Master

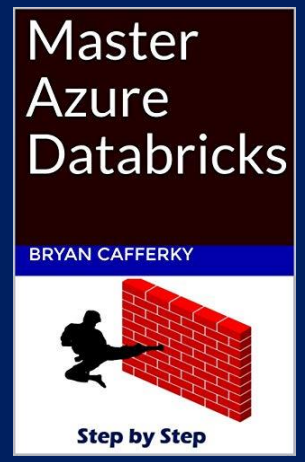
databricks + APACHE Spark™

Lesson 1
Introduction

Bryan Cafferky
Data & AI Enabler

Where Are We Going?

- What is Apache Spark?
- What is Databricks?
- Scaling Up and Out with Barry the Weightlifter
- My Kitchen Drawer and the Apache Hadoop Project
- Understanding Apache Spark and Databricks



What is Apache Spark?

- An open-source big data platform for data science
- Big Data includes massive data volume, streaming data, unstructured and semi-structured data, images, video, sound.
- Bring your own tools
- Weak support for collaboration
- Not optimized for the cloud

What is Databricks?

- Commercial product from the creators of Apache Spark
- Complete development environment for Apache Spark
- Numerous proprietary Spark enhancements
- Ideal for Data Science team collaboration
- Many development tools
- Optimized for the Cloud

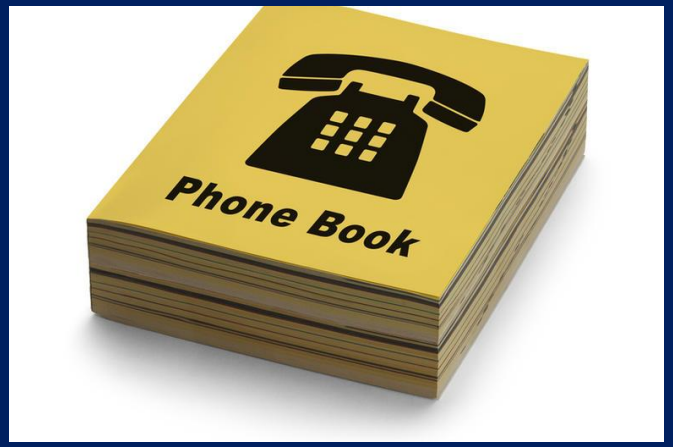
Scale Up, Scale Out, and Barry the Weightlifter

Scale Up vs. Scale Out

Scale Up

Scale Out

Scaling Out

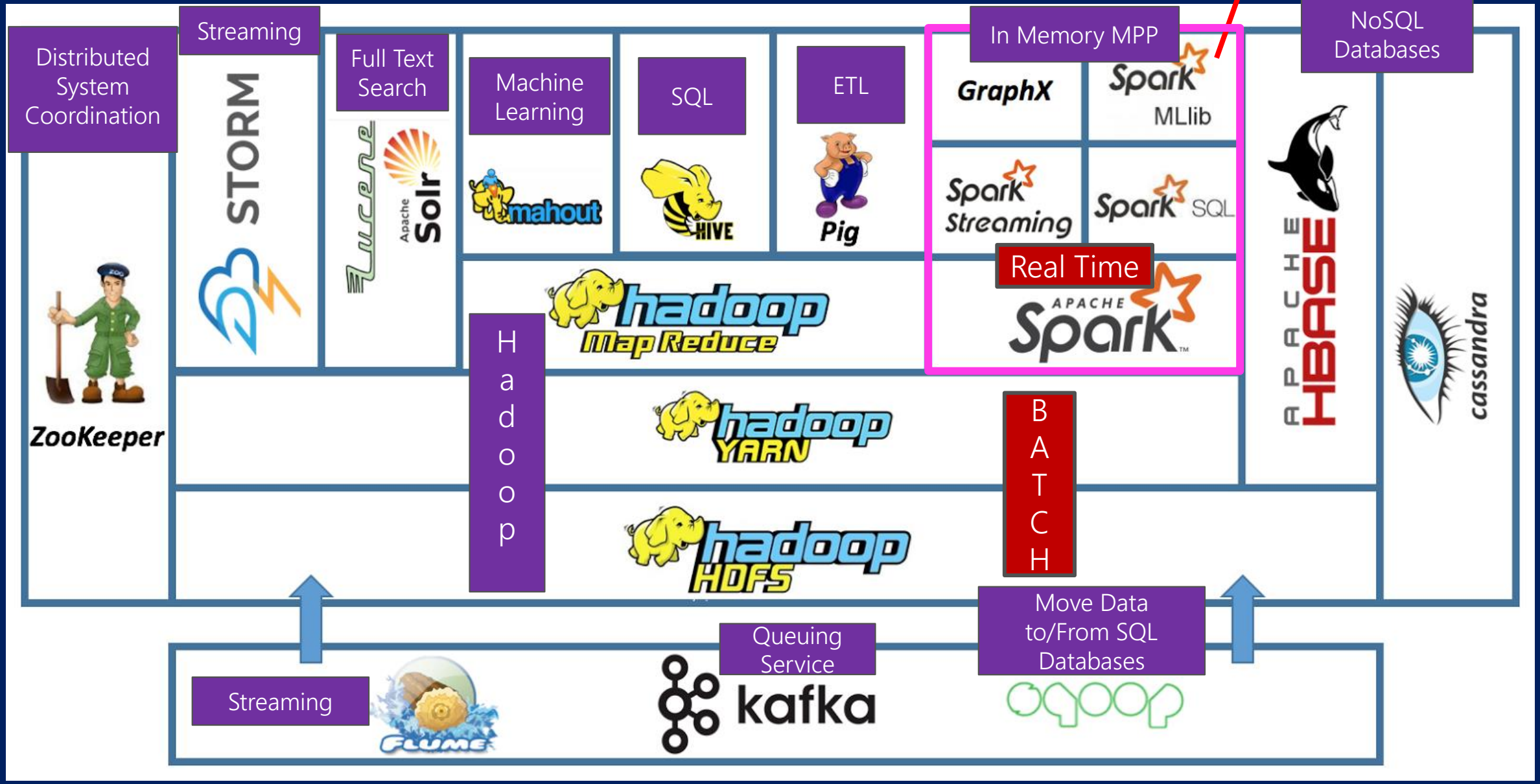


Scale Out

Apache Hadoop project

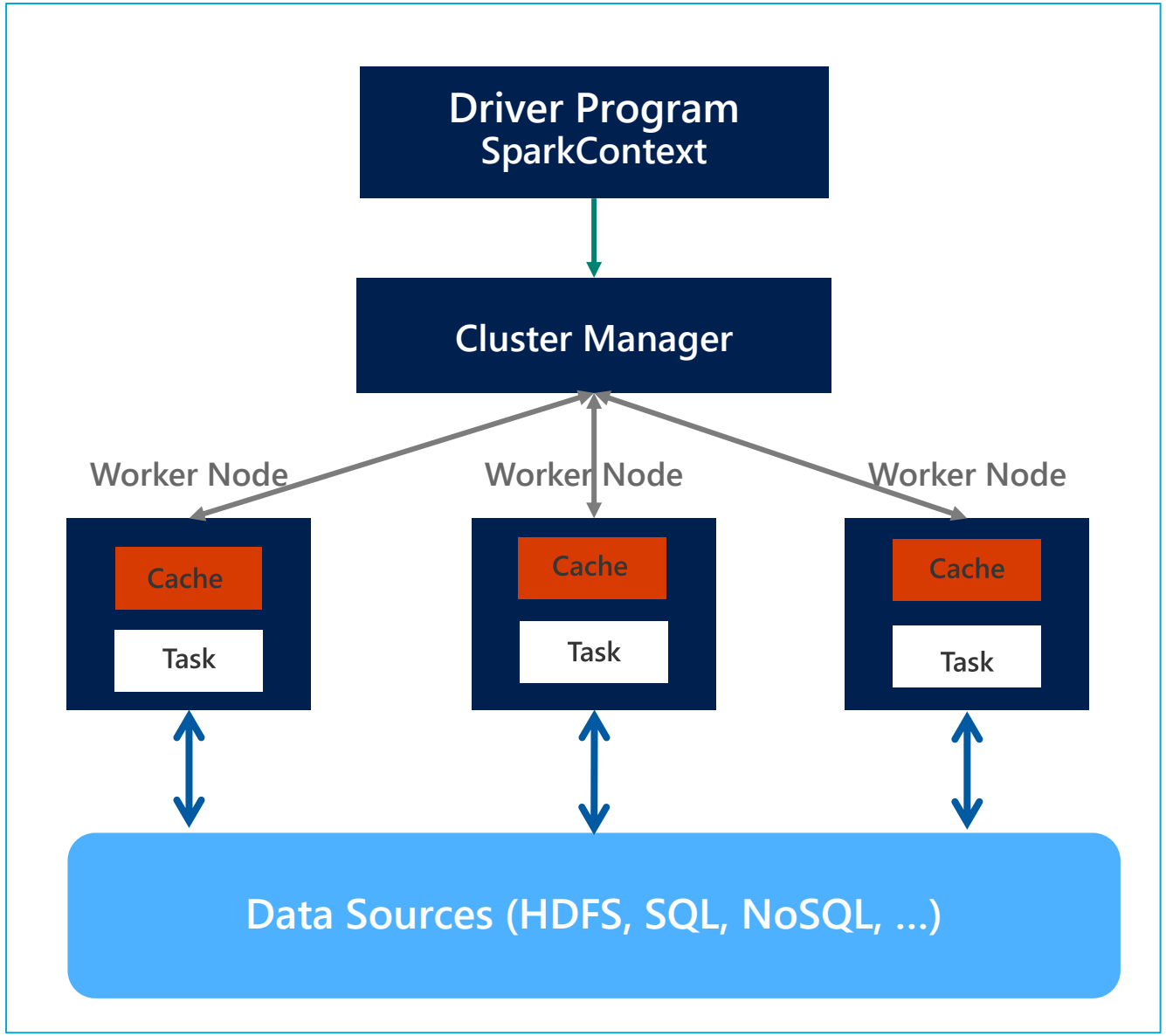
Apache Hadoop and My Kitchen Drawer

The Apache Hadoop Ecosystem



GENERAL SPARK CLUSTER ARCHITECTURE

- 'Driver' runs the user's 'main' function and executes the various parallel operations on the worker nodes.
- The results of the operations are collected by the driver
- The worker nodes read and write data from/to Data Sources including HDFS.
- Worker node also cache transformed data in memory as RDDs (Resilient Data Sets).
- Worker nodes and the Driver Node execute as VMs in public clouds (AWS, Google and Azure).

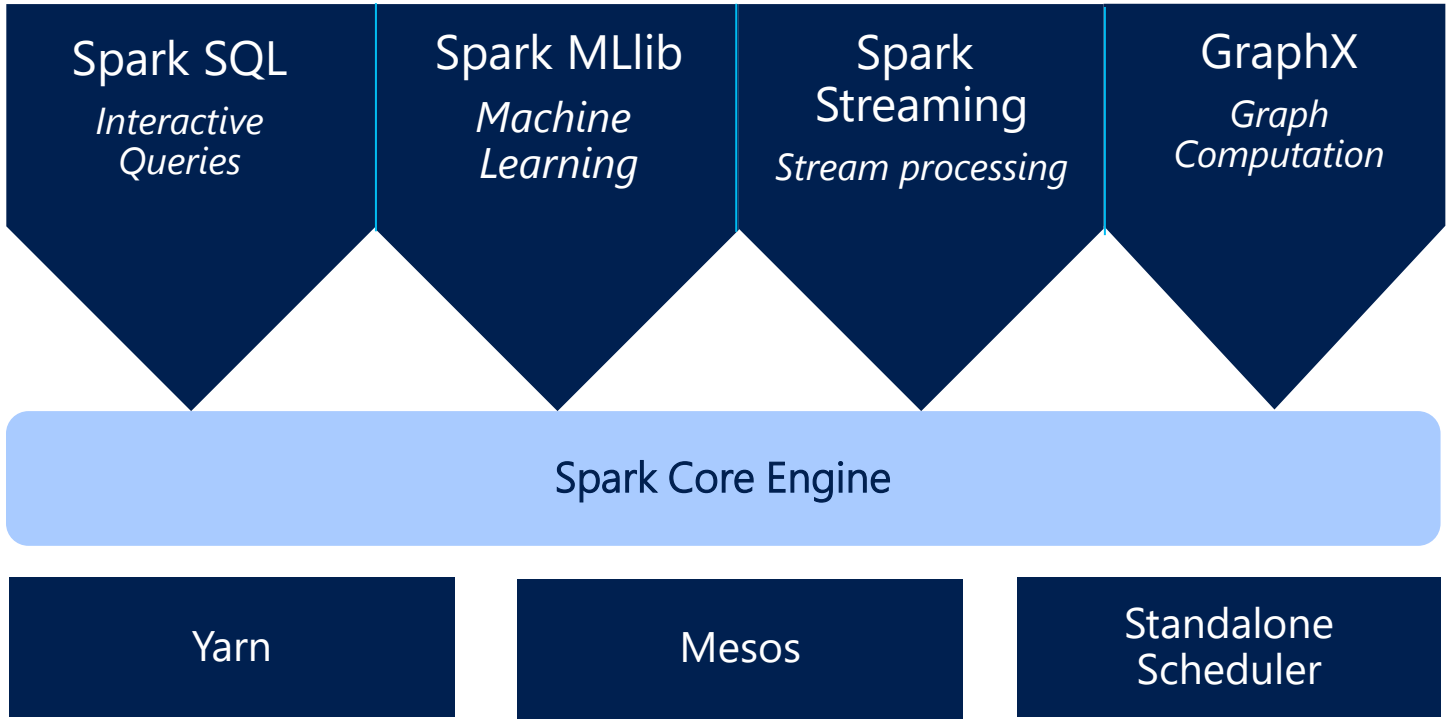


APACHE SPARK

A unified, open source, parallel, data processing framework for Big Data Analytics

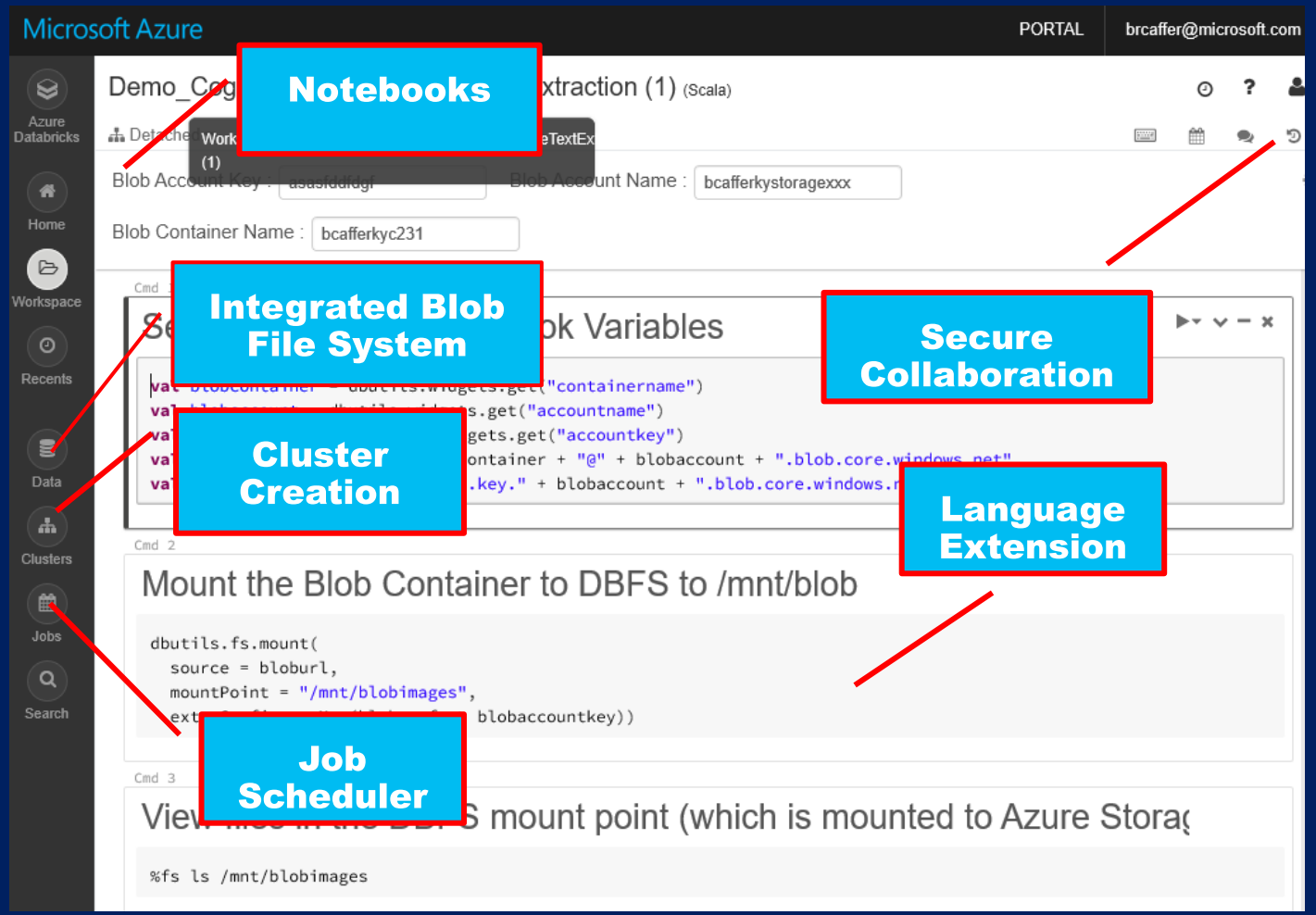
Spark Unifies:

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

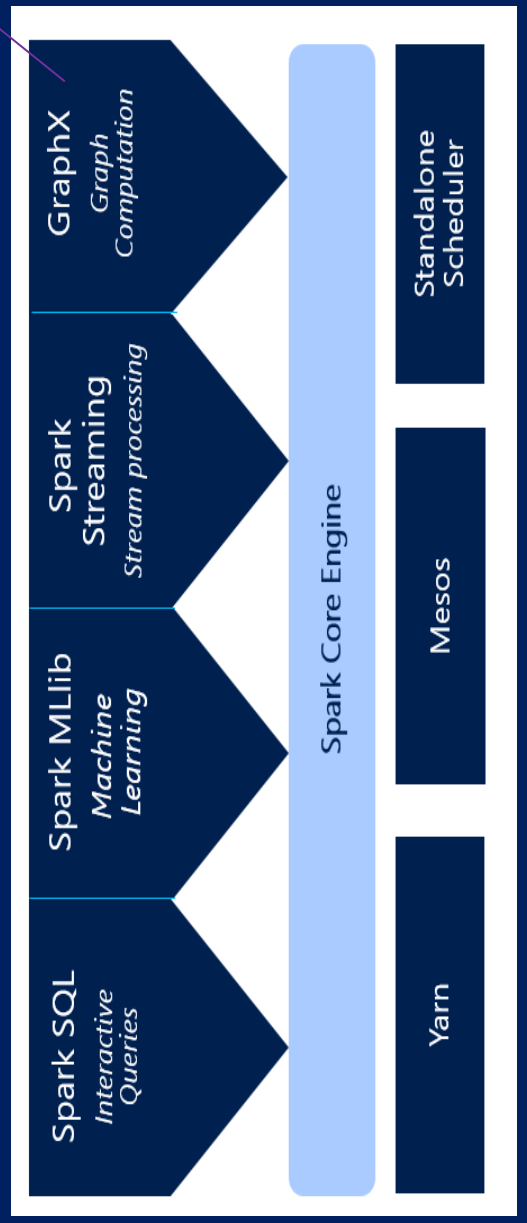


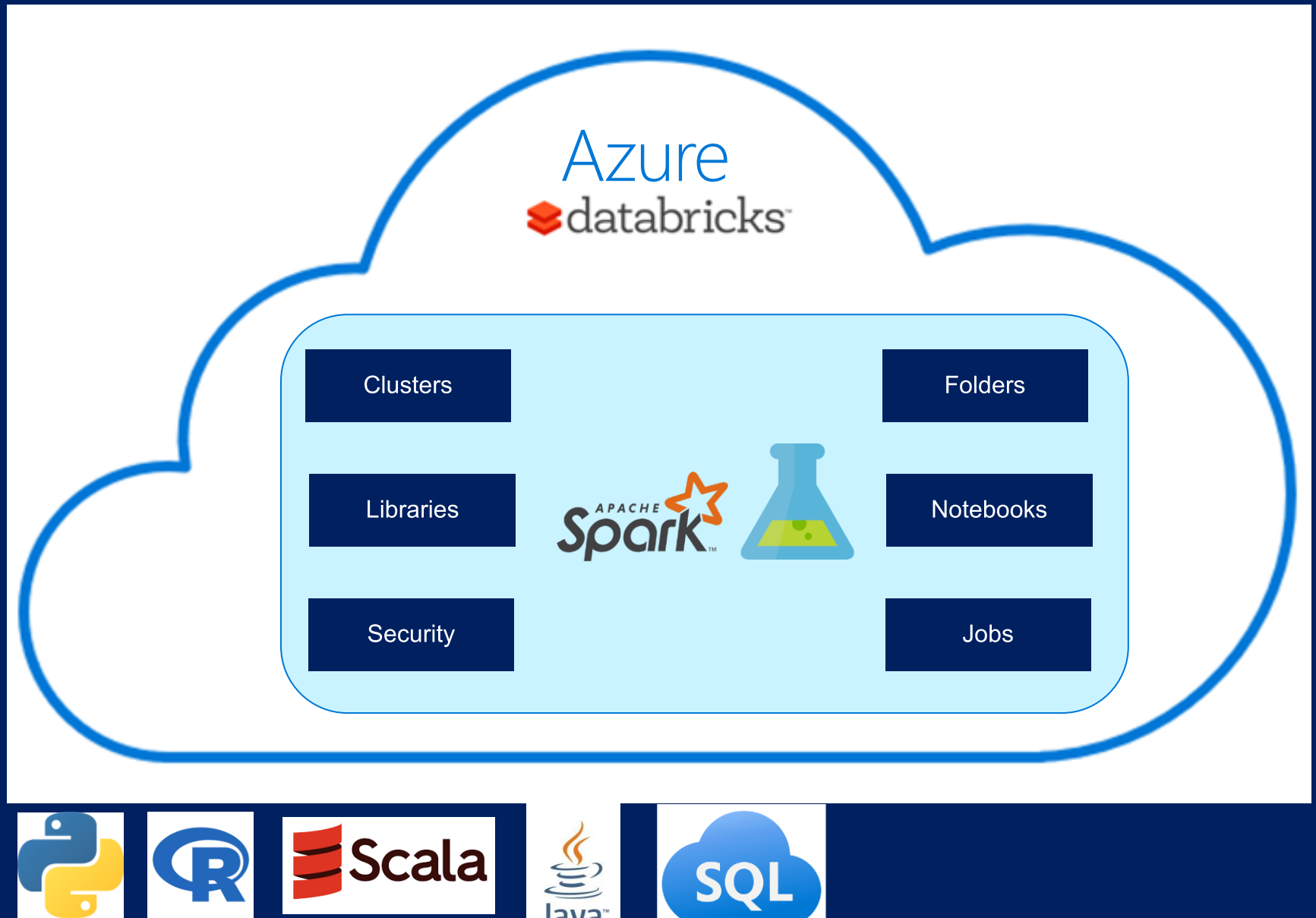
Azure Databricks

It all runs on Spark

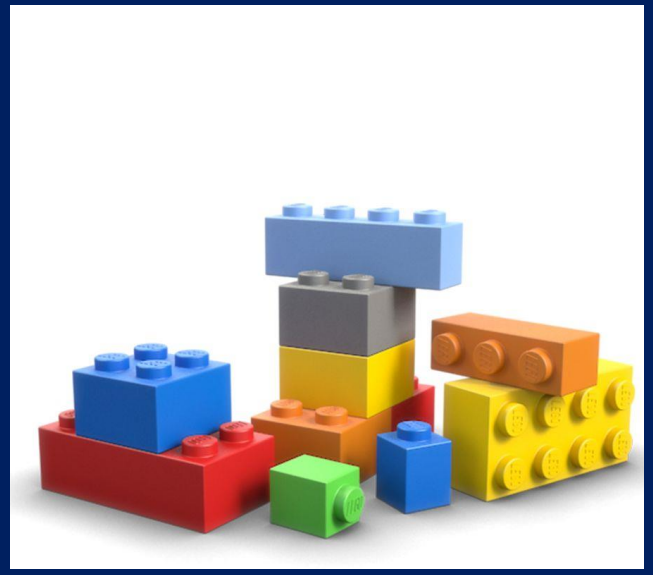


Spark Optimizations




Azure
databricks
Clusters
Folders
Libraries
Notebooks
Security
Jobs
APACHE
Spark
Scala
Java
SQL

Apache Spark vs. Azure Databricks



Apache Spark

Databricks

Wrapping Up

- What is Apache Spark?
- What is Databricks?
- Scaling Up, Scaling Out and Barry the Weightlifter
- Apache Hadoop and My Kitchen Drawer
- Understanding Spark and Databricks

Thank You!