Date: 22-12-23

**SPARK ASSIGNMENT**

**PYSPARK:**

**PySpark** is an Apache Spark library written in Python to run Python applications using Apache Spark capabilities. Using PySpark we can run applications parallelly on the distributed cluster (multiple nodes).

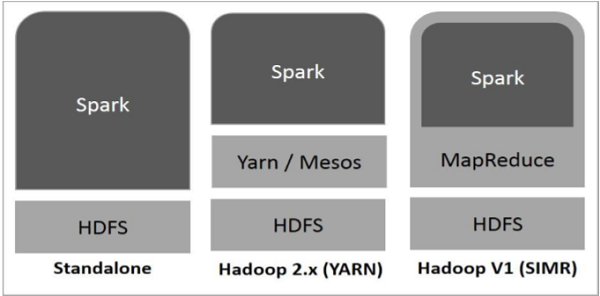
**APACHE SPARK:**

**Apache Spark** is general purpose cluster computing system. It provides high-level API in Java,**Scala**, Python, and **R**. Spark provide an optimized engine that supports general execution graph. It also has abundant high-level tools for structured data processing, machine learning, graph processing and streaming.

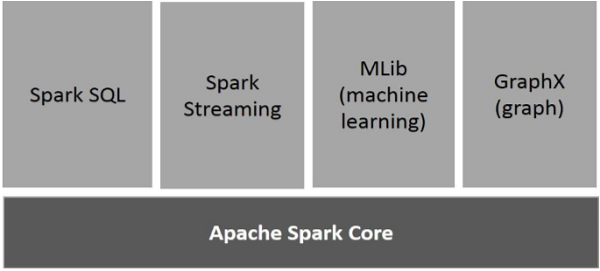
**The key features of Apache Spark Core are:**

* It is in charge of essential I/O functionalities.
* Significant in programming and observing the role of the **Spark cluster**.
* Task dispatching.
* Fault recovery.
* It overcomes the snag of**MapReduce** by using in-memory computation.

SPARK BUILT ON HADOOP:



Components of Spark



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
* Lazy evaluation
* Cache & persistence
* Inbuild-optimization when using DataFrames
* Supports ANSI SQL