Databricks

Unifying Data Science and Engineering

The beginnings of Apache Spark at UC Berkeley

AMPLab funded by tech companies:

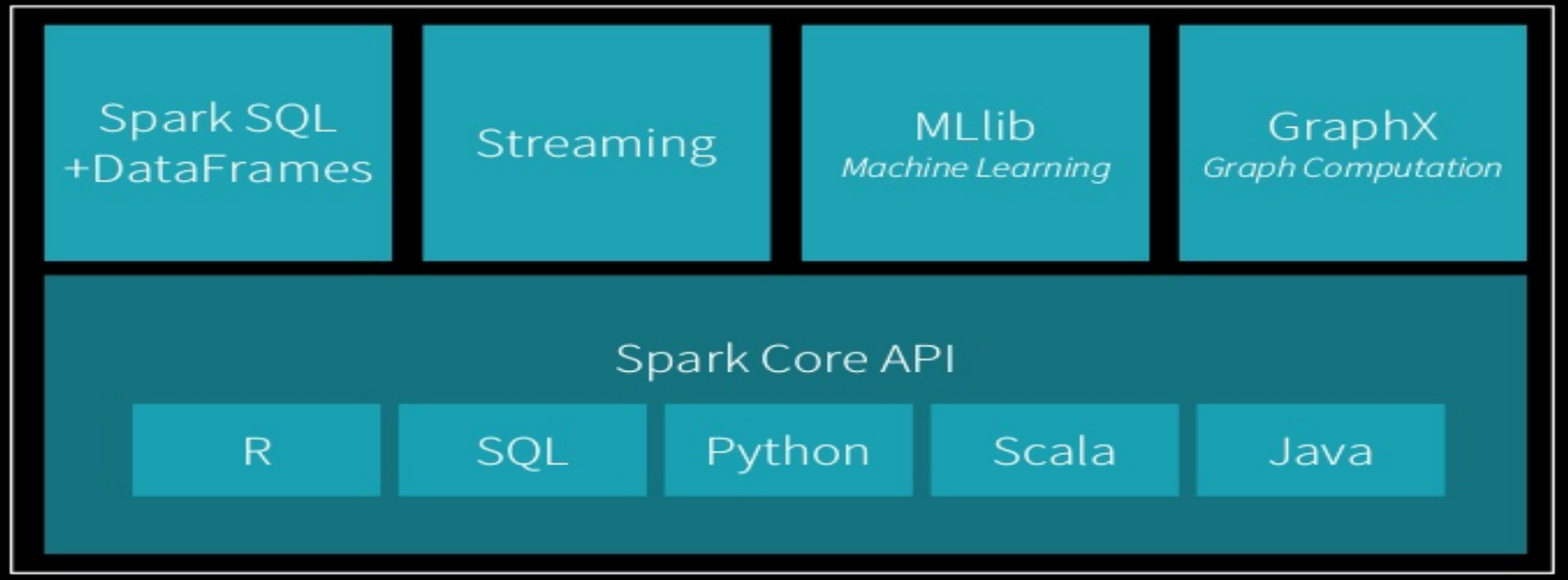
twitter facebook. Google

- Got a glimpse at their most impactful internal projects
- They were leveraging massive amounts of data
- Doing high impact machine learning/AI

We wanted to democratize Data + AI

databricks

Apache Spark



databricks

Databricks started in 2013

Bring Apache Spark to the Enterprise

databricks

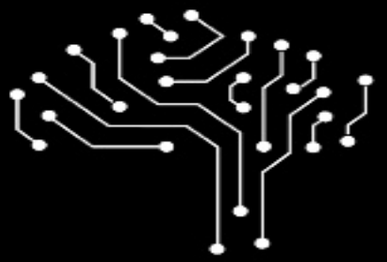
Only 1% of enterprises successful with AI

Other 99% struggle due to organizational silos

IT

Java
Parquet
TERADATA
S3
mongoDB
ORACLE
hadoop
kafka
C++
Scala

Line of Business



Keras
python
PYTORCH
learn
mxnet
TensorFlow

Data Engineers

Data Scientists

databricks

Databricks goal is to unify data science & engineering

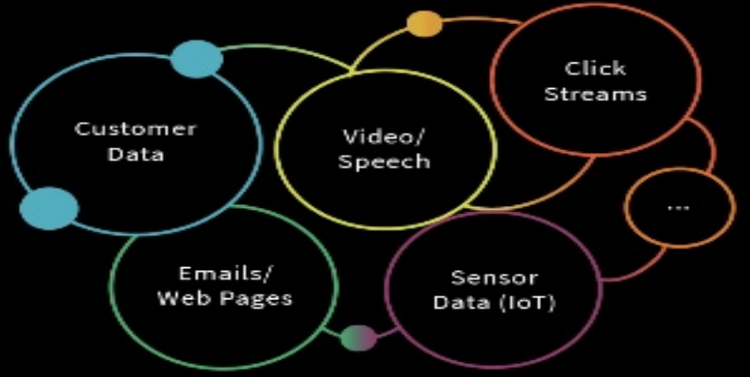
3 challenges created by data & AI divide

1. Data is not ready for AI
2. Data and AI technology silos
3. Data scientists and engineers are in silos

1 Data is not ready for AI

databricks

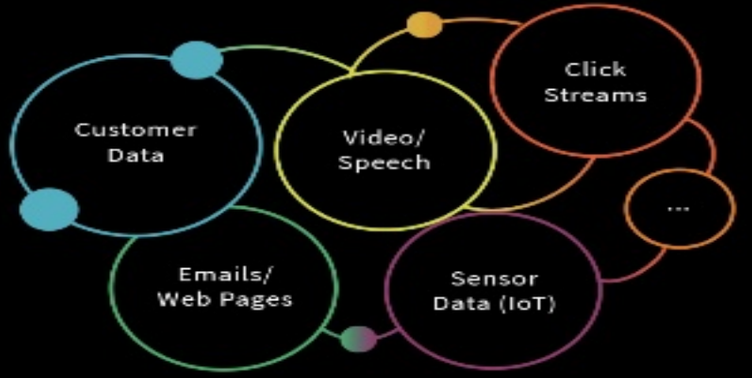
Massive data in data lakes



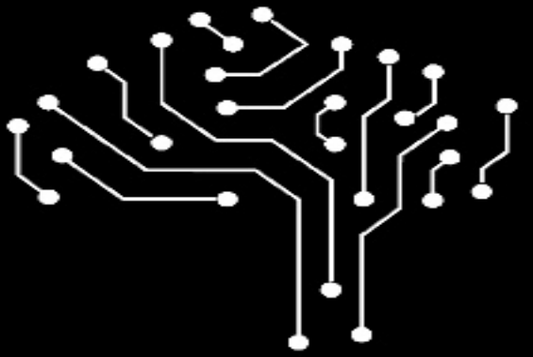
Data Lake

databricks

Vision to do AI on that data



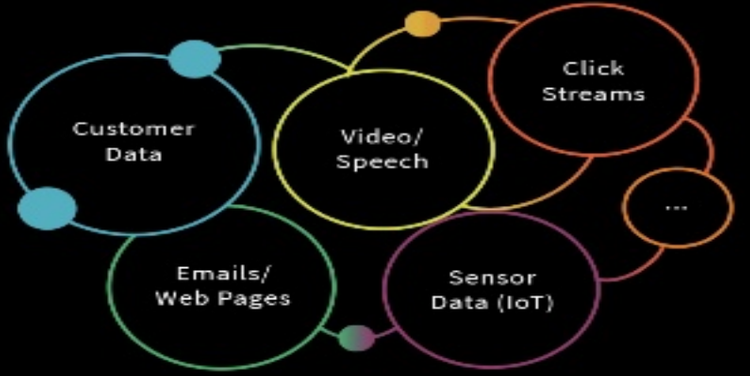
Data Lake



AI

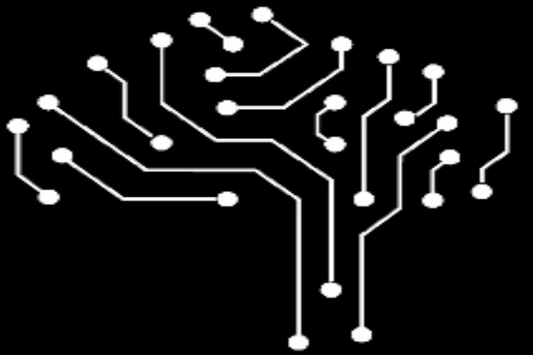
databricks

Data is not ready for AI



Data Lake

Inconsistent Data
Lack of Schema
Slow Performance and Costly



AI

databricks

Databricks Delta

Brings data reliability and performance to data lakes

Data Reliability + **Fast Analytics**

Parquet · hadoop · kafka · Azure Blob Storage · S3

databricks

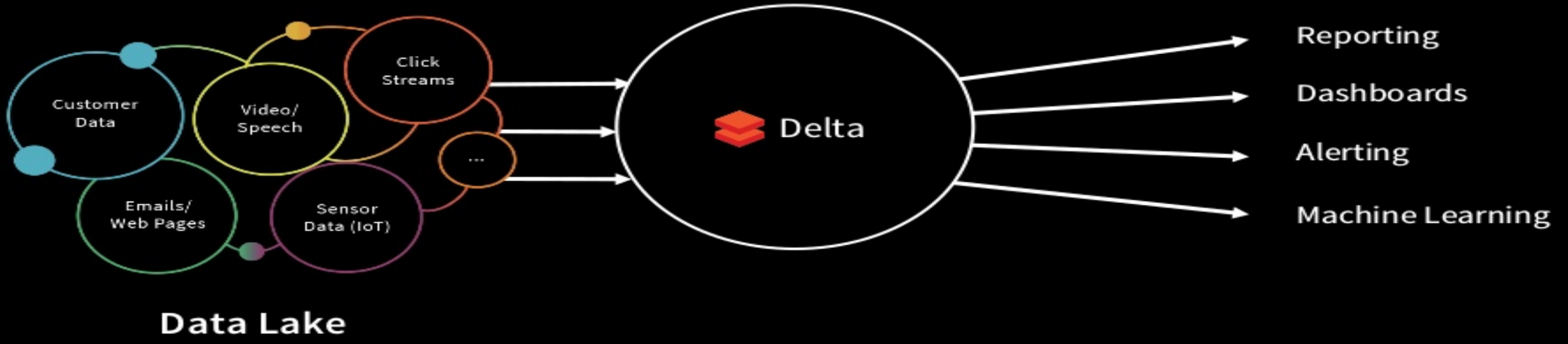
Databricks Delta: makes data ready for AI

Data Reliability

Schema Enforcement
ACID Transactions

Query Performance

Very Fast at Scale
Indexing (10-100x Faster)

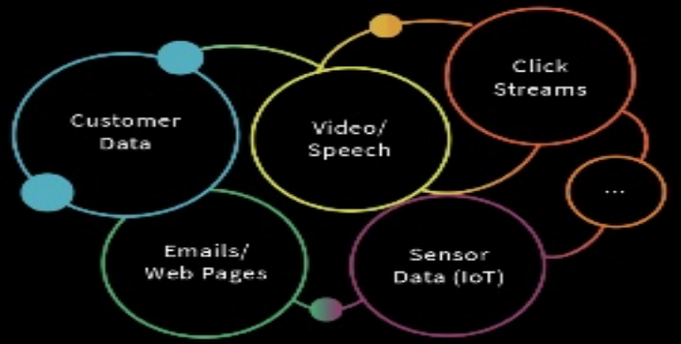


Data Lake

databricks

2 Data and AI technology silos

Data & AI technology silos

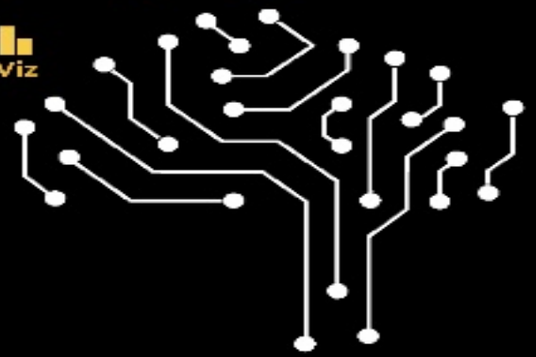


hadoop mongoDB. Java Scala
C++ ORACLE kafka TERADATA

Great for Data, but not AI

Supporting and Deployment Libraries

mleap TFServing TensorBoard SciPy
matplotlib R ANACONDA. NumPy
Pandas PyViz Azure ML



python Keras mxnet PYTORCH
TensorFlow learn H_2O.ai

Great for AI, but not for data

Data & AI technology silos

Supporting and Deployment Libraries

Databricks Runtime for ML

Ready to use clusters with built-in ML Frameworks

TensorFlow™ · scikit-learn · Keras · XGBoost · PYTORCH · Horovod · NumPy

GPU support

Azure · aws

Great for Data, but not AI

Great for AI, but not for data

3 Data scientists and engineers are in silos

Data scientists & engineers are in silos

1. **Data Prep**
   Hard to make pipelines reliable

   Apache Spark

2. **Build Model**
   Challenging to track and reproduce experiments

   GitHub CONDA TensorFlow Excel learn XGBoost

3. **Deploy Model**
   Have to ensure reliability, SLAs, and quality

   kubernetes docker

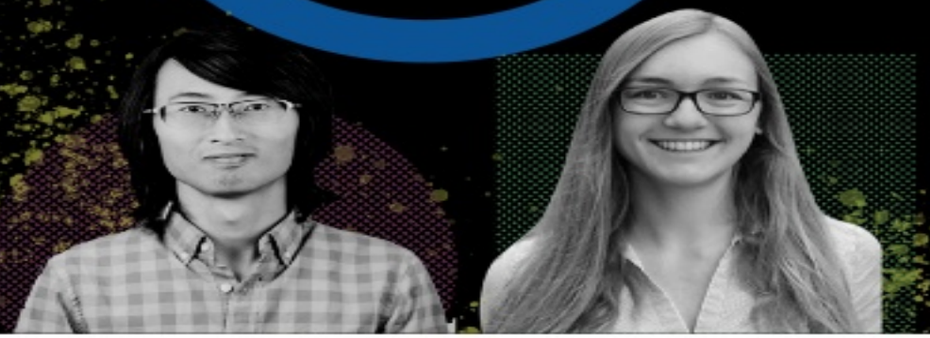
Data Engineers

Data Scientists

databricks

Databricks MLflow: unifies data scientists & engineers

Data Engineers



Data Scientists

databricks

Databricks MLflow: unifies data scientists & engineers

1 **Data Prep**

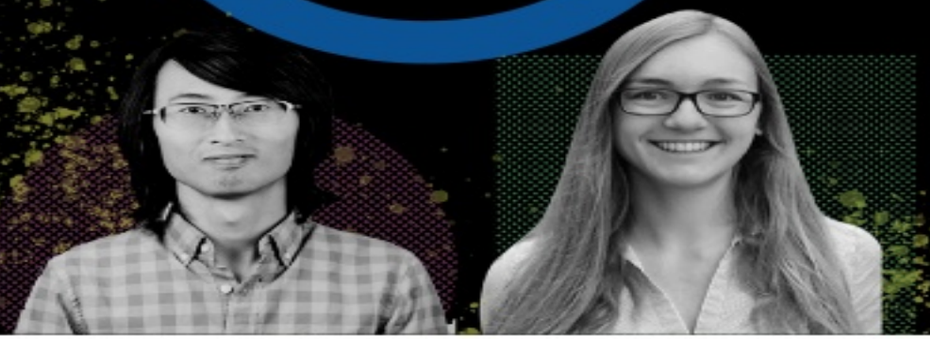
Build reliable data pipelines
Track the datasets
Databricks Delta

3 **Deploy Model**

Deploy models in production, track their quality
MLflow Serving

2 **Build Model**

Track Experiments
Reproduce experiments
MLflow Project & Tracker
Databricks Runtime for ML



Data Engineers

Data Scientists

databricks

Databricks MLflow: unifies data scientists & engineers

1 Data Prep

Build reliable data pipelines

Track the datasets

Databricks Delta

2 Build Model

Track Experiments

Reproduce experiments

Databricks Runtime for ML

MLflow Project & Tracker

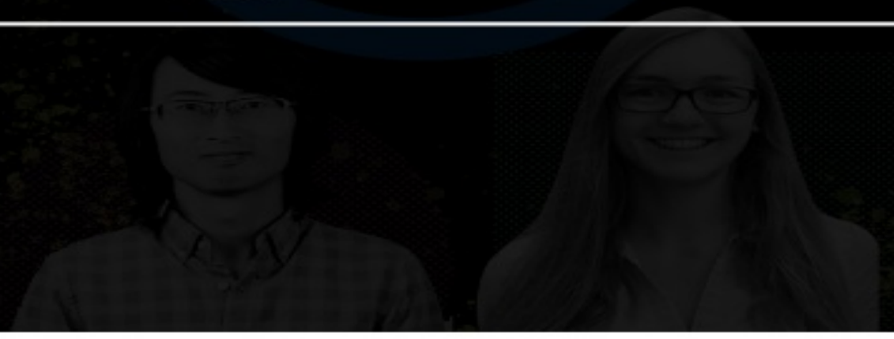
3 Deploy Model

Deploy models in production, track their quality

MLflow Serving

Announcing: **time travel**

Delta + mlflow



Data Engineers

Data Scientists

databricks


REGENERON


Hotels.com™

Databricks makes AI possible

for the other 99% by

unifying data science and engineering


Databricks Unified Analyitcs
demo by Michael Armbrust
databricks