AZURE **DATABRICKS**

- INTRODUCTION

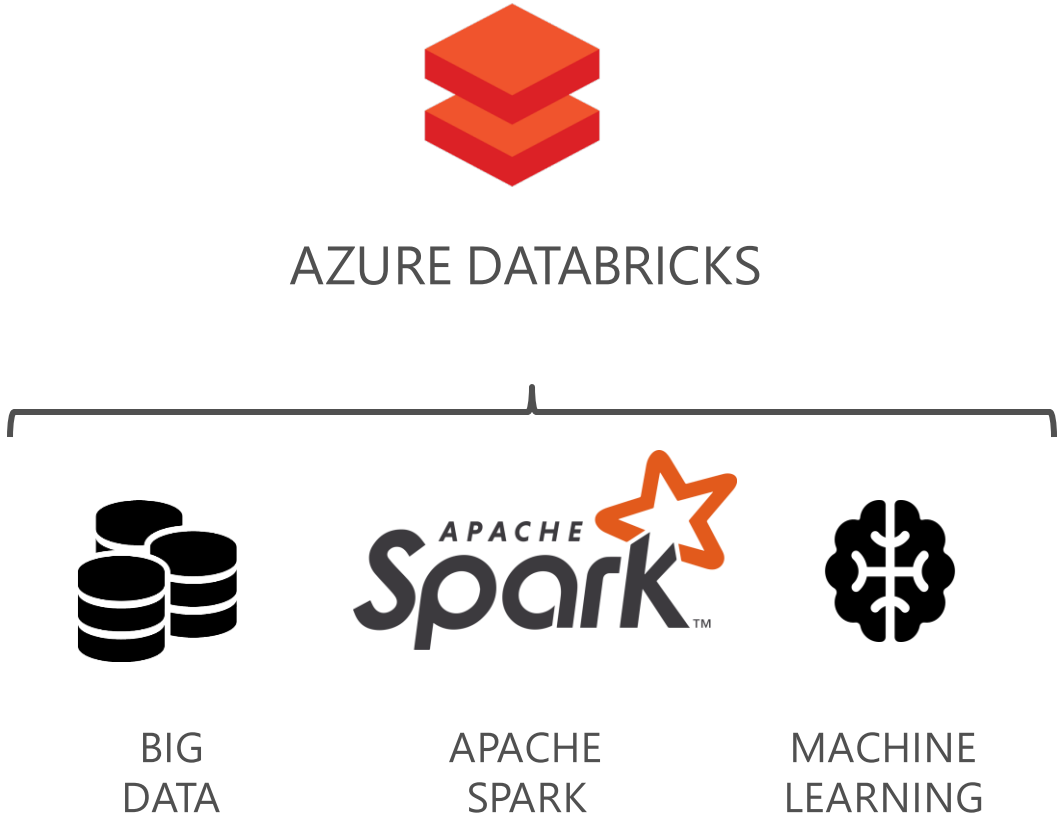
ADAM CZYŻEWSKI

DOMINIK HUSS

AZURE DATABRICKS

INTRODUCTION

- Founded in late 2013
- By the creators of Apache Spark, original team from UC Berkeley AMPLab
- Largest code contributor code to Apache Spark
- Provides certifications such as Databricks Certified Application, Databricks Certified Distribution and Databricks Certified Developer
- Main Product: The Unified Analytics Platform
- In Oct 2017, introduced Databricks Delta (currently in private preview).



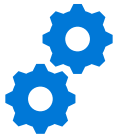
AZURE DATABRICKS

IDEA

AZURE
DATABRICKS

COLLABORATION

SECURITY



EASY CLUSTER
CREATION

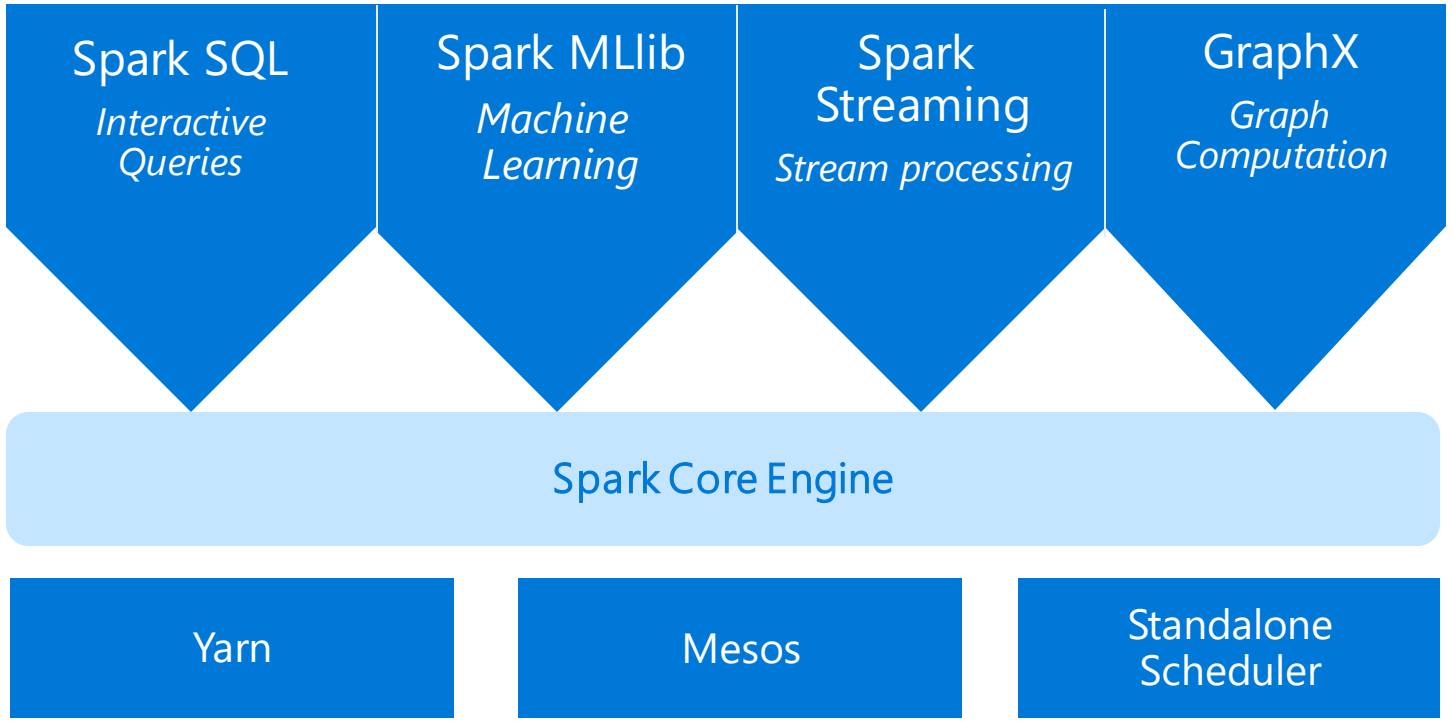
NOTEBOOKS

APACHE SPARK

AN UNIFIED, OPEN SOURCE, PARALLEL, DATA PROCESSING FRAMEWORK FOR BIG DATA ANALYTICS

Spark Unifies:

- Batch Processing
- Interactive SQL
- Real-time processing
- Machine Learning
- Deep Learning
- Graph Processing



APACHE SPARK

BENEFITS

Performance

Using in-memory computing, Spark is considerably faster than Hadoop (100x in some tests).

Can be used for batch and real-time data processing.

Developer Productivity

Easy-to-use APIs for processing large datasets.

Includes 100+ operators for transforming.

Unified Engine

Integrated framework includes higher-level libraries for interactive SQL queries, Stream Analytics, ML and graph processing.

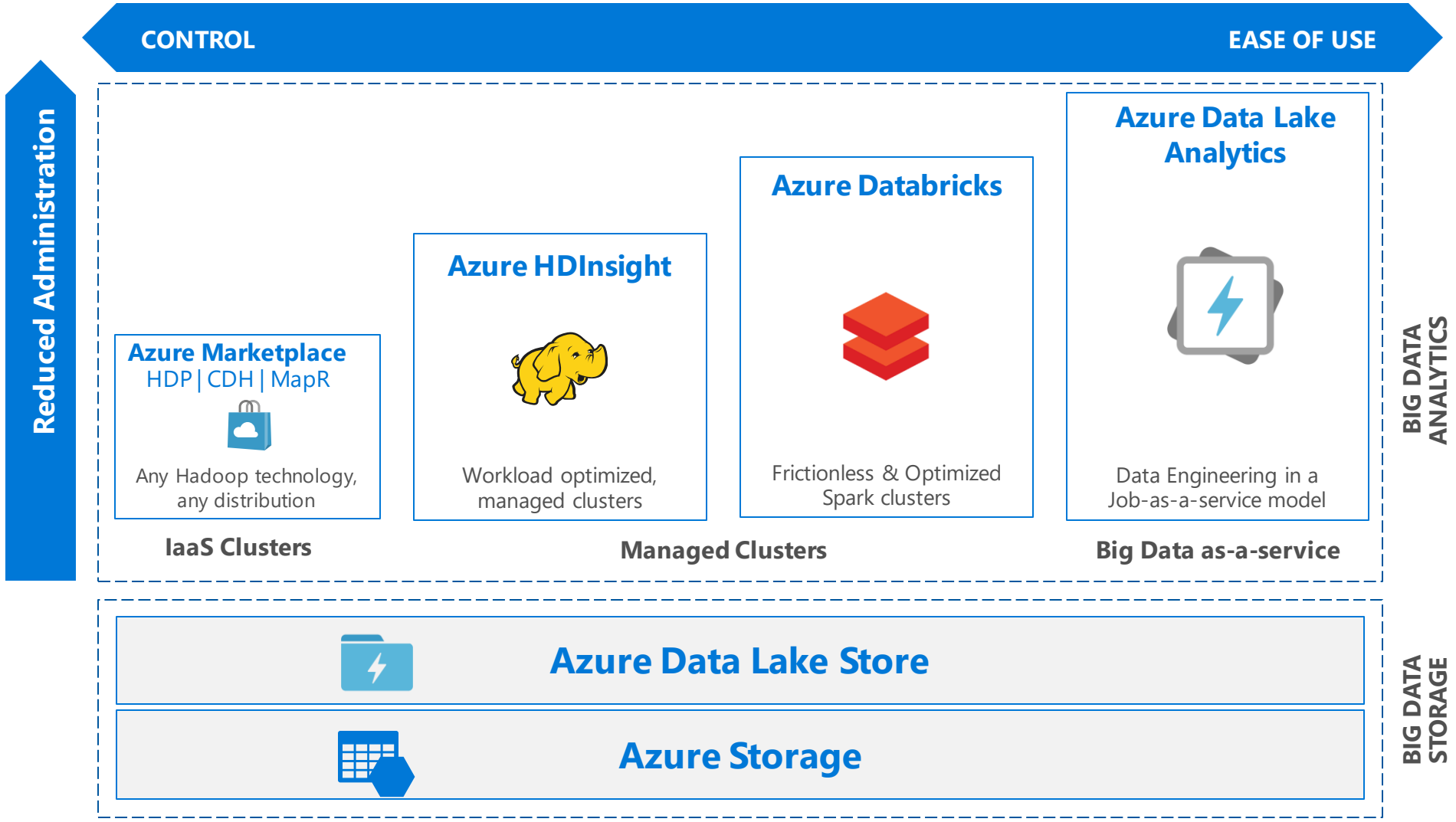
A single application can combine all types of processing

Ecosystem

Spark has built-in support for many data sources, rich ecosystem of ISV applications and a large dev community.

Available on multiple public clouds (AWS, Google and Azure) and multiple on-premises distributors

AZURE DATABRICKS

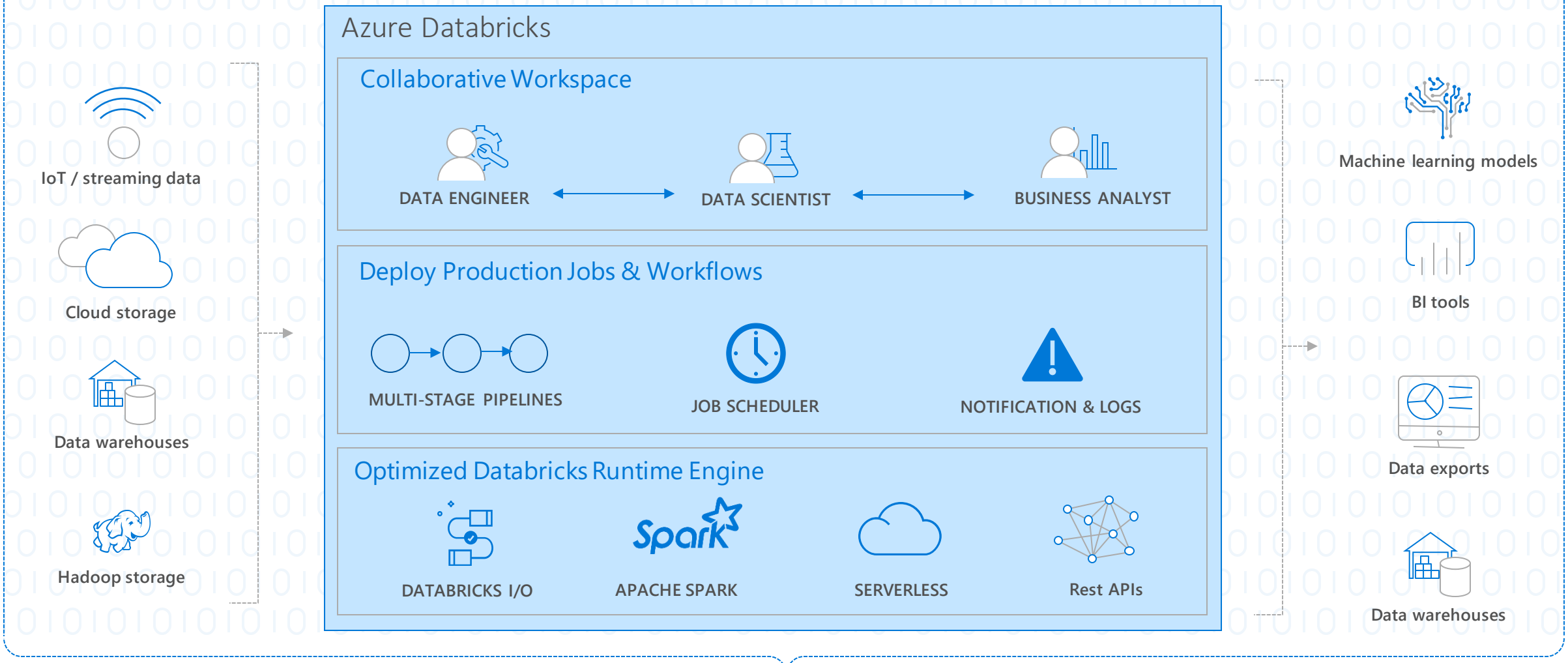


AZURE DATABRICKS

INTRO

- Azure Databricks is a **first party** service on Azure.
- Azure Databricks is integrated seamlessly with Azure services:
  - Azure Portal: Service an be launched directly from Azure Portal
  - Azure Storage Services: Directly access data in Azure Blob Storage and Azure Data Lake Store
  - Azure Active Directory: For user authentication, eliminating the need to maintain two separate sets of uses in Databricks and Azure.
  - Azure SQL DW and Azure Cosmos DB: Enables you to combine structured and unstructured data for analytics
  - Apache Kafka for HDInsight: Enables you to use Kafka as a streaming data source or sink
  - Azure Billing: You get a single bill from Azure
  - Azure Power BI: For rich data visualization
- Eliminates need to create a separate account with Databricks.

AZURE DATABRICKS



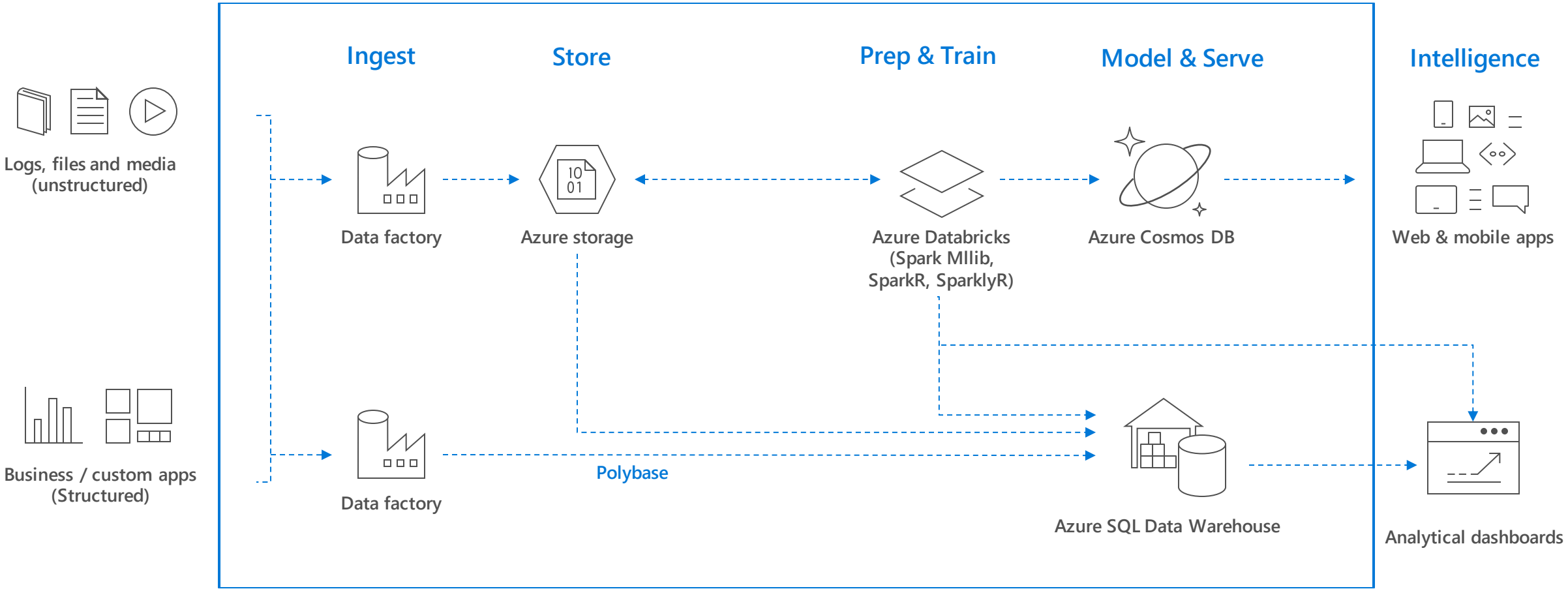
Enhance Productivity

Build on secure & trusted cloud

Scale without limits

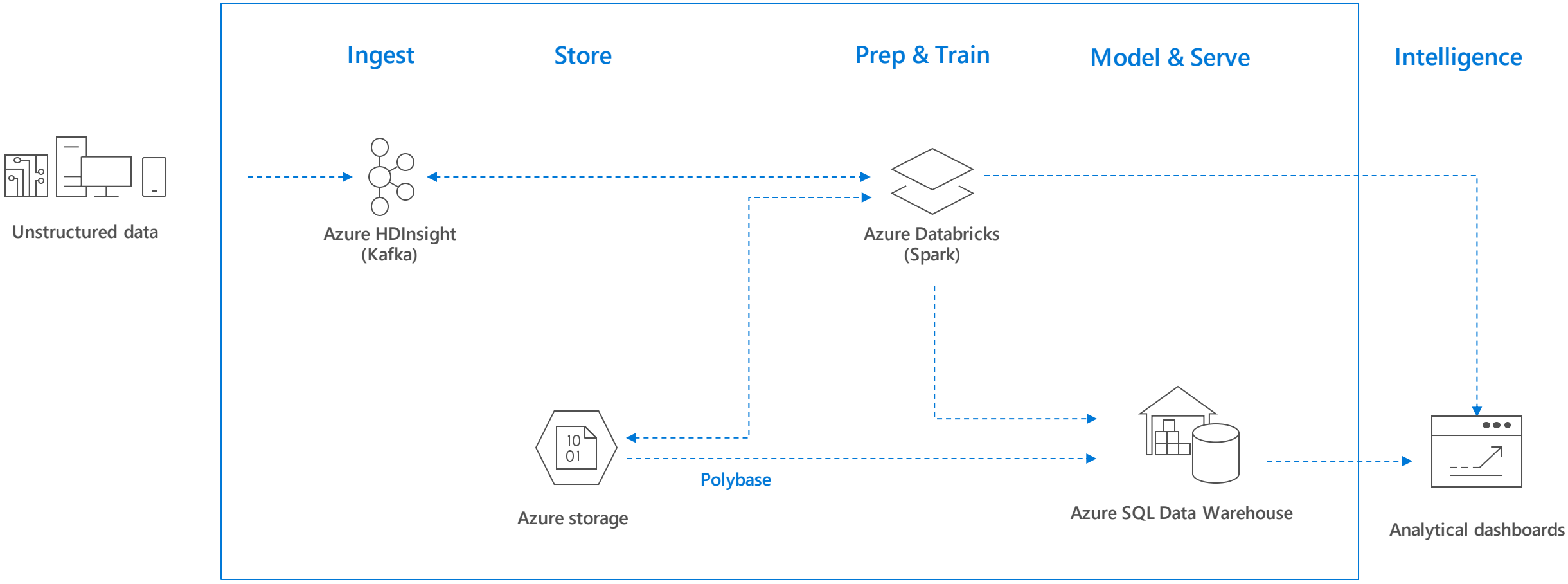
AZURE DATABRICKS

ADVANCED ANALYTICS ON BIG DATA



AZURE DATABRICKS

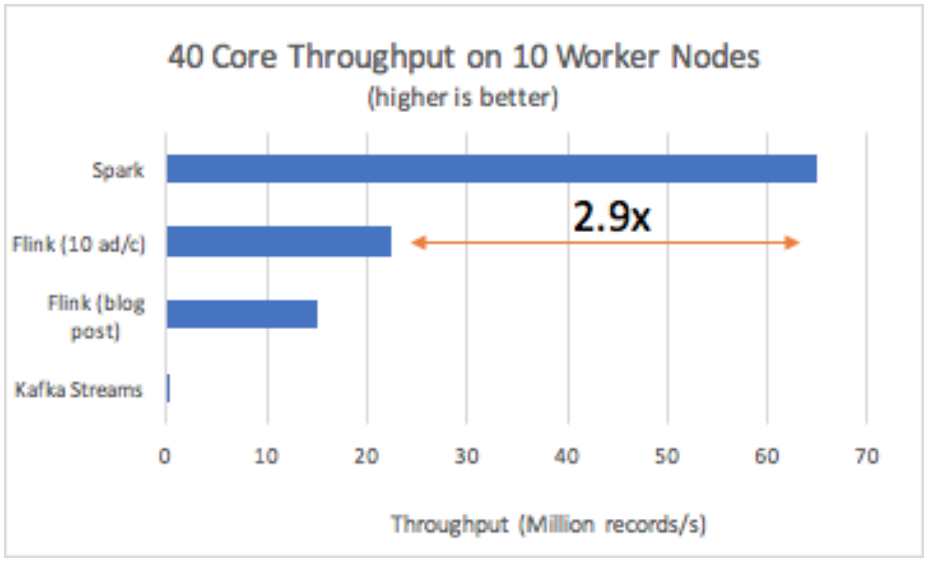
REAL-TIME ANALYTICS



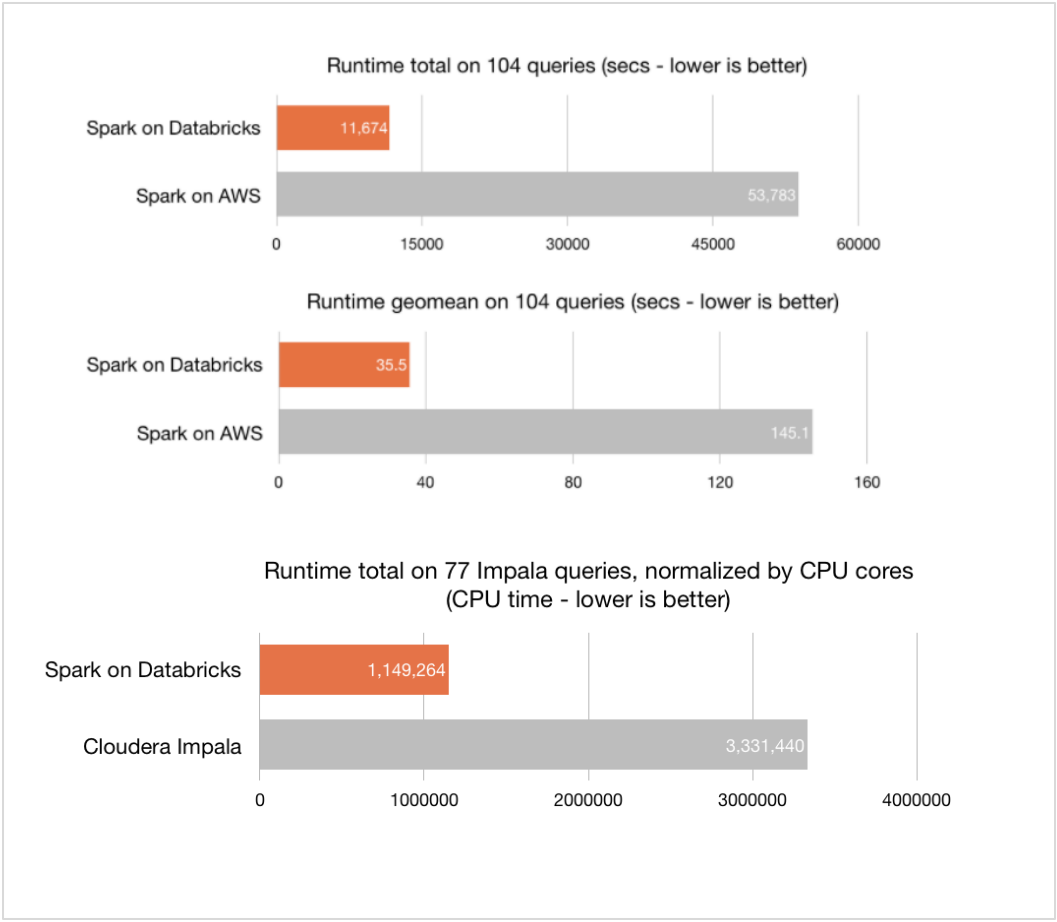
AZURE DATABRICKS

BENCHMARKS

Streaming



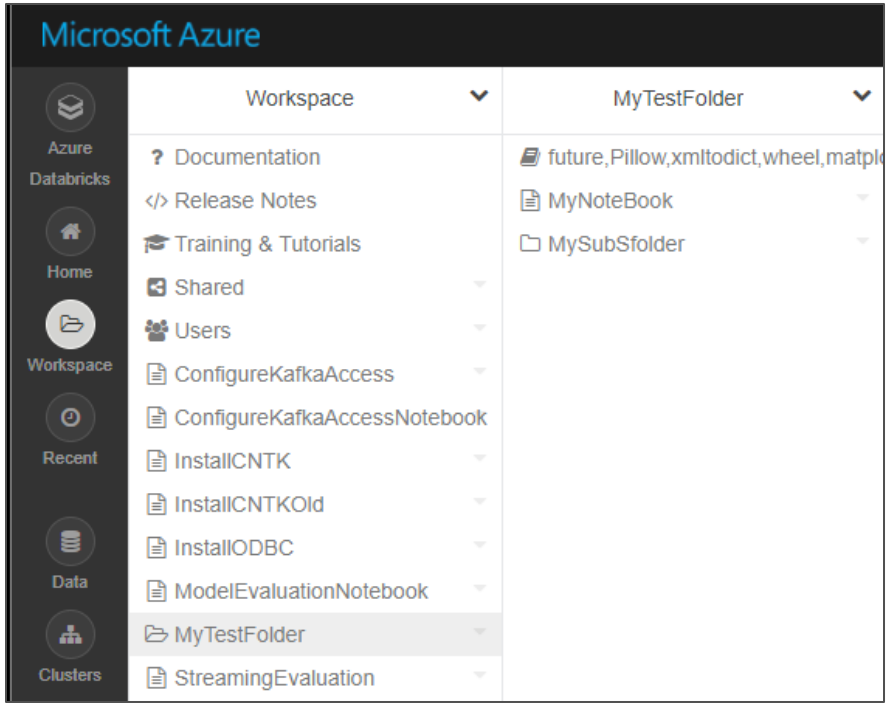
SQL (TPC-DS)

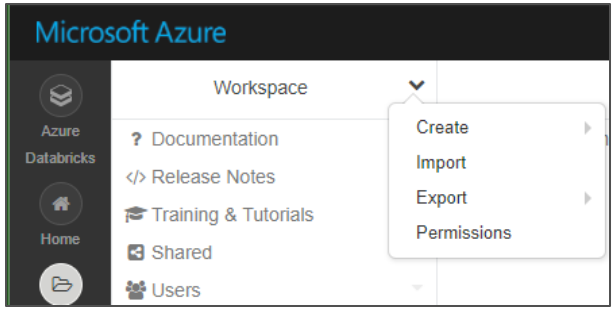


AZURE DATABRICKS

WORKSPACE

- Workspaces—sort of like Directories— are a convenient way to organize an user's Notebook, Libraries and Dashboards.
- Everything in a workspace is organized into hierarchical folders. Folders can hold Libraries, Notebooks, Dashboard or more (sub) folders.
  - Icons indicate the type of the object contained in a folder
- Every user has one directory that is private and unshared.
  - By default, the workspace and all its contents are available to users.
- Fine grained access control can be defined on workspaces to enable *secure collaboration with colleagues*.

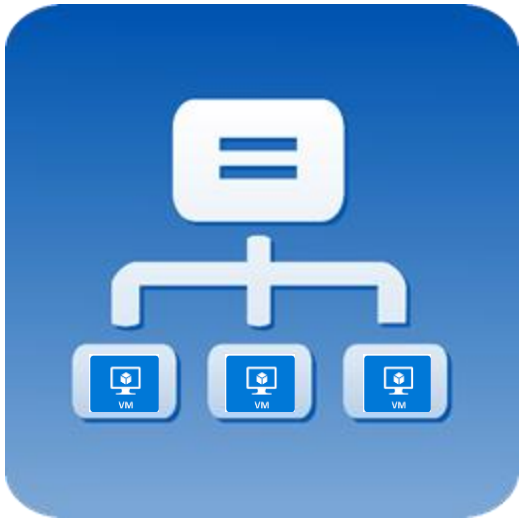




AZURE DATABRICKS

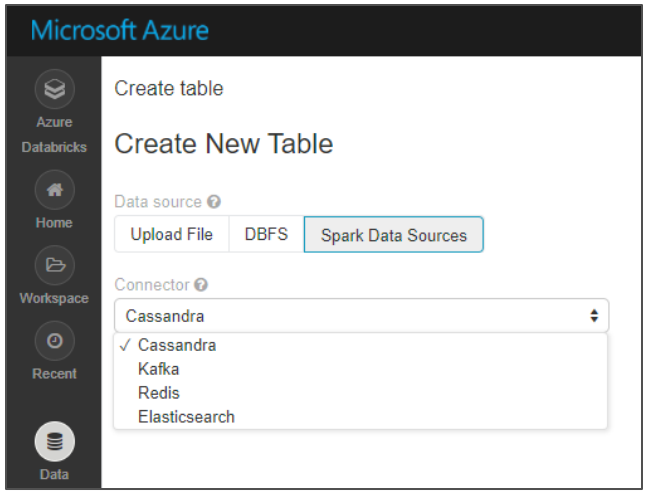
CLUSTERS

- Azure Databricks clusters are the set of **Azure Linux VMs** that host the Spark Worker and Driver Nodes
- Your Spark application code (i.e. Jobs) runs on the provisioned clusters.
- **Azure Databricks clusters are launched in your subscription**—but are managed through the Azure Databricks portal.
- Azure Databricks provides a **comprehensive set of graphical wizards** to manage the complete lifecycle of clusters—from creation to termination.

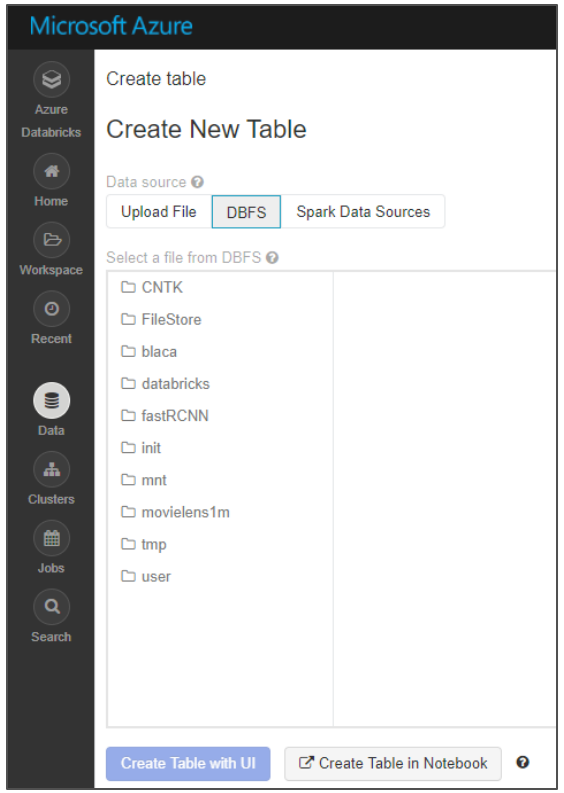


AZURE DATABRICKS

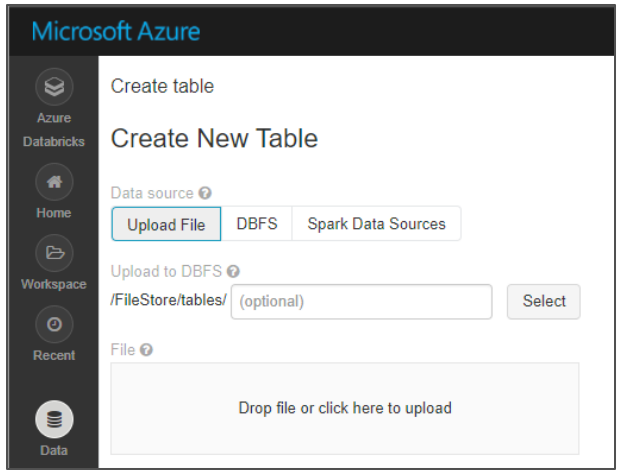
DATA



From Spark Data Sources



From data in DBFS



From local files (in CSV, JSON or Avro formats)

AZURE DATABRICKS

NOTEBOOKS

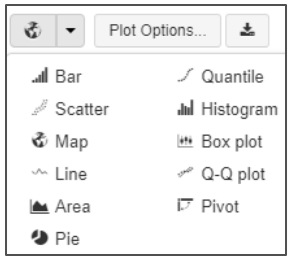
Normally a notebook is associated with a specific language. However, with Azure Databricks notebooks, you can mix multiple languages in the same notebook. This is done using the language magic command:

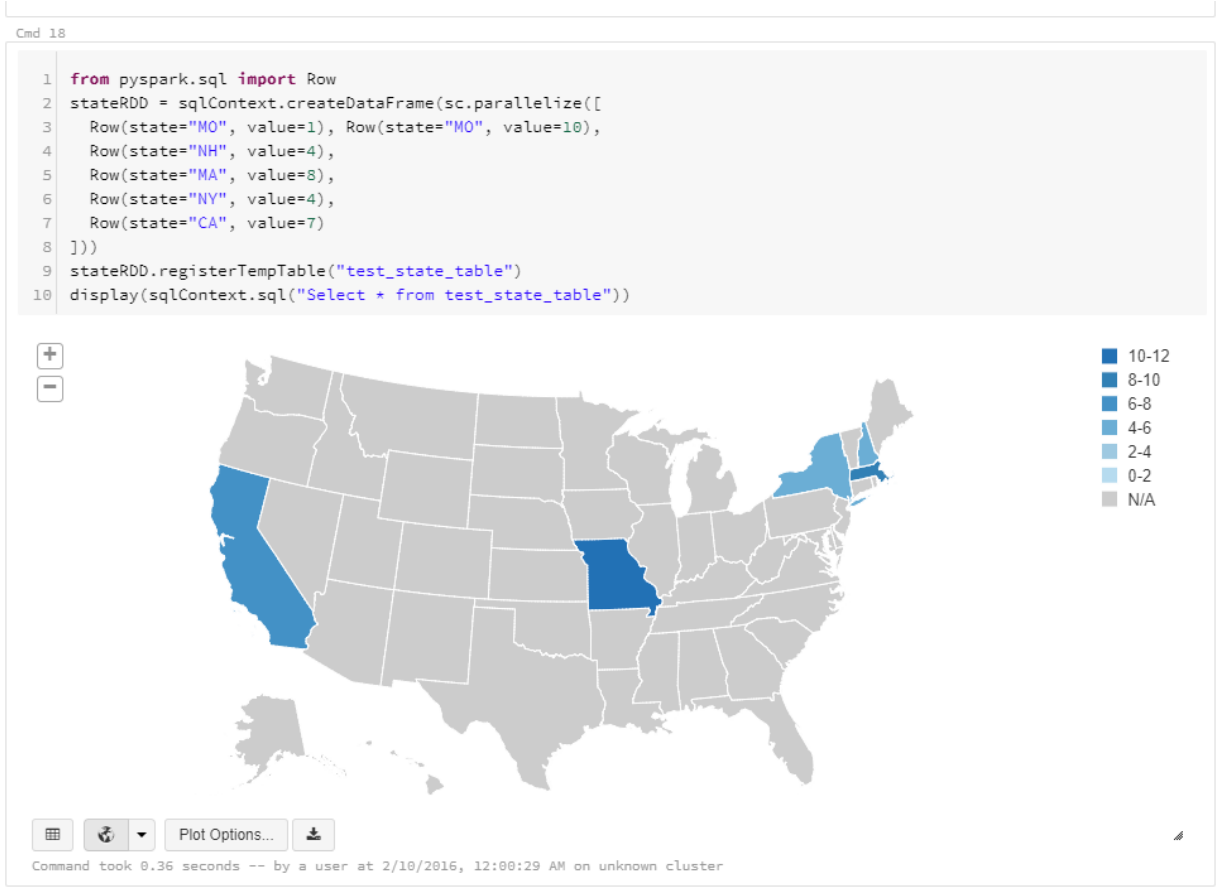
- %python Allows you to execute python code in a notebook (even if that notebook is not python)
- %sql Allows you to execute sql code in a notebook (even if that notebook is not sql).
- %r Allows you to execute r code in a notebook (even if that notebook is not r).
- %scala Allows you to execute scala code in a notebook (even if that notebook is not scala).
- %sh Allows you to execute shell code in your notebook.
- %fs Allows you to use Databricks Utilities - dbutils filesystem commands.
- %md To include rendered markdown

AZURE DATABRICKS

NOTEBOOK - VISUALISATION

- All notebooks, *regardless of their language*, support Databricks visualizations.
- When you run the notebook the visualizations are rendered inside the notebook in-place
- The visualizations are written in HTML.
  - You can save the HTML of the entire notebook by exporting to HTML.
  - If you use Matplotlib, the plots are rendered as images so you can just right click and download the image
- You can change the plot type just by picking from the selection



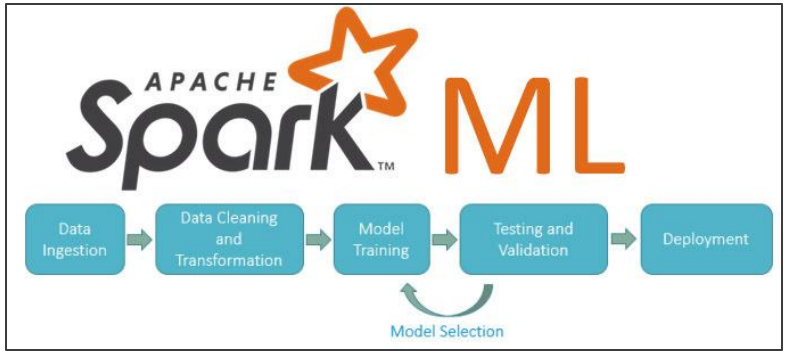


AZURE DATABRICKS

MACHINE LEARNING

- Microsoft Machine Learning Library for Apache Spark (MMLSpark) lets you easily create scalable machine learning models for large datasets. It includes integration of SparkML pipelines with the Microsoft Cognitive Toolkit and OpenCV, enabling you to:
- Spark MLlib comes pre-installed on Azure Databricks
- 3rd Party libraries supported include: H20 Sparkling Water, SciKit-learn and XGBoost
- Supports Deep Learning Libraries/frameworks including:
  - Microsoft Cognitive Toolkit (CNTK).
    - Article explains how to install CNTK on Azure Databricks.
  - TensorFlowOnSpark
  - BigDL

TensorFlow



AZURE DATABRICKS

MACHINE LEARNING - SPARKML

Spark ML Algorithms

Classification and Regression	• Linear Models (SVMs, logistic regression, linear regression) • Naïve Bayes • Decision Trees • Ensembles of trees (Random Forest, Gradient-Boosted Trees) • Isotonic regression
Clustering	• k-means and streaming k-means • Gaussian mixture • Power iteration clustering (PIC) • Latent Dirichlet allocation (LDA)
Collaborative Filtering	• Alternating least squares (ALS)
Dimensionality Reduction	• SVD • PCA
Frequent Pattern Mining	• FP-growth • Association rules
Basic Statistics	• Summary statistics • Correlations • Stratified sampling • Hypothesis testing • Random data generation