Microsoft

Data Analytics Pipelines with Spark and Azure Databricks

20 November 2018

Lace Lofranco
Senior Software Engineer
Microsoft

Survey

Session objective

At the end of this session, you should:

- Know the key capabilities of Spark and the Azure Databricks platform
- Have an understanding of building advanced analytics workloads with Spark on Azure Databricks

Agenda

Spark Fundamentals

Unified Computing Engine

Azure Databricks

Managed Apache Spark, Integrations with Azure Services

Demo

Recommendation System

Spark Fundamentals

hadoop
Map Reduce

HIVE

STORM

TEZ

mahout

Apache Flink

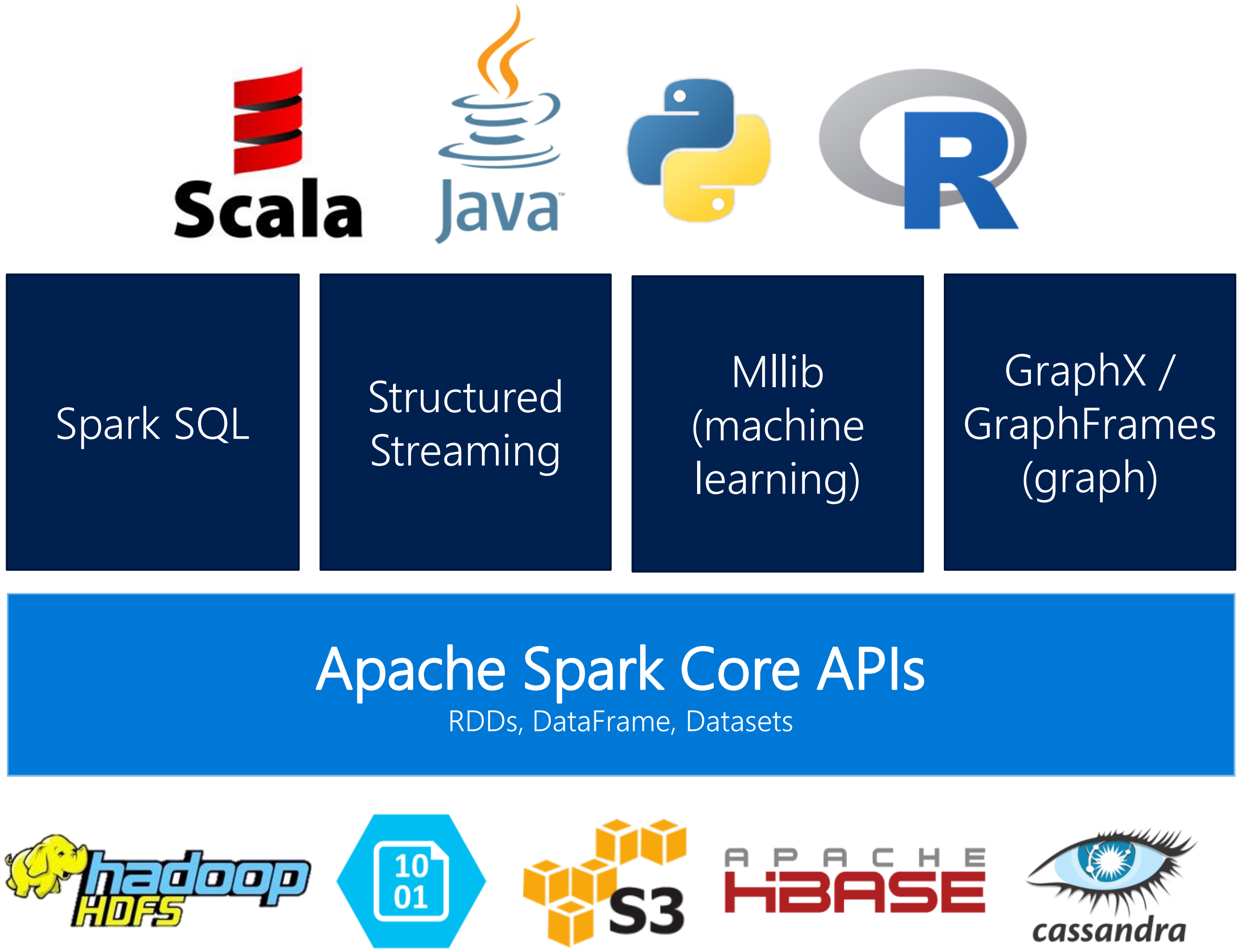
FLUME

APACHE
GIRAPH

sqoop

Apache Spark

a **unified computing engine** and a set of libraries for parallel data processing on computer clusters



Spark SQL

Structured Streaming

Mllib (machine learning)

GraphX / GraphFrames (graph)

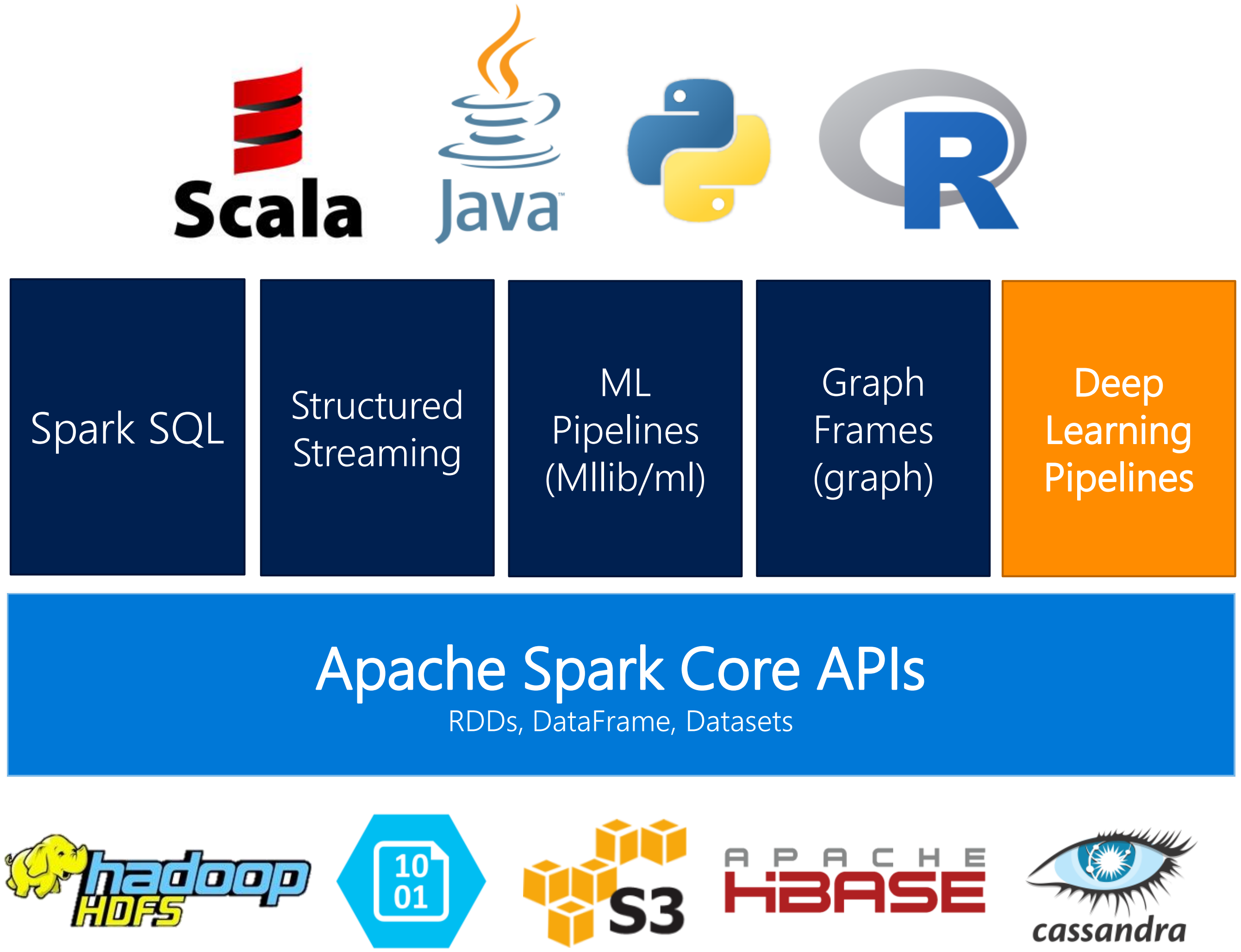
Apache Spark Core APIs

RDDs, DataFrame, Datasets

Spark: The Definitive Guide, Matei Zaharia, Bill Chambers

Apache Spark

a **unified computing engine** and a set of libraries for parallel data processing on computer clusters



Spark SQL

Structured Streaming

ML Pipelines (Mllib/ml)

Graph Frames (graph)

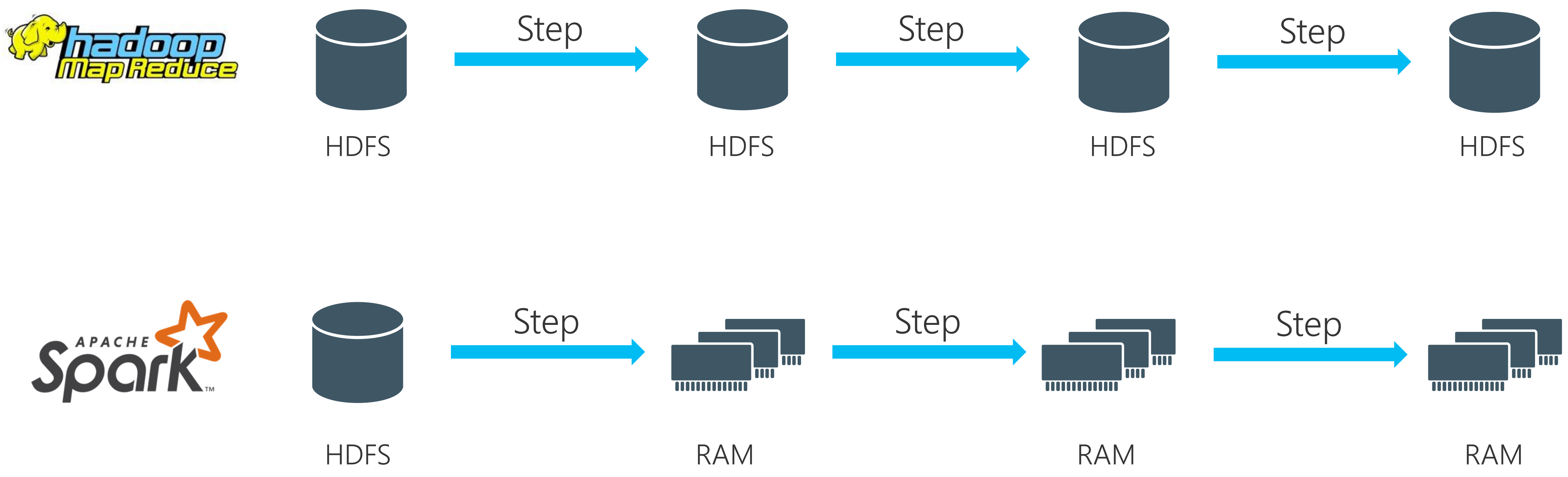
Deep Learning Pipelines

Apache Spark Core APIs

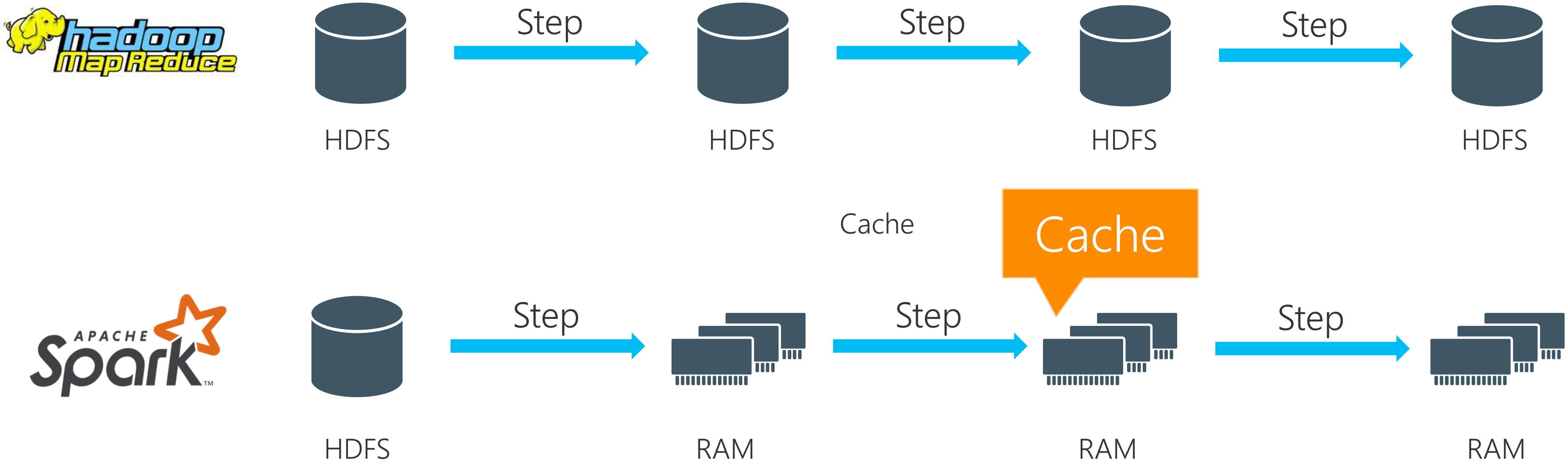
RDDs, DataFrame, Datasets

Spark: The Definitive Guide, Matei Zaharia, Bill Chambers

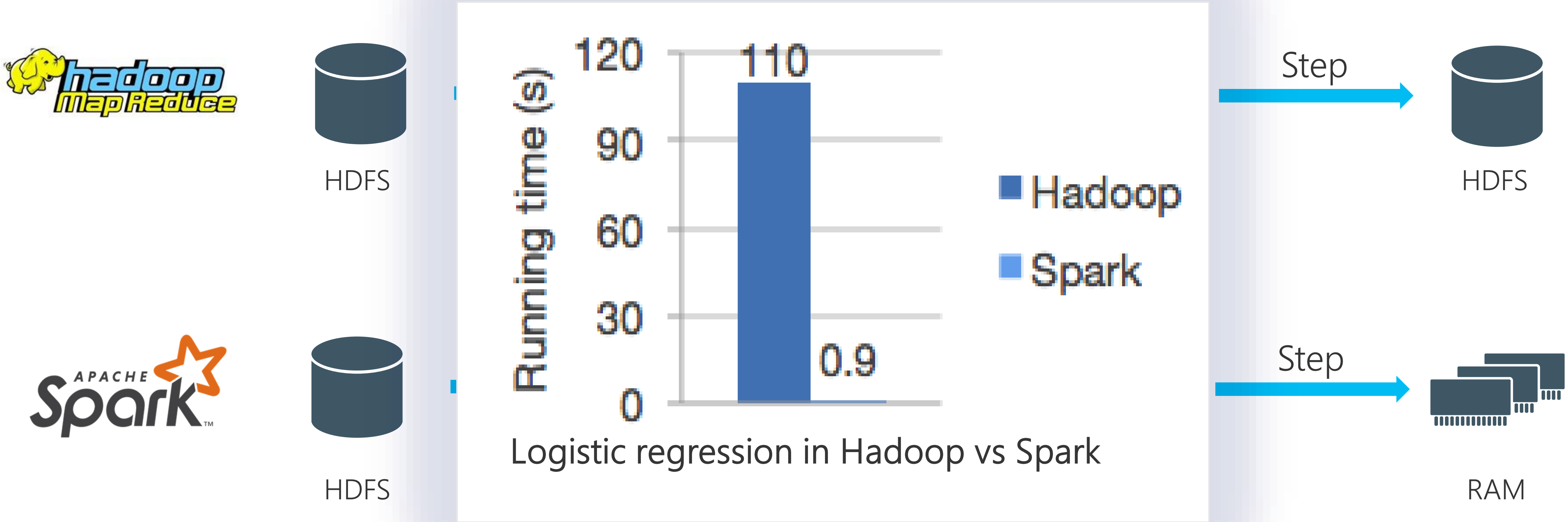
Why Spark is fast



Why Spark is fast



Why Spark is fast



Logistic regression in Hadoop vs Spark

Source: http://spark.apache.org/

Apache Spark: APIs

RDDs

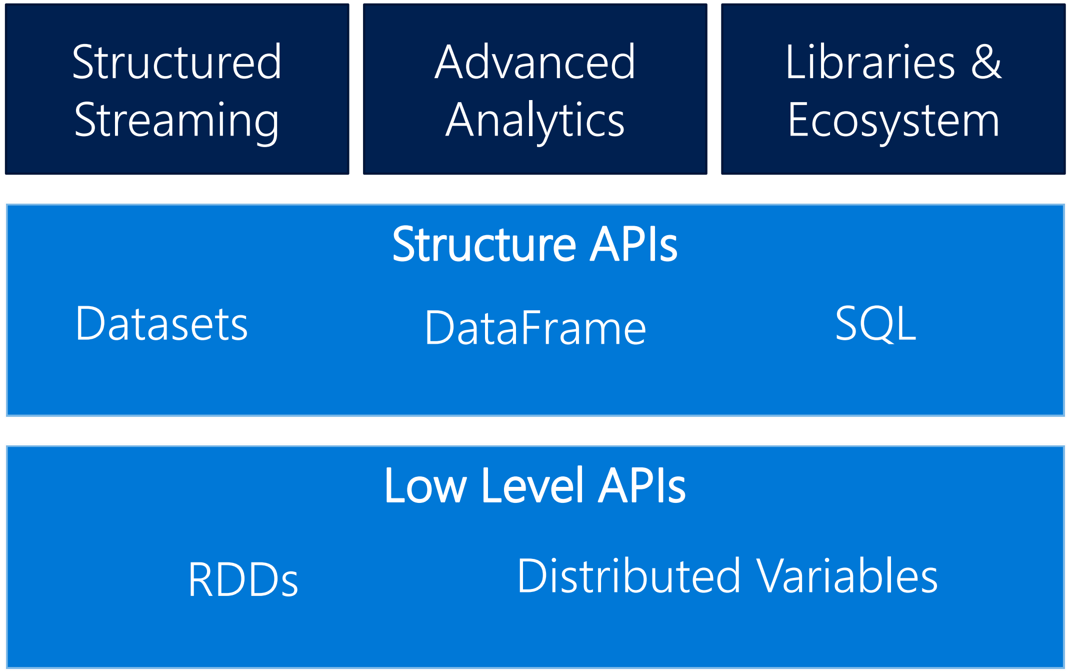
Core building block of data processing pipelines

DataFrames

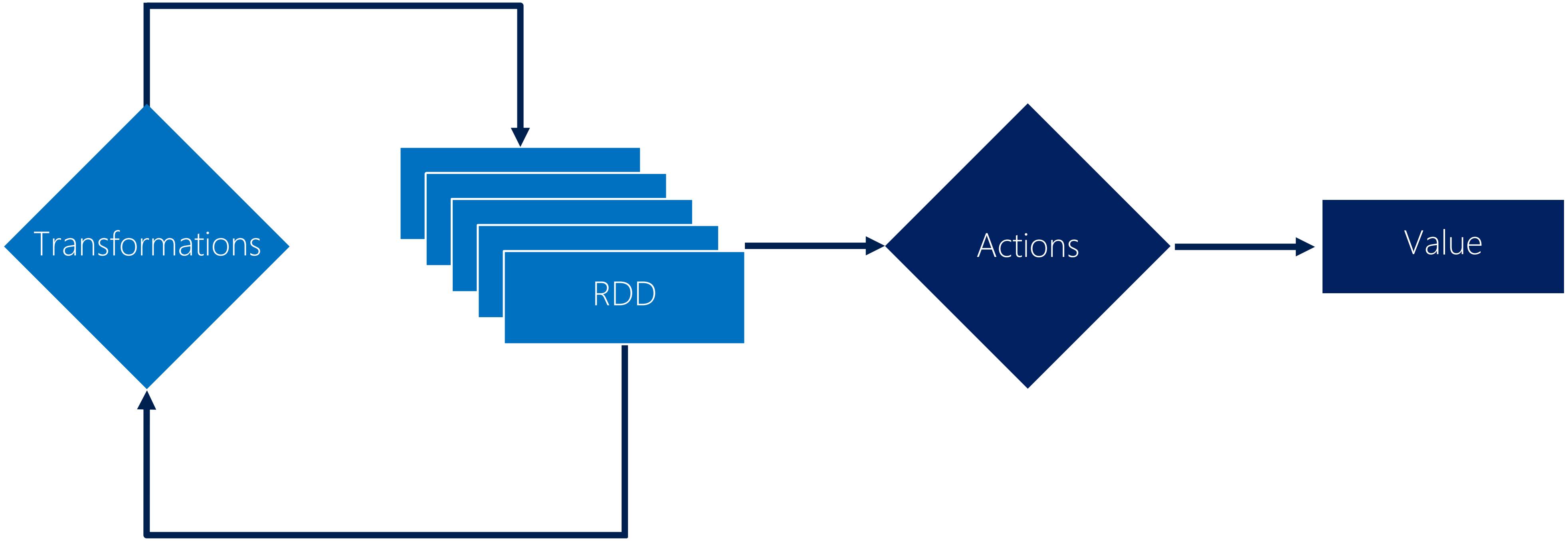
High level APIs that take advantage of query optimizer

Datasets

Data Frames with user objects and custom code



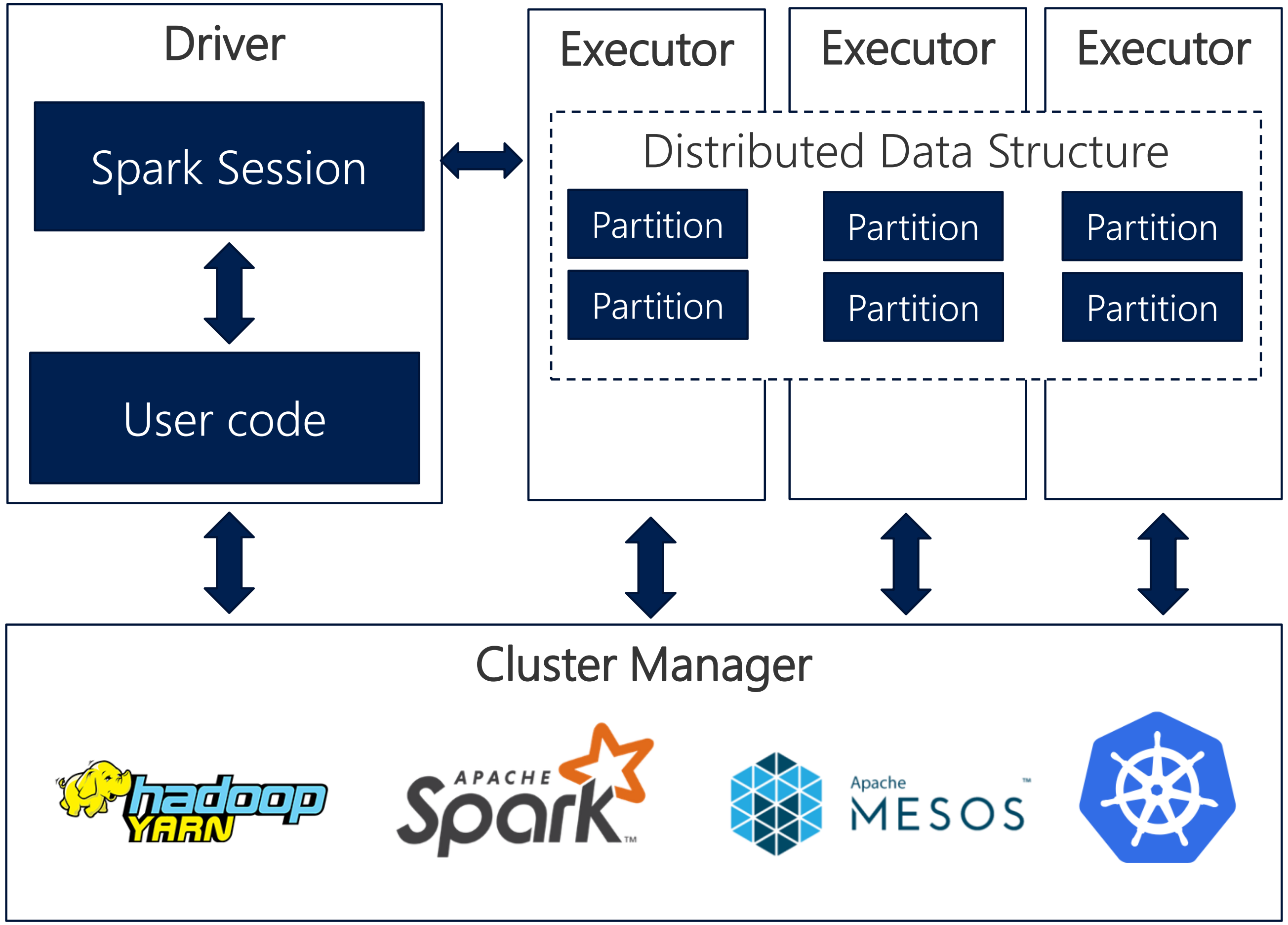
Transformations and Actions



Transformations and Actions

Transformations	Actions
`select`	`show`
`distinct`	`count`
`groupBy`	`collect`
`sum`	`save`
`orderBy`	`first`
`filter`	
`limit`	
`summarize`	
… `and much more`	

Inside a Spark Application



Azure Databricks

Spark as a managed service on Azure

Azure Databricks

Managed Apache Spark platform optimized for Azure

First party service

- Not an Azure Marketplace or 3rd party hosted service

Azure Integration

- Azure Active Directory
- Azure data connectors
- Azure Billing
- Power BI

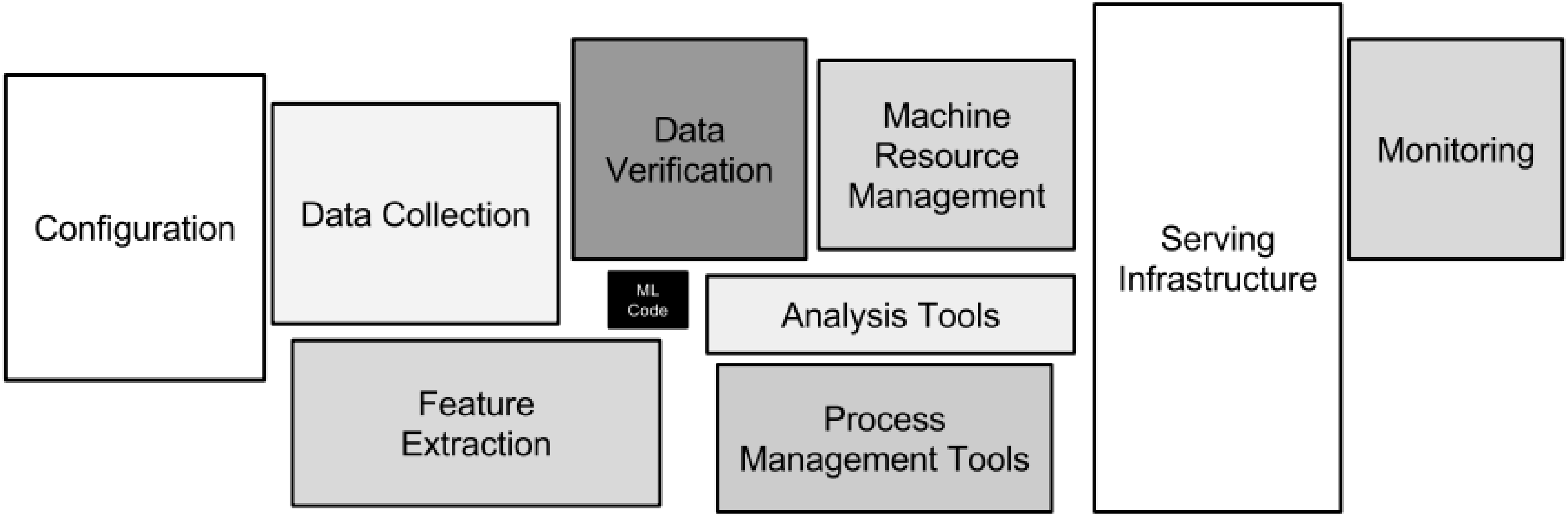
databricks™

Microsoft Azure

Demo

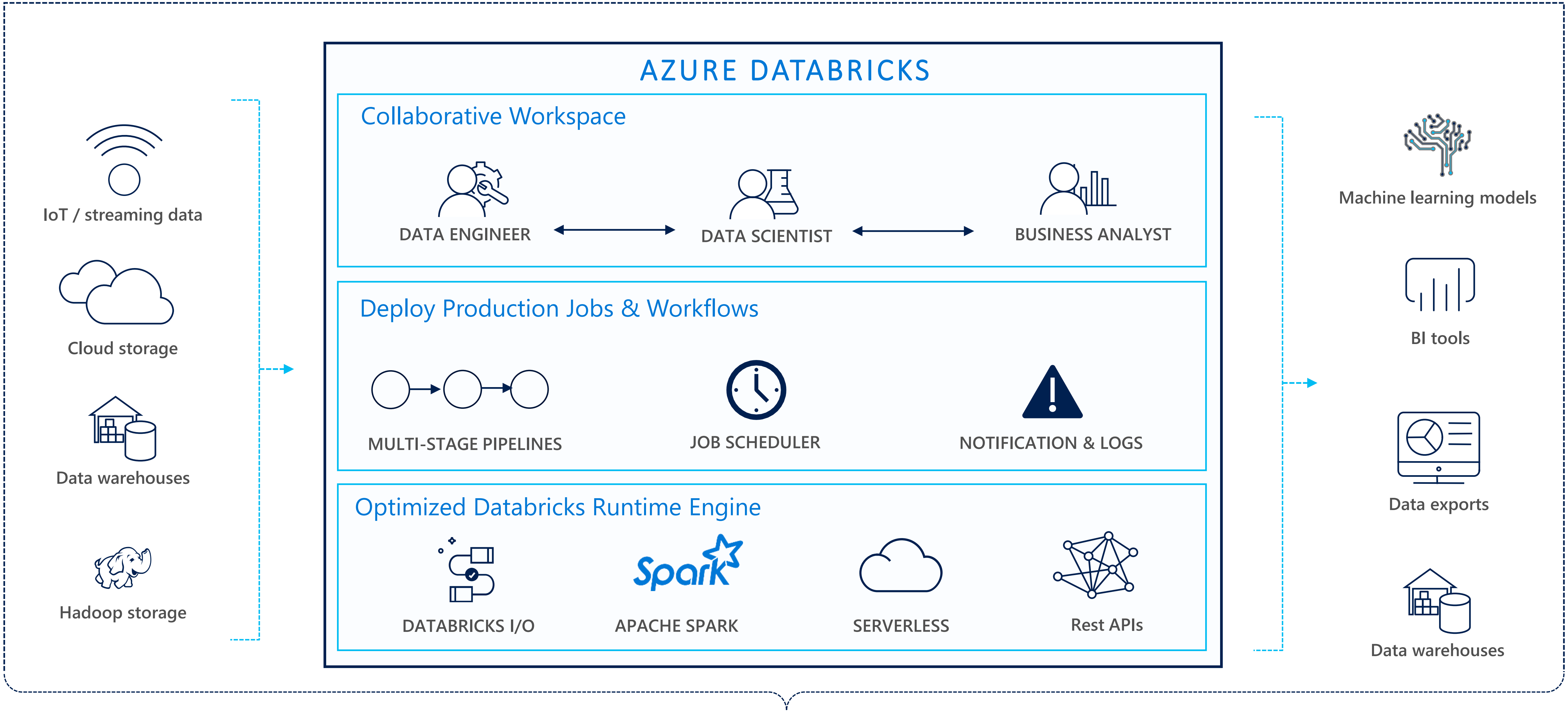
Hello Azure Databricks!

Hidden Technical Debt in ML Systems



https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

Azure Databricks



IoT / streaming data

Cloud storage

Data warehouses

Hadoop storage

AZURE DATABRICKS

Collaborative Workspace

DATA ENGINEER ↔ DATA SCIENTIST ↔ BUSINESS ANALYST

Deploy Production Jobs & Workflows

MULTI-STAGE PIPELINES

JOB SCHEDULER

NOTIFICATION & LOGS

Optimized Databricks Runtime Engine

DATABRICKS I/O

APACHE SPARK

SERVERLESS

Rest APIs

Machine learning models

BI tools

Data exports

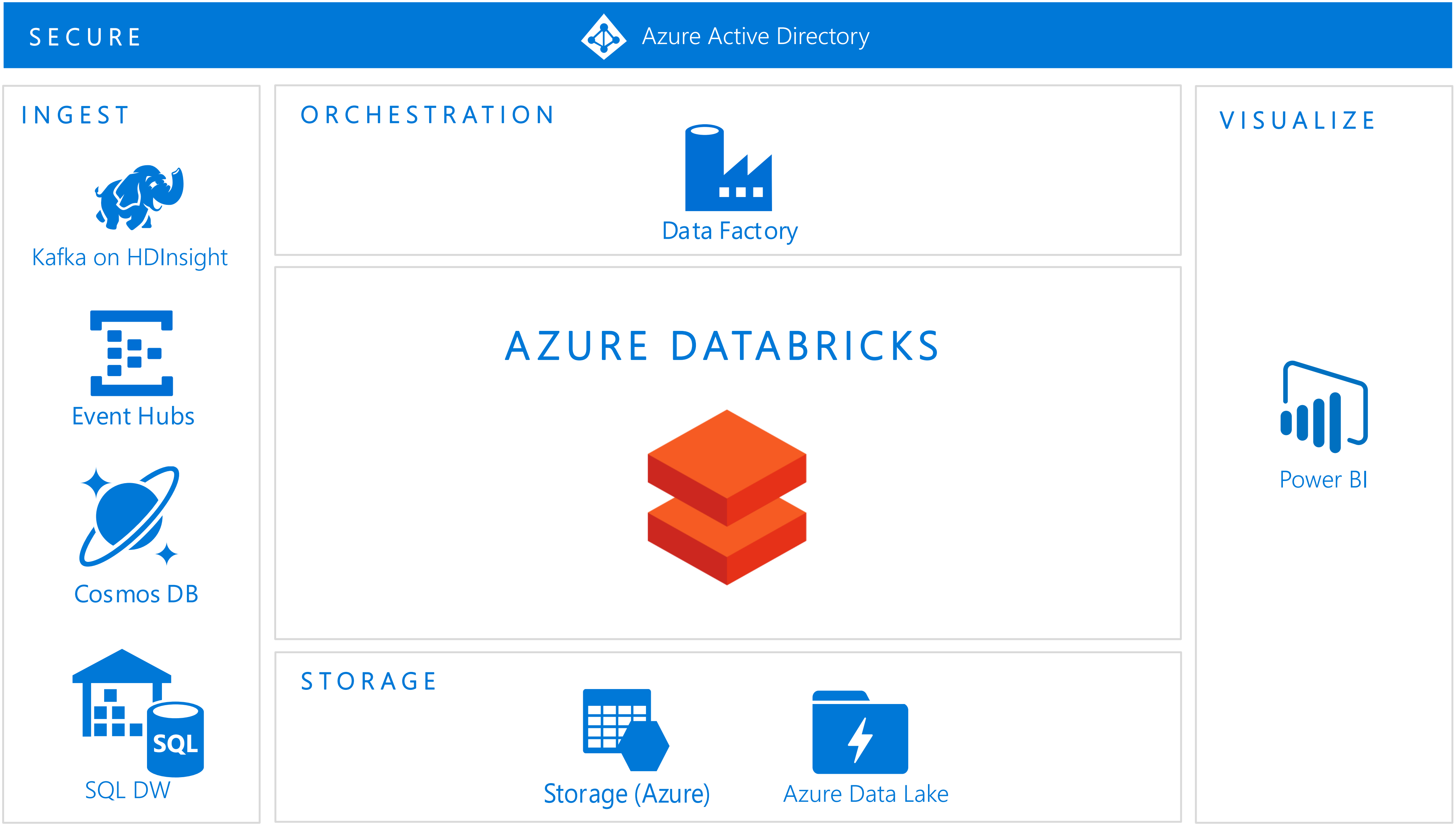
Data warehouses

Enhance Productivity

Build on secure & trusted cloud

Scale without limits

Azure Integration



SECURE

Azure Active Directory

INGEST

Kafka on HDInsight

Event Hubs

Cosmos DB

SQL DW

ORCHESTRATION

Data Factory

AZURE DATABRICKS

STORAGE

Storage (Azure)

Azure Data Lake

VISUALIZE

Power BI

Databricks Core Concepts

Workspaces

Clusters

Notebooks

Jobs

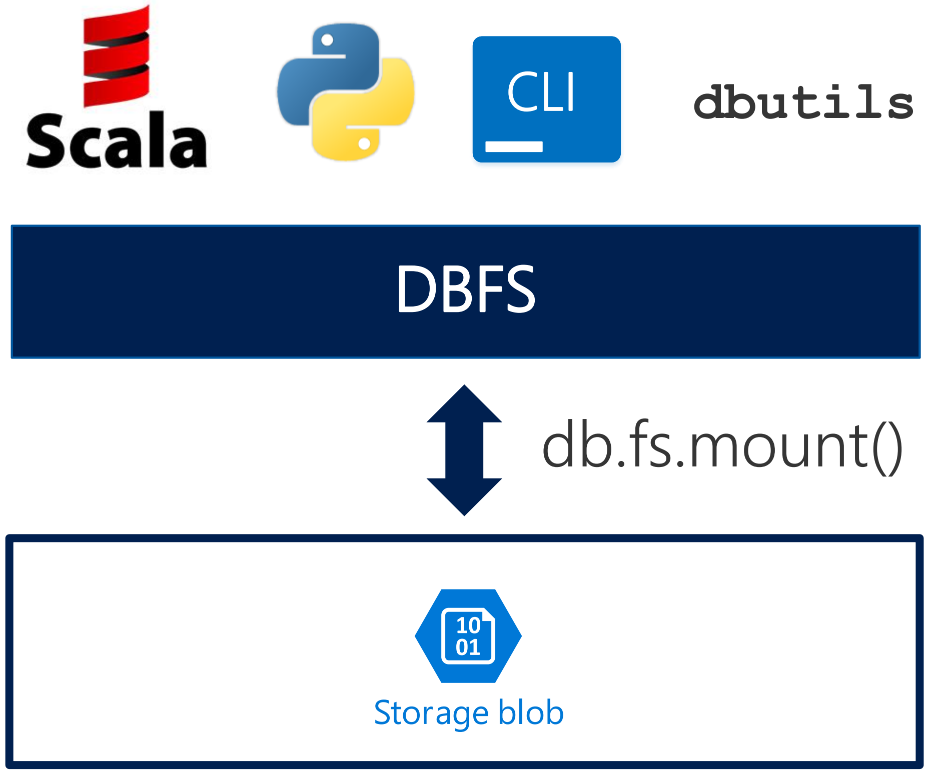
Libraries

Tables

Secrets

Databricks File System (DBFS)

- Distributed file system that is a layer over Azure Blob Storage
- Data is persisted even after cluster termination
- Data can be cached locally on the SSD of the worker nodes
- Available in Python and Scala and accessible via DBFS CLI



Demo

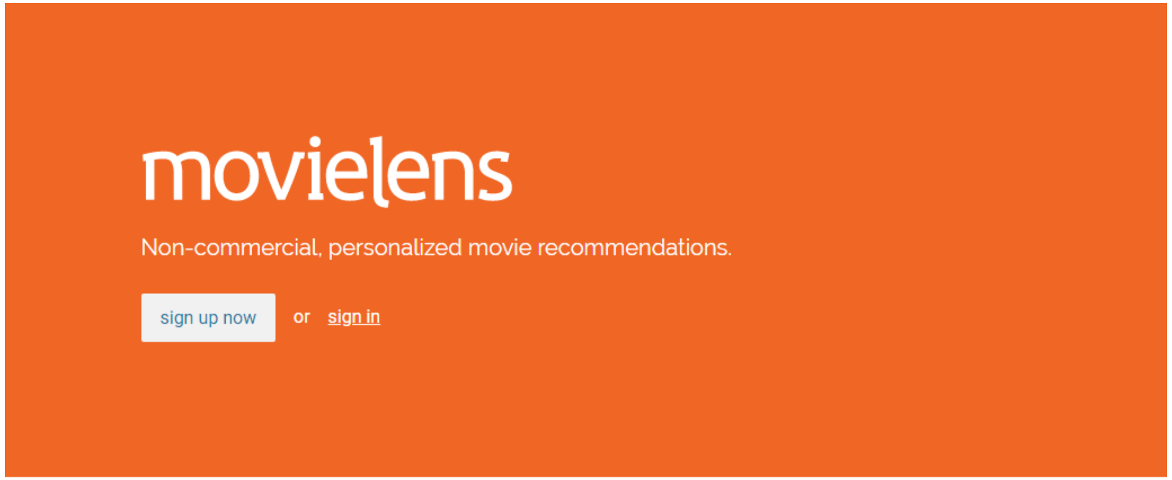
Mount Blob Storage in DBFS

Movie Recommendation System

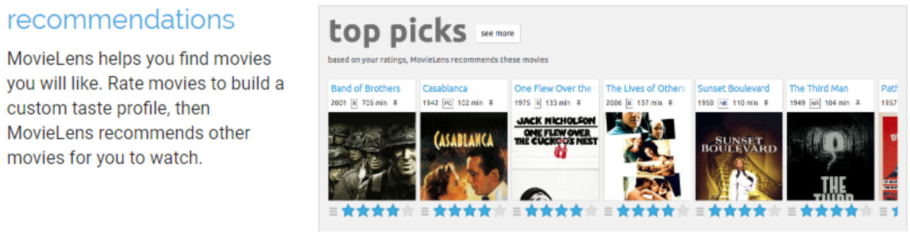
MovieLens Dataset

26M ratings and 750K tag applications applied to 45K movies by 270K users

https://movielens.org/



Non-commercial, personalized movie recommendations.



recommendations

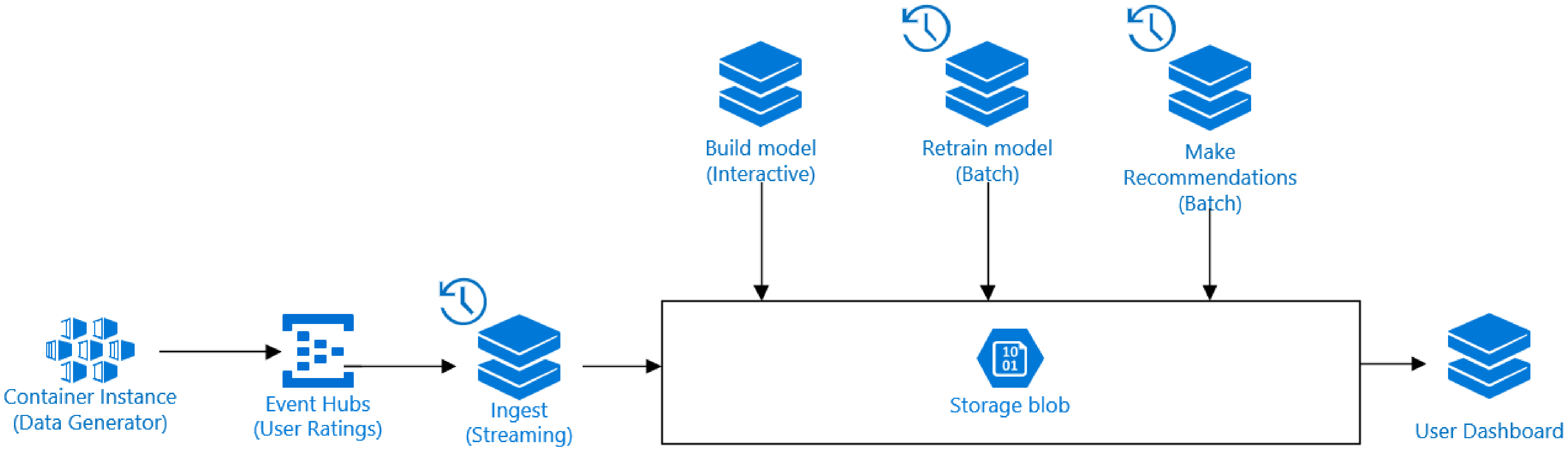
MovieLens helps you find movies you will like. Rate movies to build a custom taste profile, then MovieLens recommends other movies for you to watch.

top picks

based on your ratings, MovieLens recommends these movies

F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19 (December 2015), 19 pages. DOI=http://dx.doi.org/10.1145/2827872

Demo Architecture



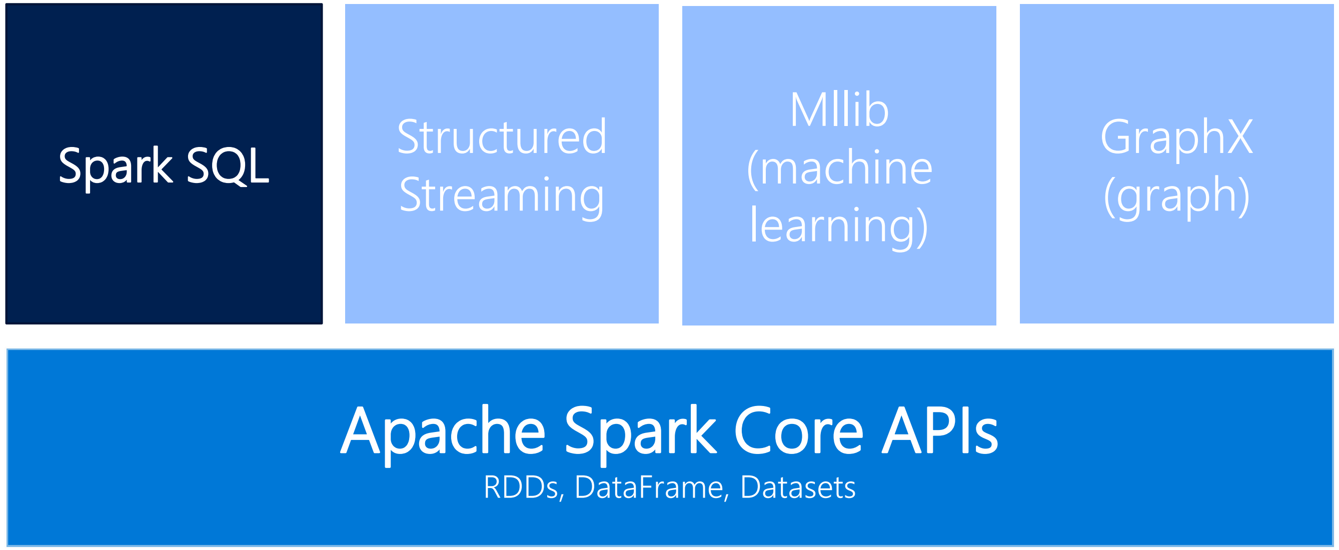
Spark SQL

Spark's interface for working with structured and semi-structured data

Built on the DataFrame & Datasets API

Hive Integration

Provides JDBC/ODBC access



Spark SQL

Structured Streaming

Mllib (machine learning)

GraphX (graph)

Apache Spark Core APIs

RDDs, DataFrame, Datasets

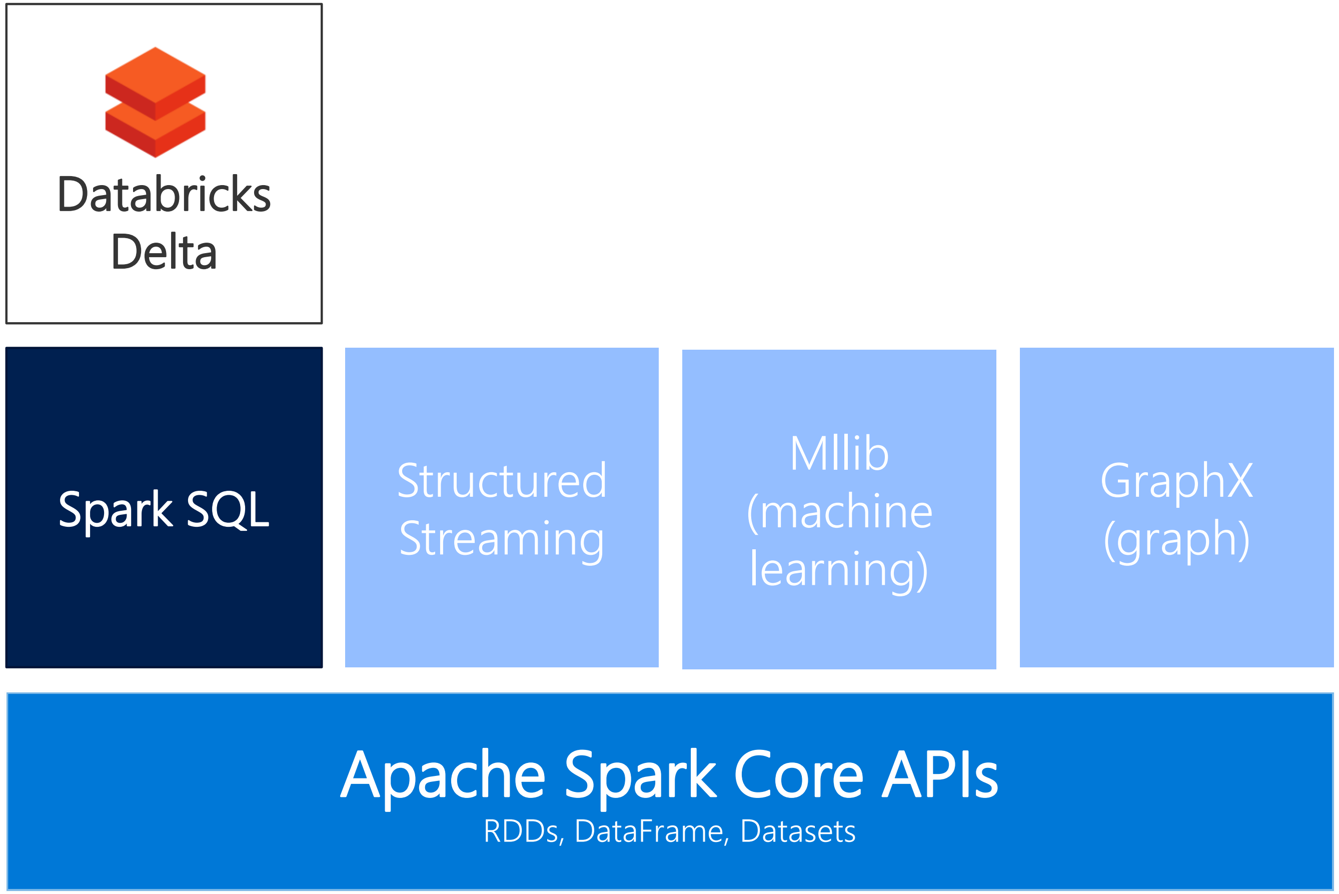
Databricks Delta

Powerful transactional storage layer using Spark & DBFS

Provides ACID transactions

Fast read access with automatic file management and table statistics

In Preview



Databricks Delta

Spark SQL

Structured Streaming

Mllib (machine learning)

GraphX (graph)

Apache Spark Core APIs

RDDs, DataFrame, Datasets

Demo

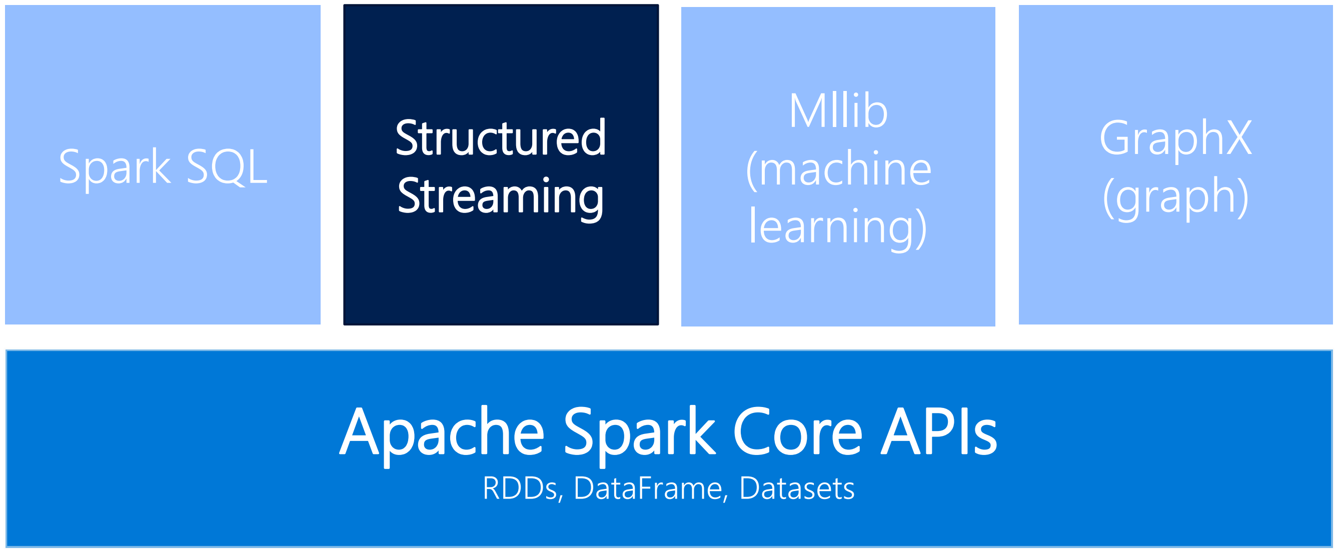
Create and query Tables with Spark SQL

Spark Structured Streaming

Scalable and fault-tolerant stream processing engine

Successor of Spark Streaming (DStreams API)

Same code for Batch and Streaming



Spark SQL

Structured Streaming

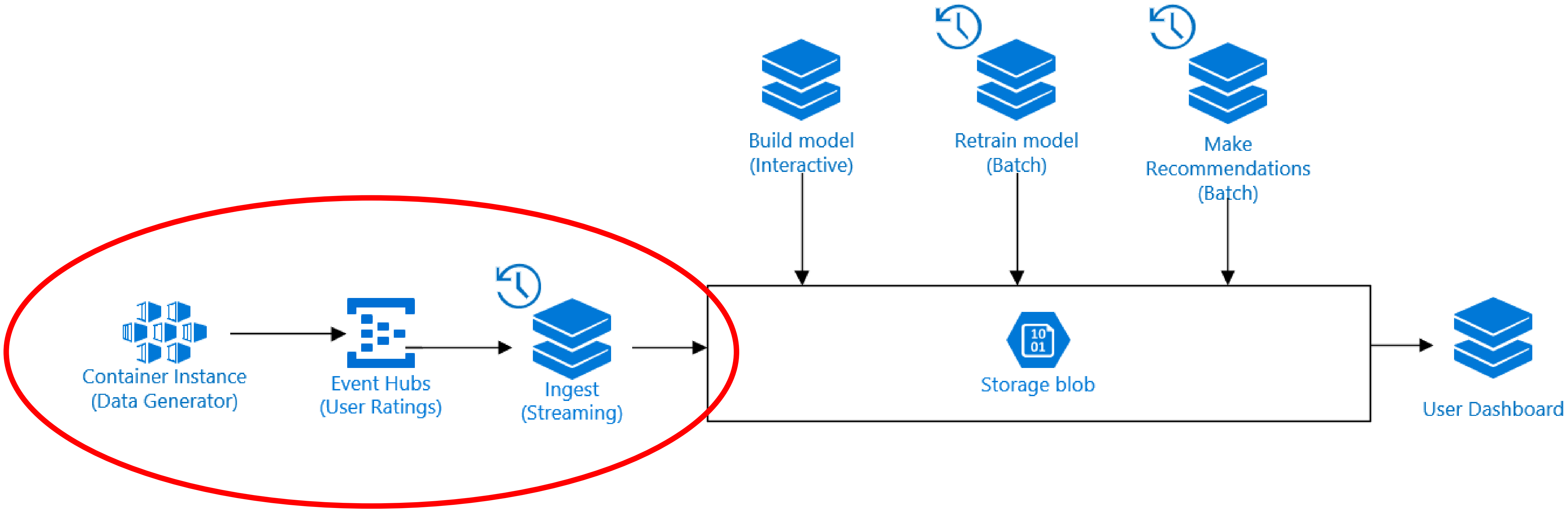
Mllib (machine learning)

GraphX (graph)

Apache Spark Core APIs

RDDs, DataFrame, Datasets

Demo Architecture



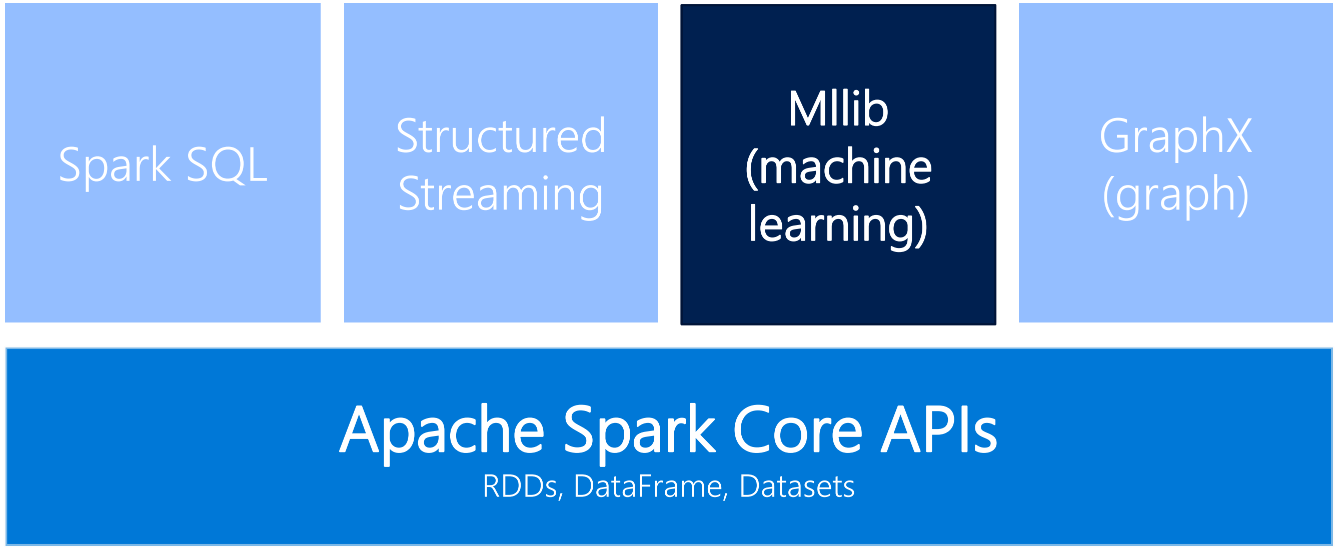
Demo

Ingest ratings data from Event Hubs with Spark Structured Streaming

Spark MLlib

Scalable Machine Learning library on Spark

- Common ML algorithms
  - classification, regression, clustering, & collaborative filtering
- Featurization
  - Feature extraction, Transformation, dimensionality reduction
- ML Pipelines
  - Combine Transformers and Estimators



Spark SQL

Structured Streaming

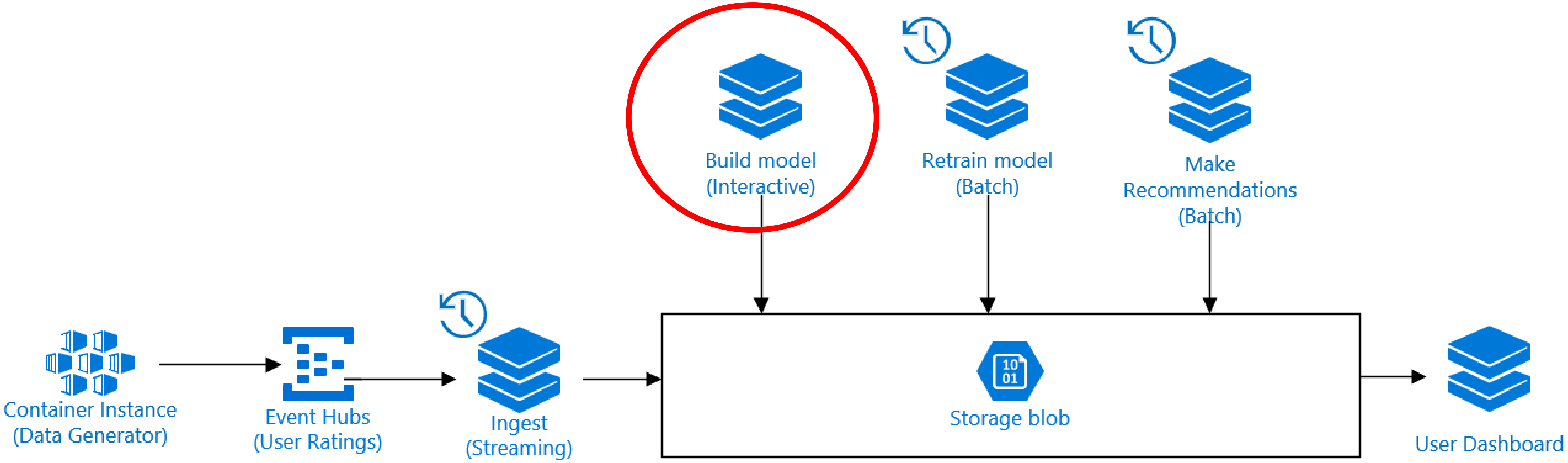
Mllib (machine learning)

GraphX (graph)

Apache Spark Core APIs

RDDs, DataFrame, Datasets

Demo Architecture



Demo

Build collaborative filtering recommendation model with Spark ML

Productionizing Machine Learning Workloads

In Spark…

1. Batch inference
2. Structured Streaming

Out of Spark…

Export model

- Mleap, MLFlow Models, AzureML Service

Containerized Web Service

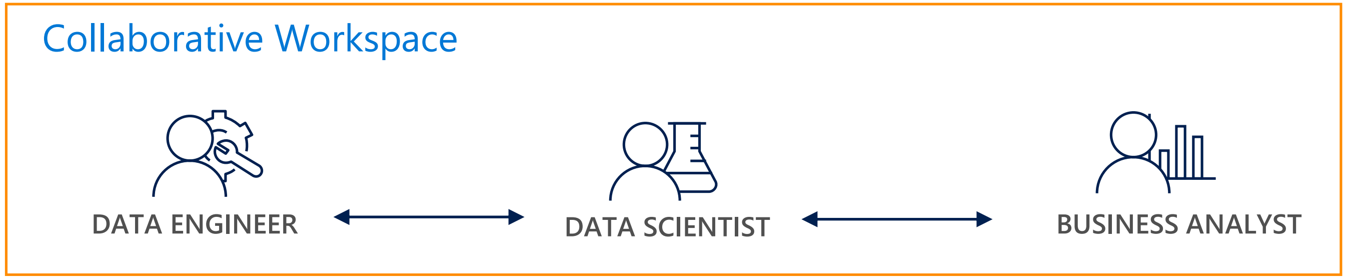
Productionizing Machine Learning Workloads

ML persistence

- Sparks support saving multi-stage models built by Data Scientist in Python/R and loading in Scala/Java

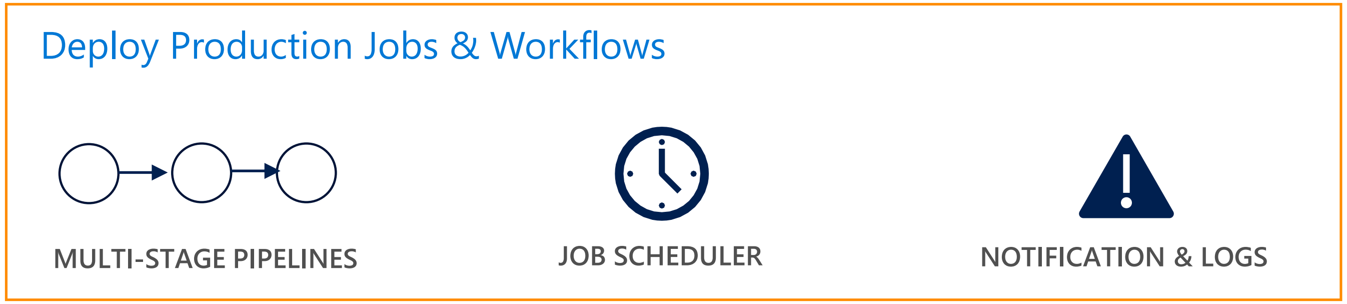
Schedule pipelines with Jobs

Notification and alerting



Collaborative Workspace

DATA ENGINEER ↔ DATA SCIENTIST ↔ BUSINESS ANALYST



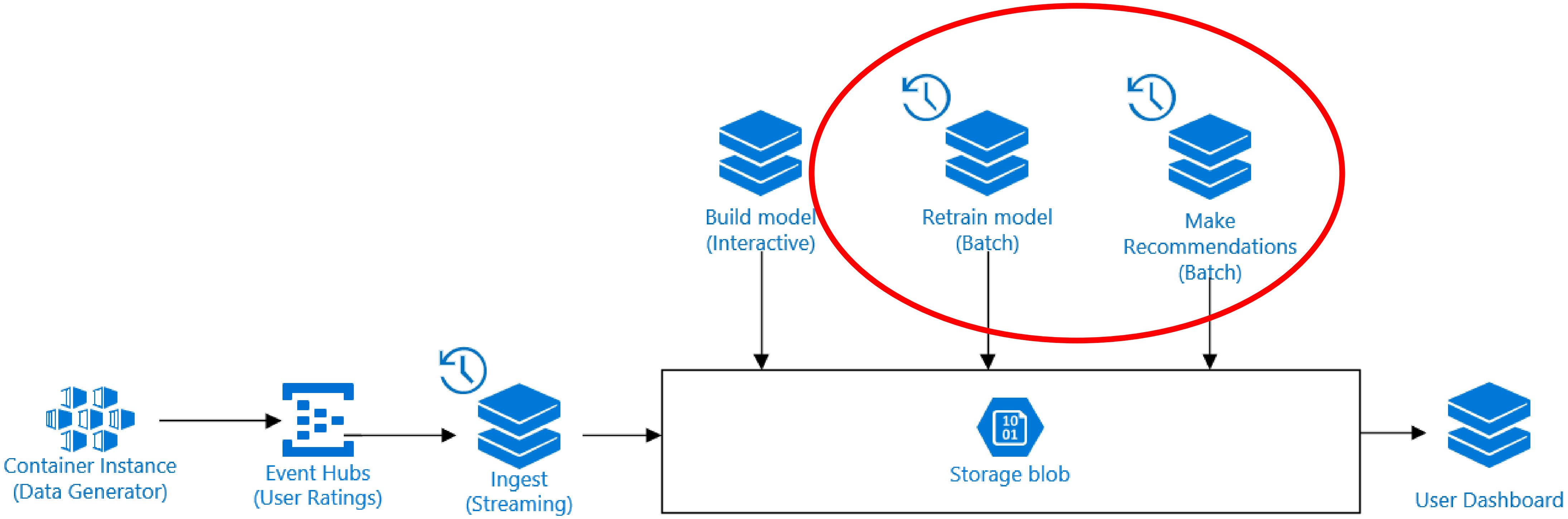
Deploy Production Jobs & Workflows

MULTI-STAGE PIPELINES

JOB SCHEDULER

NOTIFICATION & LOGS

Demo Architecture



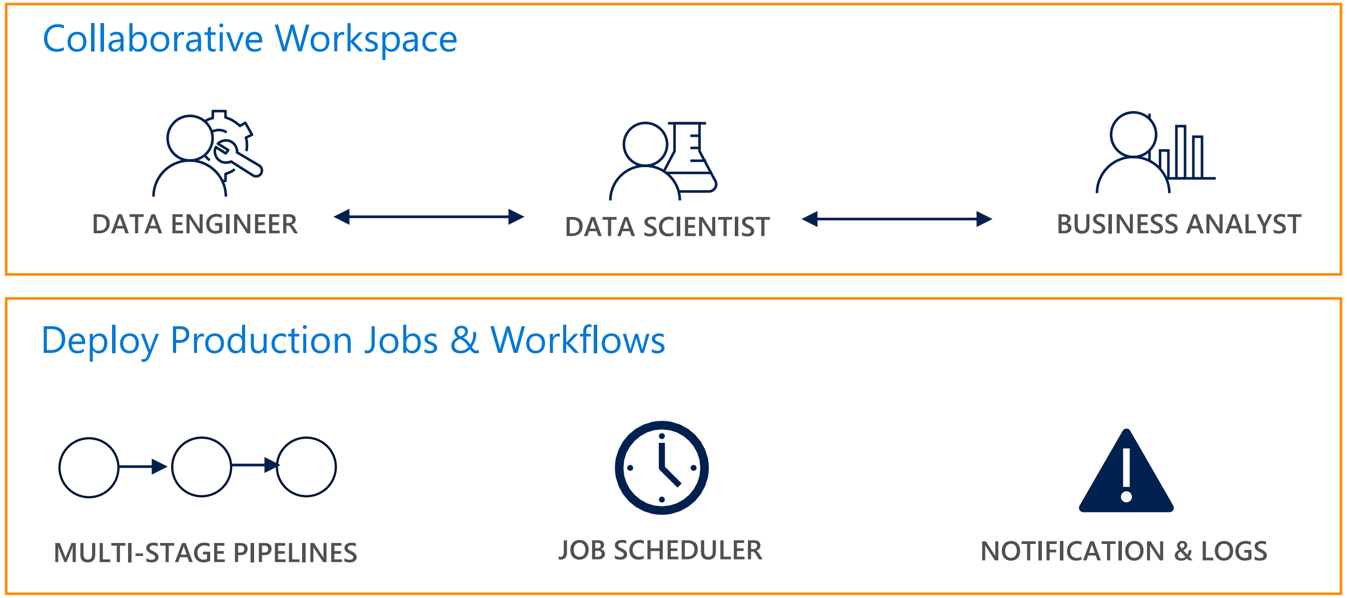
Demo

Productionize workflow with Spark Jobs

Visualize with Dashboards

Convert Notebooks into Dashboards

Parameterize Notebooks using Widgets



Collaborative Workspace

DATA ENGINEER ↔ DATA SCIENTIST ↔ BUSINESS ANALYST

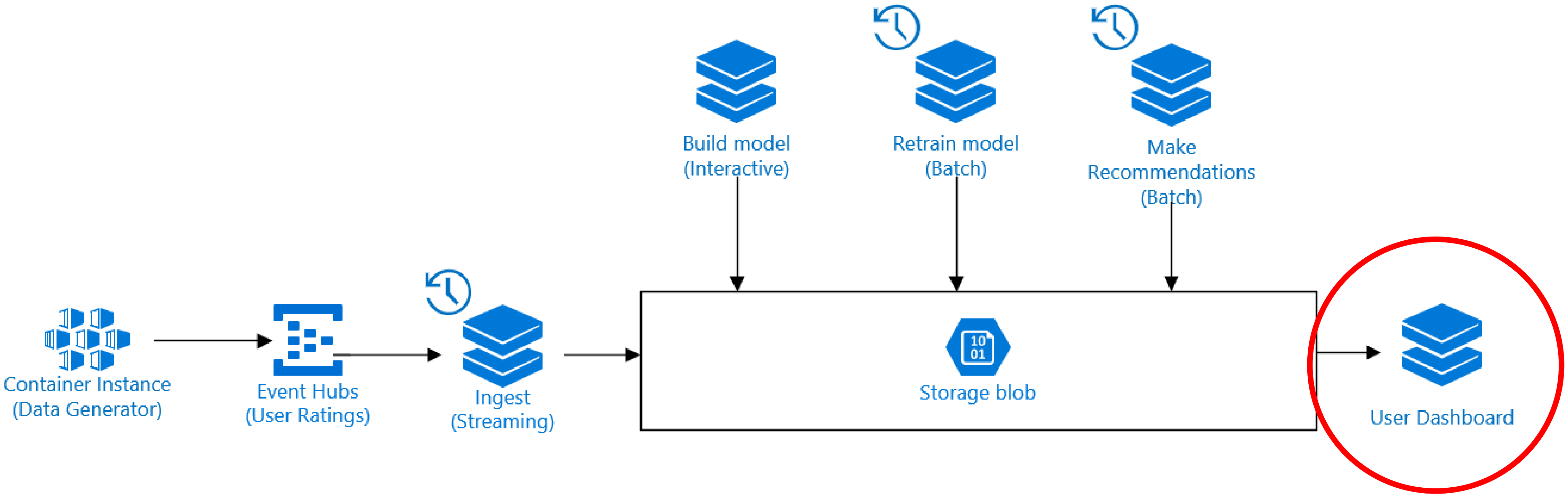
Deploy Production Jobs & Workflows

MULTI-STAGE PIPELINES

JOB SCHEDULER

NOTIFICATION & LOGS

Demo Architecture



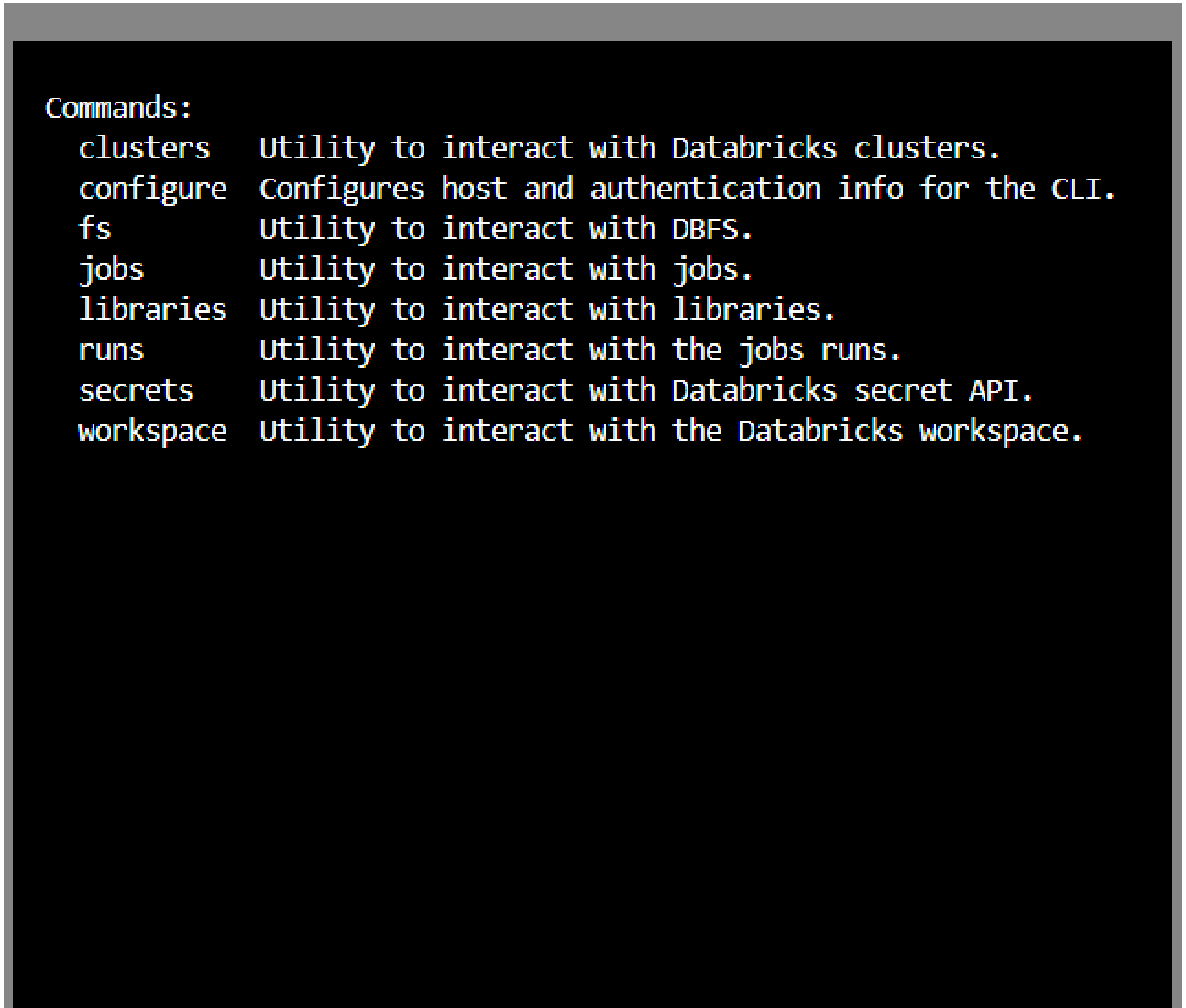
Demo

User Recommendation Dashboard

Databricks Developer Tooling

Databricks CLI

Databricks REST API



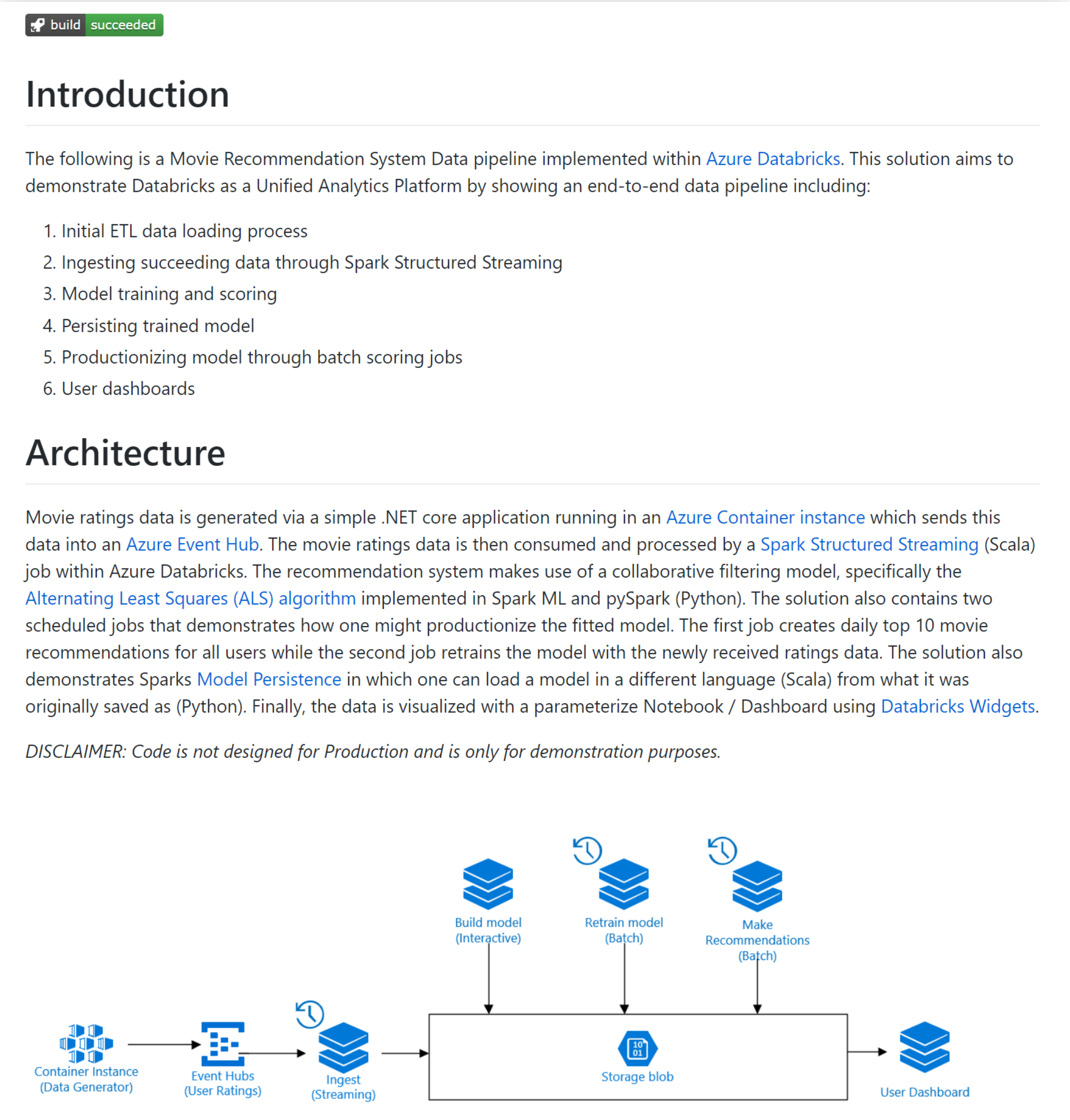
```
Commands:
  clusters   Utility to interact with Databricks clusters.
  configure  Configures host and authentication info for the CLI.
  fs         Utility to interact with DBFS.
  jobs       Utility to interact with jobs.
  libraries  Utility to interact with libraries.
  runs       Utility to interact with the jobs runs.
  secrets    Utility to interact with Databricks secret API.
  workspace  Utility to interact with the Databricks workspace.
```

Try the demo!

https://github.com/devlace/azure-databricks-recommendation-system

To deploy…

```
docker run -it
devlace/azdatabricksrecommend
```



Introduction

The following is a Movie Recommendation System Data pipeline implemented within Azure Databricks. This solution aims to demonstrate Databricks as a Unified Analytics Platform by showing an end-to-end data pipeline including:

1. Initial ETL data loading process
2. Ingesting succeeding data through Spark Structured Streaming
3. Model training and scoring
4. Persisting trained model
5. Productionizing model through batch scoring jobs
6. User dashboards

Architecture

Movie ratings data is generated via a simple .NET core application running in an Azure Container instance which sends this data into an Azure Event Hub. The movie ratings data is then consumed and processed by a Spark Structured Streaming (Scala) job within Azure Databricks. The recommendation system makes use of a collaborative filtering model, specifically the Alternating Least Squares (ALS) algorithm implemented in Spark ML and pySpark (Python). The solution also contains two scheduled jobs that demonstrates how one might productionize the fitted model. The first job creates daily top 10 movie recommendations for all users while the second job retrains the model with the newly received ratings data. The solution also demonstrates Sparks Model Persistence in which one can load a model in a different language (Scala) from what it was originally saved as (Python). Finally, the data is visualized with a parameterize Notebook / Dashboard using Databricks Widgets.

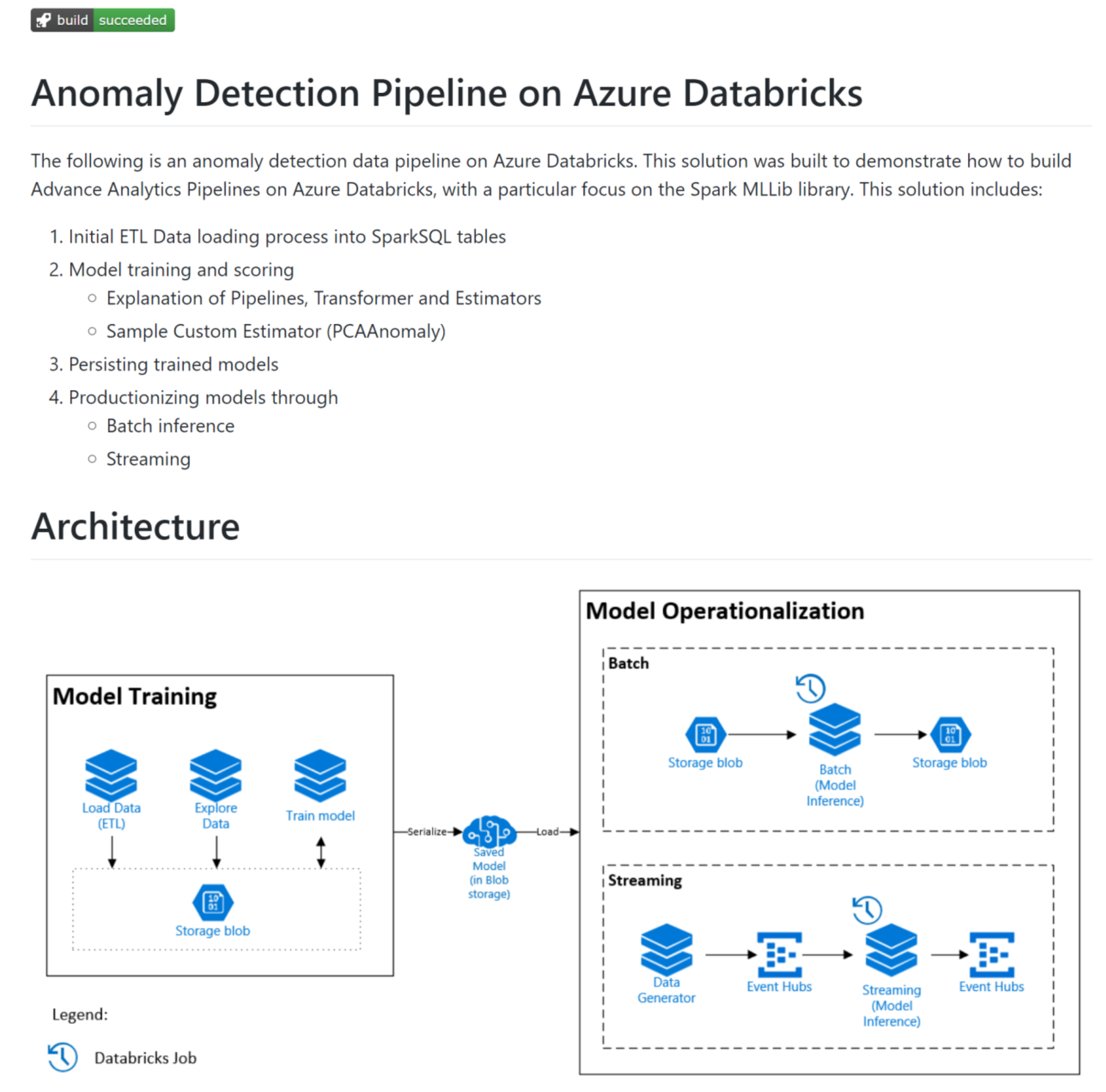
DISCLAIMER: Code is not designed for Production and is only for demonstration purposes.

Other Databricks Demos…

https://github.com/devlace/azure-databricks-anomaly

To deploy…

```
docker run –it
devlace/azdatabricksanomaly
```



build succeeded

Anomaly Detection Pipeline on Azure Databricks

The following is an anomaly detection data pipeline on Azure Databricks. This solution was built to demonstrate how to build Advance Analytics Pipelines on Azure Databricks, with a particular focus on the Spark MLLib library. This solution includes:

1. Initial ETL Data loading process into SparkSQL tables
2. Model training and scoring
   - Explanation of Pipelines, Transformer and Estimators
   - Sample Custom Estimator (PCAAnomaly)
3. Persisting trained models
4. Productionizing models through
   - Batch inference
   - Streaming

Architecture

More resources

Official Apache Spark website

Azure Databricks Documentation

[Book] Spark: The Definitive Guide

Microsoft

Thank you!

Lace Lofranco
Senior Software Engineer, Microsoft
lace.lofranco@microsoft.com
Twitter: @LaceLofranco
Github: https://github.com/devlace