Microsoft

Building Advanced Analytics Pipelines with Azure Databricks

20 Sept 2018

Lace Lofranco
Senior Software Engineer, Microsoft

Survey

Session objective

At the end of the this session, you should:

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

Agenda

Apache Spark Fundamentals

Unified Computing Engine

Azure Databricks

Managed Apache Spark, Integrations with Azure Services

Demo

Anomaly Detection System

Spark Fundamentals

hadoop
Map Reduce

HIVE

STORM

TEZ

mahout

Apache Flink

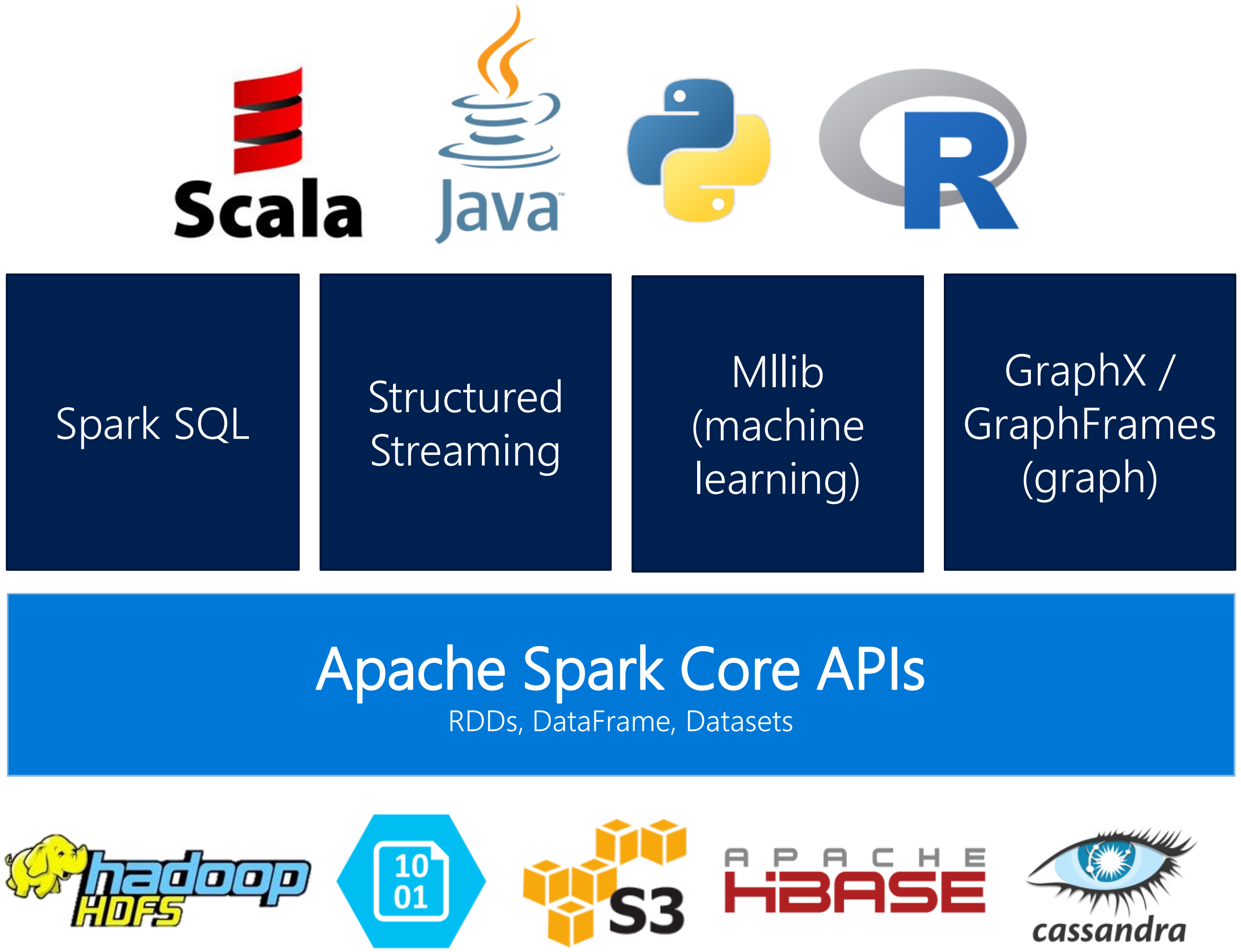
FLUME

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)

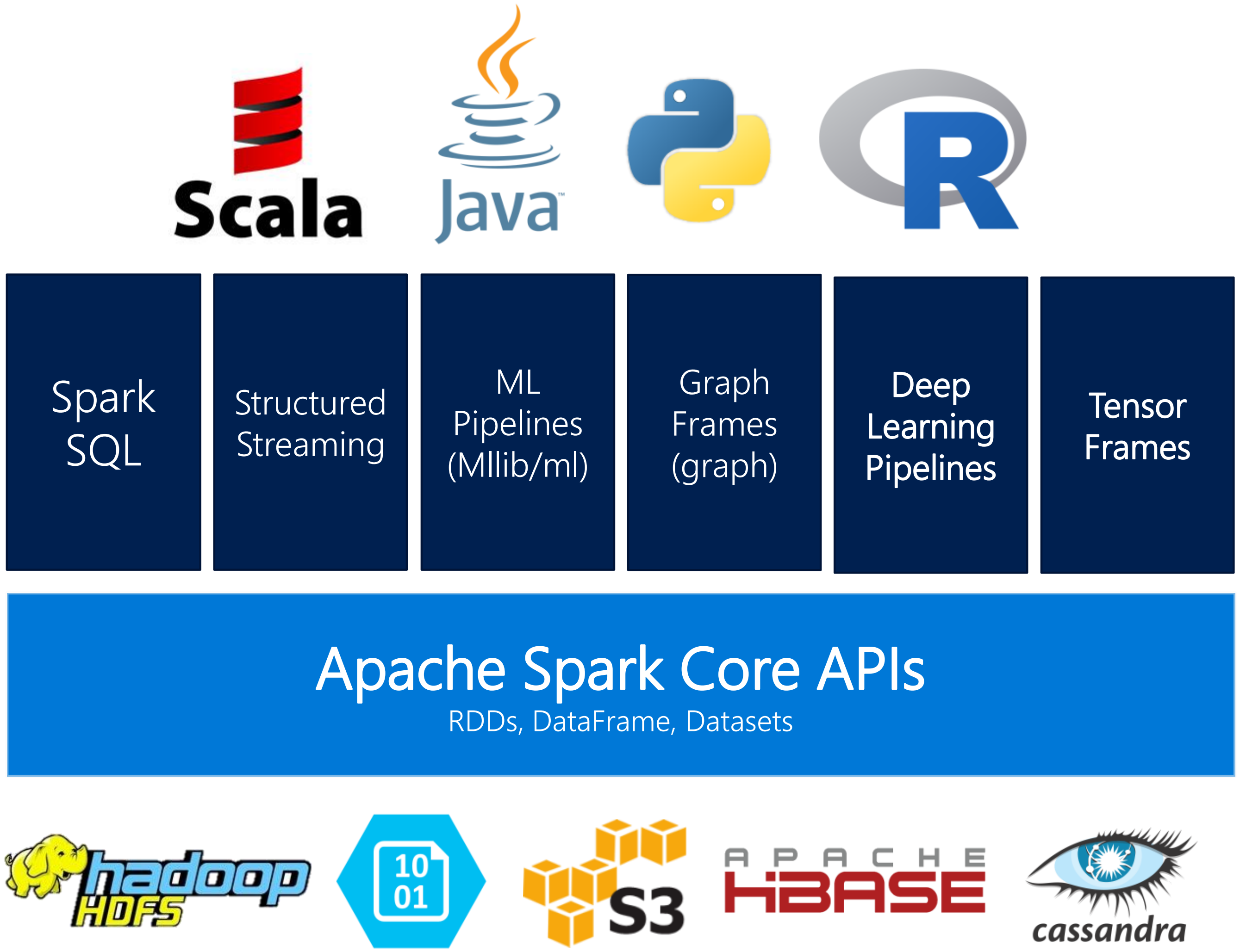
Apache Spark Core APIs

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

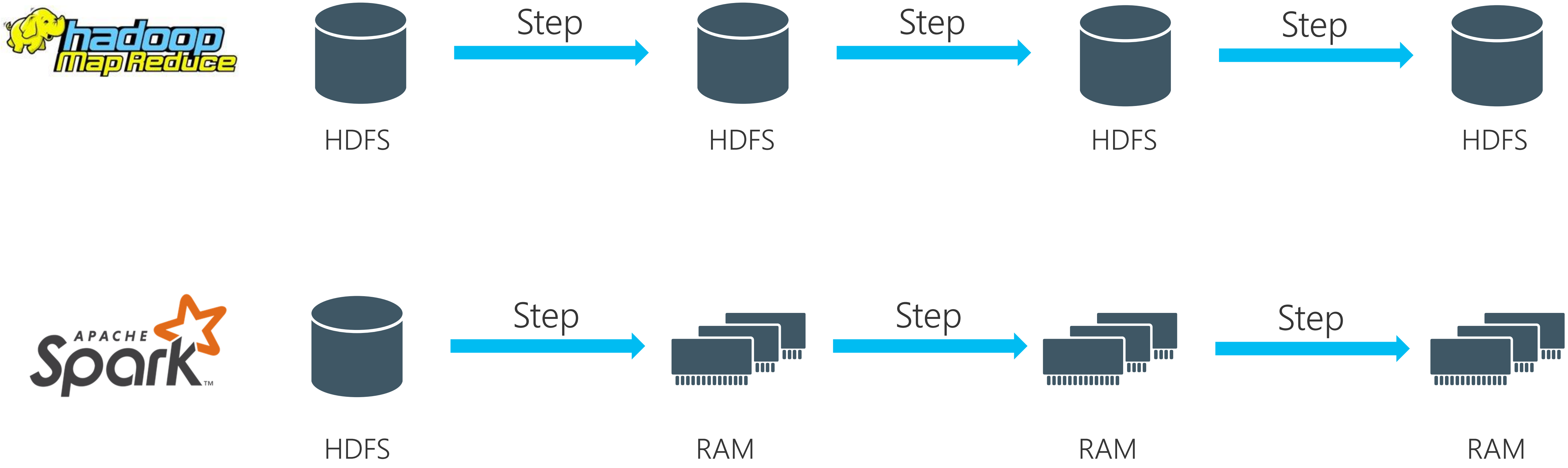
Tensor Frames

Apache Spark Core APIs

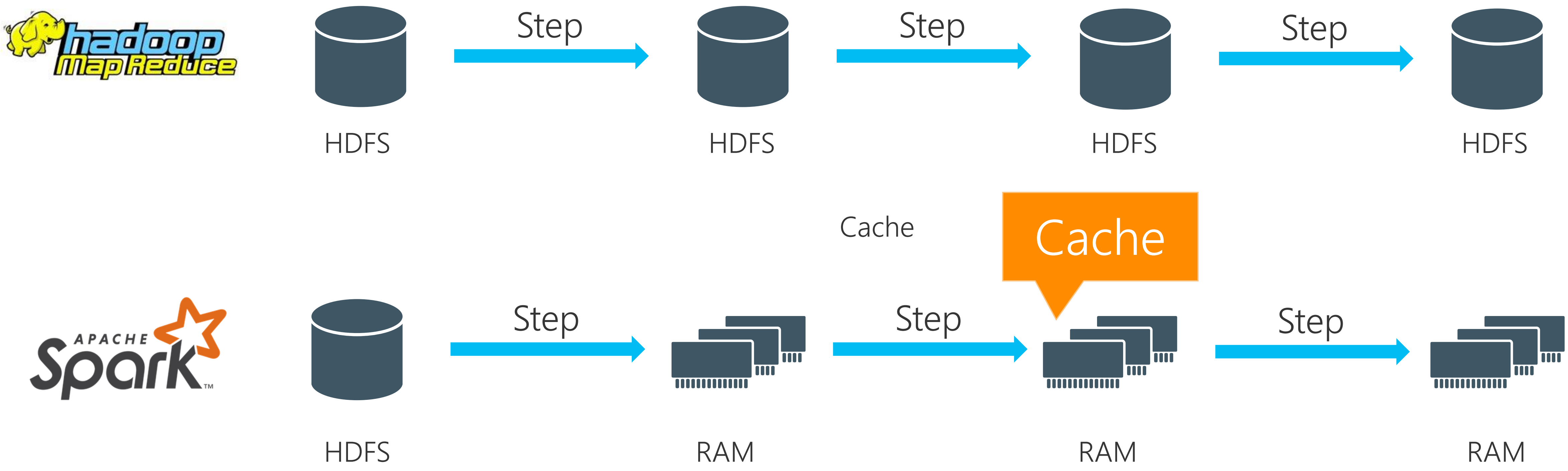
RDDs, DataFrame, Datasets

Spark: The Definitive Guide, Matei Zaharia, Bill Chambers

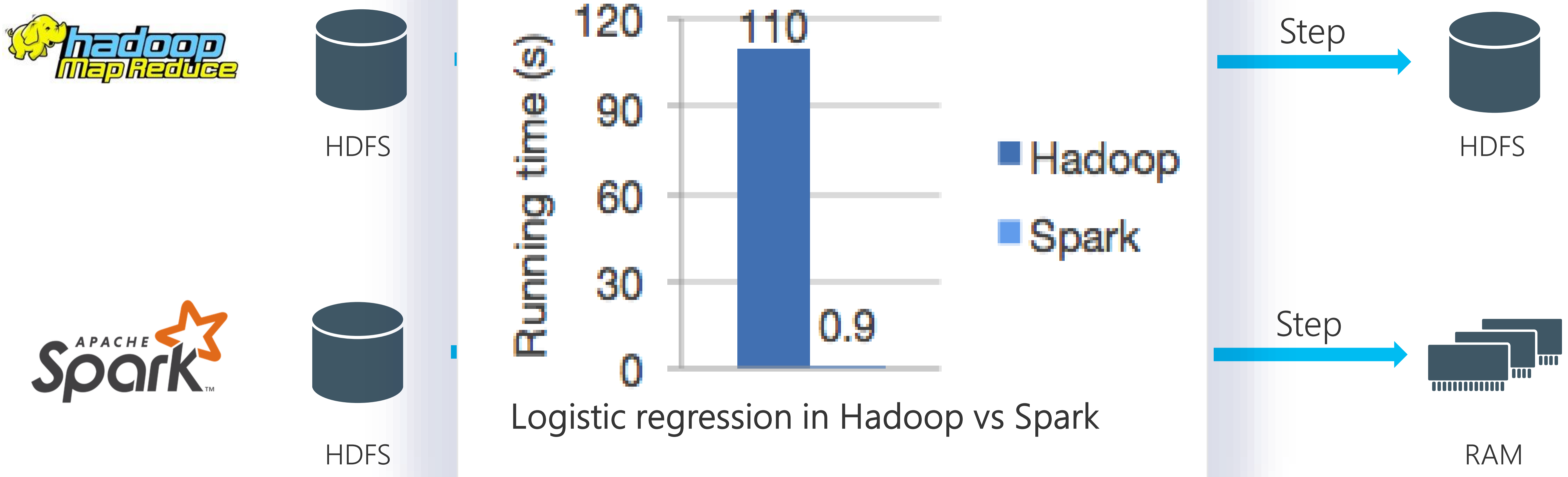
Why Spark is fast



Why Spark is fast



Why Spark is fast



Logistic regression in Hadoop vs Spark

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

Apache Spark: APIs

RDDs

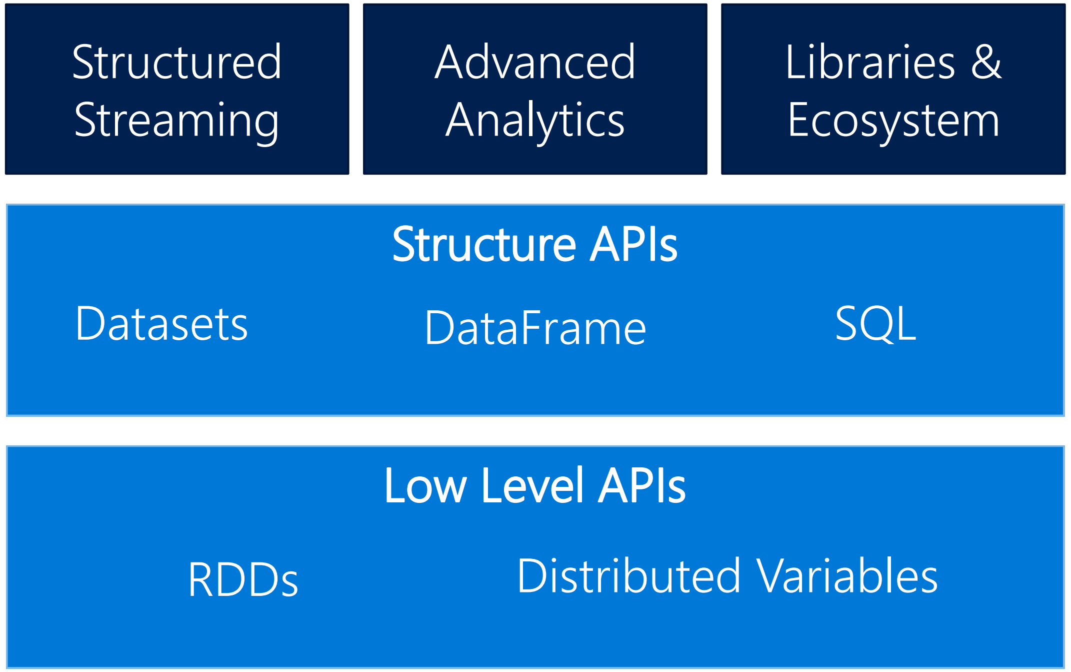
- Core building block of data processing pipelines

DataFrames

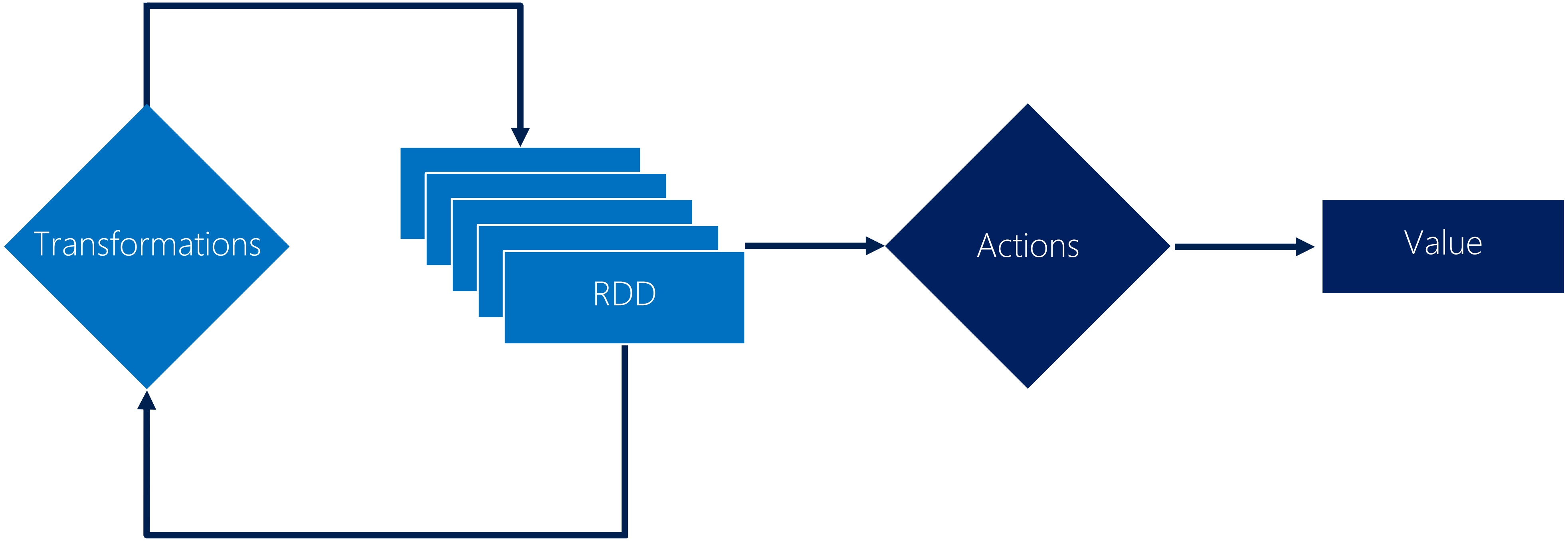
- High level APIs that take advantage of query optimizer

Datasets

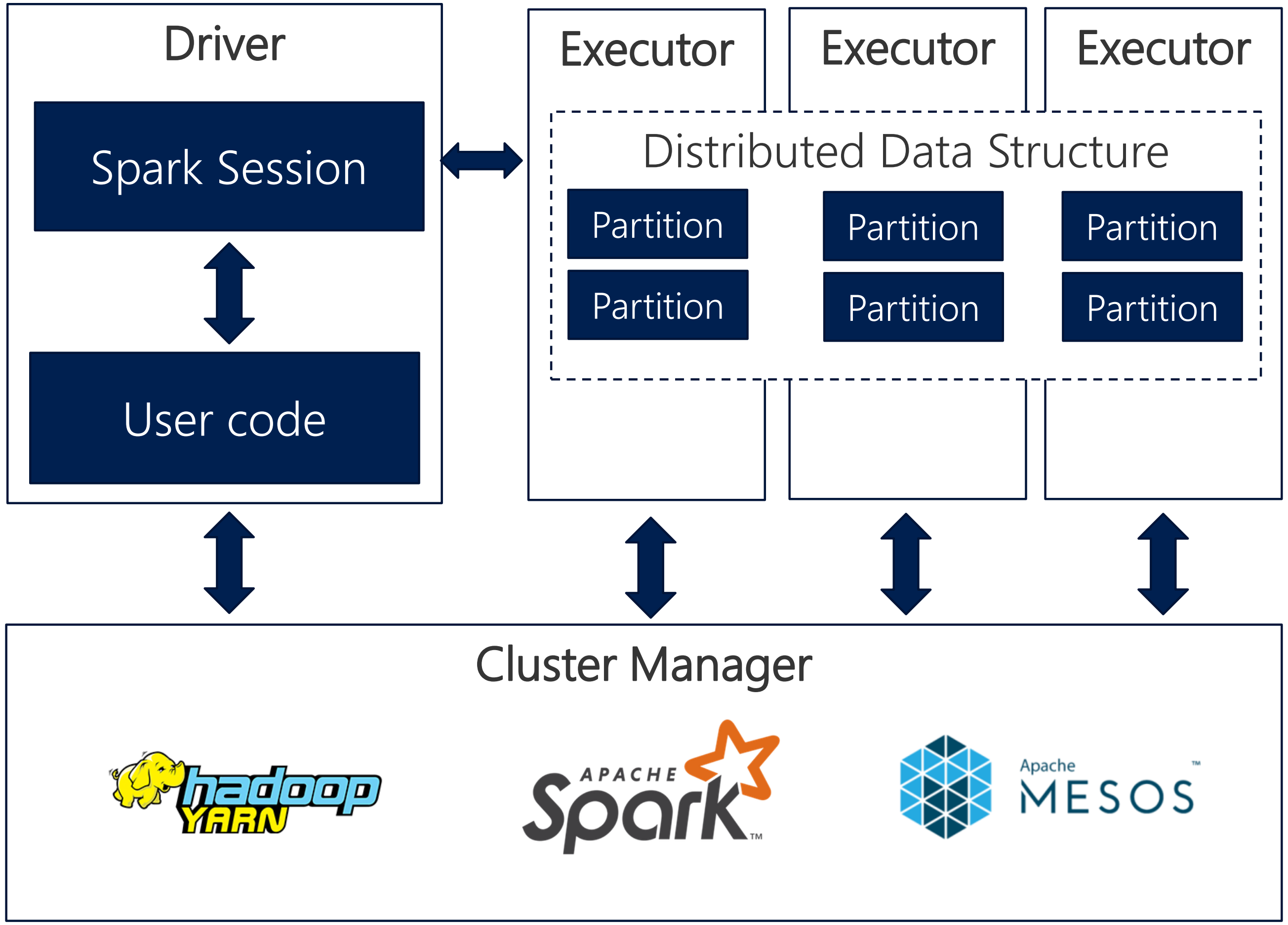
- Data Frames with user objects and custom code



Transformations and Actions



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

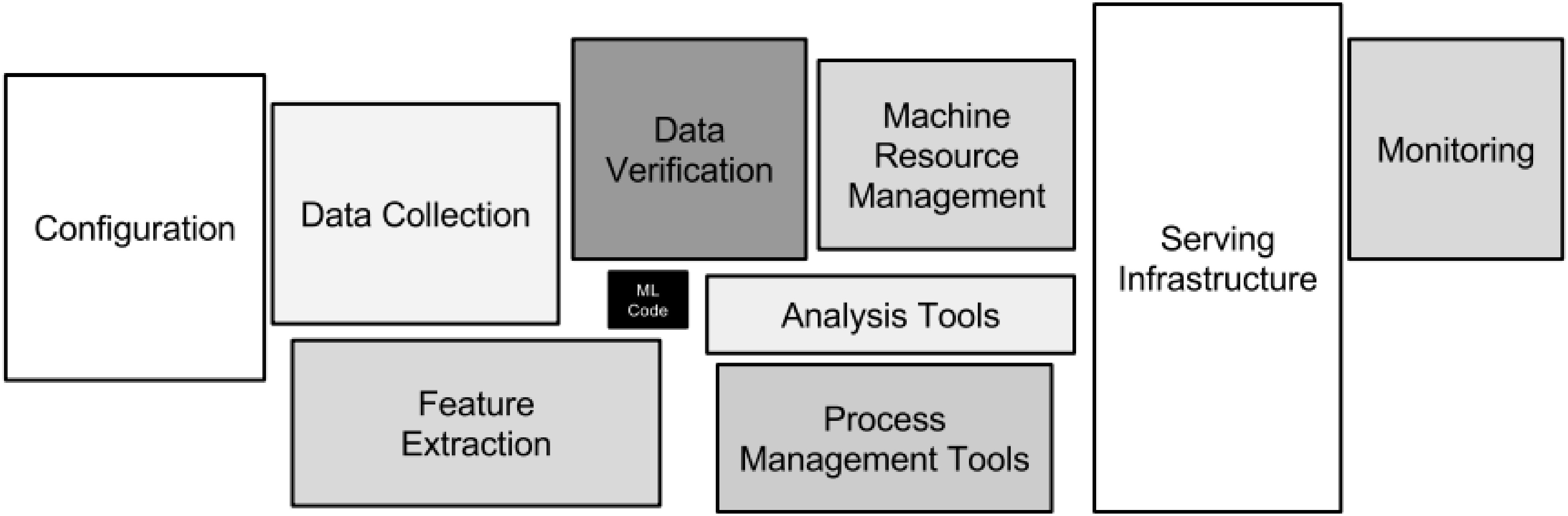
Azure Integration

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

Demo

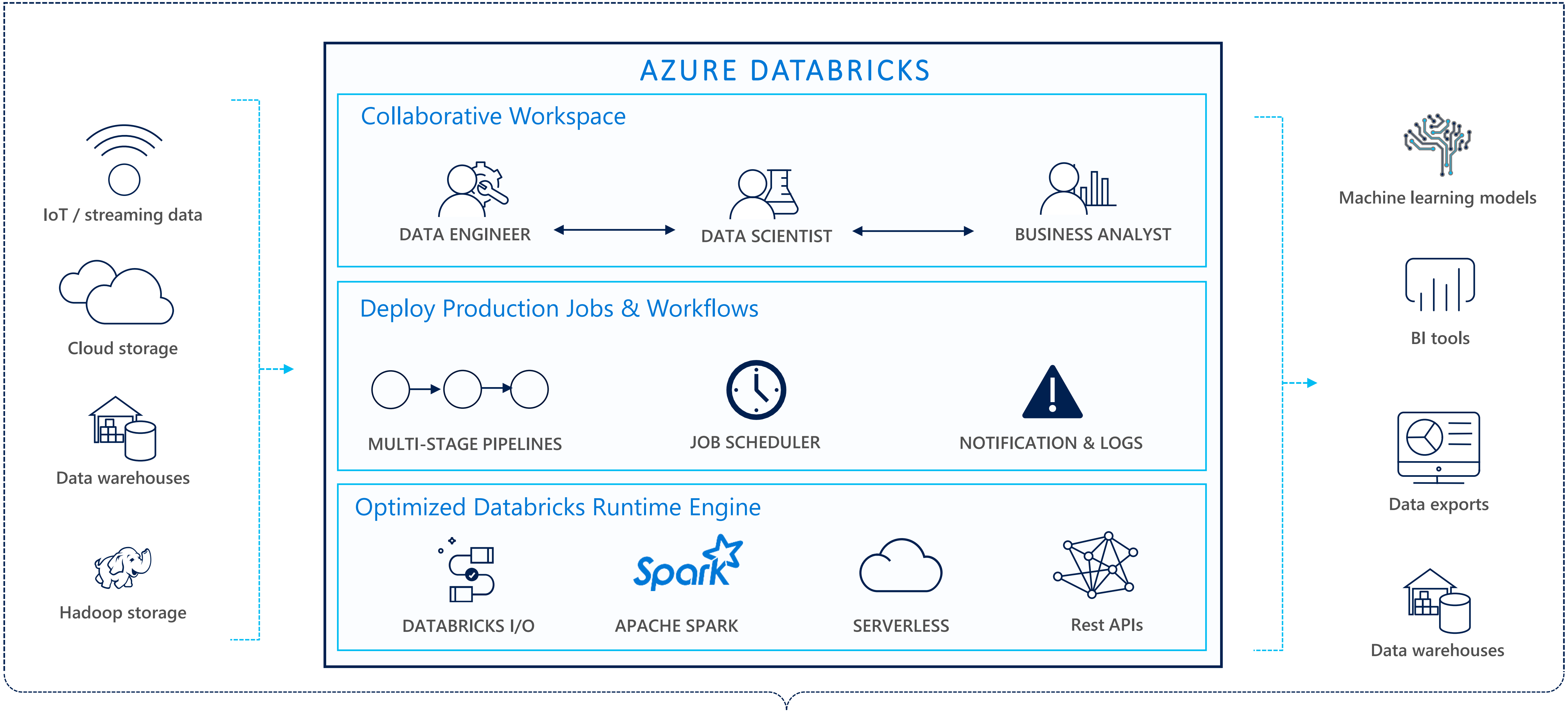
Hello Azure Databricks!

Hidden Technical Debt in ML Systems



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

Azure Databricks

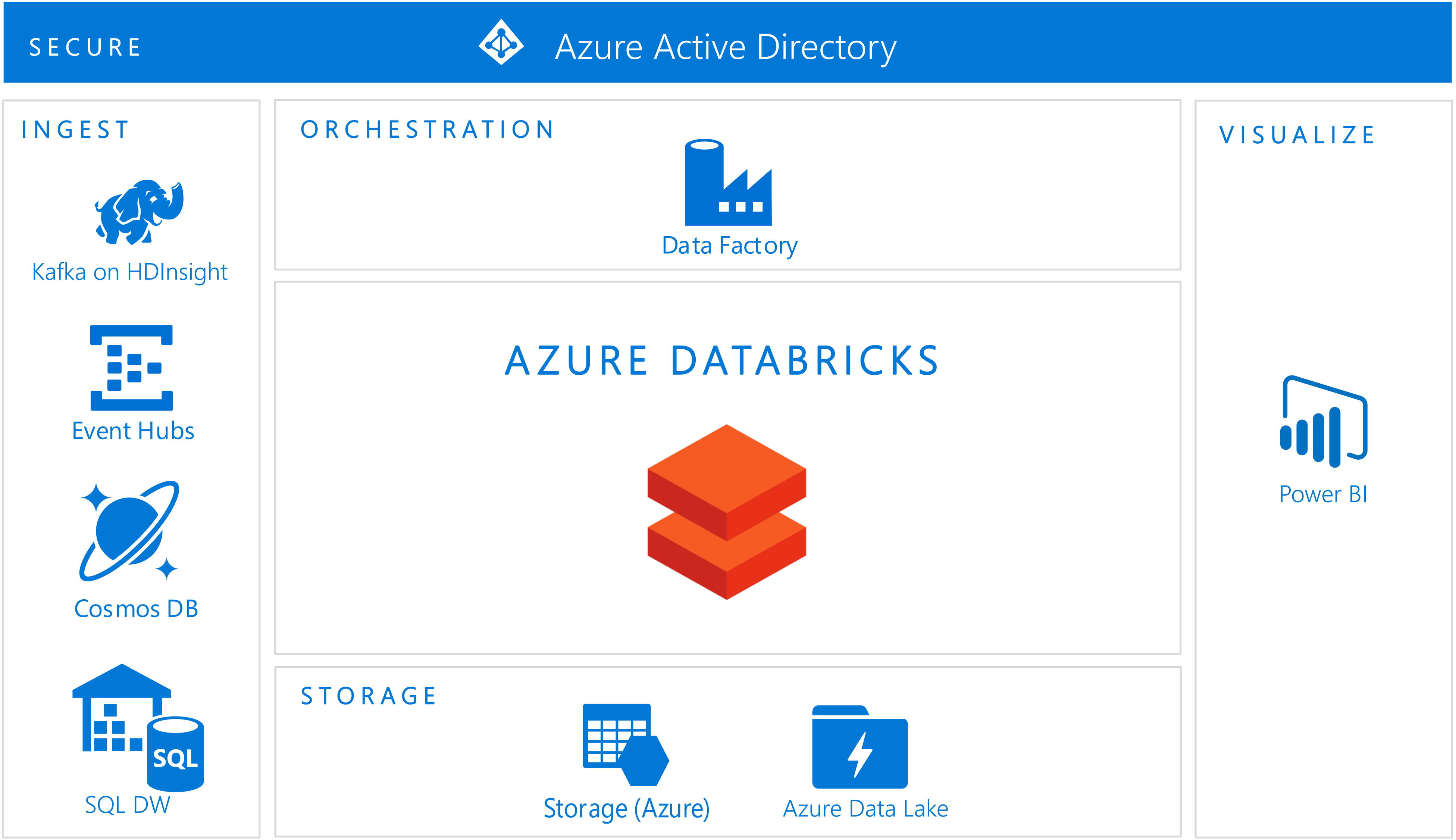


Enhance Productivity

Build on secure & trusted cloud

Scale without limits

Azure Integration



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

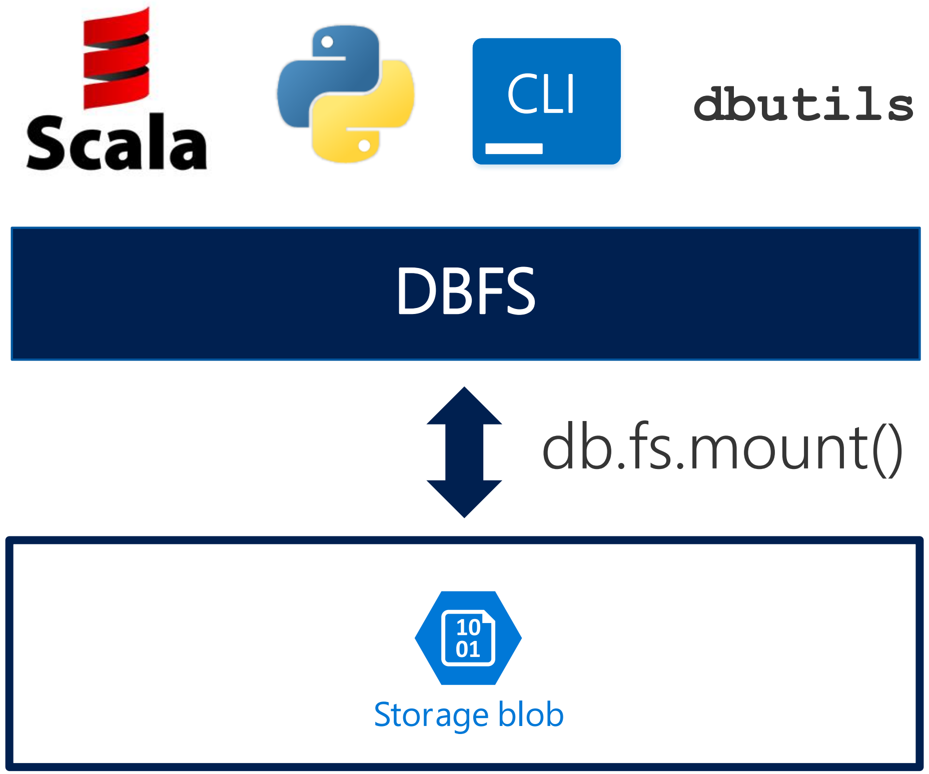
Libraries

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

Anomaly Detection – Network Intrusion

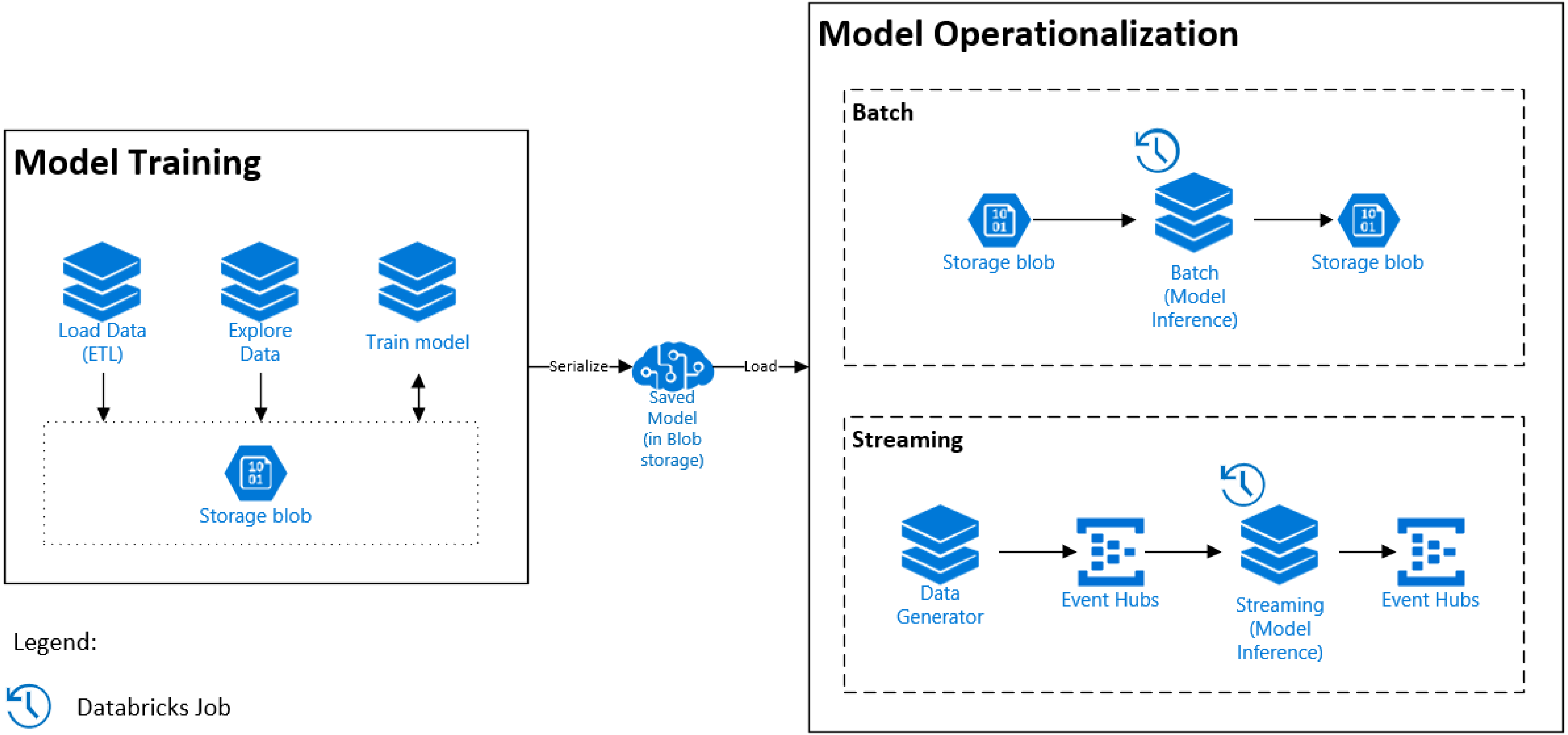
KDD Cup 1999 Data

DARPA Intrusion Detection Evaluation Program

TCP dump data with ‘normal’ connections and ‘attacks’

http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

Demo Architecture



Legend:

◷ Databricks Job

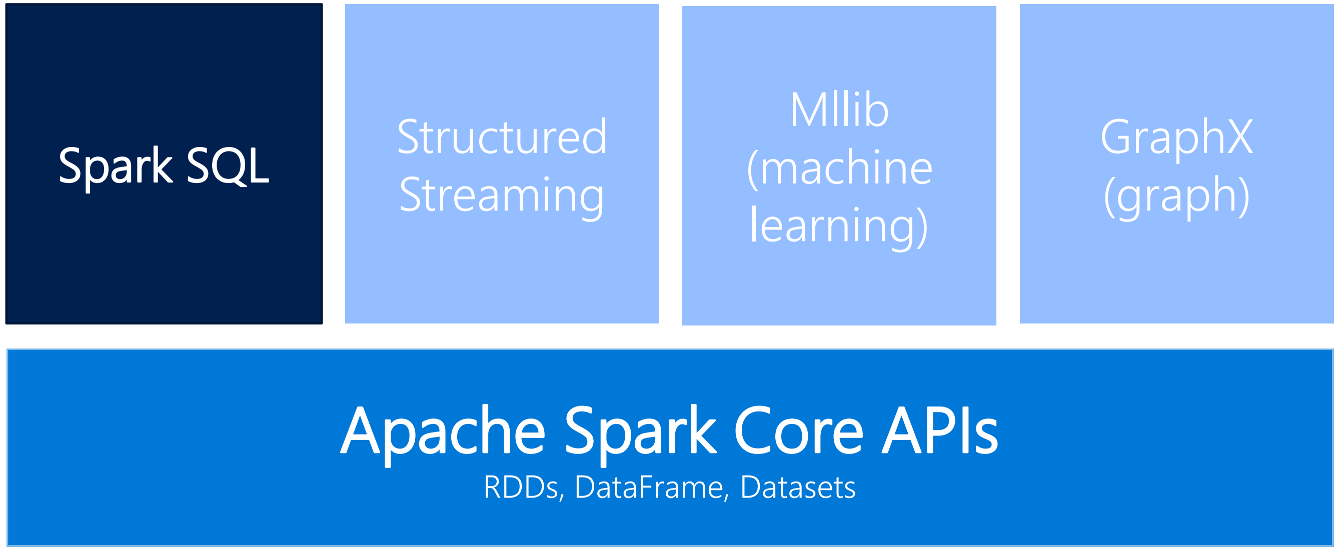
Spark SQL

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

Built on the DataFrame & Datasets API

Hive Integration

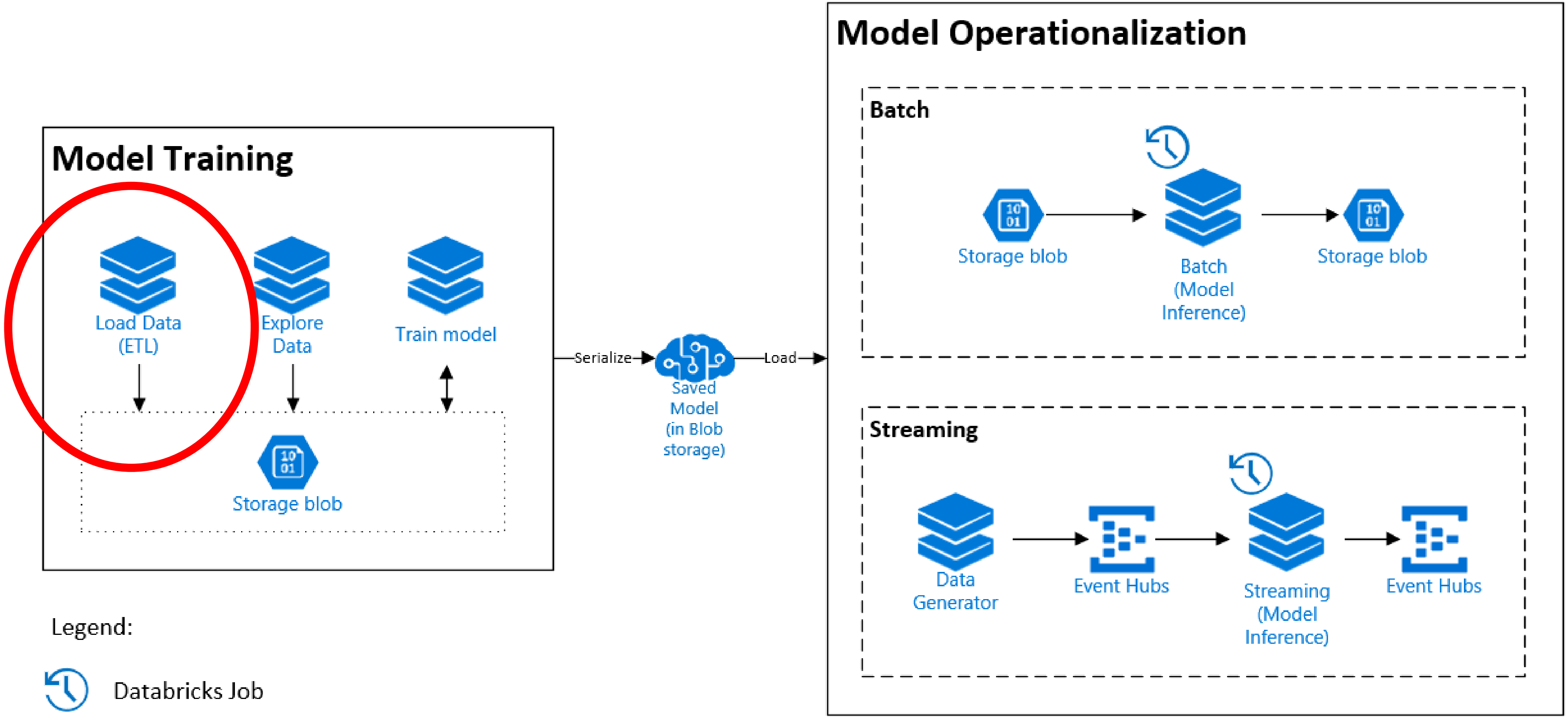
Provides JDBC/ODBC access



Demo

ETL with SparkSQL

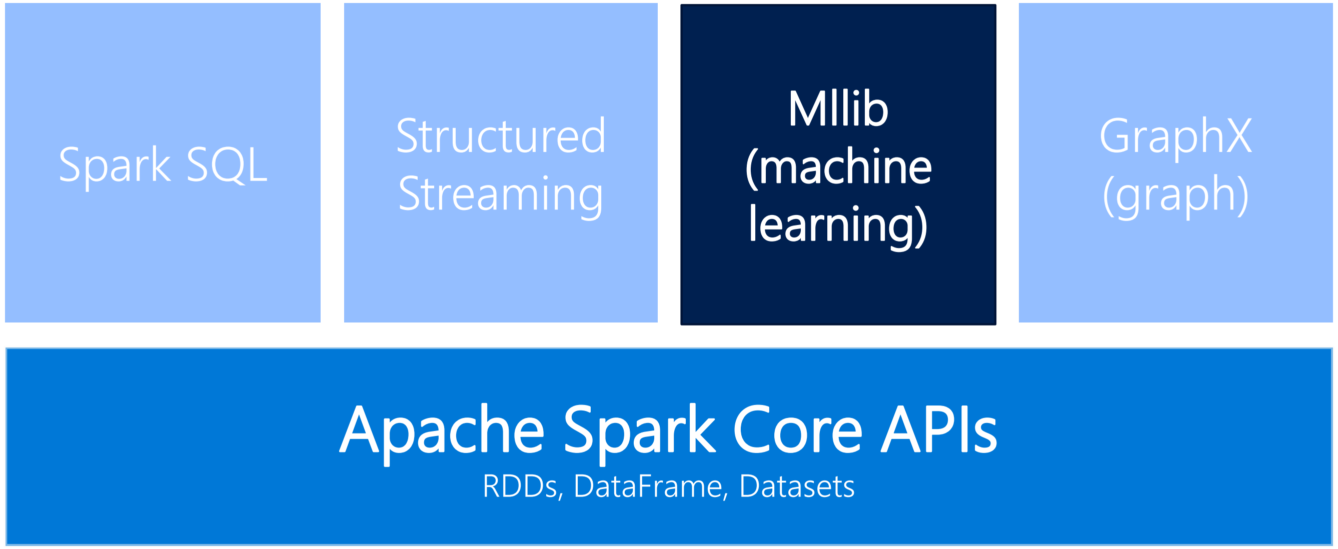
Demo Architecture



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

Models and Features

Feature Extractors

TF-IDF, Word2Vec, CountVectorizer

Feature Transformers

Tokenizer, PCA, StringIndexer, OneHotEncoder, VectorAssember, Normalizer, StandardScaler, SQLTransformer, QuantileDiscretizer, and *more*.

Feature Selectors

VectorSlicer, Rformula, ChiSqSelector

Locality Sensitive Hashing (LSH)

Approx. Similarity Join, Nearest Neighbor Search, Bucketed Random Projection

Classification / Regression

GLMs, Decision tree, Random Forest, Gradient-boosted Tree, Linear SVM, Naïve Bayes

Clustering

K-means, Latent Dirichlet Allocation (LDA), Gaussian Mixture Model

Collaborative Filtering

Alternating Least Square (ALS)

Frequent Pattern Mining

FP-Growth

Model Selection

CrossValidation, Regression/ClassificationEvaluator

Spark MLlib Concepts

DataFrame

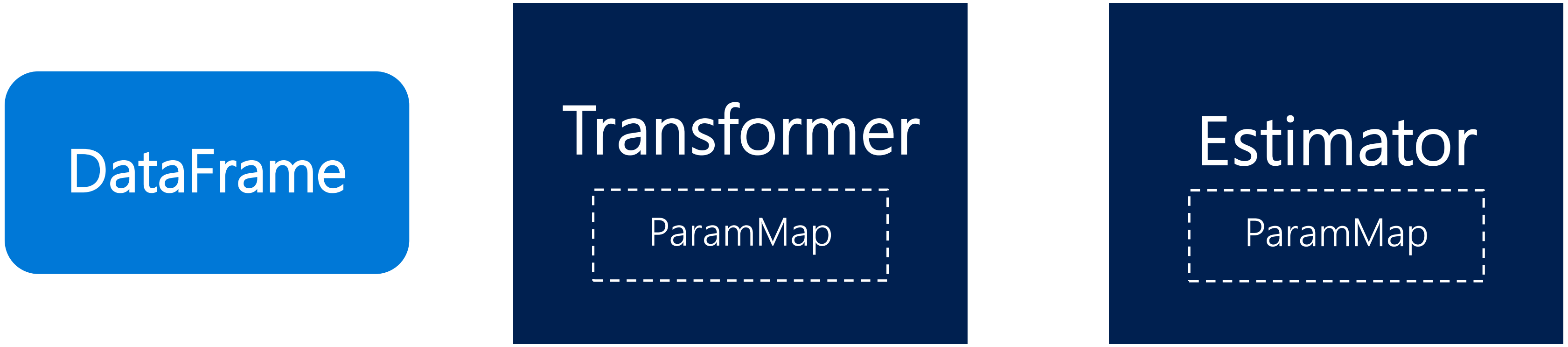
Spark MLlib Concepts

DataFrame

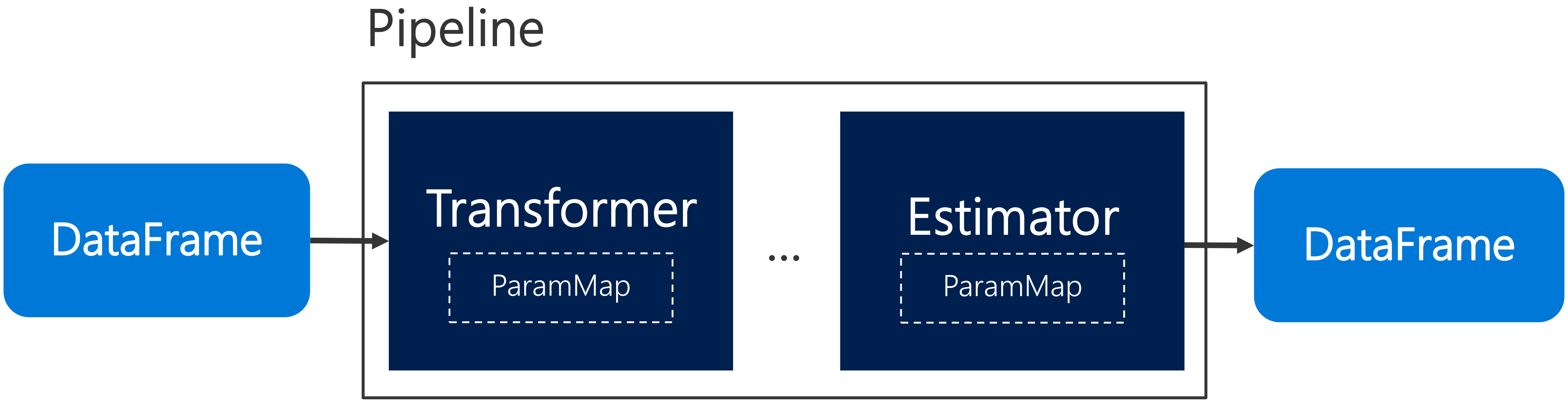
Transformer

Estimator

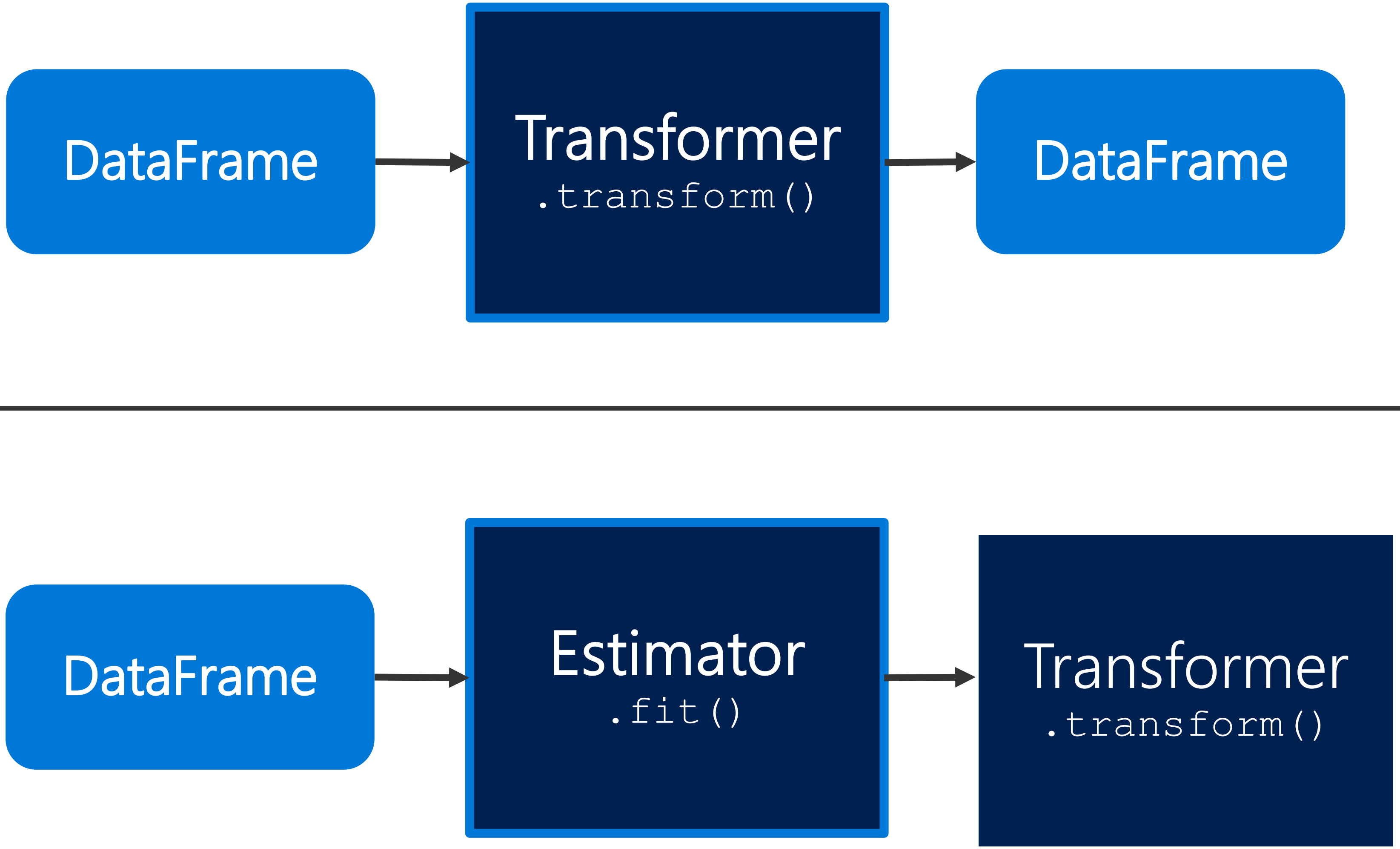
Spark MLlib Concepts



Spark MLlib Concepts



Estimators and Transformers



Custom Transformers and Estimators

Spark MLlib is extensible

Microsoft Machine Learning for Spark (MMLSpark)

Deep learning and data science tools on Spark

https://github.com/Azure/mmlspark

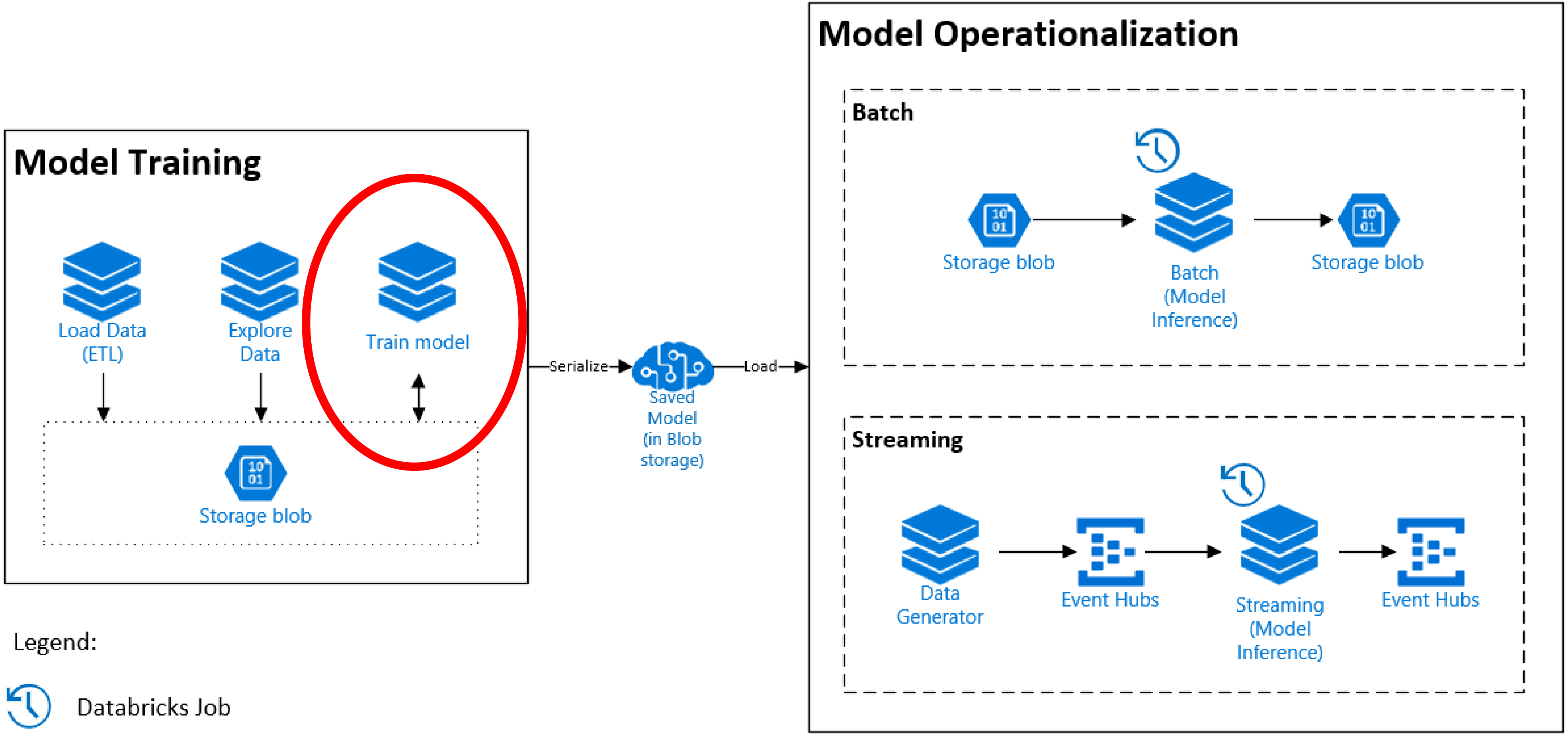
Third-party Spark packages

https://spark-packages.org/

Demo

Train an Anomaly Detection model

Demo Architecture



Legend:

🕘 Databricks Job

Productionizing Machine Learning Workloads

In Spark…

1. Batch inference
2. Structured Streaming

Out of Spark…

Export model

- Mleap, MLFlow Models

Containerized Web Service

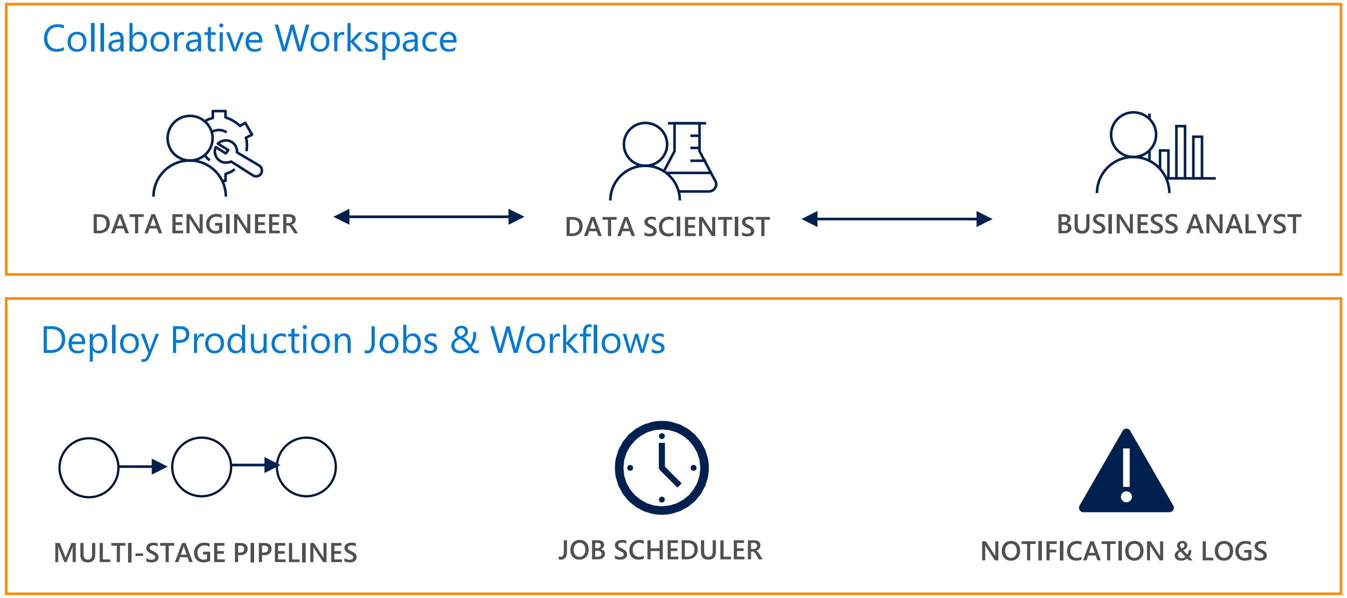
Productionizing Machine Learning Workloads

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

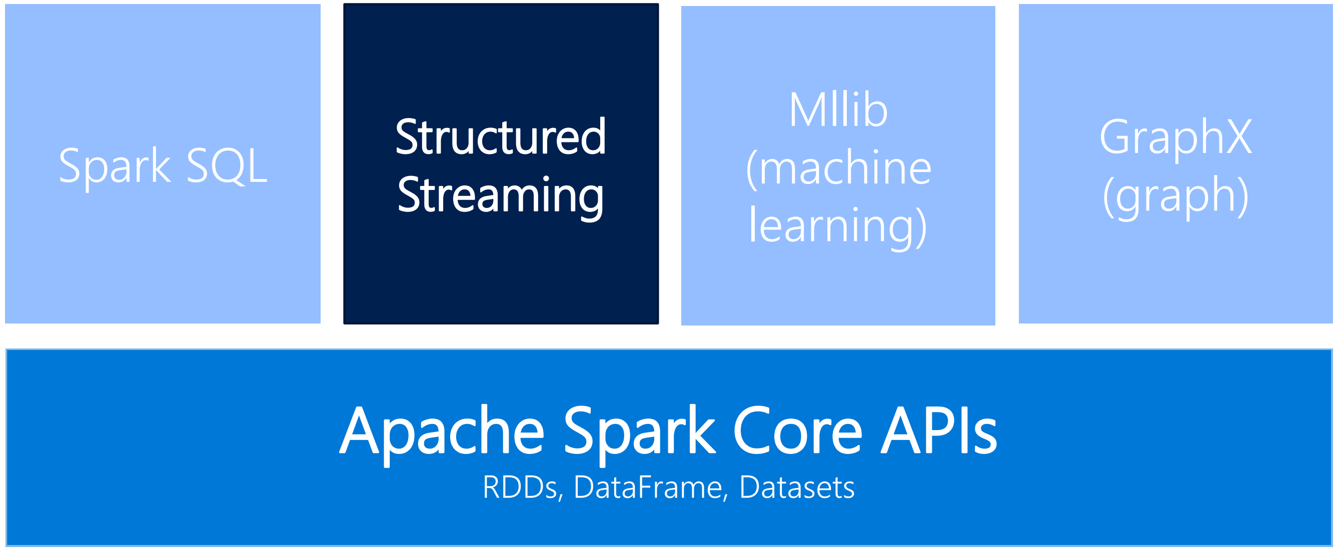


Spark Structured Streaming

Scalable and fault-tolerant stream processing engine

Successor of Spark Streaming (DStreams API)

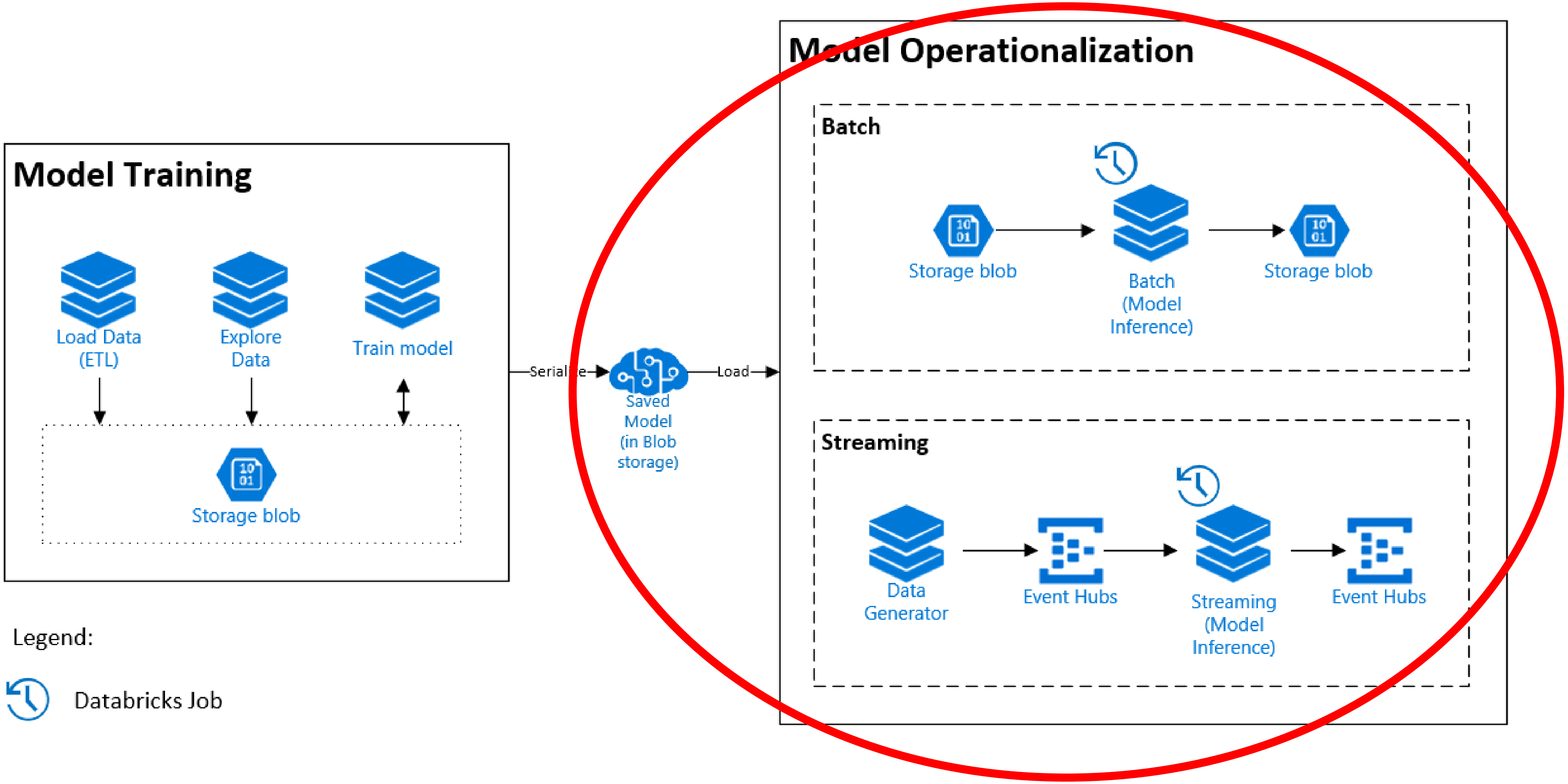
Same code for Batch and Streaming



Demo

Productionize workflow with Spark Jobs

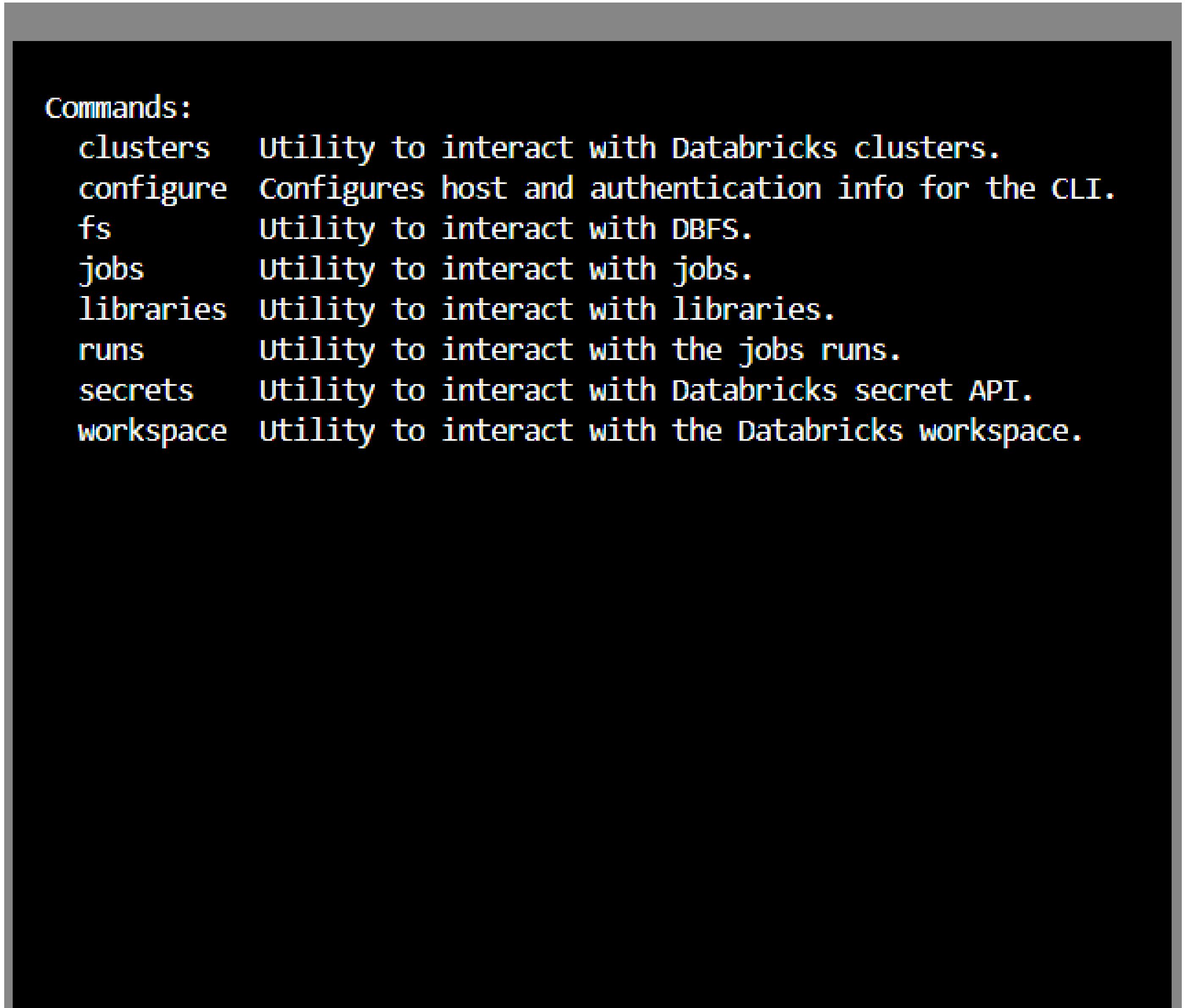
Demo Architecture



Databricks Developer Tooling

Databricks CLI

Databricks REST API

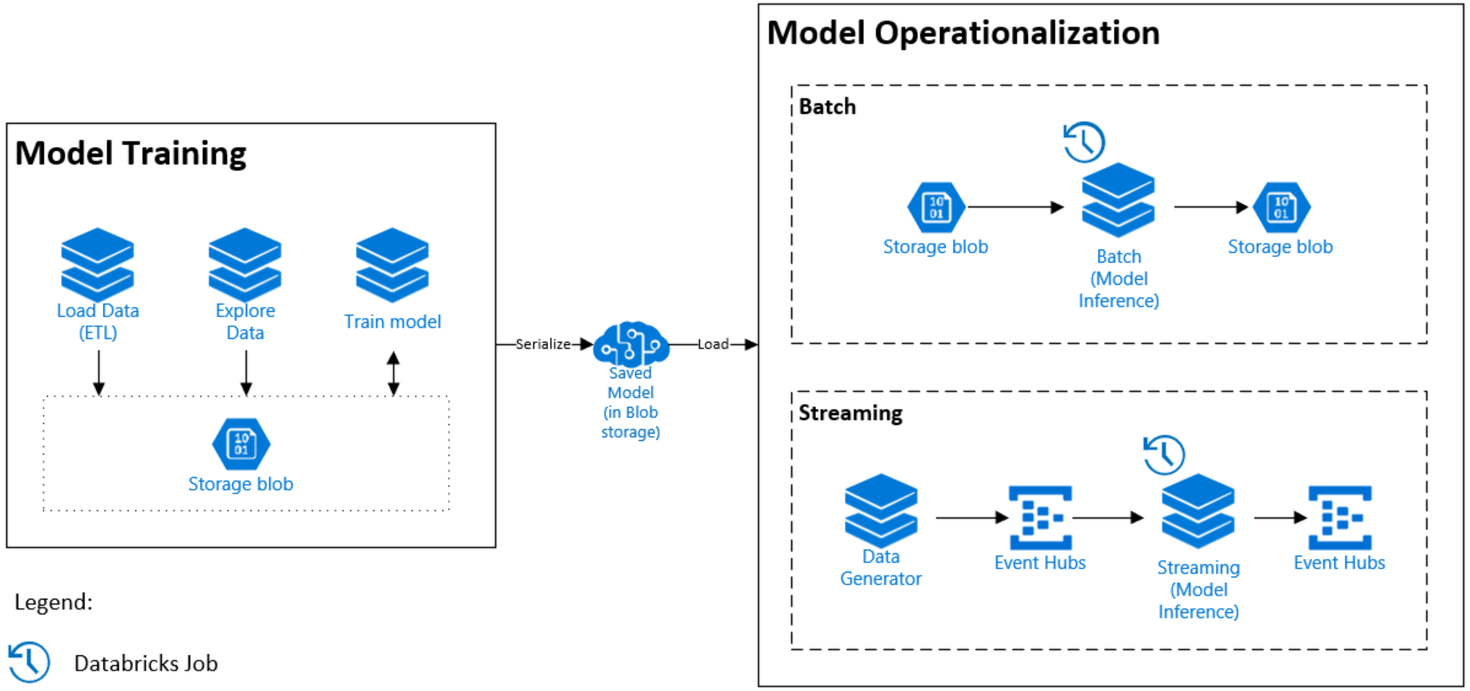


```
Commands:
  clusters   Utility to interact with Databricks clusters.
  configure  Configures host and authentication info for the CLI.
  fs         Utility to interact with DBFS.
  jobs       Utility to interact with jobs.
  libraries  Utility to interact with libraries.
  runs       Utility to interact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

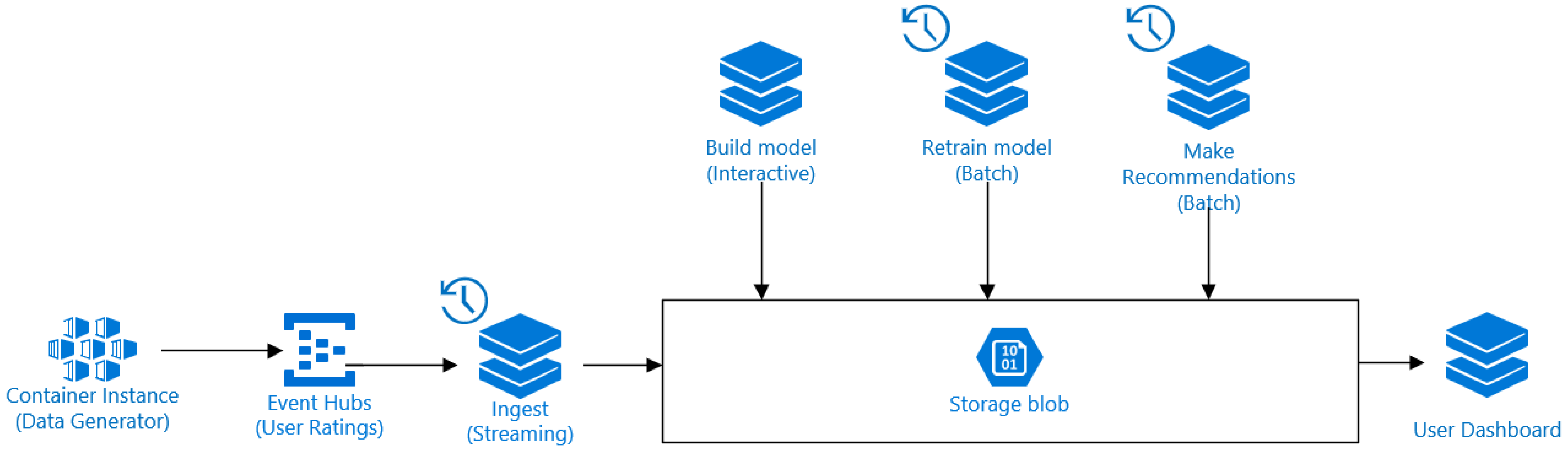
To deploy... `docker –it devlace/azdatabricksanomaly`



Other Databricks Demos...

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

To deploy... **`docker -it devlace/azdatabricksrecommend`**



More resources

Official Apache Spark website

Azure Databricks Documentation

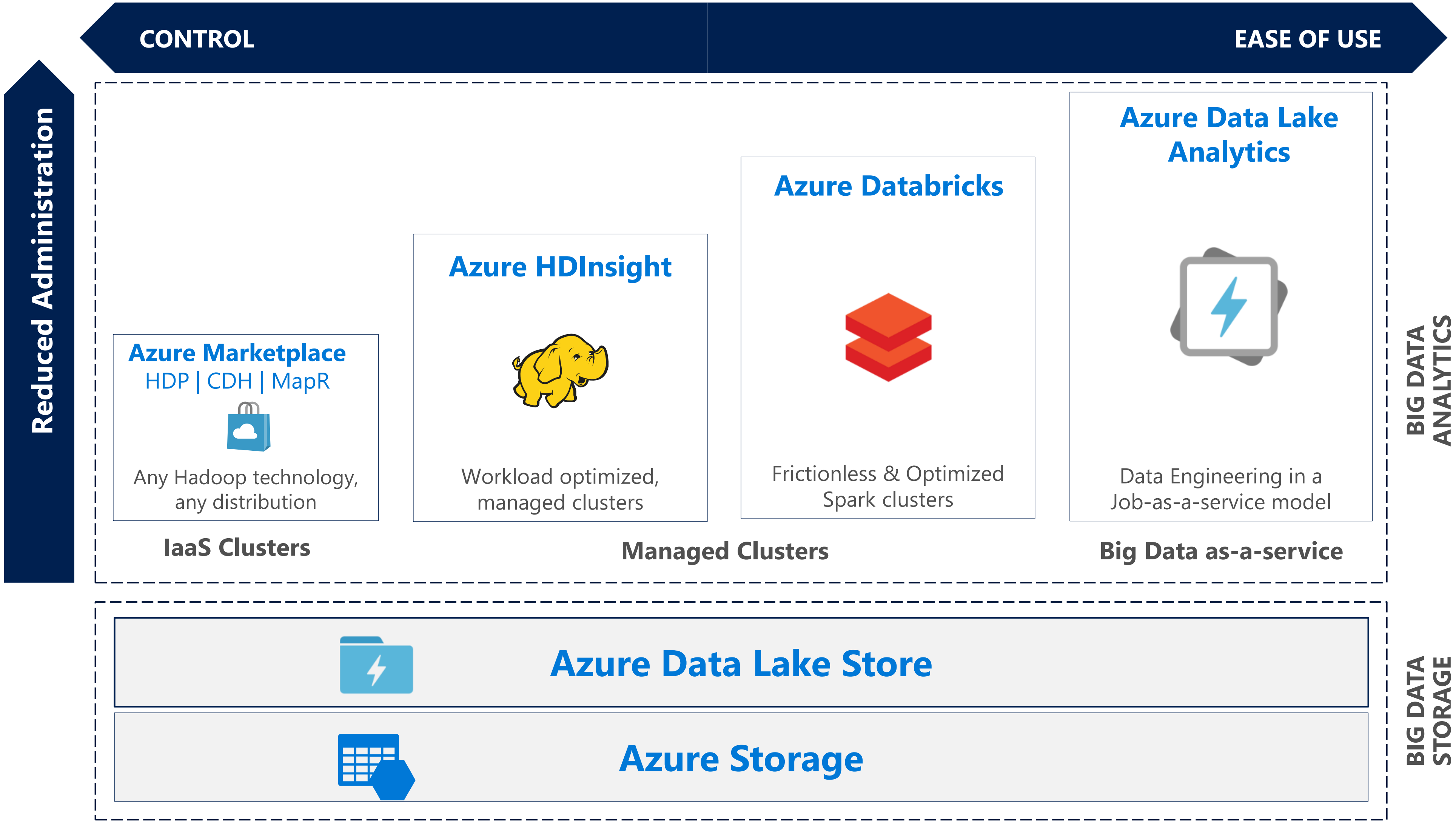
[Book] Spark: The Definitive Guide

Microsoft

Thank you!

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

Different Big Data Solutions



CONTROL

EASE OF USE

Reduced Administration

Azure Marketplace
HDP | CDH | MapR

Any Hadoop technology, any distribution

IaaS Clusters

Azure HDInsight

Workload optimized, managed clusters

Azure Databricks

Frictionless & Optimized Spark clusters

Managed Clusters

Azure Data Lake Analytics

Data Engineering in a Job-as-a-service model

Big Data as-a-service

BIG DATA ANALYTICS

Azure Data Lake Store

Azure Storage

BIG DATA STORAGE