Microsoft

2018 Global Azure BOOTCAMP

Spark as a service with Azure Databricks

21 April 2018

Lace Lofranco
Senior Software Development Engineer
Microsoft

Session objective

At the end of the this session, you should:

- Know the key capabilities of the Azure Databricks platform and its integration with Azure services
- Have a basic understanding of building advance analytics workloads with Spark on Azure Databricks

Hello, I'm Lace

- **Senior Software Development Engineer** in the **Commercial Software Engineering** Team in **Microsoft**
- Focus on Big Data analytics, data engineering, and machine learning
- Organizer of Melbourne Azure Nights Meetup

Microsoft

AZURE NIGHTS

Agenda

Spark Fundamentals

Unified Computing Engine

Azure Databricks

Managed Apache Spark, Integrations with Azure Services

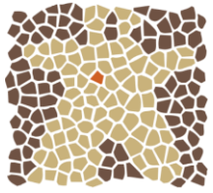
Demo

Recommendation System

Spark Fundamentals

STORM

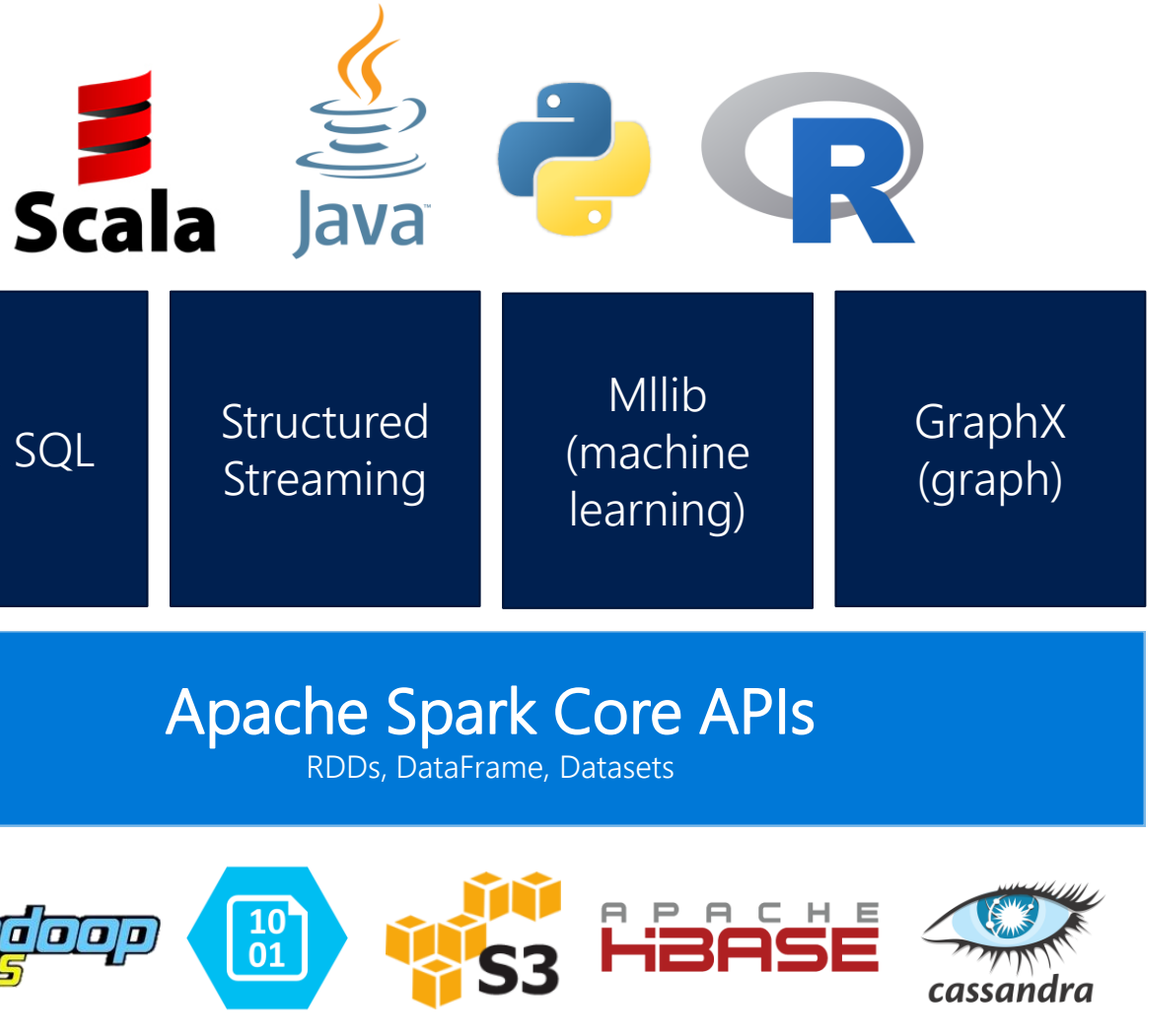
Apache Flink



APACHE
GIRAPH

Apache Spark

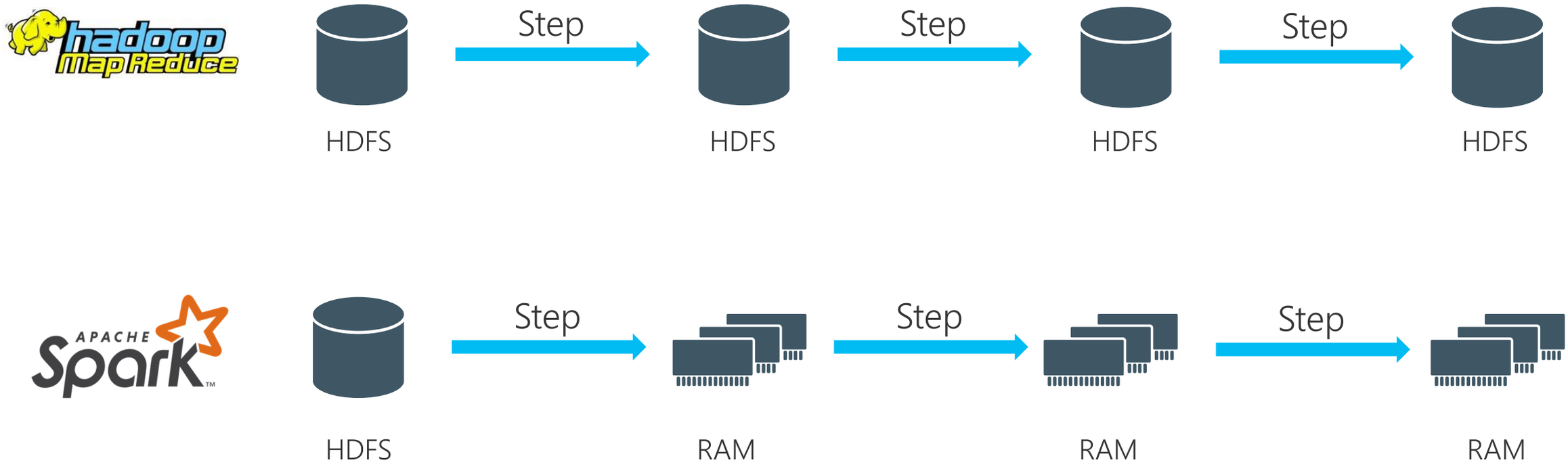
a **unified computing engine** and a set of libraries for parallel data processing on computer clusters



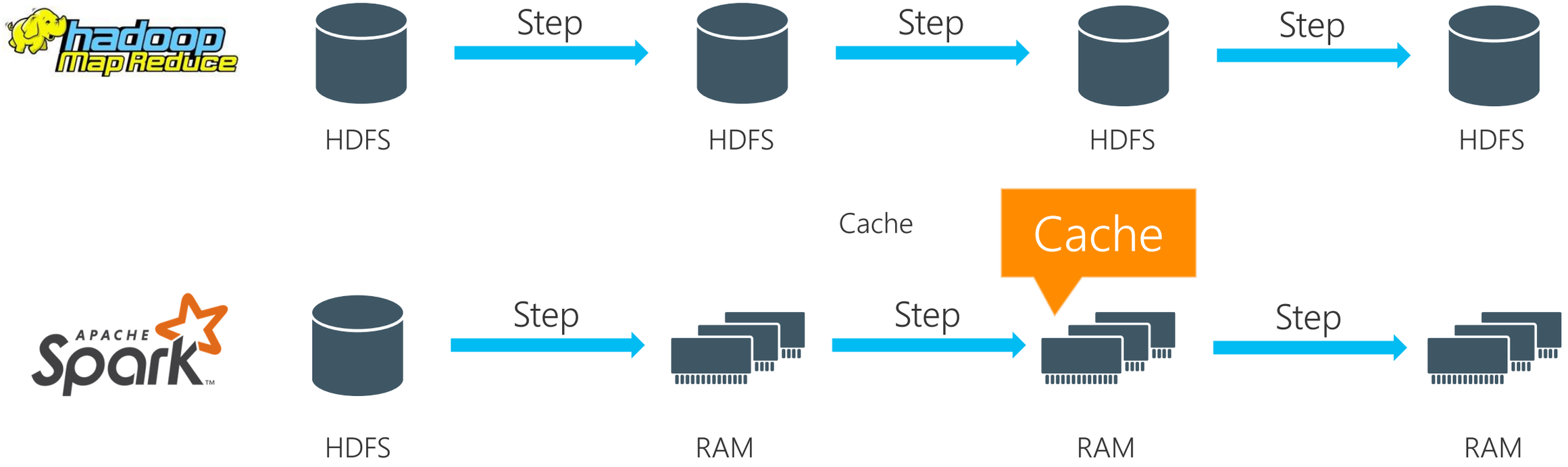
APACHE Spark™

Spark: The Definitive Guide, Matei Zaharia, Bill Chambers

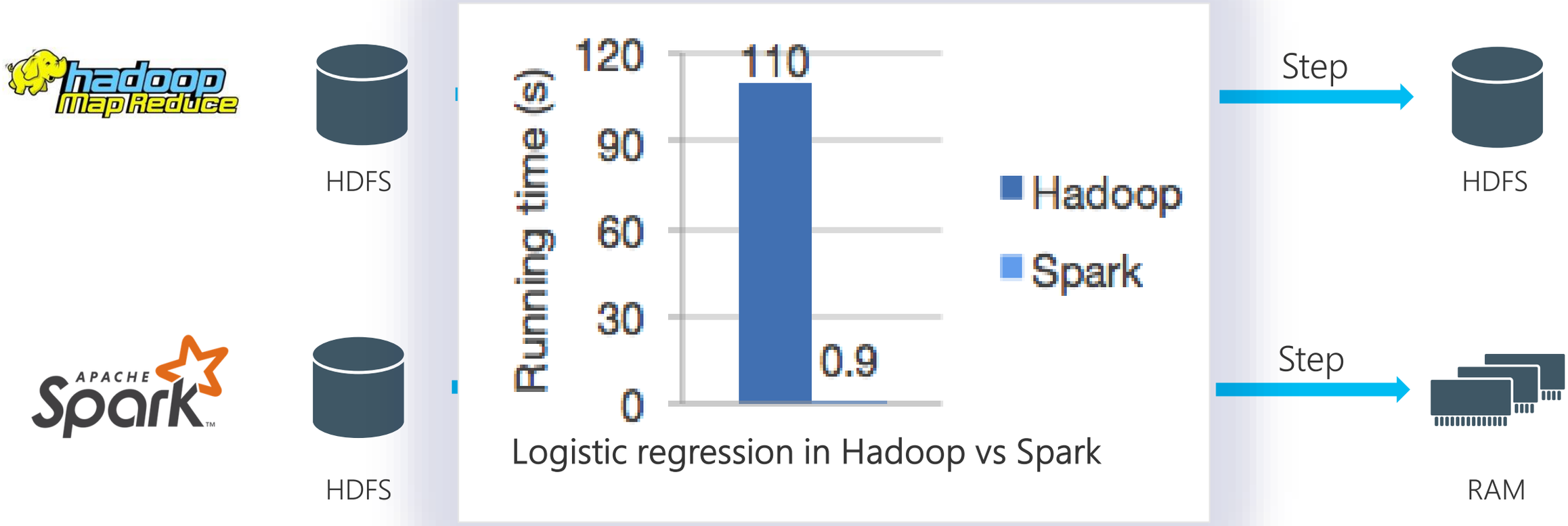
Why Spark is fast



Why Spark is fast



Why Spark is fast



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

Apache Spark: APIs

RDDs

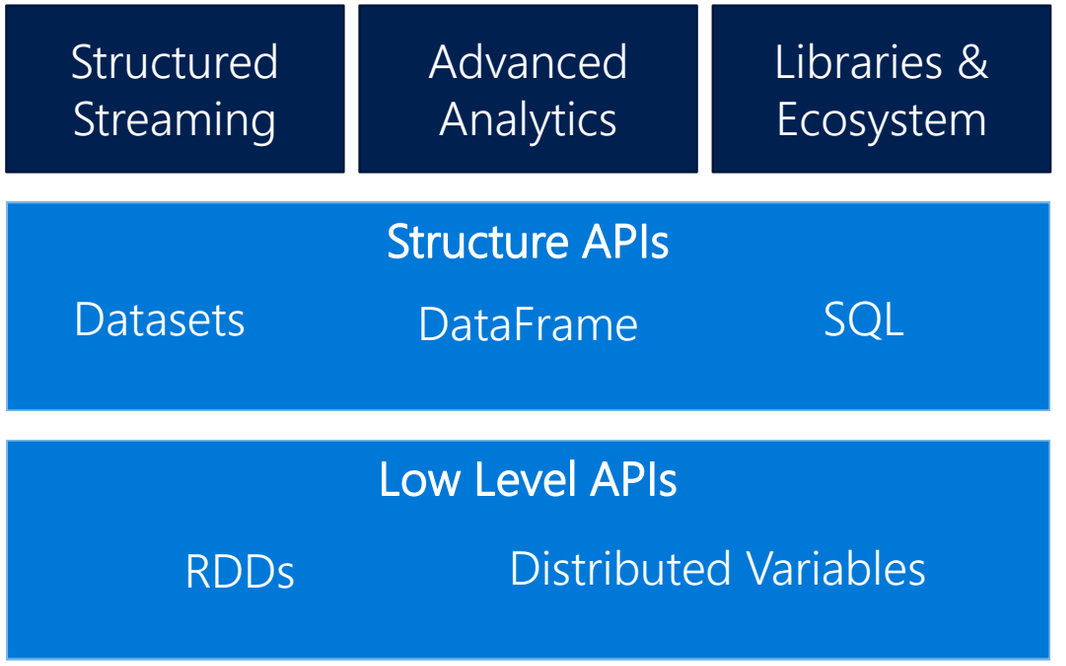
Core building block of data processing pipelines

DataFrames

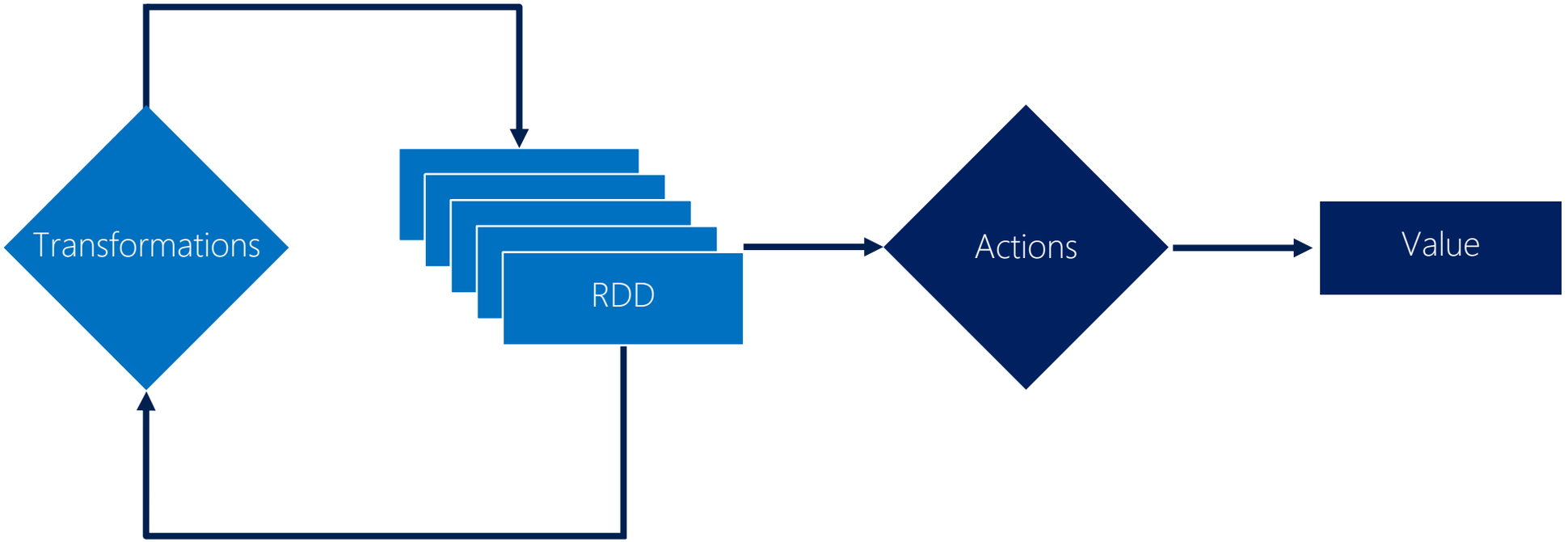
High level APIs that take advantage of query optimizer

Datasets

Data Frames with user objects and custom code



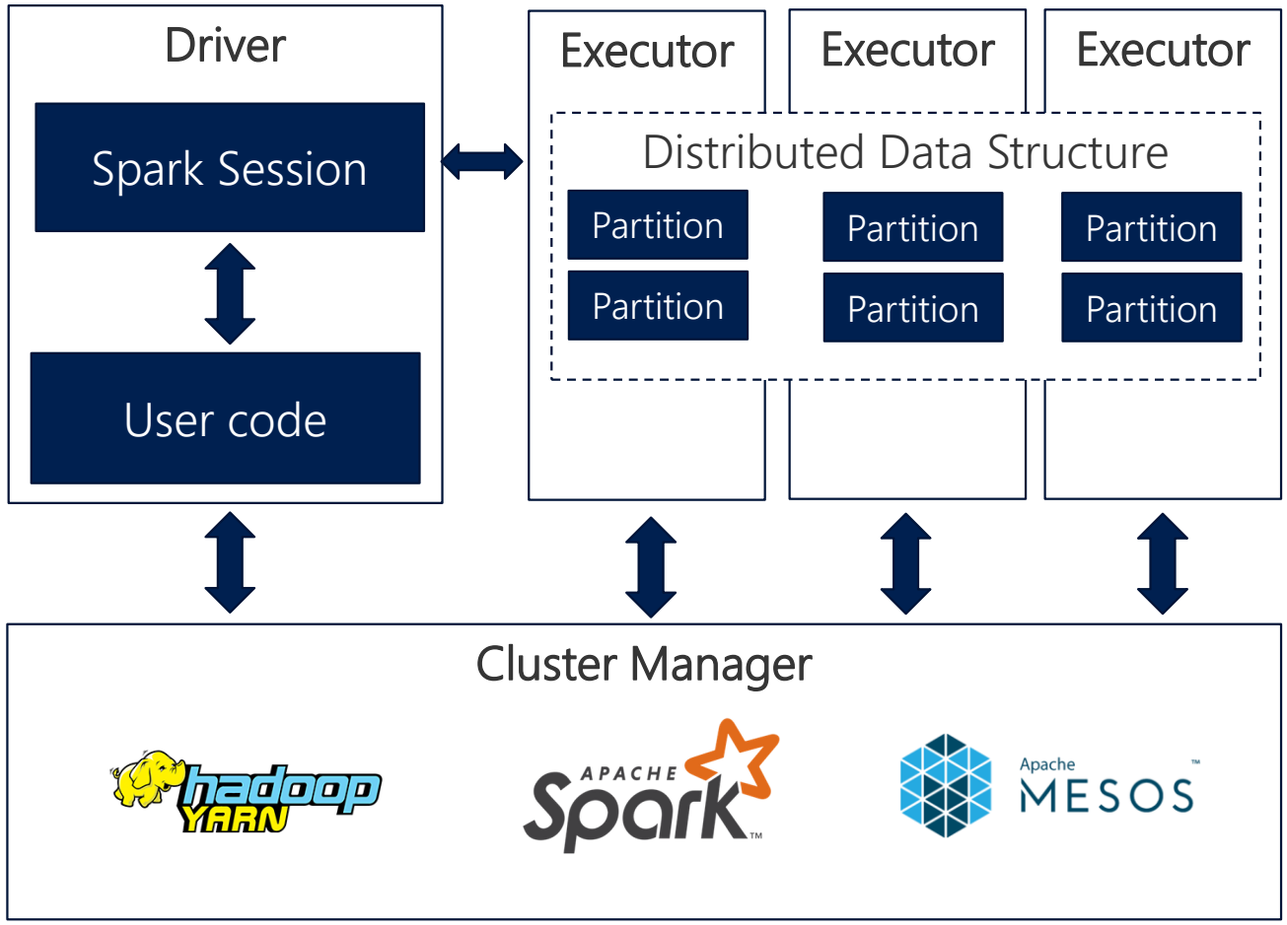
Transformations and Actions



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

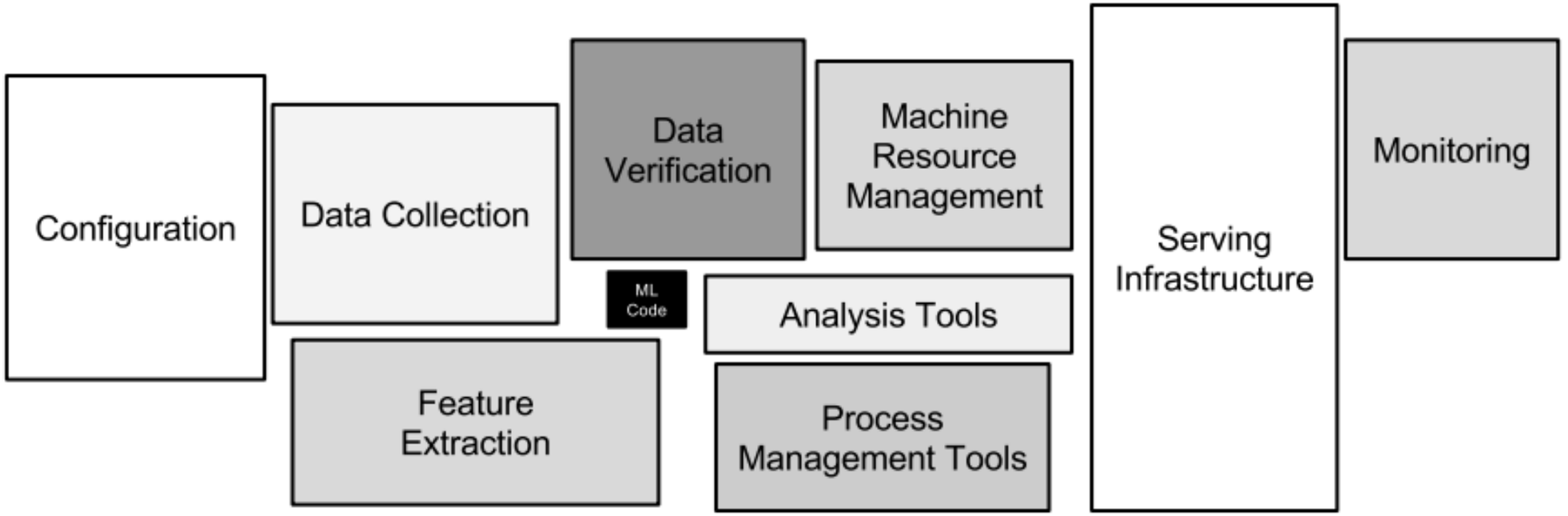
Azure Integration

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

Demo

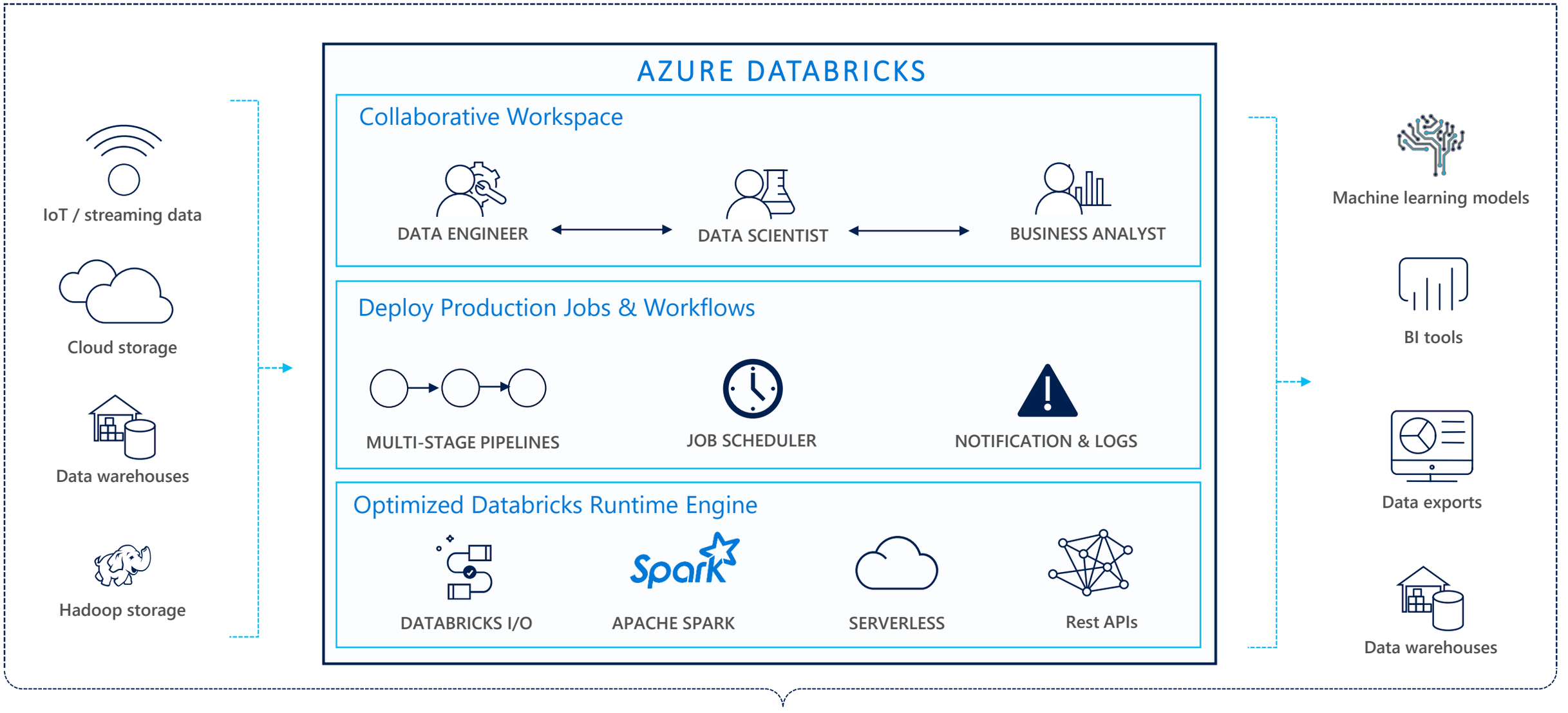
Hello Azure Databricks!

Hidden Technical Debt in ML Systems



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

Azure Databricks

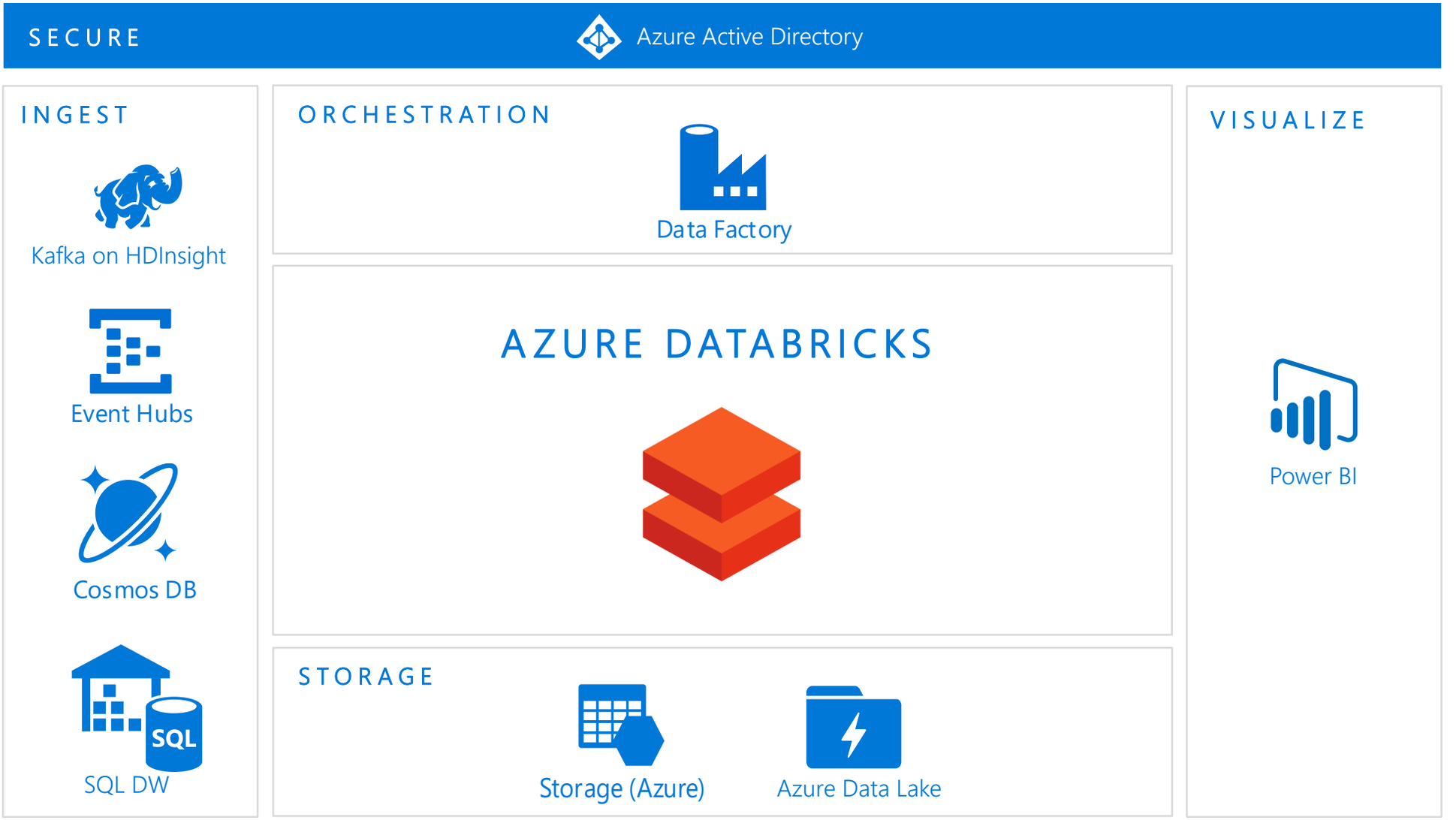


Enhance Productivity

Build on secure & trusted cloud

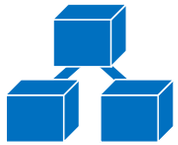
Scale without limits

Azure Integration

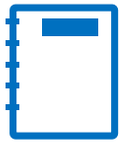


Databricks Core Concepts

Workspaces



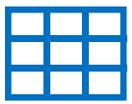
Clusters



Notebooks

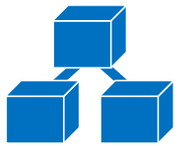
Jobs

Libraries



Tables

Databricks Core Concepts

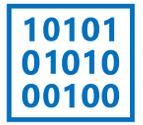


Clusters

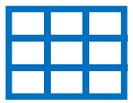
Workspaces

Notebooks

Jobs



Libraries

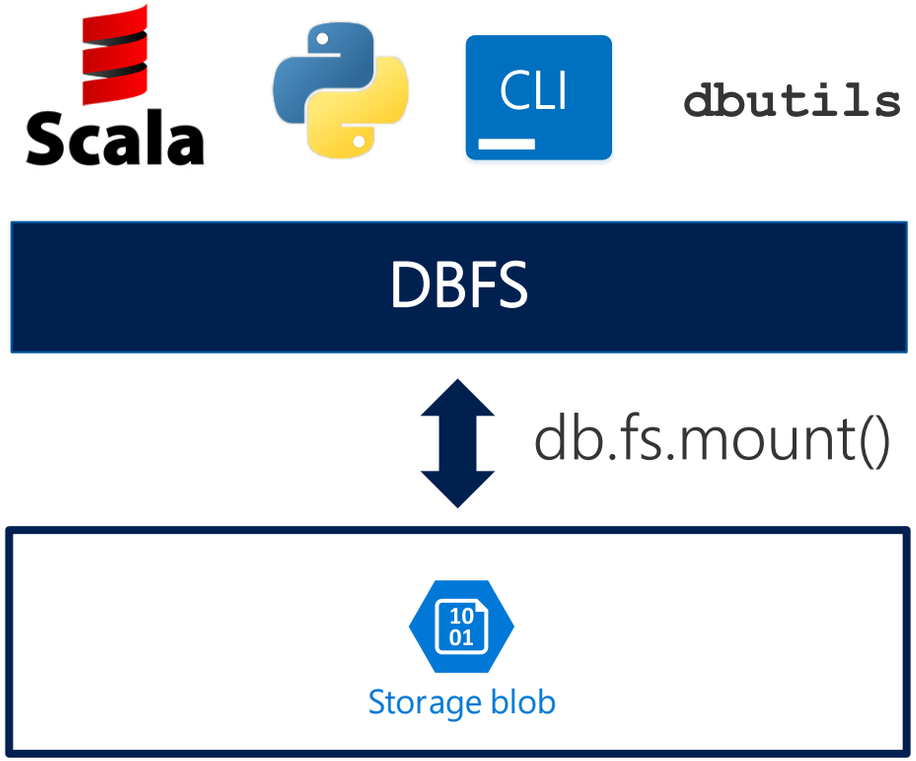


Tables

Secrets
(Preview)

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

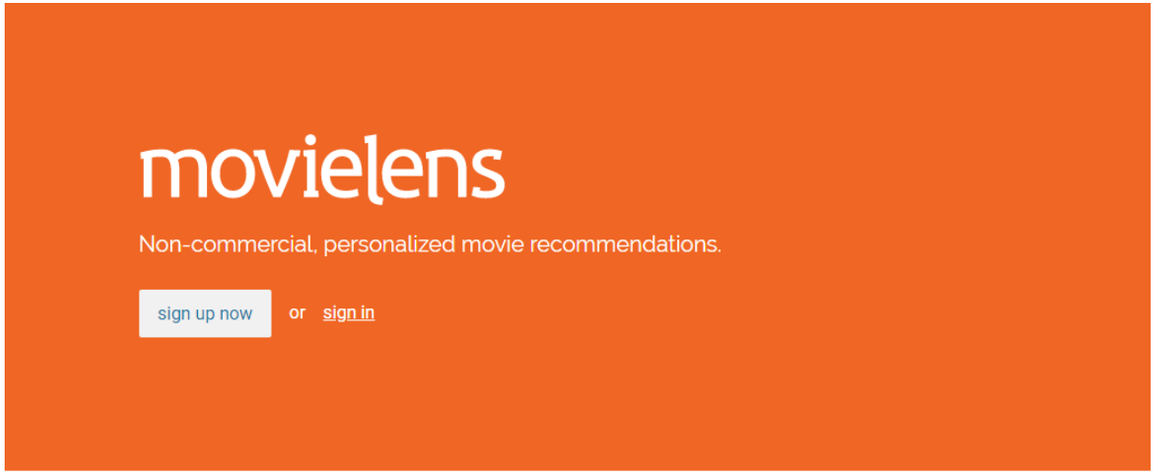
Mount Blob Storage in DBFS

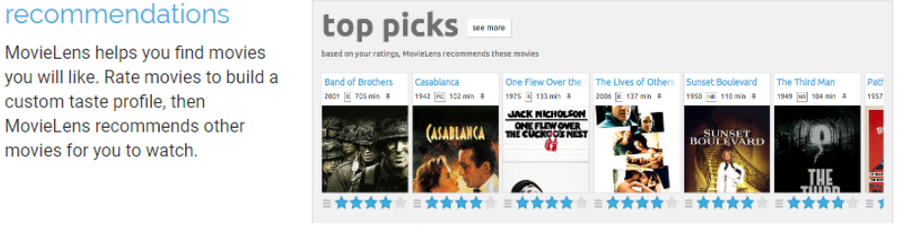
Movie Recommendation System

MovieLens Dataset

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

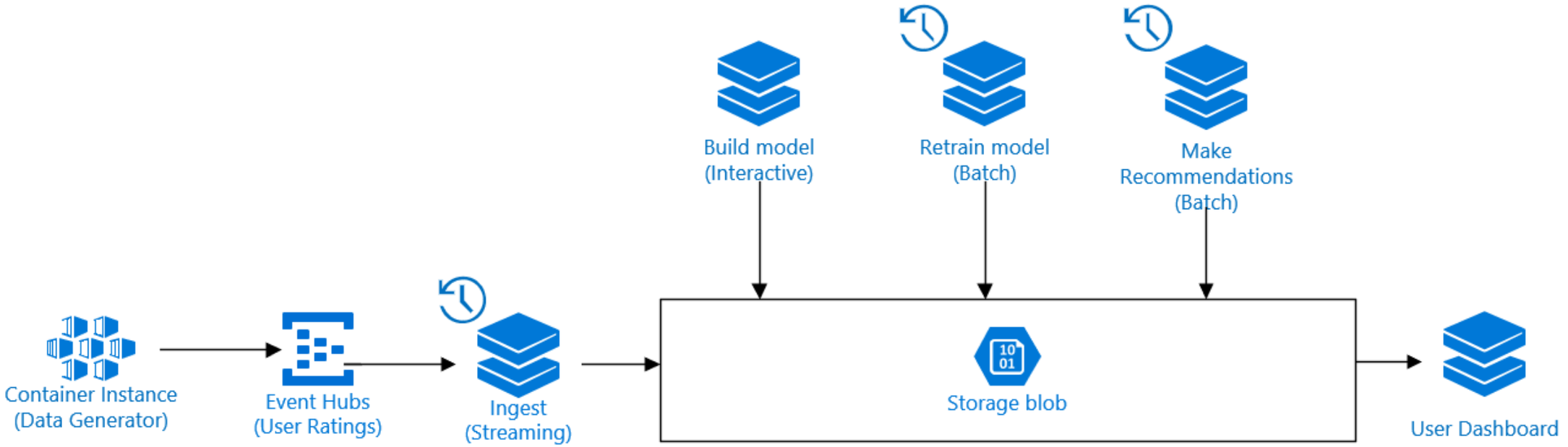
https://movielens.org/





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



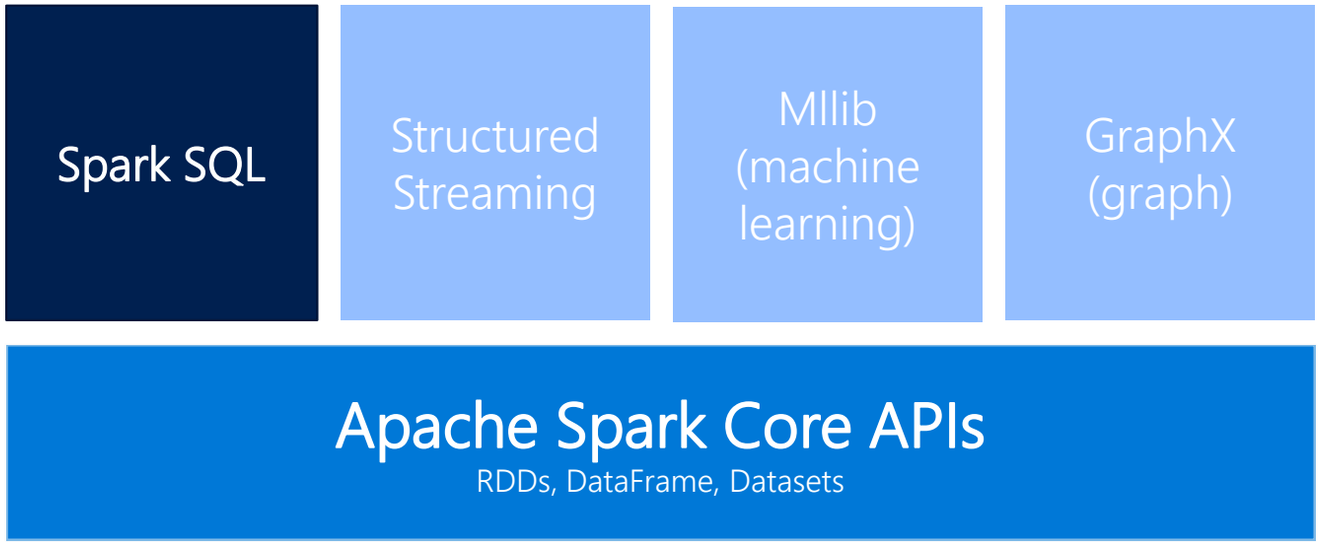
Spark SQL

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

Built on the DataFrame & Datasets API

Hive Integration

Provides JDBC/ODBC access



Demo

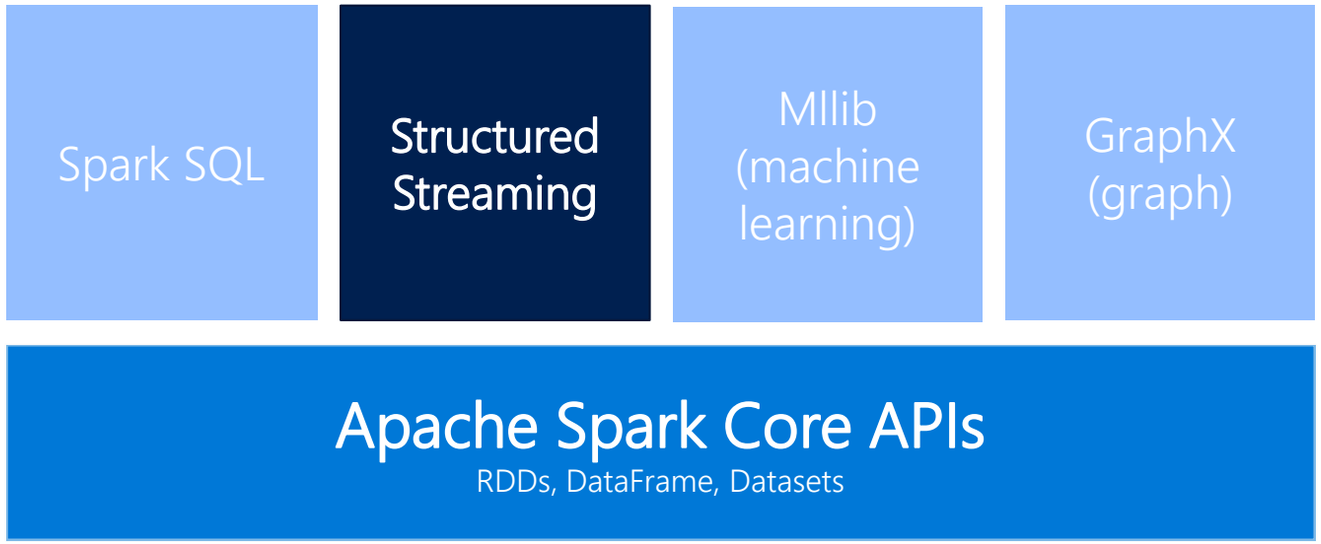
Create and query Tables with Spark SQL

Spark Structured Streaming

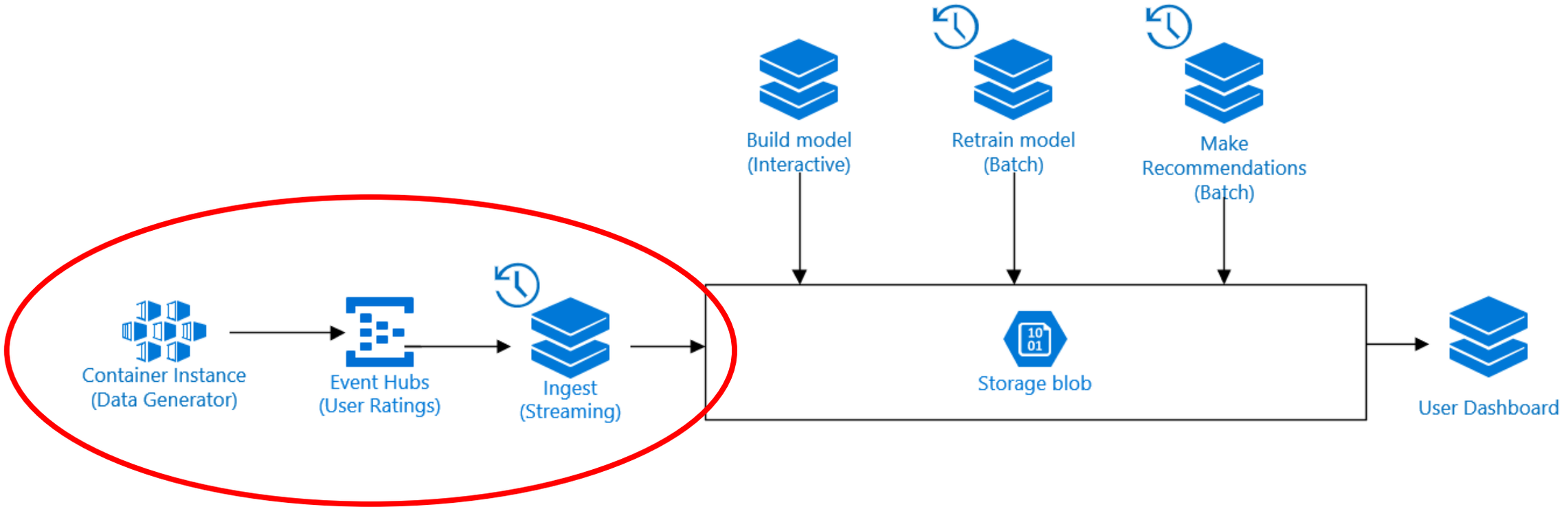
Scalable and fault-tolerant stream processing engine

Successor of Spark Streaming (DStreams API)

Same code for Batch and Streaming



Demo Architecture



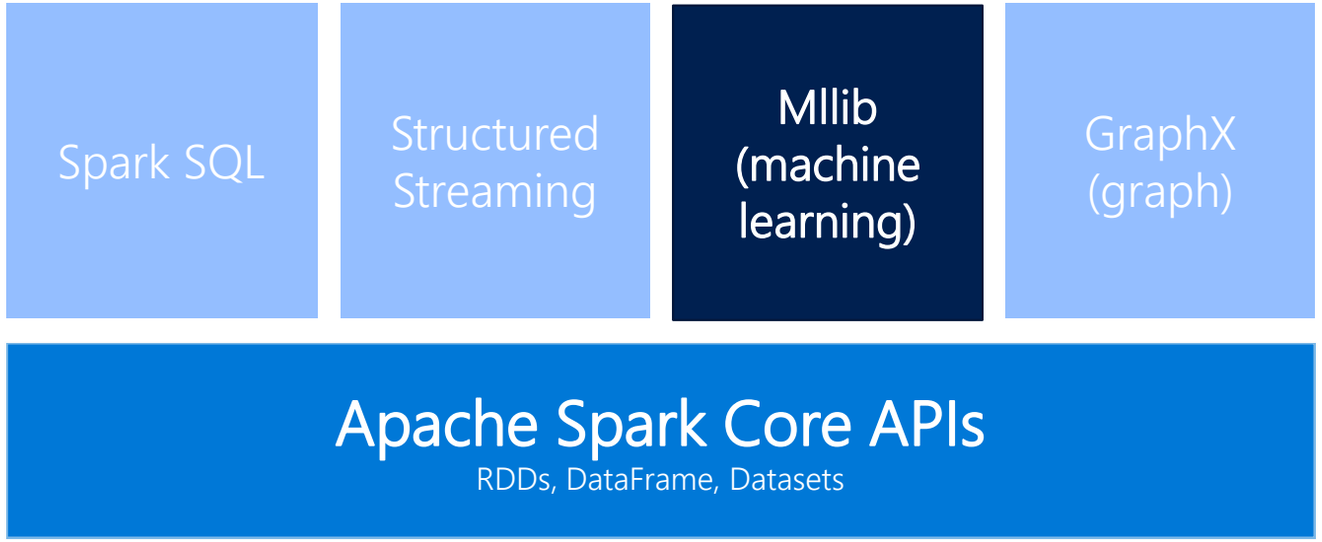
Demo

Ingest ratings data from Event Hubs with Spark Structured Streaming

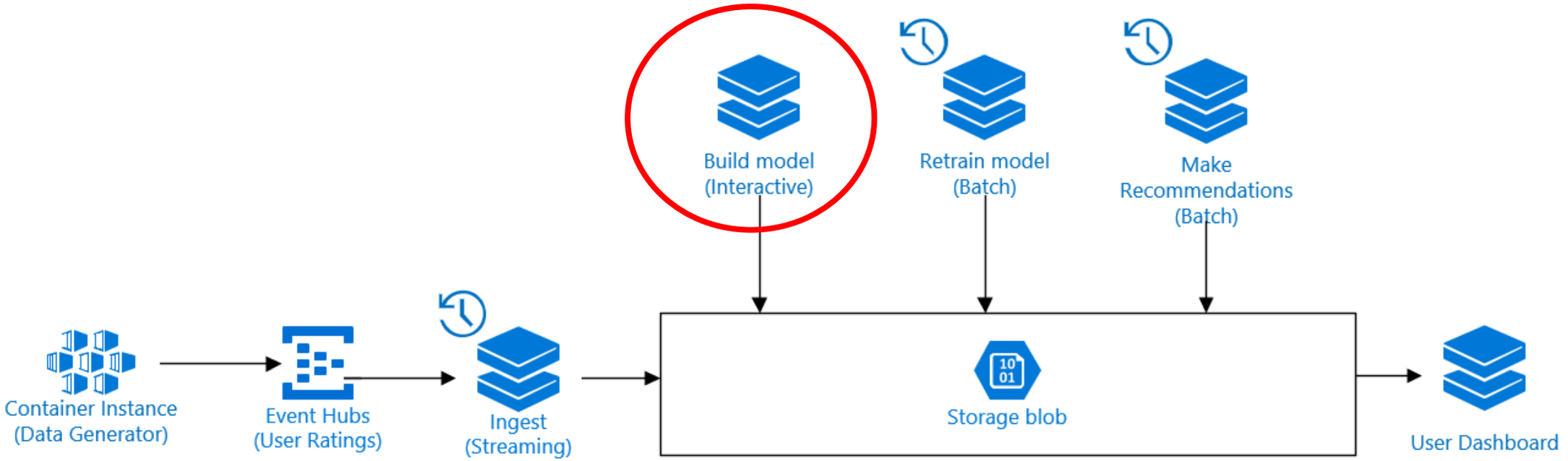
Spark MLlib

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



Demo Architecture



Demo

Build collaborative filtering recommendation model with Spark ML

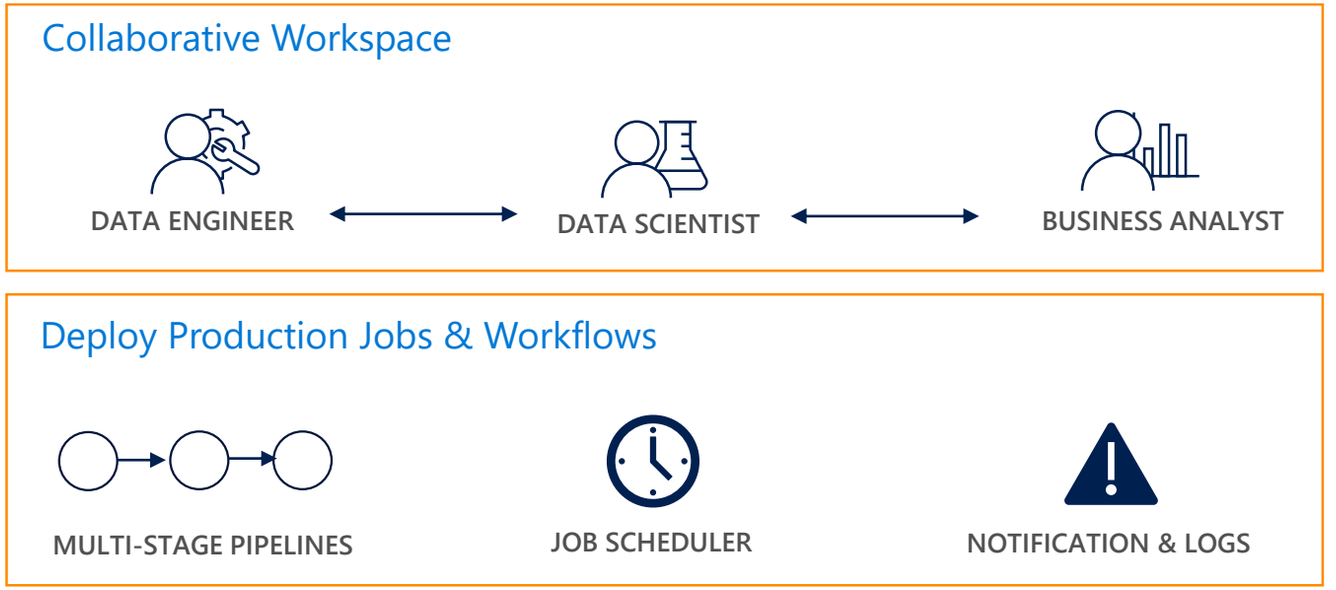
Productionizing Machine Learning Workloads

ML persistence

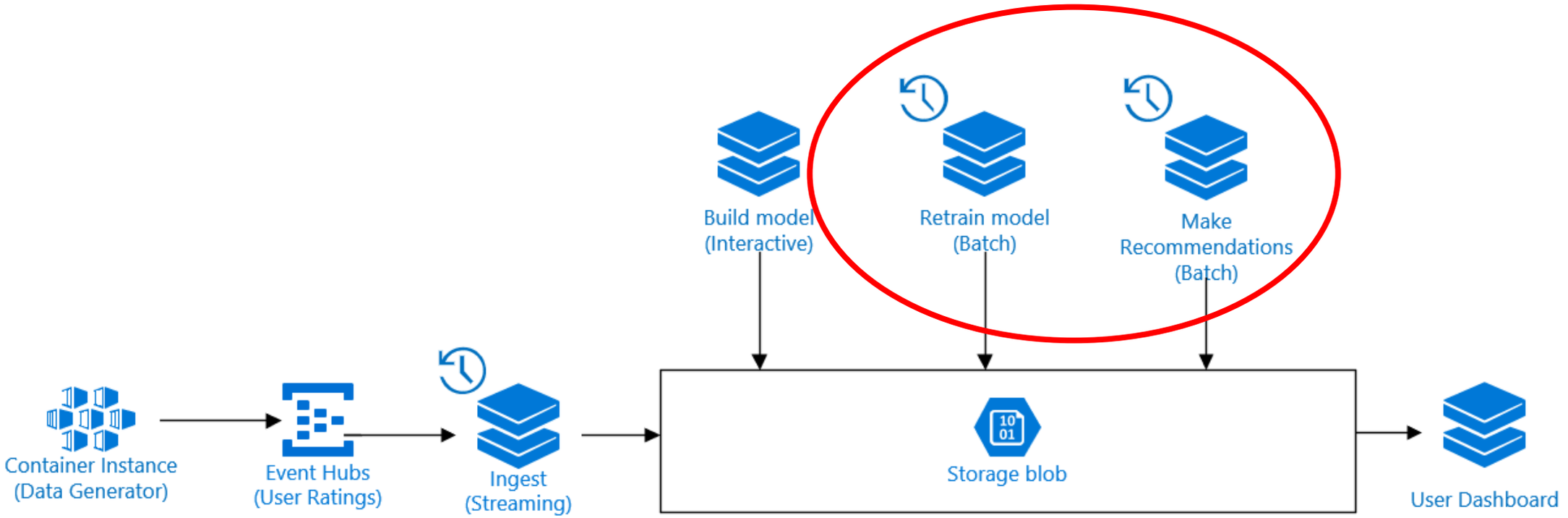
- Sparks support saving multi-stage models built by Data Scientist in Python/R and loading in Scala/Java

Schedule pipelines with Jobs

Notification and alerting



Demo Architecture



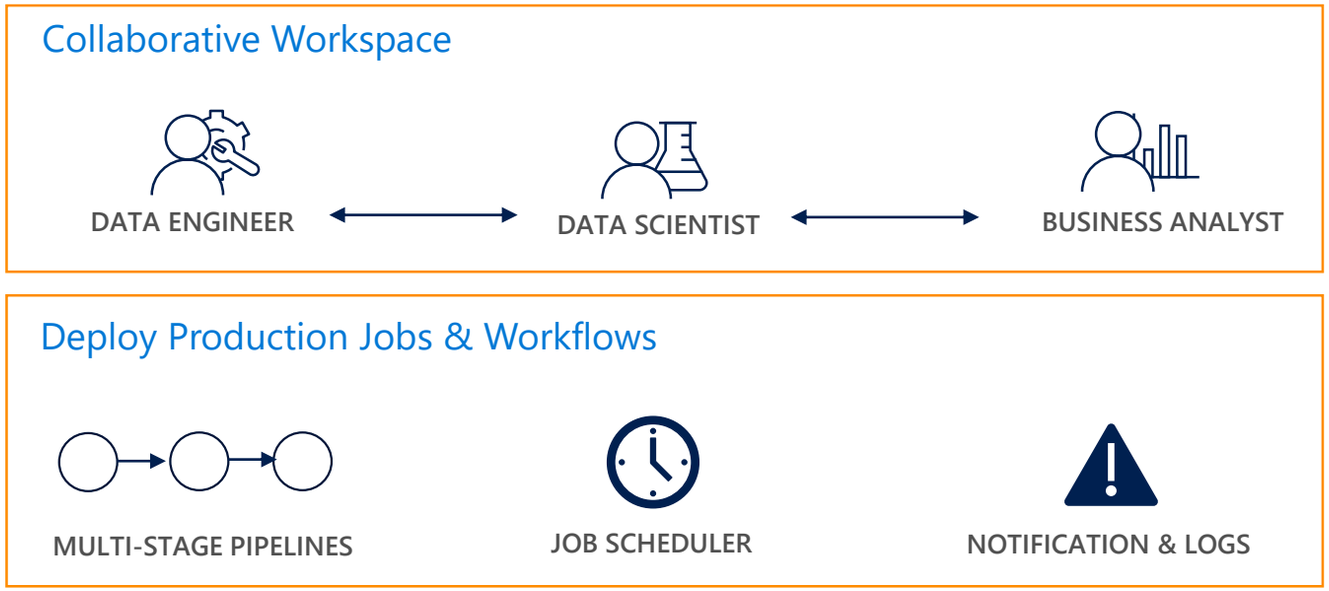
Demo

Productionize workflow with Spark Jobs

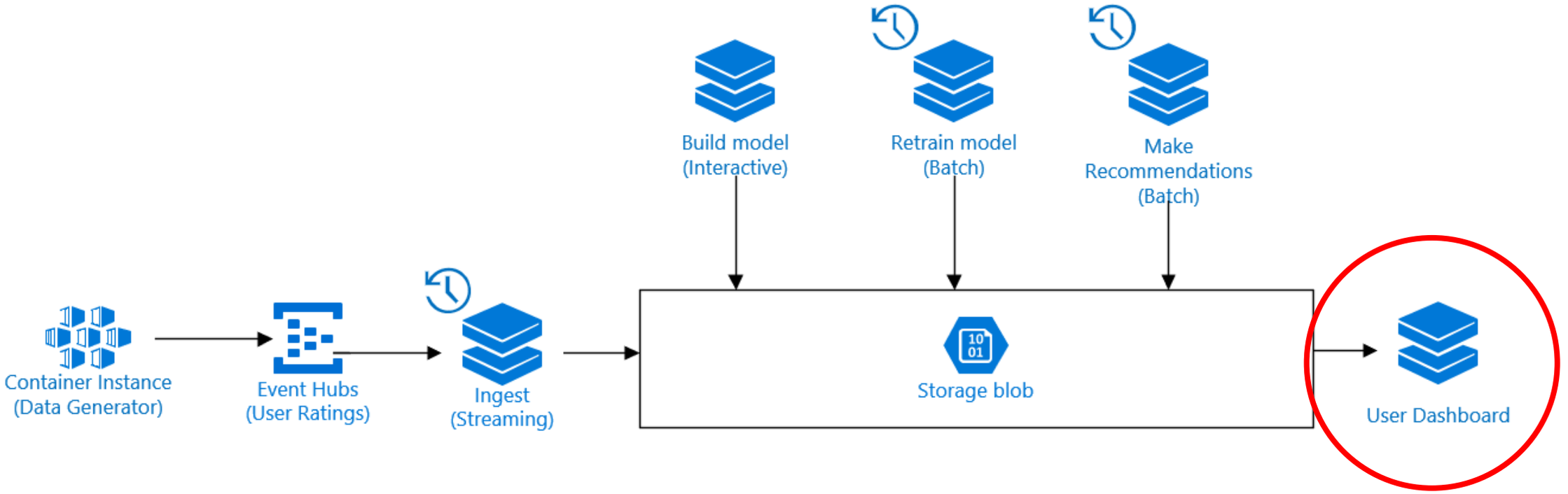
Visualize with Dashboards

Convert Notebooks into Dashboards

Parameterize Notebooks using Widgets



Demo Architecture



Demo

User Recommendation Dashboard

Try the demo!

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

To deploy...

```
make deploy
```

To download requirements...

```
make requirements
```

More resources

Official Apache Spark website

Azure Databricks Documentation

[Book] Spark: The Definitive Guide

Microsoft

Thank you!

Lace Lofranco
Senior Software Development Engineer
Commercial Software Engineering
Microsoft
lace.lofranco@microsoft.com

Different Big Data Solutions