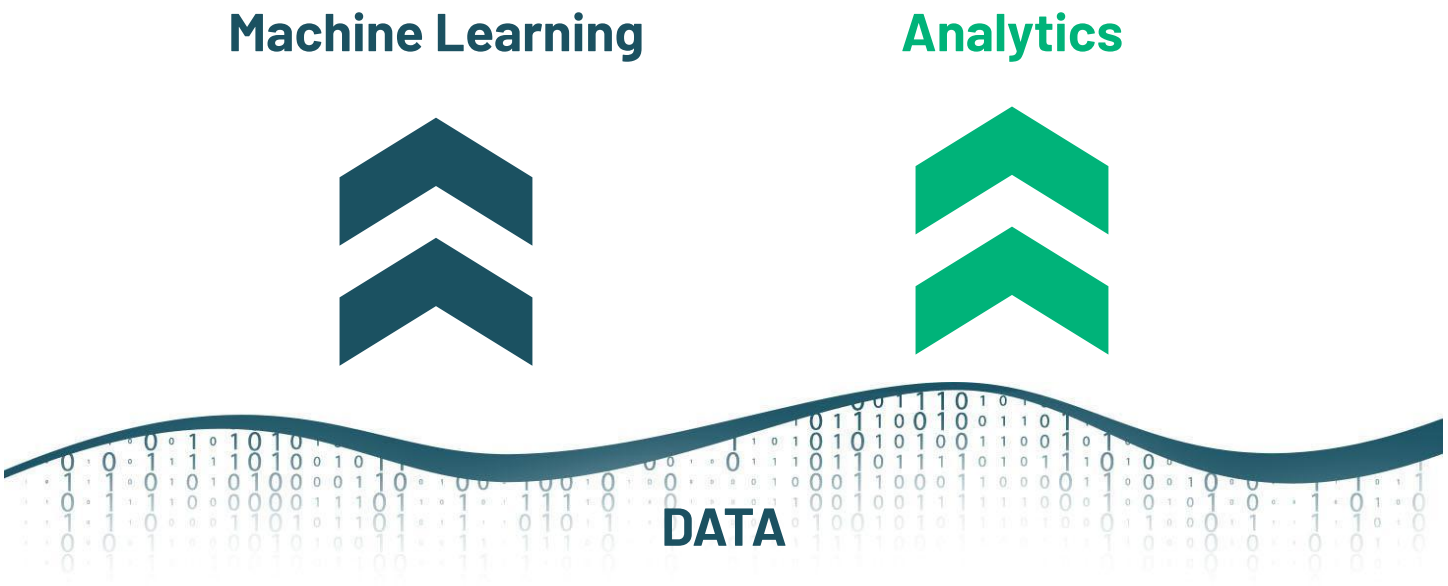
More than 180 contributors and

1.6M+

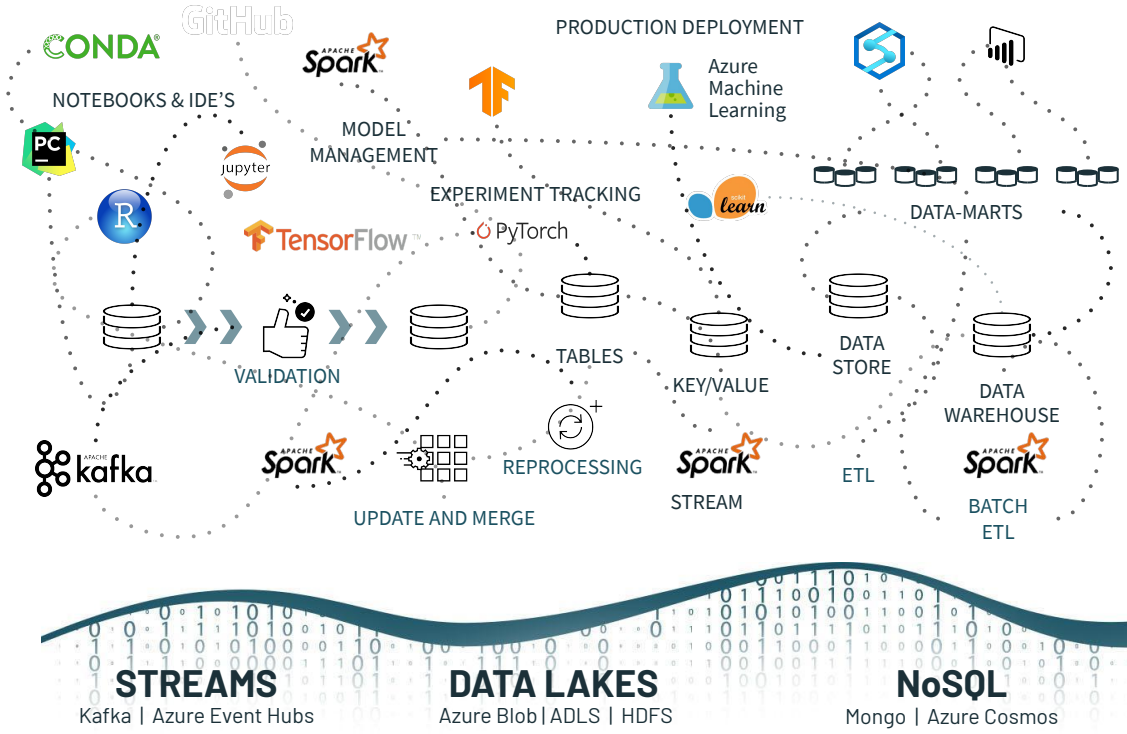
monthly downloads

Unlocking business value

These companies combined their massive data with machine learning and analytics capabilities



Most organizations fail to unlock business value due to data, technology and people silos

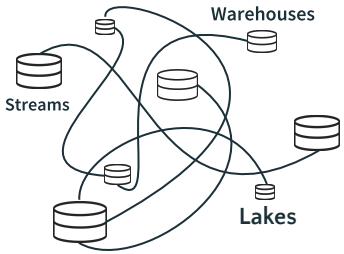


databricks

Unlocking business value: Four challenges

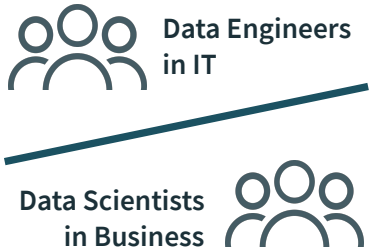
1

Data is messy, siloed and slow



2

ML is hard, Production is harder

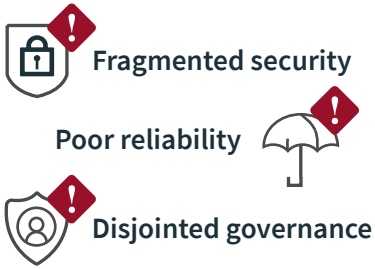


3

BI is limited to a fraction of data

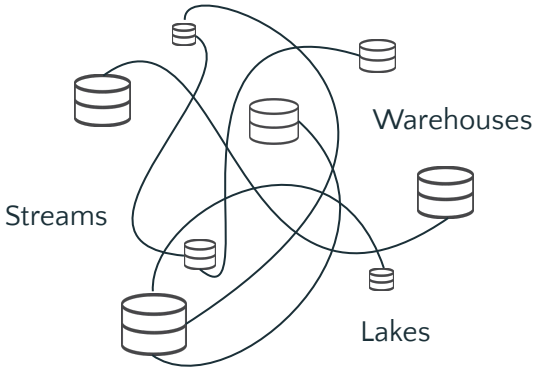
4

Lack of enterprise readiness



Make all your data ready for analytics and ML

1 Data is messy, siloed and slow



Unified Data Service

Build open, reliable, fast data lakes with all your data

Big Data

Business Data

Applications

DELTA LAKE™

Open High Quality Fast

Your Existing Data Lake

Azure Data Lake Storage

hadoop

Azure Blob Storage

APACHE Spark

Unified Engine

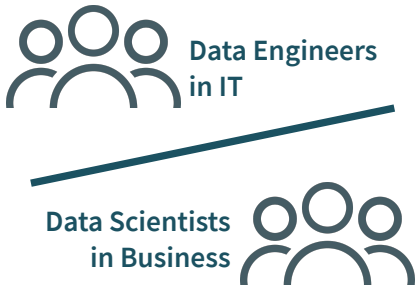
BI Reporting

Machine Learning

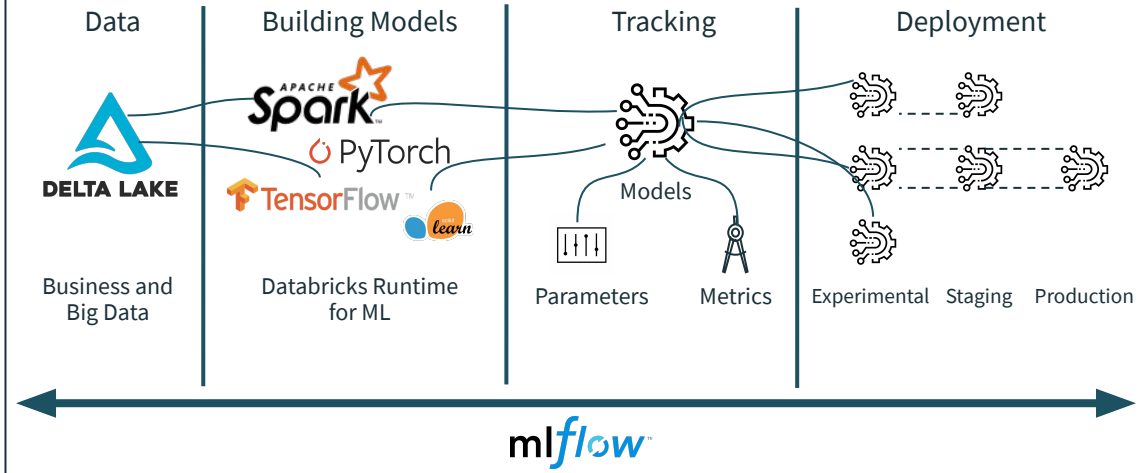
databricks

Unify data and ML across the full lifecycle

2 ML is hard, Production is harder



Data Science and ML Workspace



Standardize ML lifecycle from experimentation to production

databricks

Enable analytics directly on all your source data

3 BI is limited to a fraction of data

BI Integrations for data lake

Reports | Dashboards | Ad hoc data science

DELTA LAKE

APACHE Spark

All your data with high quality and great performance

Applications: salesforce, Marketo, zendesk, Google Analytics

Data stores: mongoDB, ORACLE, MySQL, redis, kafka, cassandra

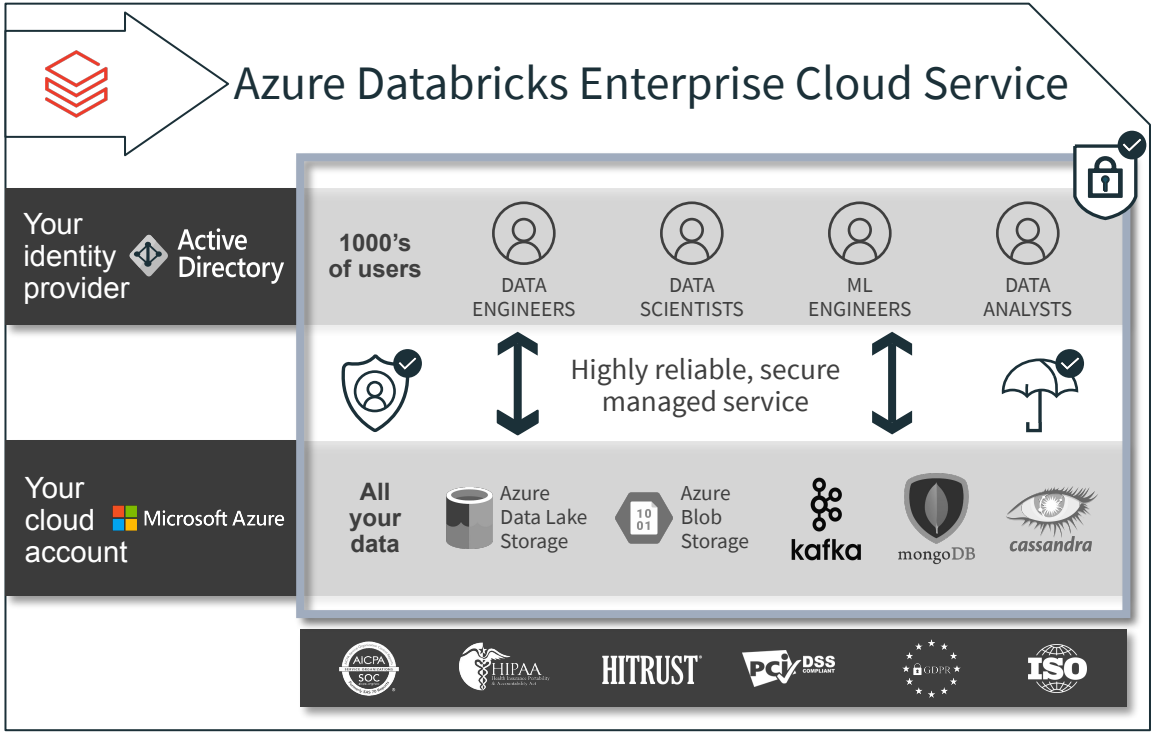
Files

databricks

Leverage cloud native platform for enterprise grade solution

4 Lack of enterprise readiness

- Fragmented security
- Poor reliability
- Disjointed governance



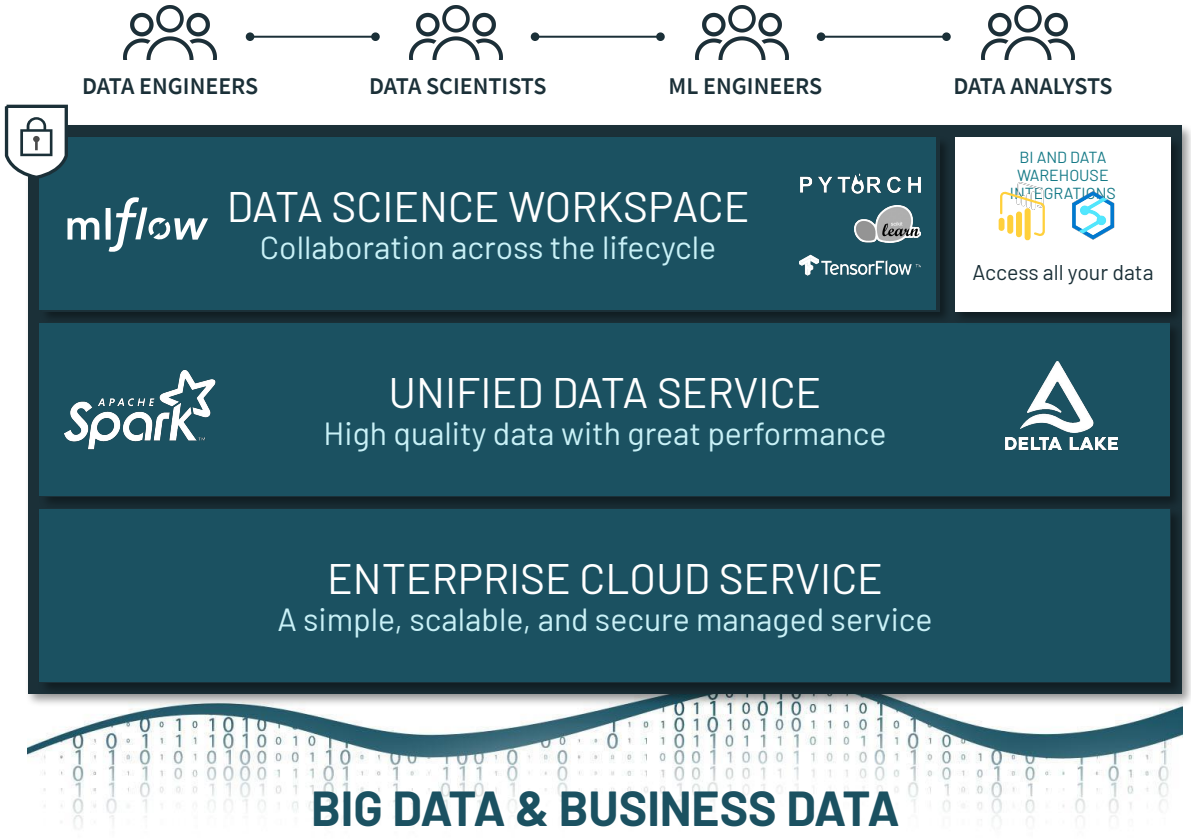
databricks

Databricks Unified Data Analytics Platform

Data science, ML, and analytics on one cloud platform

Access all business and big data in **open data lake**

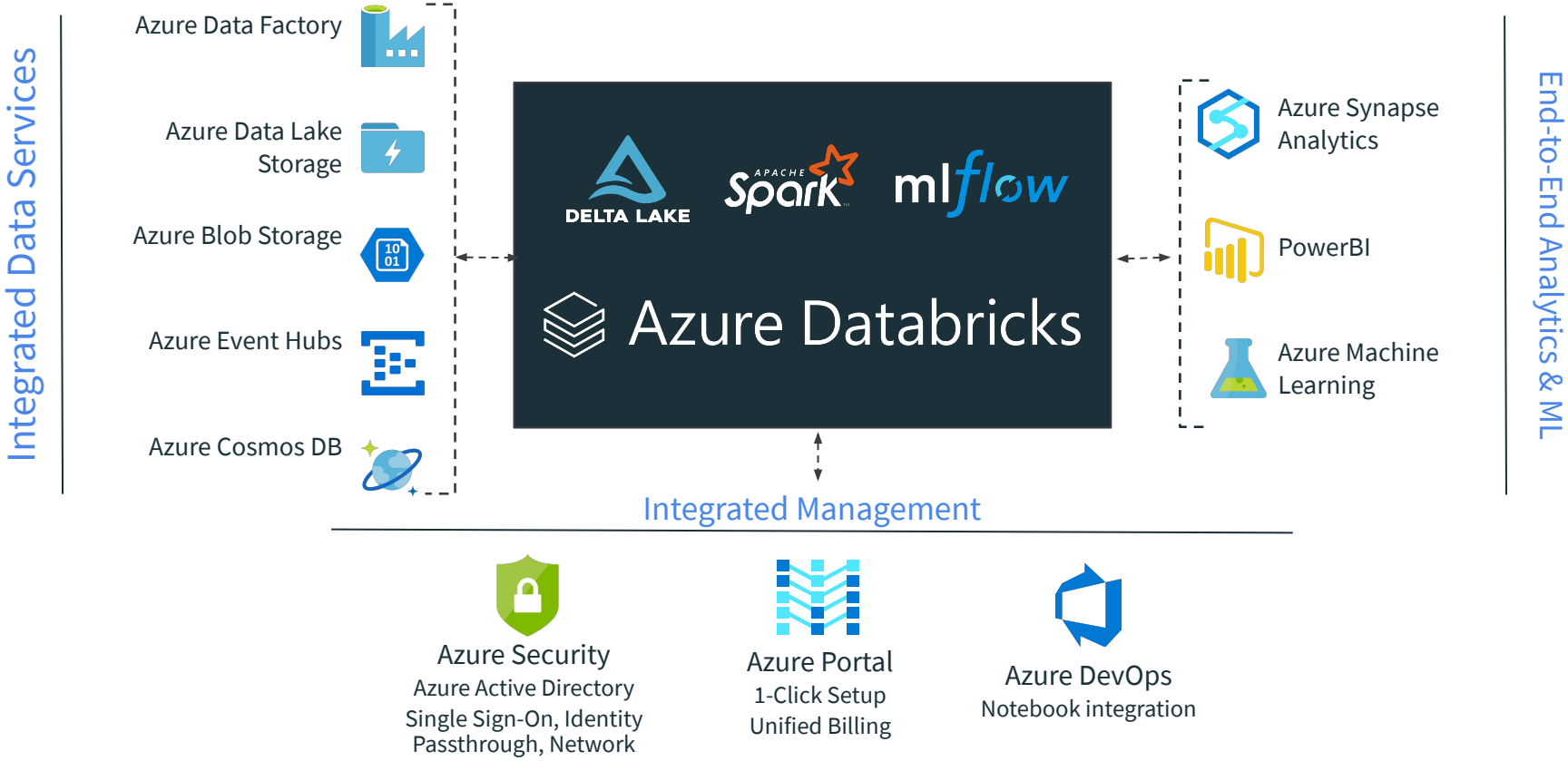
Securely integrates with your **cloud ecosystem**



databricks

Azure Databricks

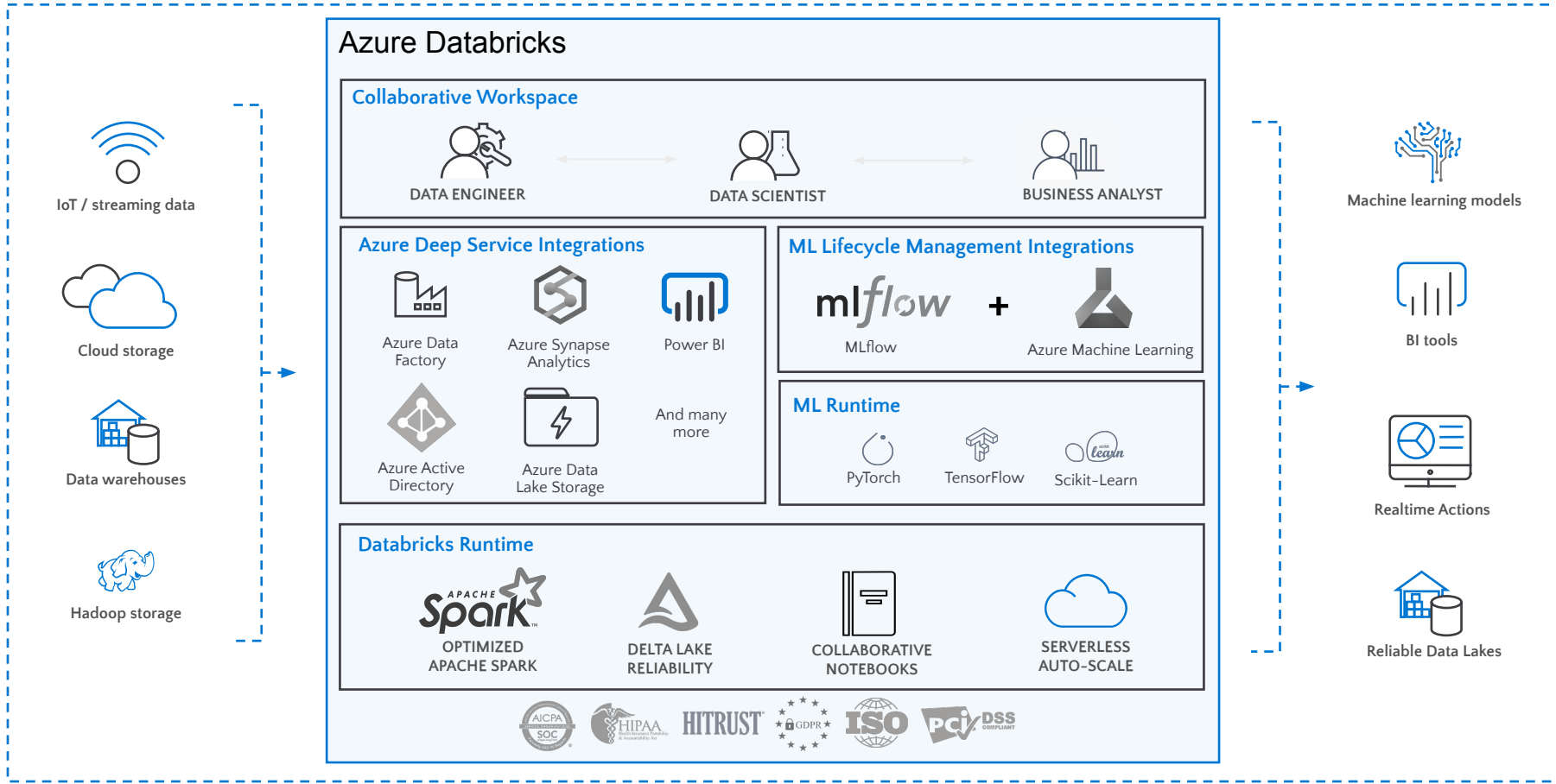
First-Party Service, Natively Integrated into Azure



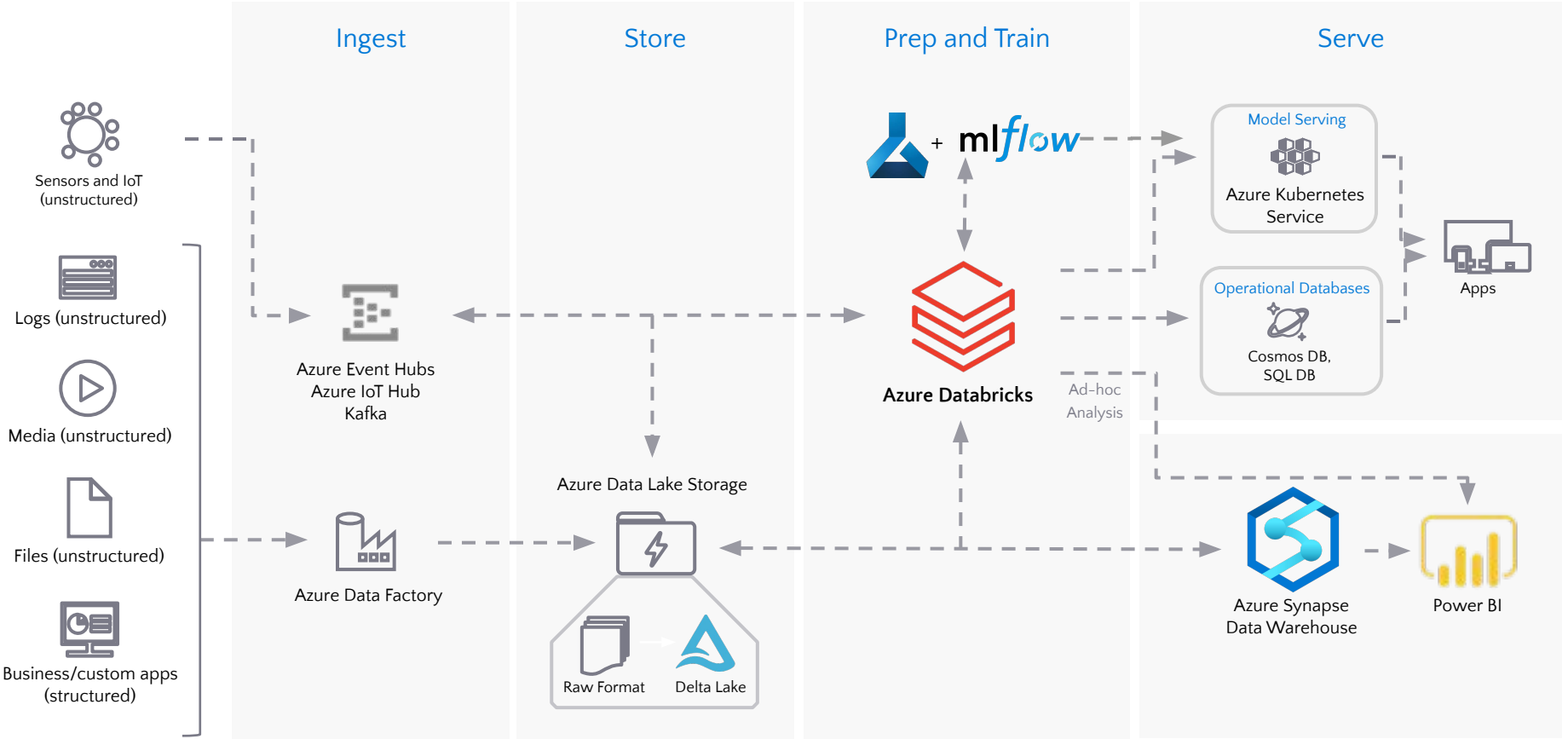
Overview of Data Engineering with Azure Databricks

databricks

Azure Databricks – Introduction



End-To-End Modern Data Architecture

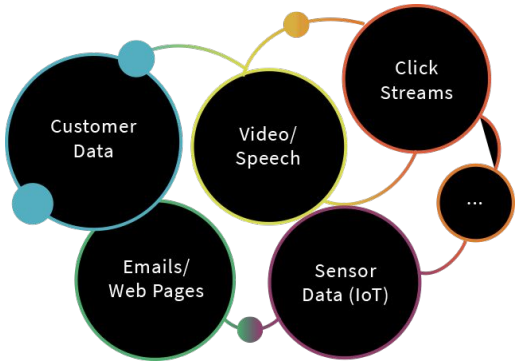


What is Delta Lake?

The Promise of the Data Lake

Azure Databricks

1. Collect Everything



2. Store it all in the Data Lake

Typically ADLS

3. Data Science & Machine Learning

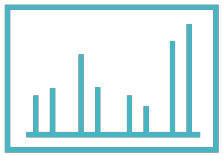
- Recommendation Engines
- Risk, Fraud Detection
- IoT & Predictive Maintenance
- Genomics & DNA Sequencing

What does a typical **data lake** project look like?

Evolution of a Cutting-Edge Data Lake

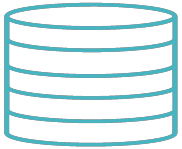
Azure Databricks

Events



Streaming Analytics

?

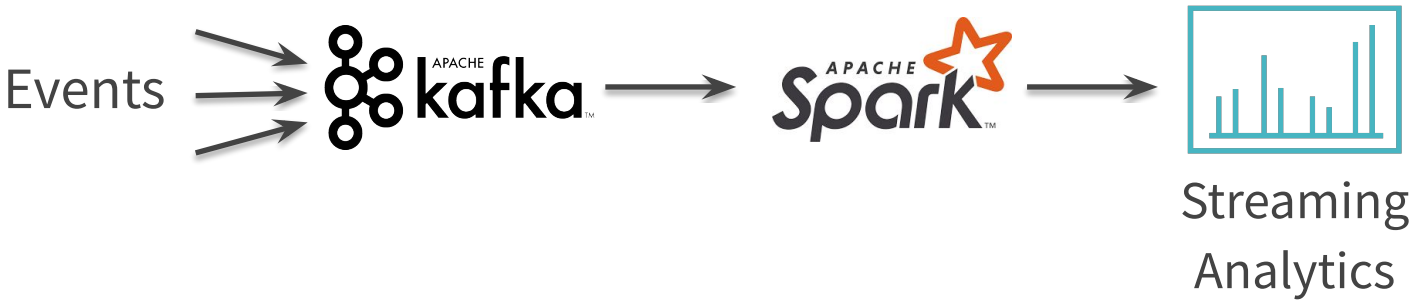


Data Lake

AI & Reporting

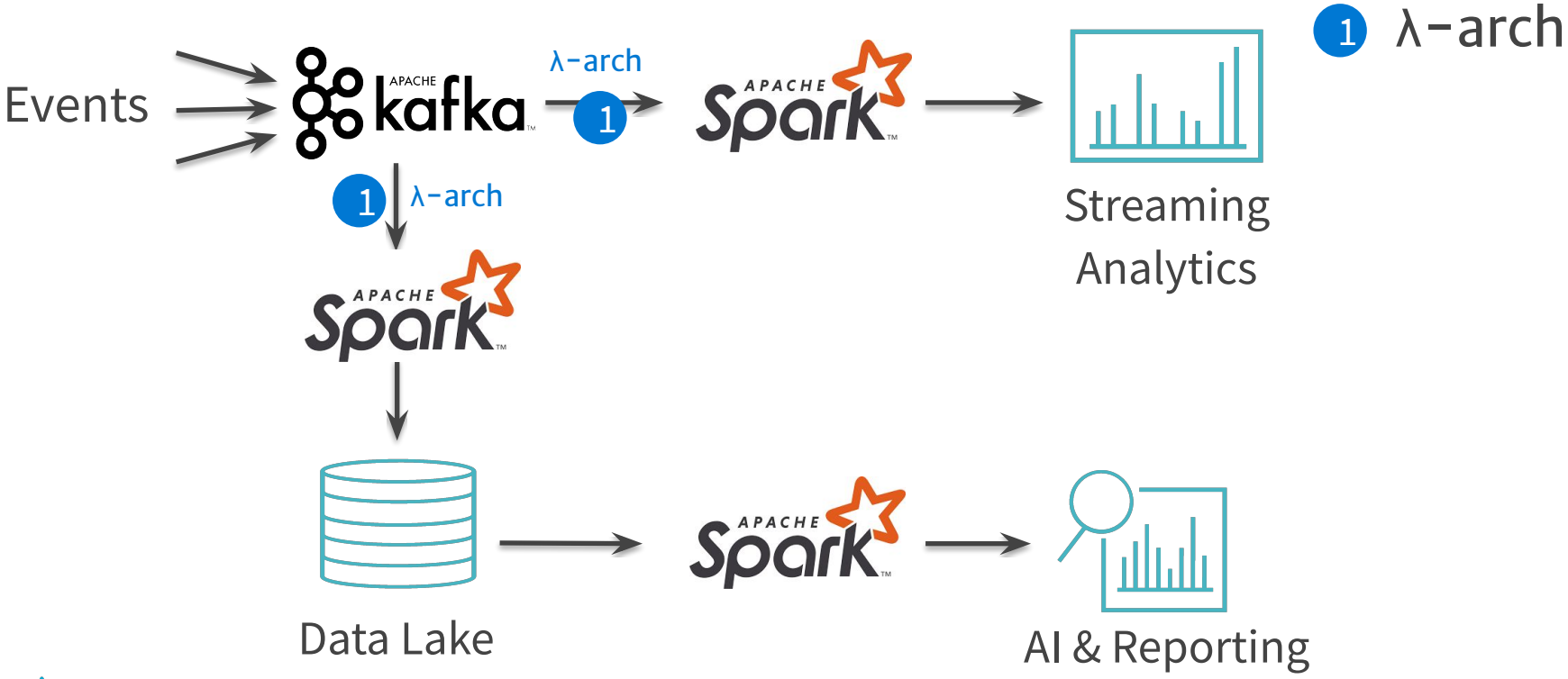
Evolution of a Cutting-Edge Data Lake

Azure Databricks



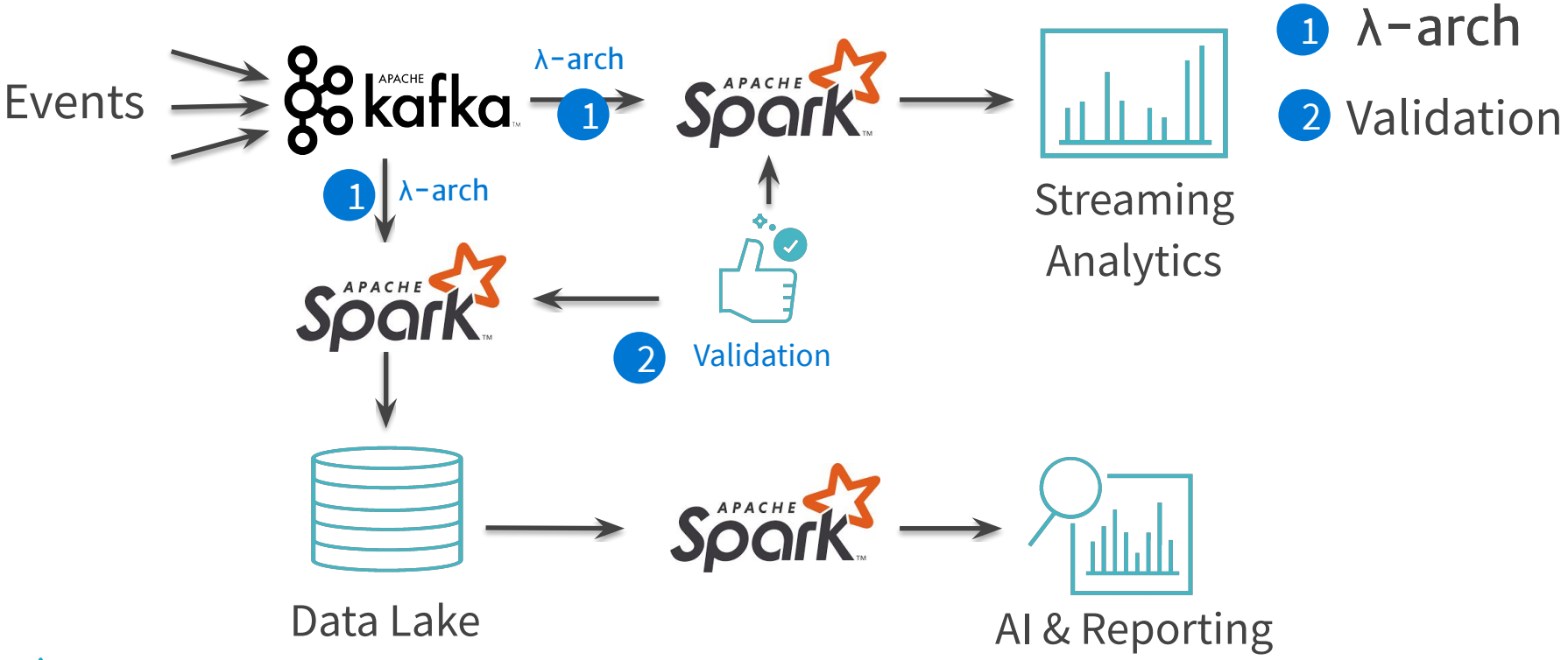
Challenge #1: Historical Queries?

Azure Databricks



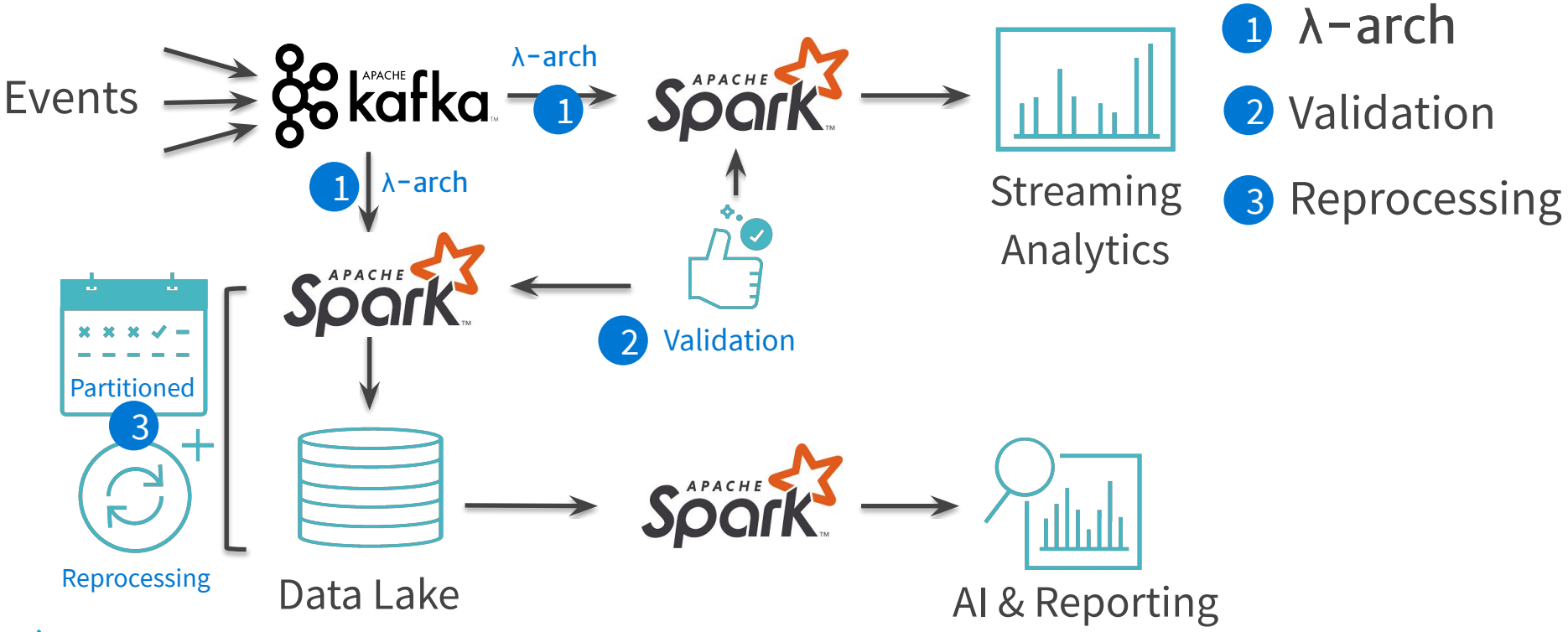
Challenge #1: Historical Queries?

Azure Databricks



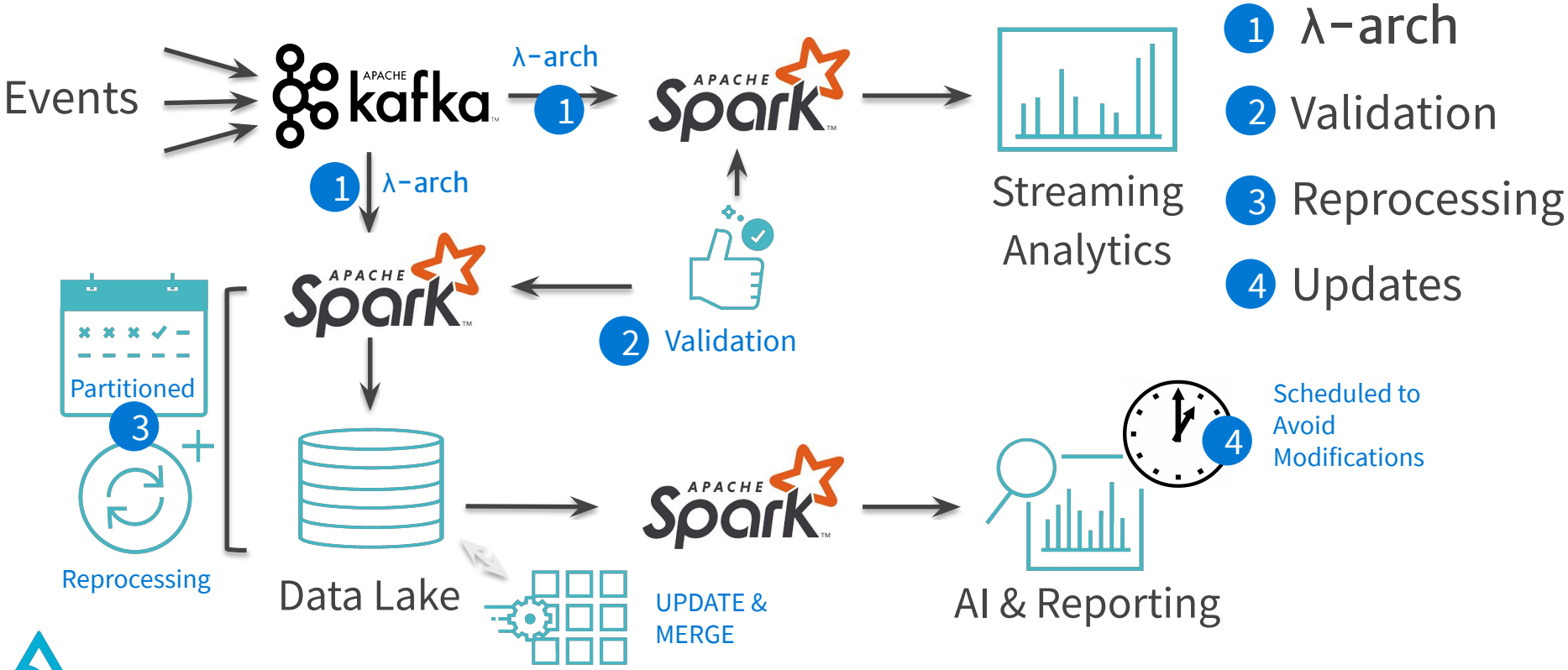
Challenge #1: Historical Queries?

Azure Databricks



Challenge #1: Historical Queries?

Azure Databricks



Data reliability challenges with data lakes

Failed production jobs leave data in corrupt state requiring tedious recovery

Lack of schema enforcement creates inconsistent and low quality data

Lack of consistency makes it almost impossible to mix appends and reads, batch and streaming

New Standard for Building Data Lakes

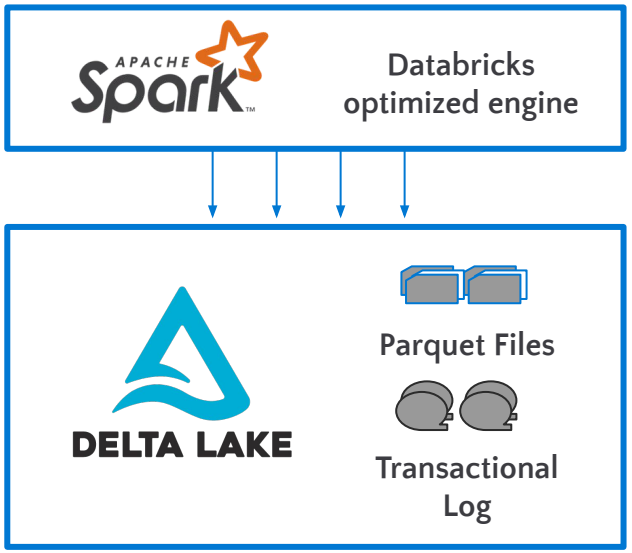
Open Source & Open Format

Adds Reliability & Performance

Fully Compatible with Spark APIs

Azure Databricks optimizes performance

Azure Databricks



Highly Performant
queries at scale

Key Features

- Indexing
- Compaction
- Data skipping
- Caching

DELTA LAKE - Easy to use

Azure Databricks

BEFORE

```
CREATE TABLE ...
USING parquet
...
```

or

```
dataframe
    .write
    .format("parquet")
    .save("/data")
```

AFTER

```
CREATE TABLE ...
USING delta
…
```

or

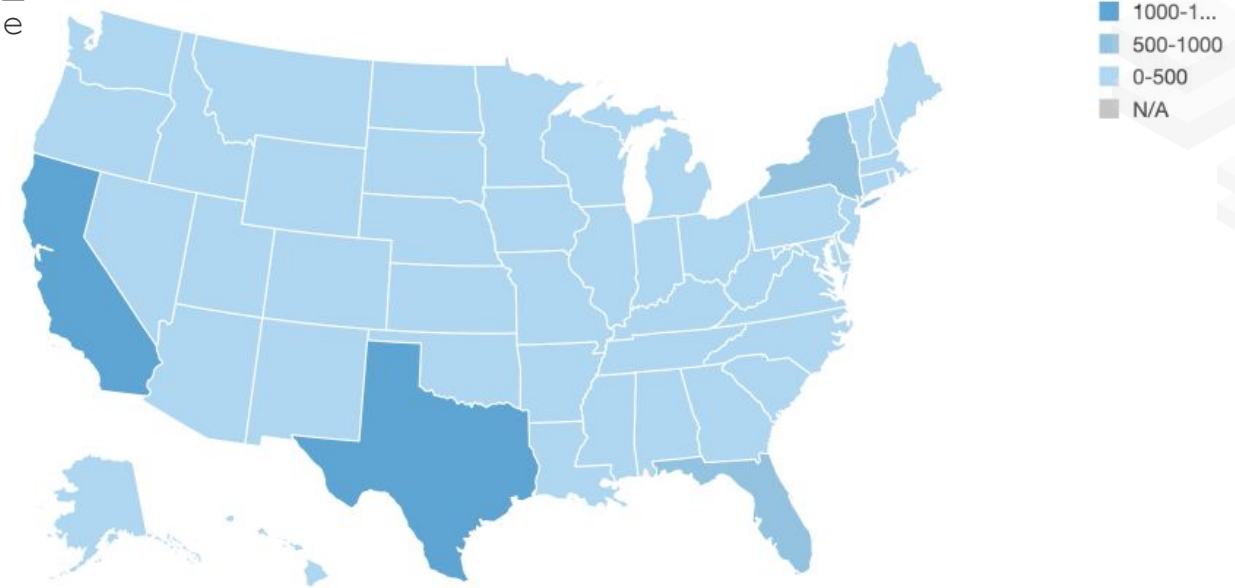
```
dataframe
    .write
    .format("delta")
    .save("/data")
```

DELTA LAKE - SQL for Data Lakes

Azure Databricks

SQL is a universal language used by data engineers, scientists, analysts for data big and small

```
SELECT addr_state, SUM('count') AS loans
FROM loan_by_state_delta
GROUP BY addr_state
```



DELTA LAKE – Reliable Data Lakes at Scale on Azure

Data Versioning | ACID Transactions | Optimized Layouts | Fast Streaming | Efficient Upserts | Schema Enforcement

ACID Transaction Guarantees

- Atomic, Consistent, Isolated, Durable

Versioned parquet files

- Delta transaction log keeps track of all operations

Efficient Upserts

- *MERGE, DELETE, UPDATE*

Time Travel

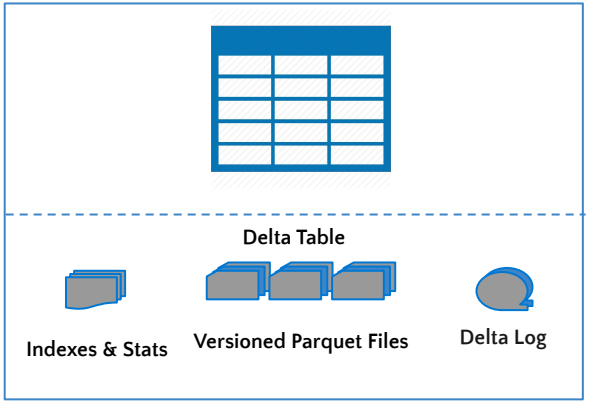
- Audit history, Pipeline Debugging, Data Reproducibility

Open Source

THE APACHE SOFTWARE FOUNDATION + THE LINUX FOUNDATION

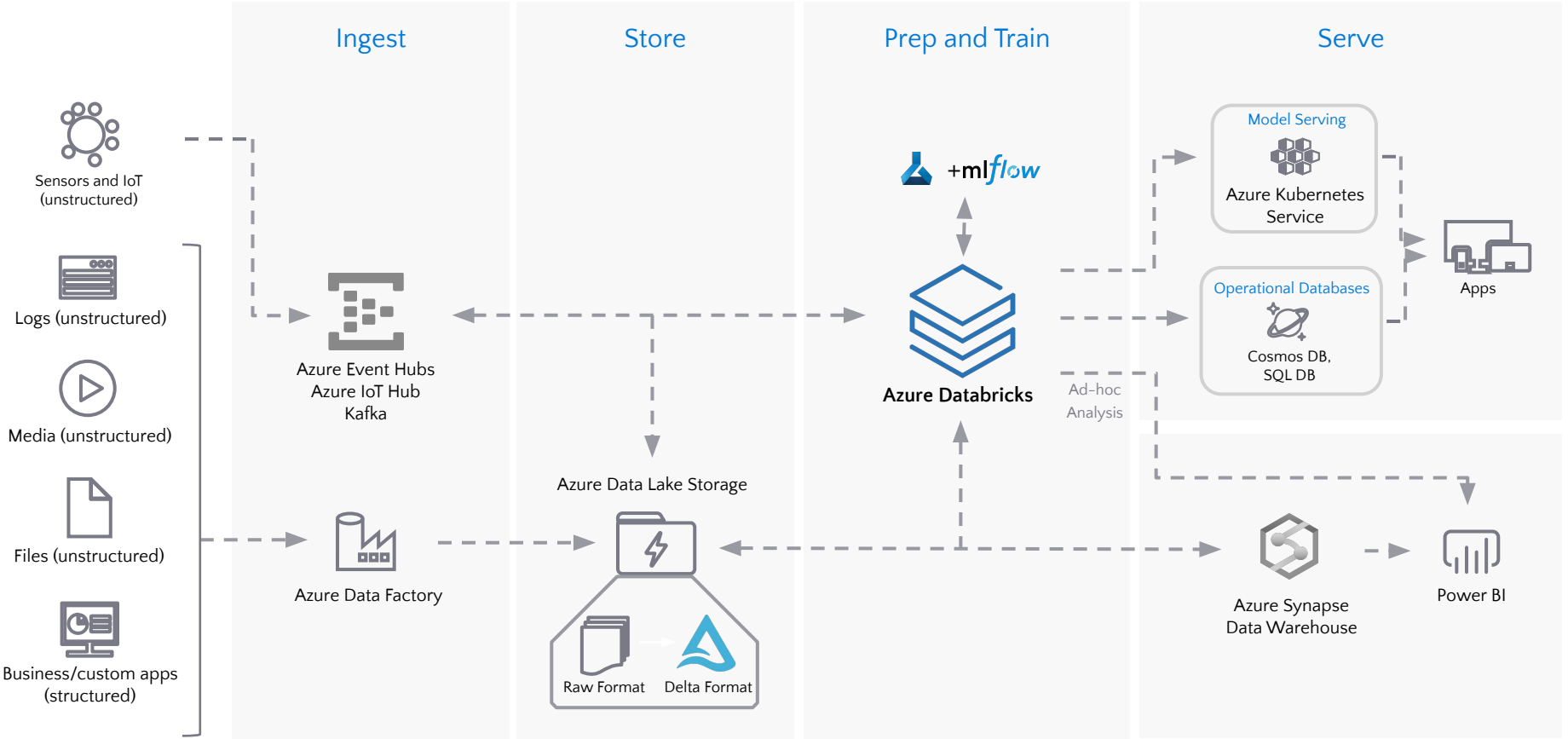
Delta Table =

Parquet + Transaction Log + Indexes/Stats



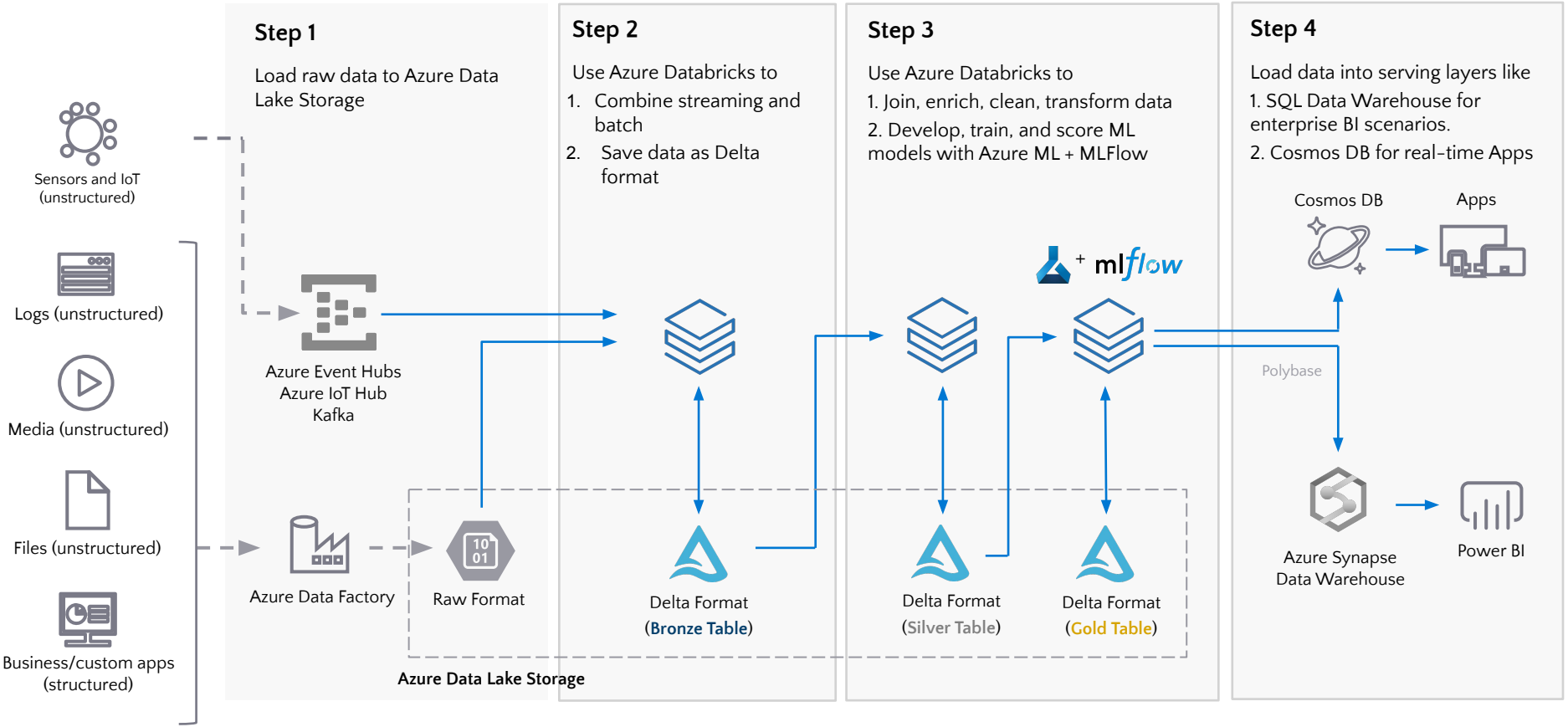
Azure Databricks – Architecture Example

Azure Databricks



Azure Databricks – Delta Lake at Scale on Azure

Azure Databricks



Overview of Data Science with Azure Databricks

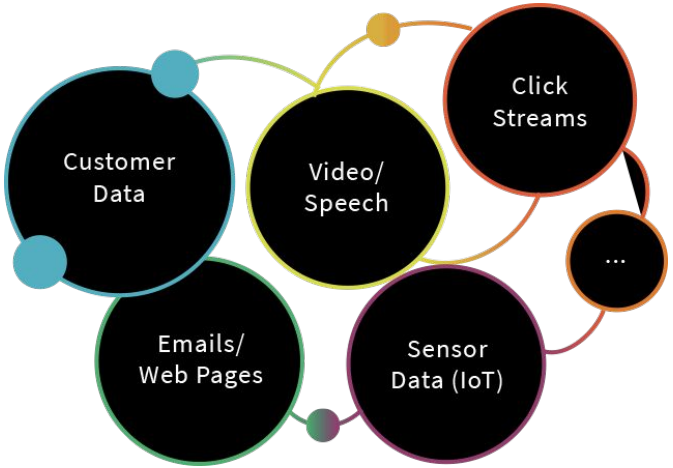
databricks

3 major challenges slowing ML projects:

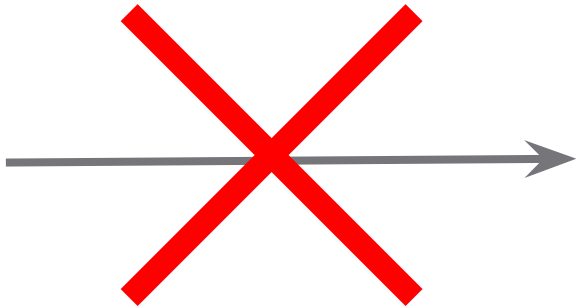
1. Data is not ready for AI
2. A Zoo of new ML Frameworks
3. Data Science & Engineering silos

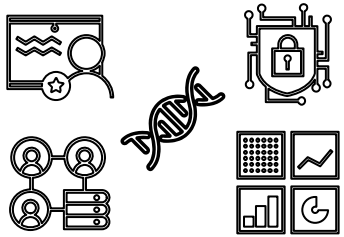
1 Data is not ready for analytics & AI

The **majority** of data projects are failing due to **unreliable data!**



Data Lake



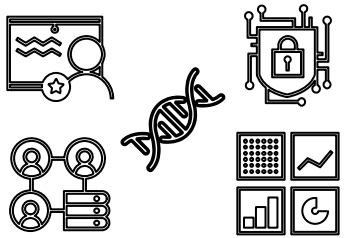


Recommendation Engines
Risk, Fraud Detection
IoT & Predictive Maintenance
Genomics & DNA Sequencing

Delta Lake: makes data ready for analytics & AI

Reliability

Performance



Recommendation Engines
Risk, Fraud Detection
IoT & Predictive Maintenance
Genomics & DNA Sequencing

2 Complexity – Zoo of ML frameworks

Machine Learning

Scikit-learn,
Spark MLlib,
H2O,
Mlpack,
Mahout
...

Deep Learning

TensorFlow,
Keras, Caffe,
PyTorch,
Theano, BigDL,
SparkDL
...

Supporting Libraries

Python, R,
Anaconda,
Numpy, Scipy,
Pandas,
Matplotlib,
PyViz
...

Serving and Monitoring

MLeap,
TF Serving,
Cassandra,
Redis,
TensorBoard
...

Databricks Runtime for ML

Ready-to-use clusters with built-in ML Frameworks

GPU support

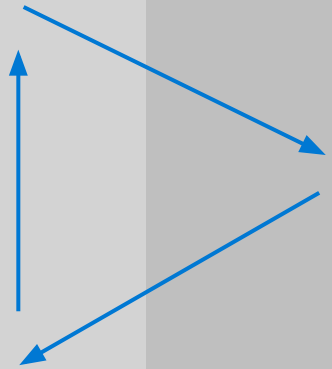
3 Data Science & Engineering silos

1 Data Prep

Hard to make pipelines reliable

2 Build Model

Challenging to track and reproduce experiments



3 Deploy Model

Have to ensure reliability, SLAs, and quality

Data Engineers

Data Scientists

Databricks MLflow: Unifies Data Scientists & Engineers

1 Data Prep

Build reliable data pipelines
Track the datasets

Delta Lake

2 Build Model

Track Experiments
Reproduce experiments

MLflow Tracking
Databricks Runtime for ML
Delta Lake Time Travel

3 Deploy Model

Deploy models in production, track their quality

MLflow Models

Data Engineers

Data Scientists

How Azure Databricks helps Data Scientists

Distributed Machine Learning

Spark MLlib for distributed models

Migrate **Single Node to distributed** with just a few lines of code changes:

- Distributed **hyperparameter search** (Hyperopt, Gridsearch)
- **PandasUDF** to distribute models over subsets of data or hyperparameters
- **Koalas**: Pandas DataFrame API on Spark

Deep Learning distributed training (**HorovodRunner**)

Access object storage as if in local storage using Databricks File System (DBFS)

Use your own tools

Multiple languages in Databricks Notebooks (Python, R, Scala, SQL)

Databricks Connect: connect external tools with Databricks (IDEs, RStudio, Jupyter...)

Python (Scikit-Learn, Pandas)

R support

Native R notebooks on Databricks

RStudio & RStudio Server integrations

Scaling and parallelizing with SparkR & SparklyR

ML Runtime Optimizations

Reliable and secure distribution of open source ML frameworks

Packages and optimizes most common ML Frameworks .

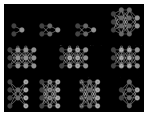


...

Built-in Optimization for Distributed Deep Learning .

Distribute and Scale any Single-Machine ML Code to 1,000's of machines.

Built-In AutoML and Experiment Tracking

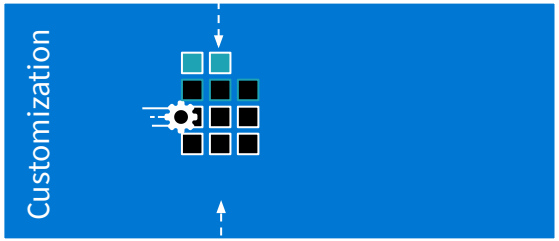


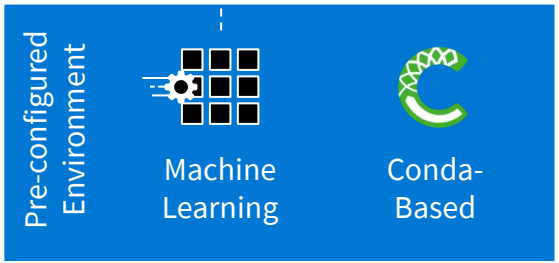

AutoML and Tracking / Visualizations with MLflow

Customized Environments using Conda .

requirements.txt

conda.yaml





databricks