Spark Assignment

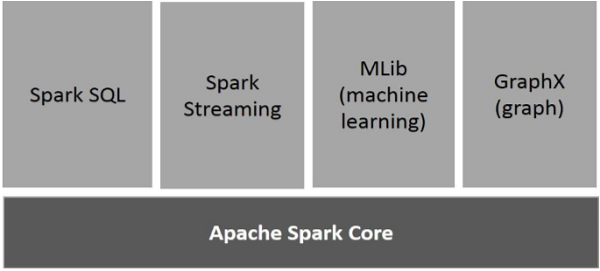
By Karthikeyan K R

Apache Spark:

* Apache Spark is general purpose cluster computing system.
* It provides high-level API in Java, Scala, Python, and R.
* The Spark can either run alone or on an existing cluster manager.

Components of Apache Spark:

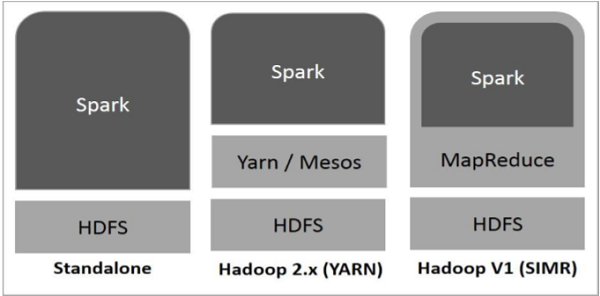
* Apache Spark- Spark Core
* Spark SQL
* Spark Streaming
* Spark MLlib
* Spark GraphX
* SparkR.



Features of Apache Spark:

* Speed
* Supports multiple languages
* Advanced Analytics

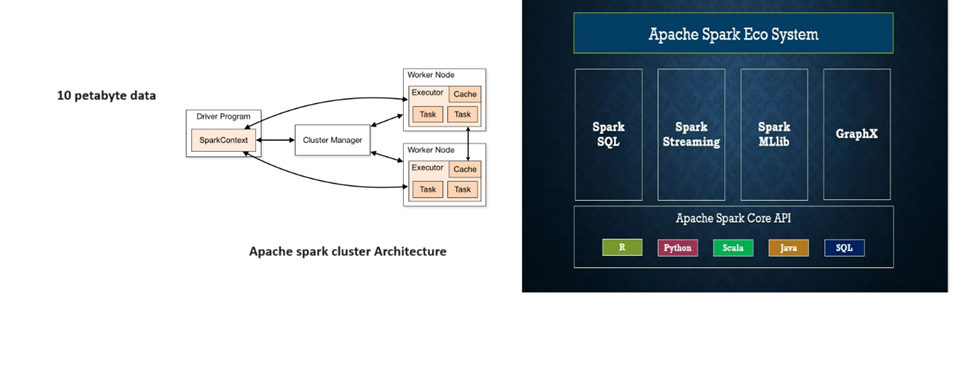
Spark Built on Hadoop:



There are three ways of Spark deployment as explained below.

* **Standalone** − Spark Standalone deployment means Spark occupies the place on top of HDFS (Hadoop Distributed File System) and space is allocated for HDFS, explicitly. Here, Spark and MapReduce will run side by side to cover all spark jobs on cluster.
* **Hadoop Yarn** − Hadoop Yarn deployment means, simply, spark runs on Yarn without any pre-installation or root access required. It helps to integrate Spark into Hadoop ecosystem or Hadoop stack. It allows other components to run on top of stack.
* **Spark in MapReduce (SIMR)** − Spark in MapReduce is used to launch spark job in addition to standalone deployment. With SIMR, user can start Spark and uses its shell without any administrative access.

Architecture of Apache Spark:



Execution of PySpark:

