



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**



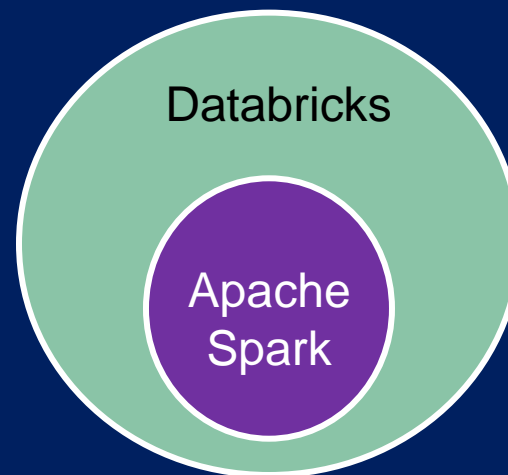
What is Apache Spark?

- **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

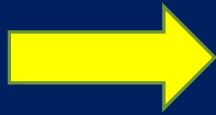


Scale Up



Scale Out

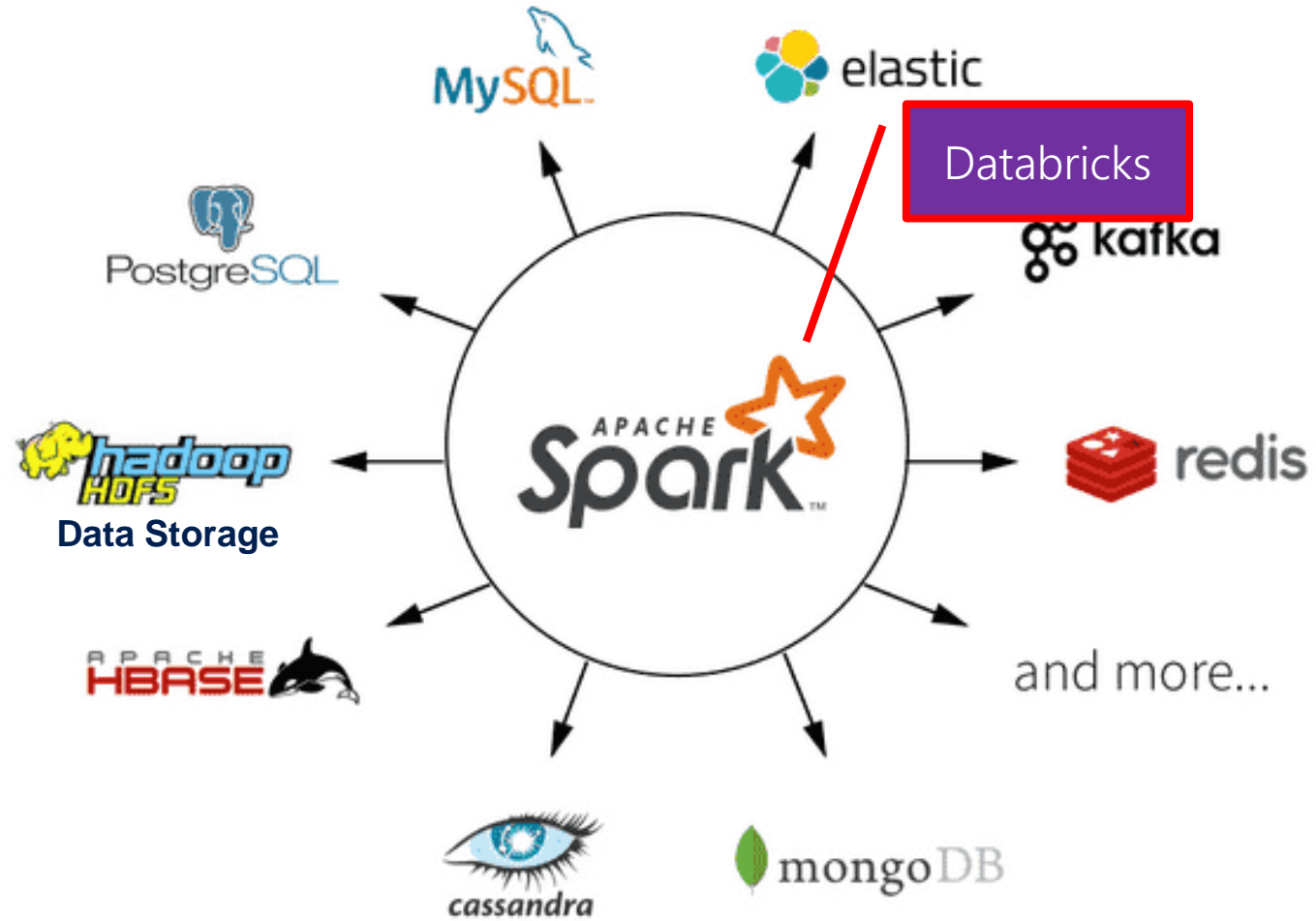
Scaling Out



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

Scale Out

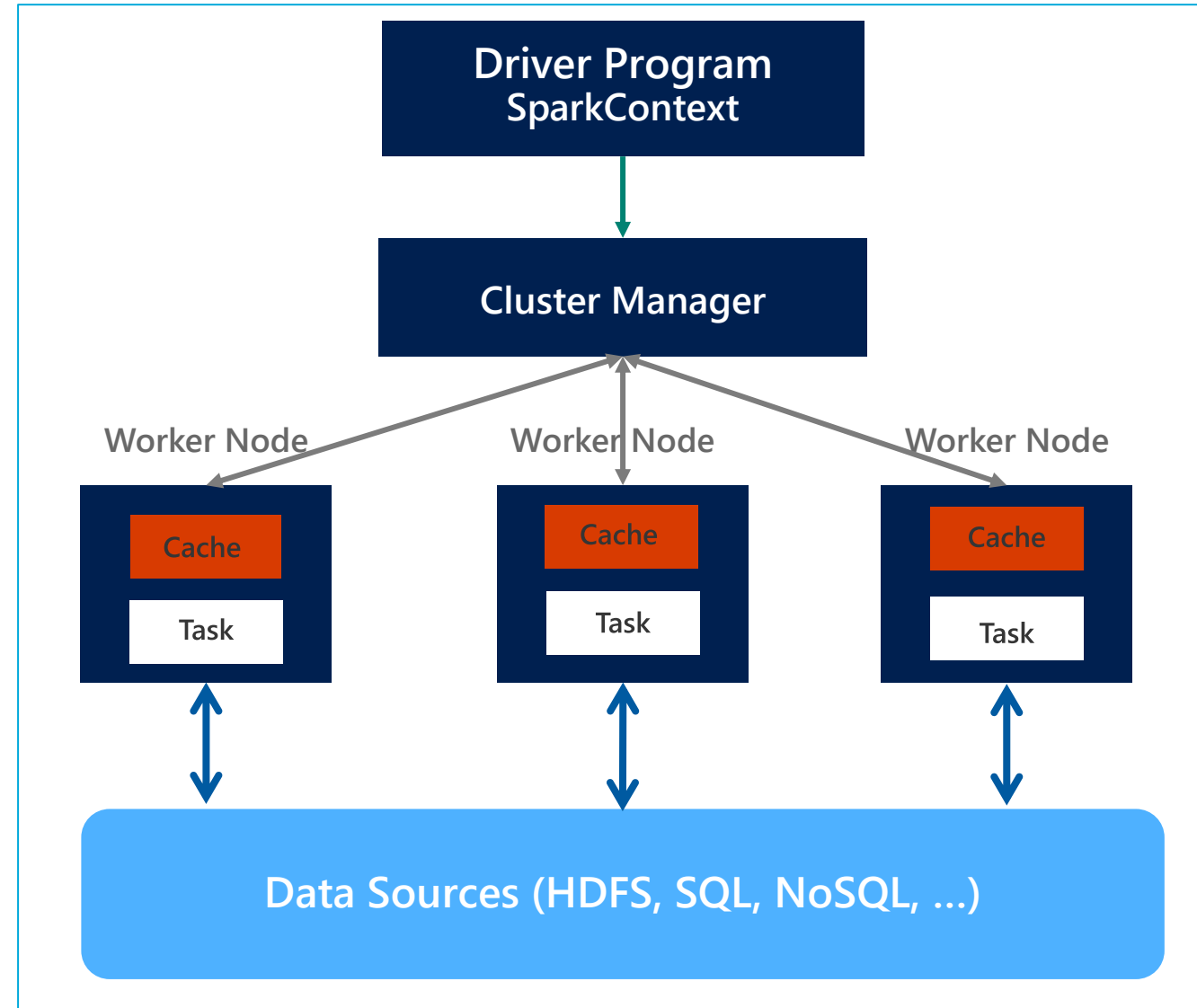
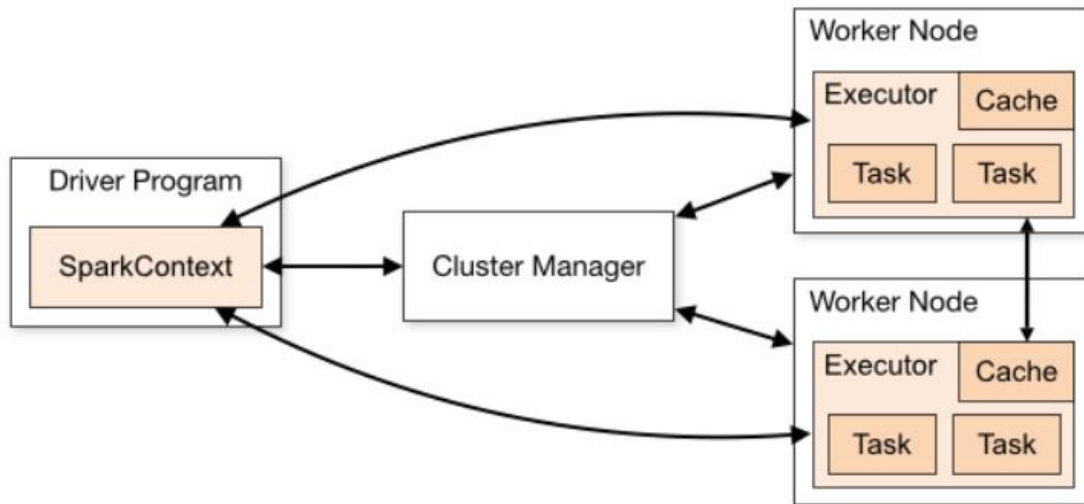
Apache Spark



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.

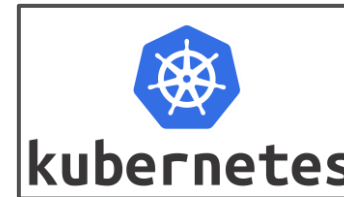
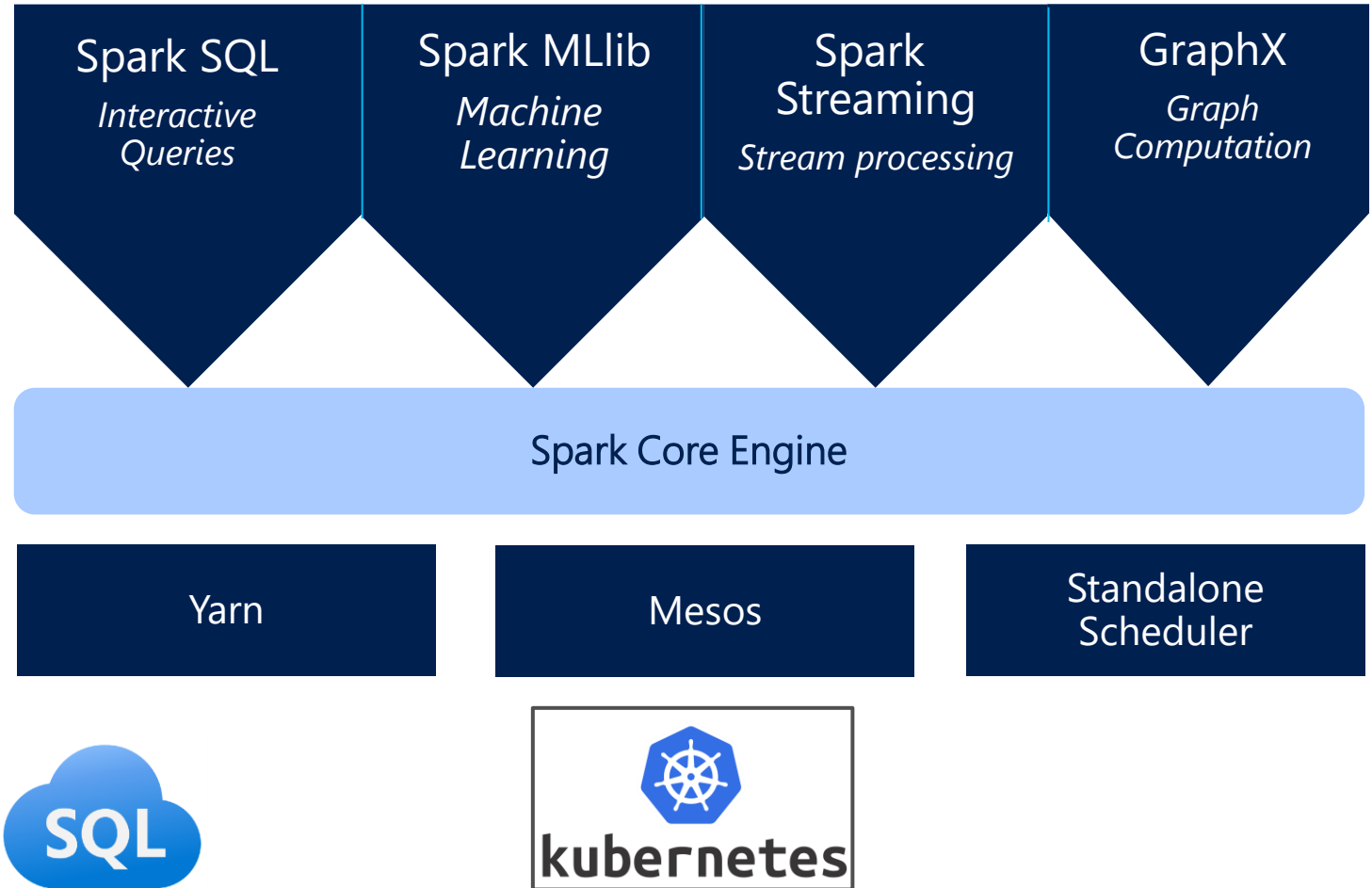


A P A C H E S P A R K

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



Microsoft Azure

databricks

Search data, notebooks, recents, and more...

CTRL + P

New

Workspace

Recents

Catalog

Workflows

Compute

Marketplace

SQL

SQL Editor

Queries

Dashboards

Genie

Alerts

Query History

SQL Warehouses

Data Engineering

Job Runs

Data Ingestion

Delta Live Tables

Machine Learning

Workspace

> Home

> Workspace

☆ Favorites

🗑️ Trash

Common Services

SQL Warehouse

ETL/ELT

Workspace > Users >

bca

⋮

Share

Create ▾

Name ↕	Type
📁 DatabricksWorkshop	Folder
📓 ADB_Dashboard	Notebook
📓 Databricks_AI_Assistant	Notebook
📓 DataDictionaryCreate	Notebook
📓 DemoNotebook01	Notebook
📄 demoquery1	Query

The Databricks Workspace User Interface



Clusters

Folders

Libraries

Notebooks

Security

Jobs/Workflows



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!