Master

databricks

2nd Edition

All New!

Lesson 1

Introduction

Where Are We Going?

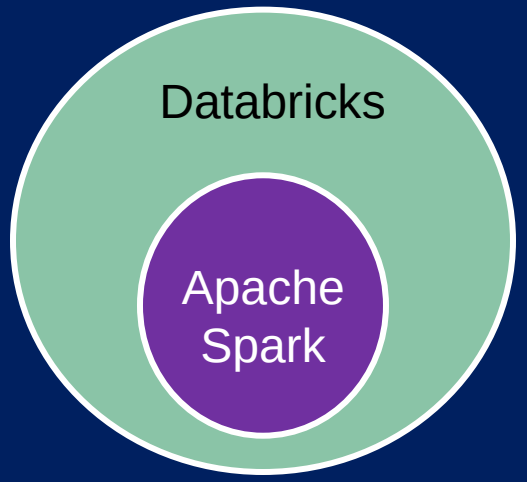
- New Series – Databricks Only, Updated Content
- What is Apache Spark?
- What is Databricks?
- Scaling Up and Out with Barry the Weightlifter
- Understanding Apache Spark and Databricks
- The Databricks Workspace & User Interface

What is Apache Spark?

- An open-source big data platform for data analytics.
- Big Data means challenging data and 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 services & features.
- Ideal for team collaboration.
- Many development tools.
- Optimized for the Cloud.
  - AWS, Azure, GCP



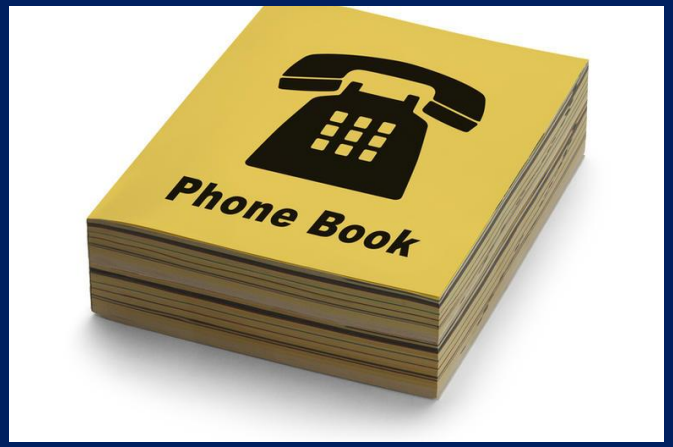
Scale Up, Scale Out, and Barry the Weightlifter

Scale Up vs. Scale Out

Scale Up

Scale Out

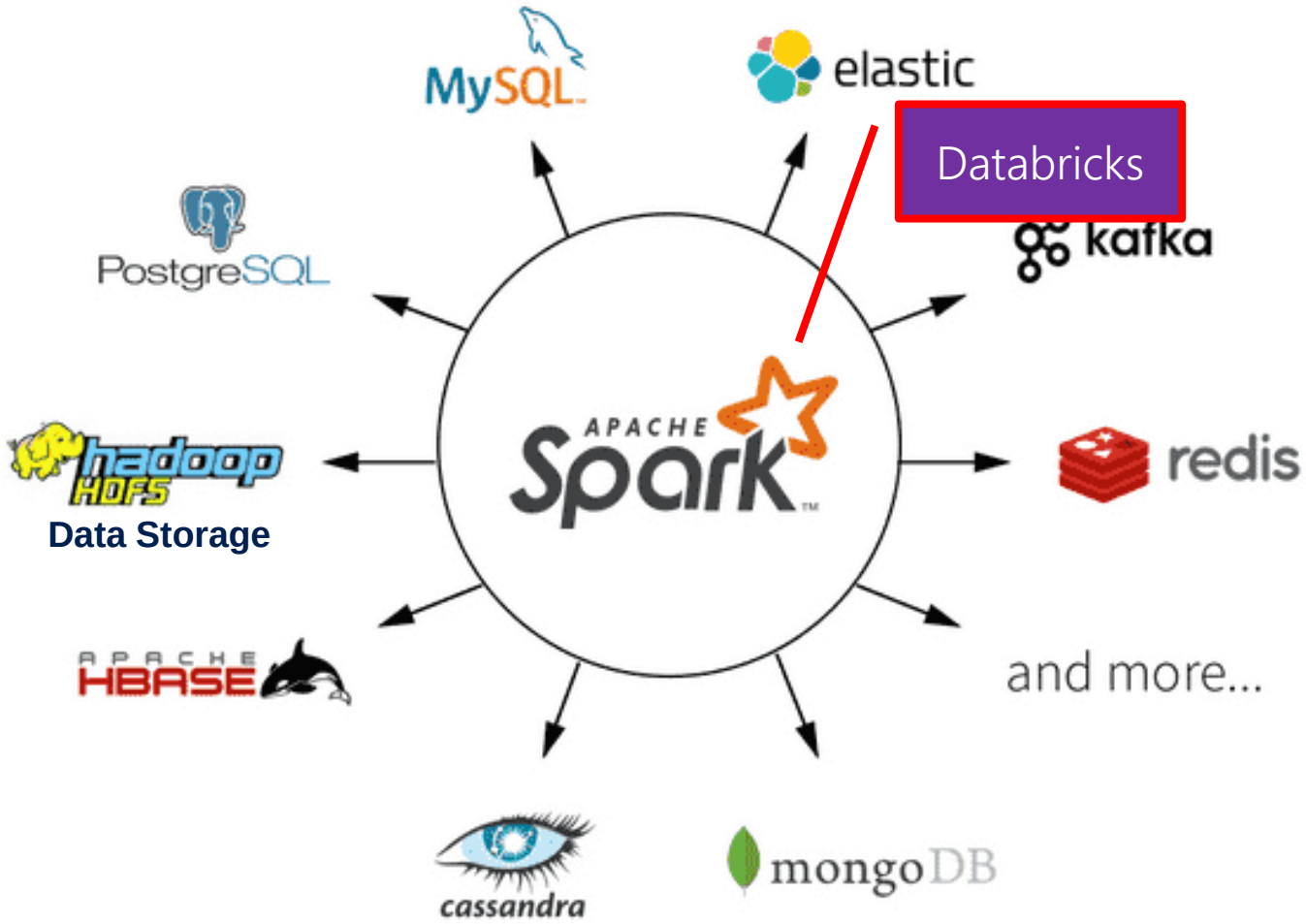
Scaling Out



Scale Out

- Spark distributes the data based on its' own optimization algorithm.

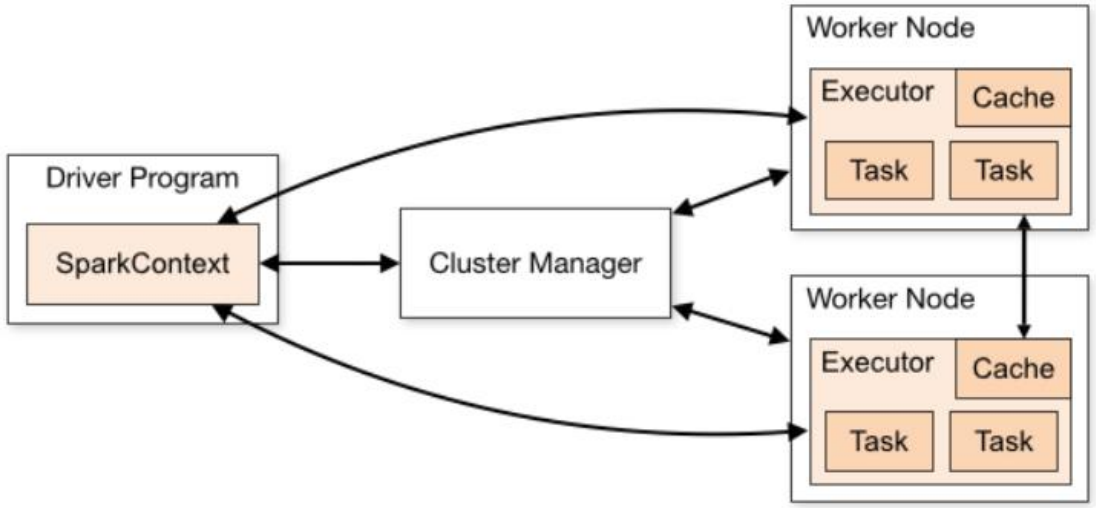
Apache Spark

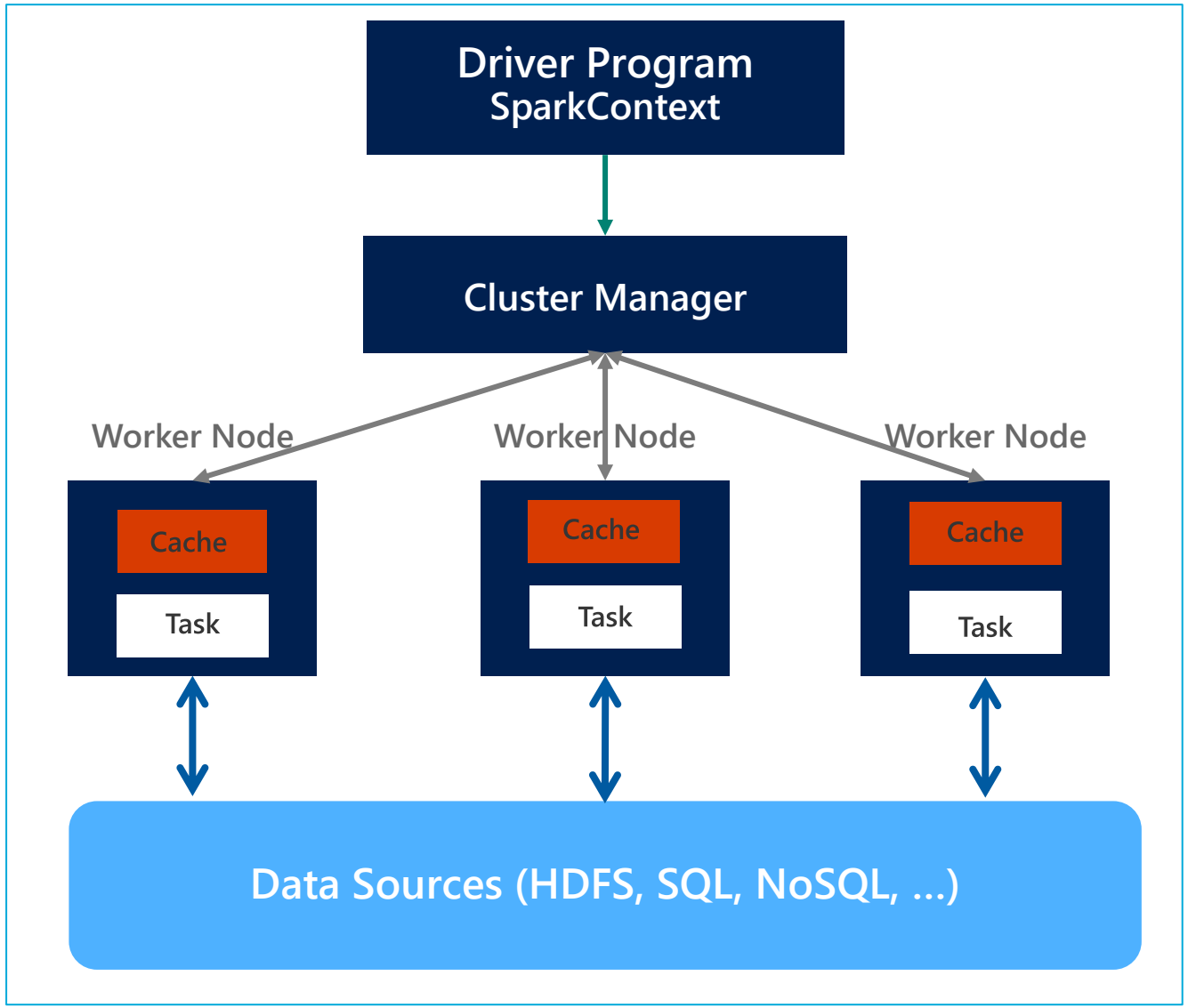


GENERAL SPARK CLUSTER ARCHITECTURE

Massive Parallel Processing

- '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.



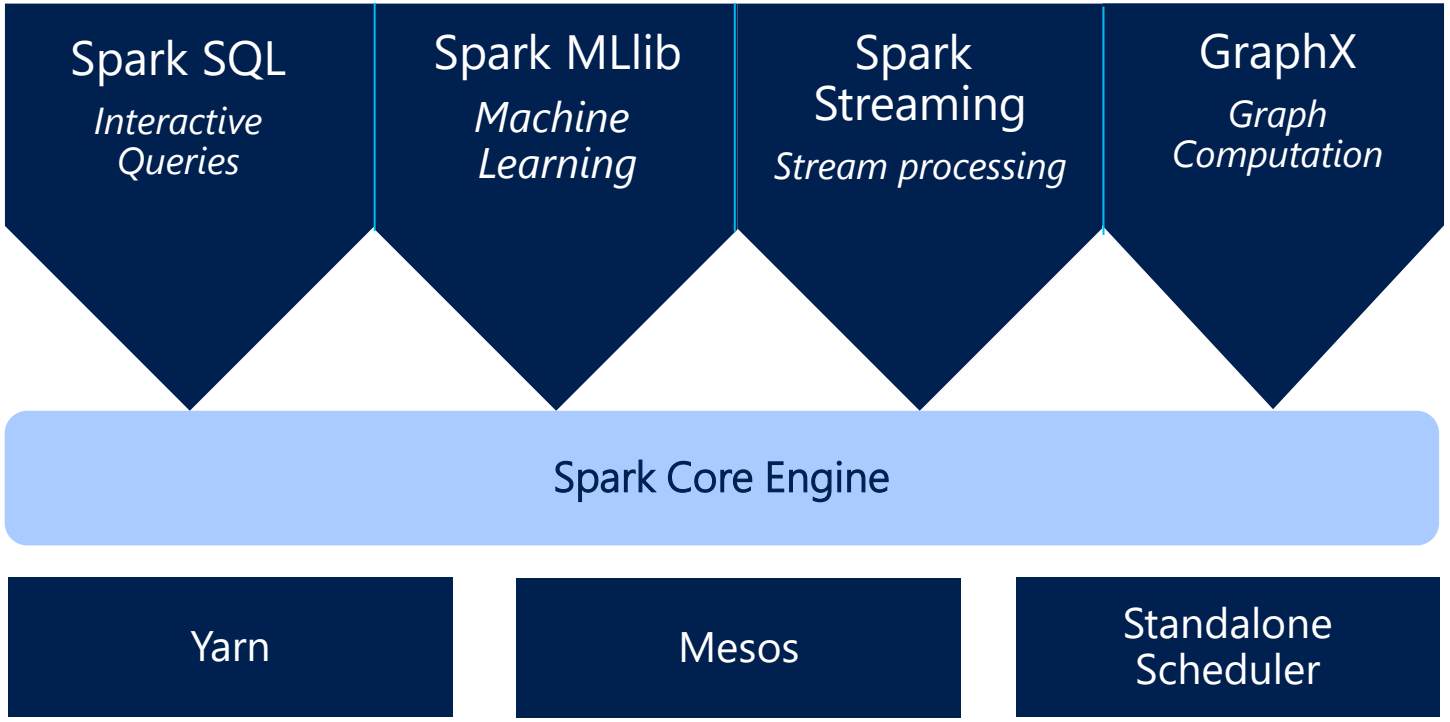


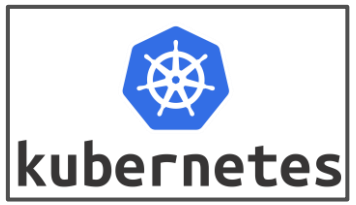
APACHE SPARK

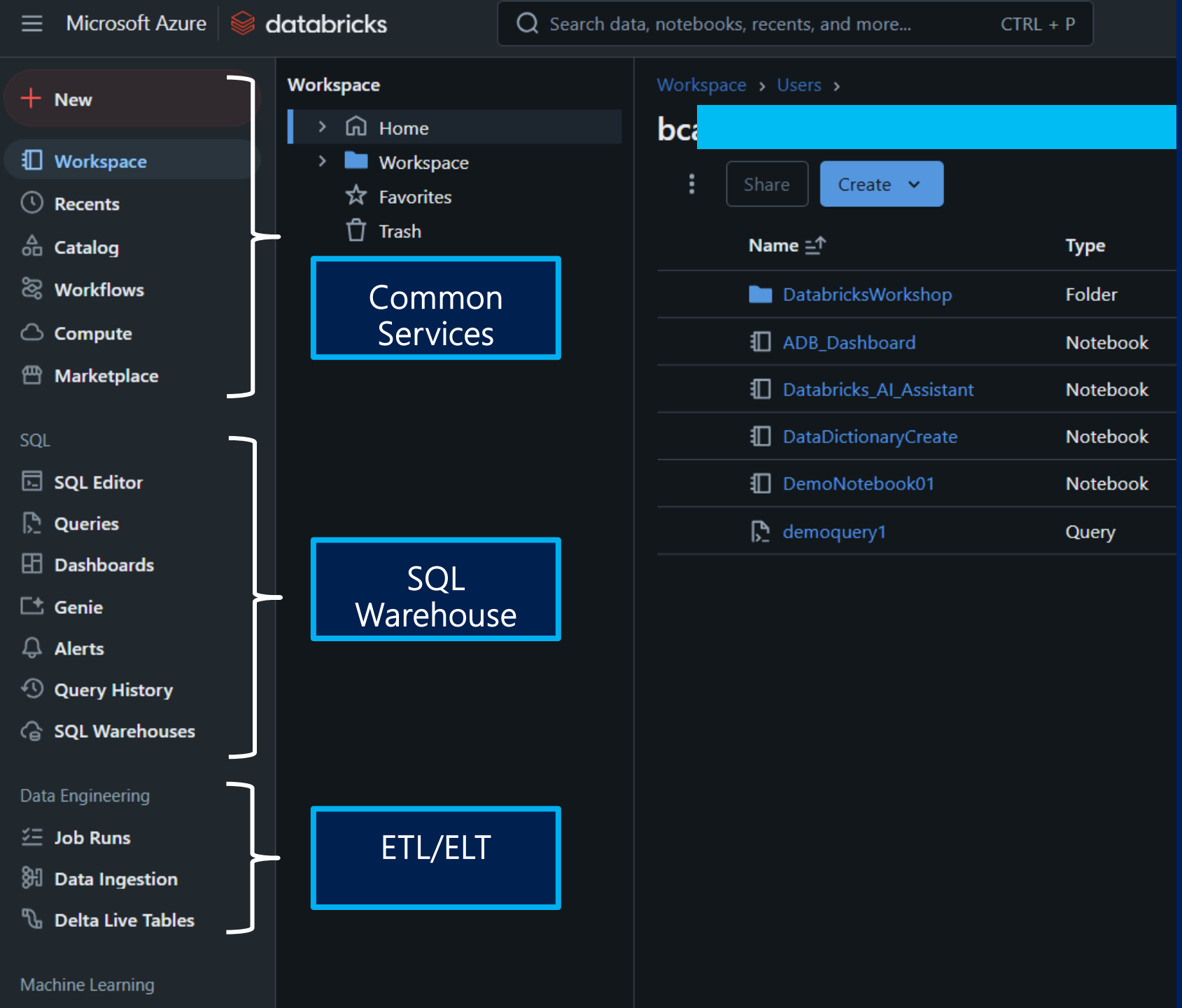
A unified, open source, parallel, data processing framework for Big Data Analytics

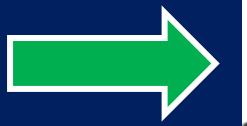
Spark Unifies:

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

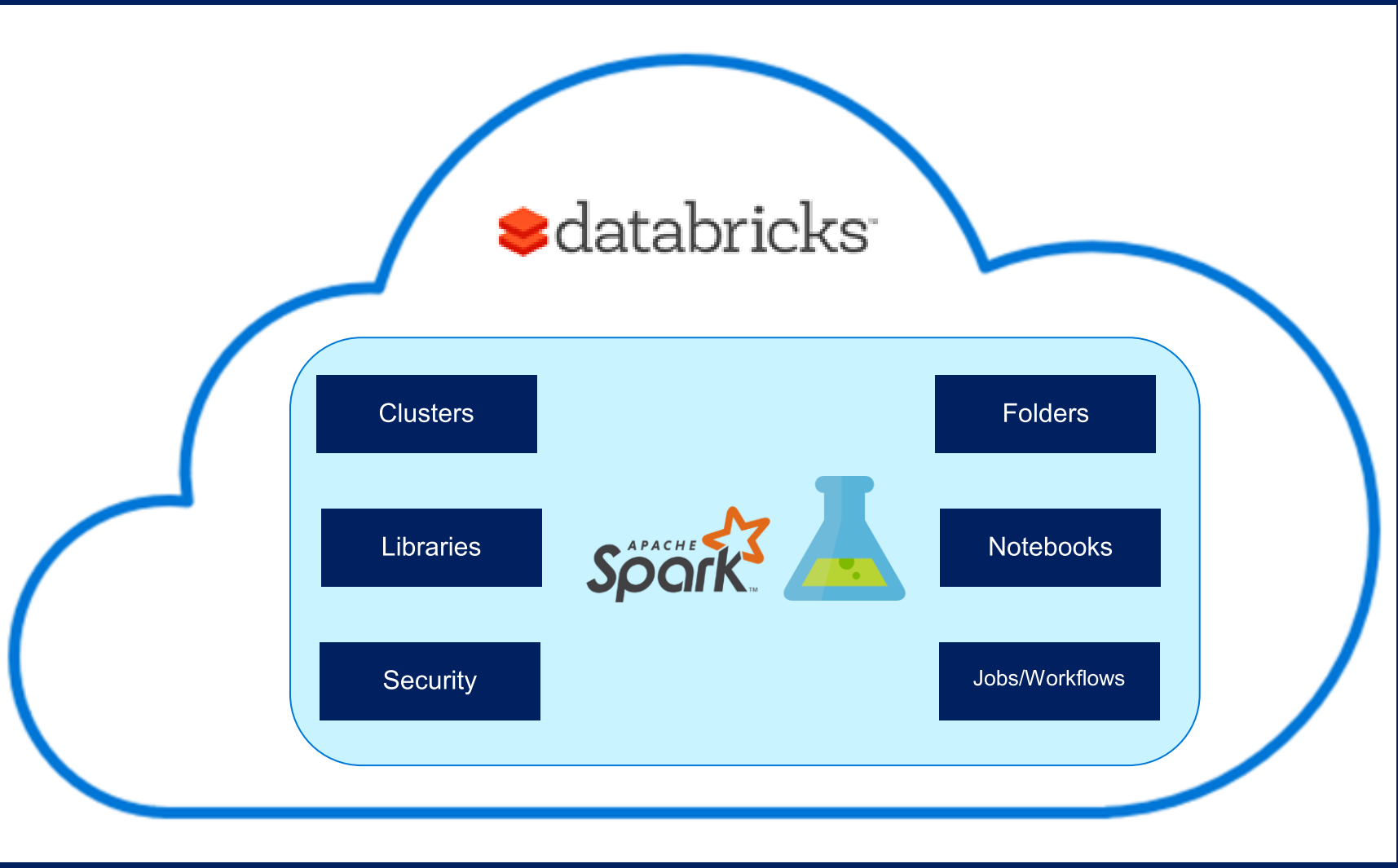








The Databricks Workspace User Interface



The Databricks Workspace Objects

Apache Spark vs. Databricks

Apache Spark

Databricks

Wrapping Up

- New Series – Databricks Only
- What is Apache Spark?
- What is Databricks?
- Scaling Up and Out with Barry the Weightlifter
- Understanding Apache Spark and Databricks
- The Databricks Workspace & User Interface

Thank You!