**Name : Hemalatha Karumuri**

**Date : 22-12-23**

**Introduction to Spark**

\*Spark was introduced by Apache Software Foundation for speeding up the Hadoop computational computing software process.

\*It is not a modified version of Hadoop

**\***It uses Hadoop in two ways – one is storage and second is processing

**\***It has its own cluster management computation and it uses Hadoop for storage purpose only.

\*It is based on Hadoop MapReduce and it extends the MapReduce model to efficiently use it for more types of computations, which includes interactive queries and stream processing.

\*The main feature of Spark is its **in-memory cluster computing** that increases the processing speed of an application.

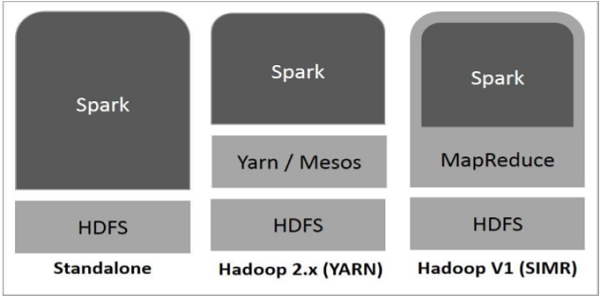
Features of Apache Spark:

\***Speed** − Spark helps to run an application in Hadoop cluster, up to 100 times faster in memory, and 10 times faster when running on disk

\***Supports multiple languages** − Spark provides built-in APIs in Java, Scala, or Python.

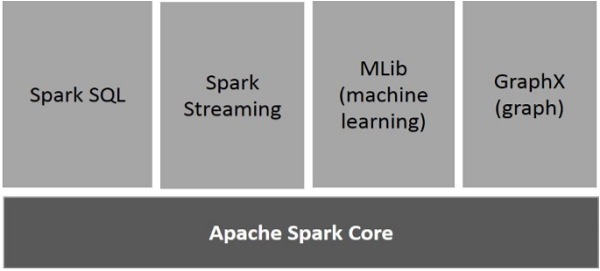
\***Advanced Analytics** − It supports SQL queries, Streaming data, Machine learning (ML), and Graph algorithms.

**How Spark can be built with Hadoop components?**



* 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.

**Components of Spark**



\*Apache Spark Core

It is the general execution engine for spark platform that all other functionality is built upon.

\*Spark SQL

Spark SQL is a component on top of Spark Core that introduces a new data abstraction called SchemaRDD, which provides support for structured and semi-structured data.

\*Spark Streaming

fast scheduling capability to perform streaming analytics.

MLlib (Machine Learning Library)

MLlib is a distributed machine learning framework above Spark because of the distributed memory-based Spark architecture

**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

**PySpark Architecture**

\*Apache Spark works in a master-slave architecture where the master is called the “Driver” and slaves are called “Workers”.

\*When you run a Spark application, Spark Driver creates a context that is an entry point to your application, and all operations (transformations and actions) are executed on worker nodes, and the resources are managed by Cluster Manager.