**Apache Spark:**

* It is an open-source, fast and general-purpose cluster-computing framework for big data processing and analytics.
* it provides APIs for programming in other languages like Python, R, and SQL.
* It has the ability to process data in-memory, which makes it significantly faster than its predecessor, Hadoop MapReduce, for certain use cases.
* Spark supports various data processing tasks, including batch processing, interactive queries, streaming analytics
* PySpark is the Python library for Apache Spark.

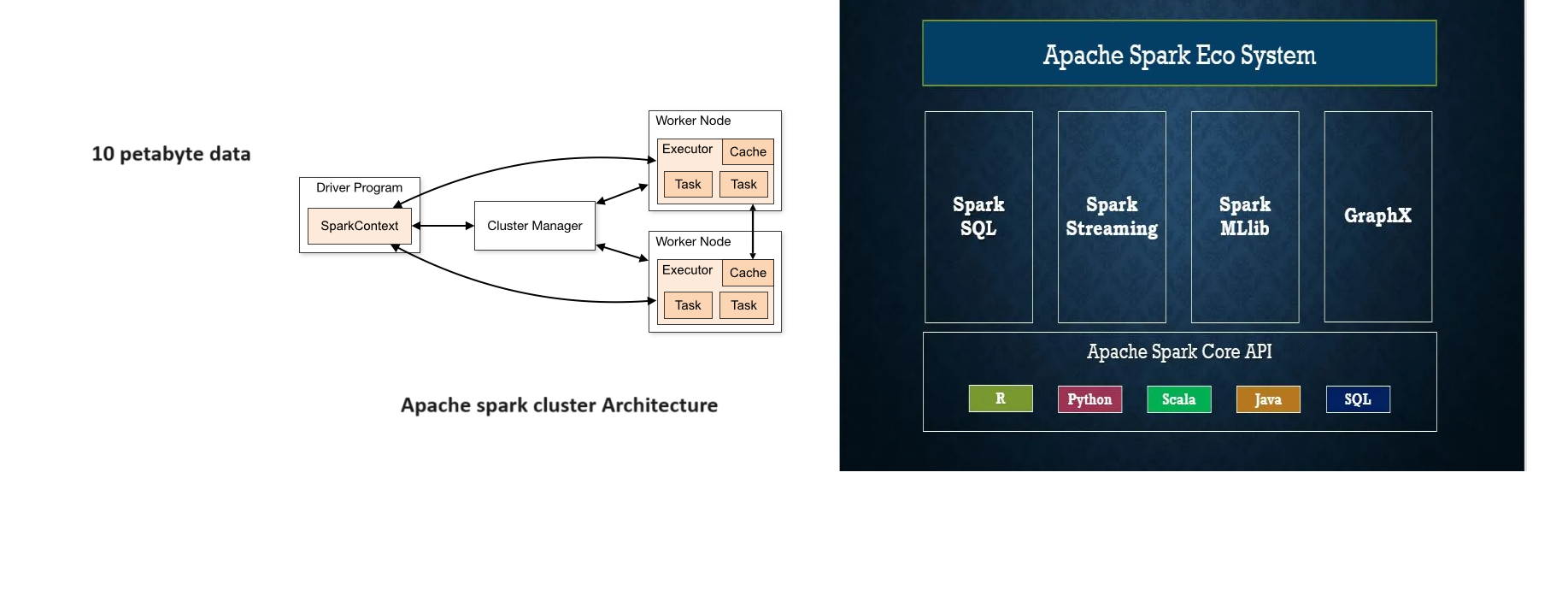
**Pyspark:**

* It is an Apache Spark library written in Python to run Python applications using Apache Spark capabilities.
* It is a Python API which is an analytical processing engine for large-scale powerful distributed data processing and machine learning applications.

**PySpark Features**

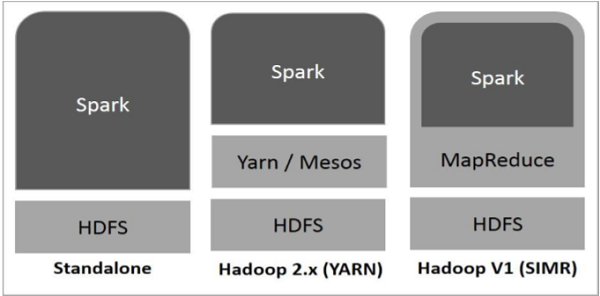
* In-memory computation
* Distributed processing using parallelize
* Can be used with many cluster managers (Spark, Yarn, Mesos e.t.c)
* Fault-tolerant
* Immutable

**PySpark Architecture:**



**Cluster Manager Types**

* Standalone – a simple cluster manager included with Spark that makes it easy to set up a cluster.
* Apache Mesos – Mesons is a Cluster manager that can also run Hadoop MapReduce and PySpark applications.
* Hadoop YARN – the resource manager in Hadoop 2. This is mostly used as a cluster manager.
* Kubernetes – an open-source system for automating deployment, scaling, and management of containerized applications.



**By doing these commands we can get the numbers from 0-100**

>>> from pyspark.sql import SparkSession

>>> Spark = SparkSession.builder.getOrCreate()

>>> myRange = Spark.range(100).toDF("number")

>>> myRange.show()

