Master

databricks + APACHE Spark™

Lesson 20 - PySpark

Introduction

By

Bryan Cafferky

YouTube

Where Are We Going?

- ✓ What is PySpark?
- ✓ Why Python?
- ✓ Python in the Spark Ecosystem

What is PySpark?

- ✓ A library that lets you leverage Spark Services
- ✓ It is Cluster Aware
- ✓ Enables Python Developers Instant Productivity on Spark

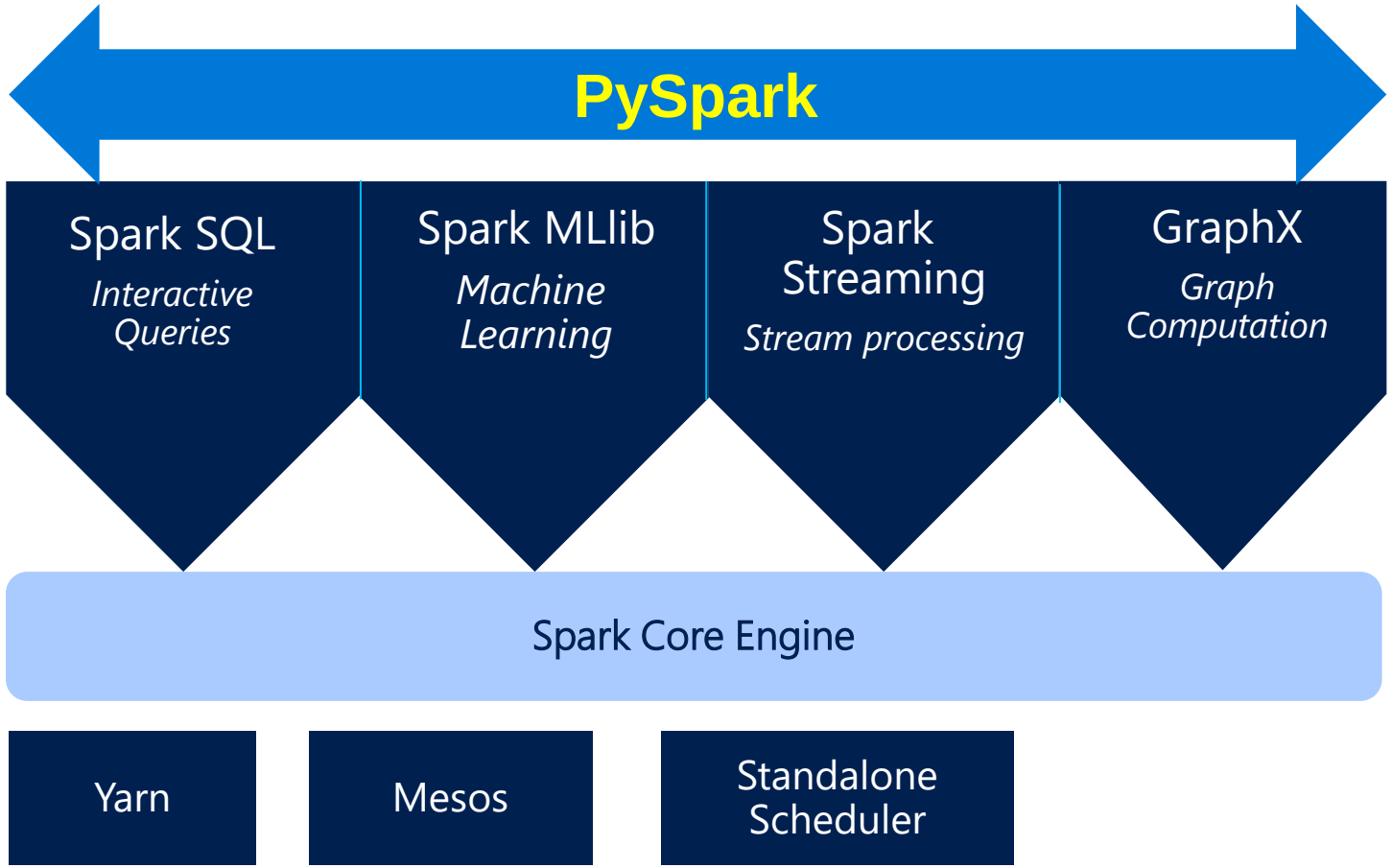
APACHE SPARK

Python is not Supported by Spark (directly)

Spark Unifies:

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

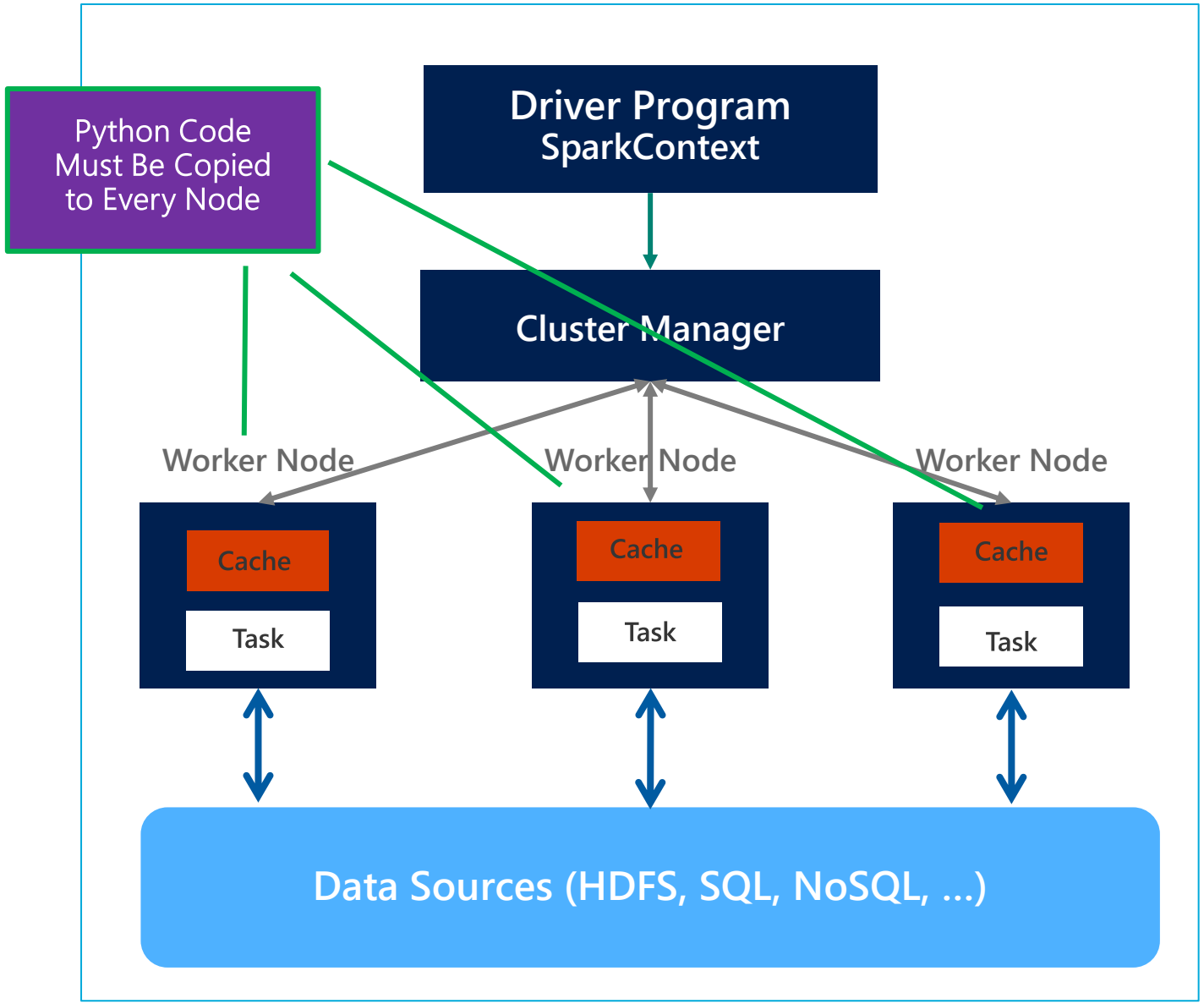
The PySpark Package Wraps the Spark APIs



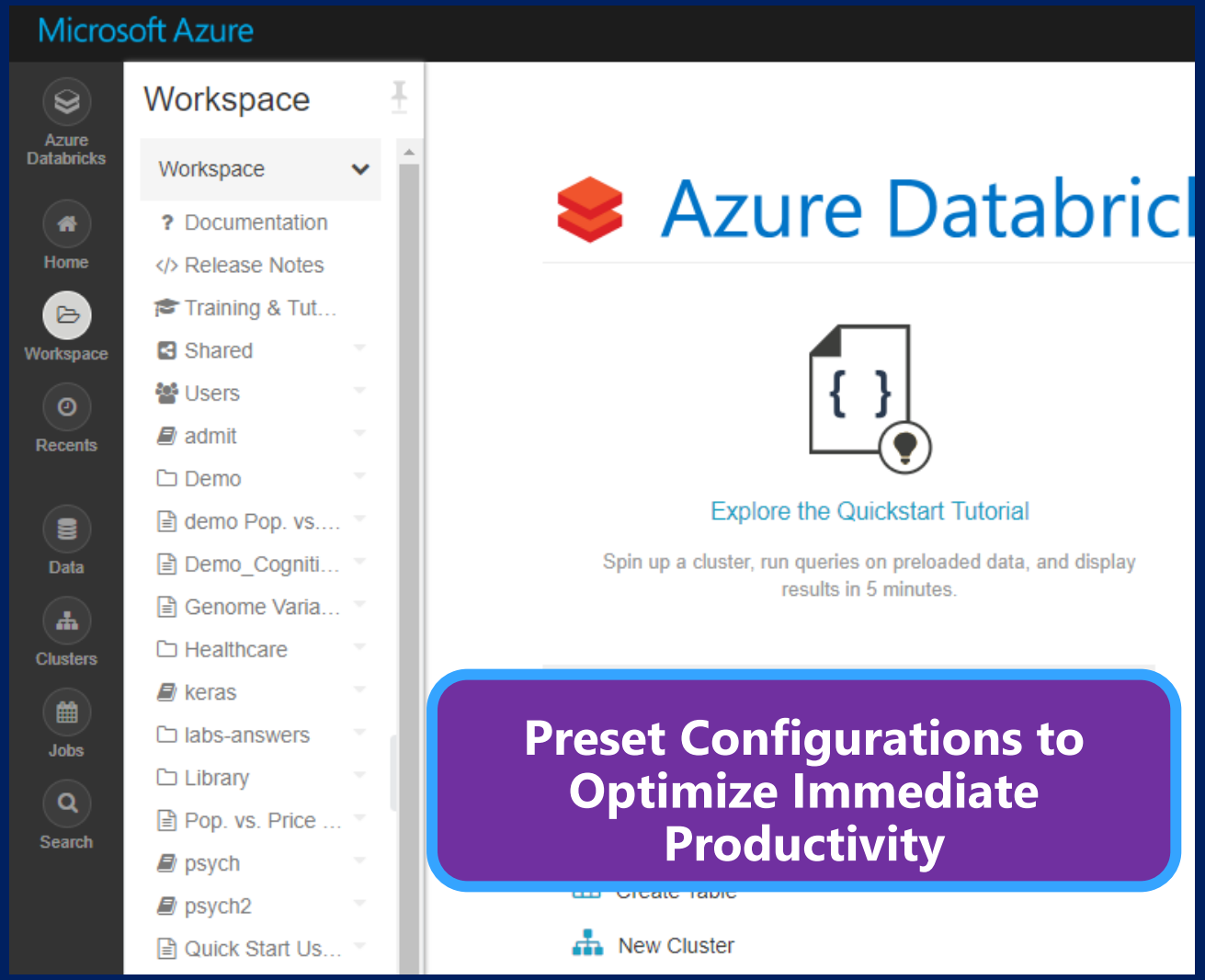
GENERAL SPARK CLUSTER ARCHITECTURE

- '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).

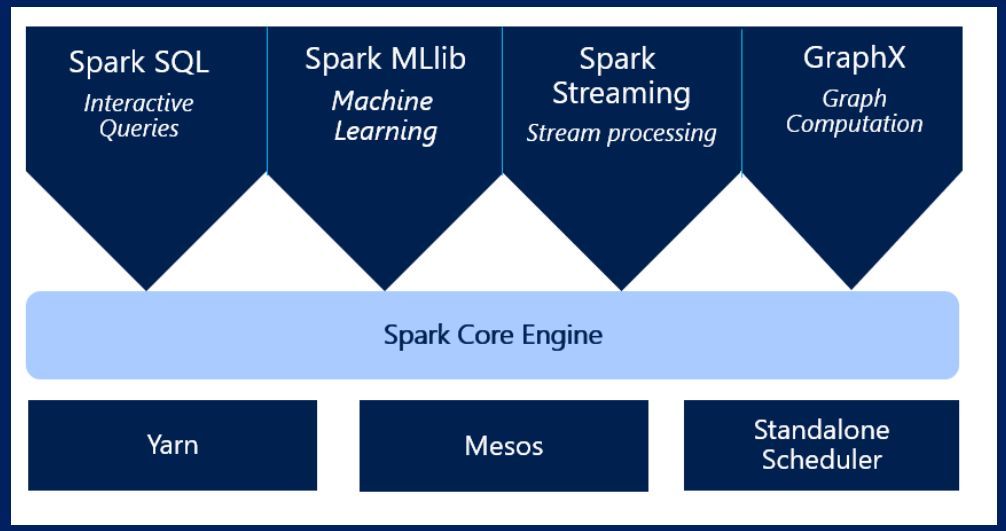
Nodes Run JVM



Databricks



Preset Configurations to Optimize Immediate Productivity



BYOT (Bring Your Own Tools)

Available on AWS, Azure, and GCP

Why Python?

- ✓ World's #1 Open Source Data Science, ML, and Analytics Language
- ✓ Vast libraries for data analysis and engineering, ML, Visualizations, and More
- ✓ Powerful and Extensible

Spark RDD to Dataframe – Win/Win

- Originally had to use Resilient Distributed Data
- Dataframe Support Added in 1.x
- Dataframes provide a Native Language Paradigm/Feel
- Easier to Read
- Performs much better!
- We will focus on the Dataframe API

What Does a Spark API Do?

- ✓ Load Data into a Spark Cluster
- ✓ Read and Manipulate Data in Spark
- ✓ Push Processing to the Spark Cluster Nodes
- ✓ Do Work on the Head Node
- ✓ Retain the Paradigm and Feel of the Calling Language

Apache Spark API

PySpark 2.3.2 documentation »

Table of Contents

Previous topic

Welcome to Spark Python API Docs!

Next topic

pyspark.sql module

This Page

Show Source

Quick search

pyspark package

Subpackages

- pyspark.sql module
- pyspark.streaming module
- pyspark.ml package
- pyspark.mllib package

Contents

PySpark is the Python API for Spark.

Public classes:

- `SparkContext`:
  Main entry point for Spark functionality.
- `RDD`:
  A Resilient Distributed Dataset (RDD), the basic abstraction in Spark.
- `Broadcast`:
  A broadcast variable that gets reused across tasks.
- `Accumulator`:
  An “add-only” shared variable that tasks can only add values to.
- `SparkConf`:
  For configuring Spark.
- `SparkFiles`:
  Access files shipped with jobs.
- `StorageLevel`:
  Finer-grained cache persistence levels.
- `TaskContext`:
  Information about the current running task, available on the workers and experimental.

https://spark.apache.org/docs/latest/api/python/pyspark.html#subpackages

Wrapping Up

- ✓ What is PySpark?
- ✓ Why Python?
- ✓ Python in the Spark Ecosystem

Thank You!