Mallu Sravana Sandhya Assessment-17 22/12/23

**Apache Spark**

* Apache Spark is an open-source unified analytics engine used for large-scale data processing
* It is designed to be fast, flexible, easy to process huge amount of data which is billions trillions of data
* It works 100 times faster than traditional machines
* It provides APIs and libraries for many programming languages like java,python, scala, R so it is known as multi language engine

**Apache Cluster Mangers**

****Standalone mode**** is a simple cluster manager incorporated with Spark. It makes it easy to setup a cluster that Spark itself manages and can run on linux, Windows, or Mac OSX. Often it is the simplest way to run Spark application in a clustered environment.

**Apache Mesos**

****Mesos**** handles the workload in distributed environment by dynamic resource sharing and isolation. It is healthful for deployment and management of applications in large-scale cluster environments. Apache Mesos clubs together the existing resource of the machines/nodes in a cluster.

**Hadoop Yarn:**

****YARN**** became the sub-project of [Hadoop](https://data-flair.training/blogs/important-big-data-terminologies-and-hadoop-concepts-you-must-know/) in the year 2012. It is also known as [MapReduce](https://data-flair.training/blogs/hadoop-mapreduce-introduction-tutorial-comprehensive-guide/) **2.0**.****YARN divides the functionality of resource manager and job scheduling into different daemons.

**Kubernetes :**

It is an open-source system for automating deployment, scaling, and management of containerized applications.

**Apache Spark Components:**

*Apache Spark COre:*

It is underlying general execution engine for spark.It provides inmemory computing and reference data sets from external storage systems

*Apache SQL:*

It is a top component in Spark core which introduces a new data absraction called Schema RDD which supports Structured and semi structured data

*Spark Streamming:*

It leverages fast scheduling capability for performing Streaming analytics

*MLlib (Machine Learning library)*

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

*Graph library*

It provides an API for expressing graph computation that can model the user-defined graphs by using Pregel abstraction

**pyspark**

Spark along with python is known as PySpark, which is used by data anaylist and data Scientists

Main features of pyspark are

* In-memory computation
* Distributed processing using parallelize
* Fault-tolerant
* Immutable
* Cache & persistence
* Inbuild-optimization when using DataFrames
* Supports ANSI SQL
* Lazy evaluation
* Can be used with many cluster managers (Spark, Yarn, Mesos e.t.c)

**Advantages**

* 100 times faster
* Multiple programming language is used
* Process data from hadoop,AWS, huge file systems
* Used to process real time data using streaming and kafka
* Has great benefits for data ingestion pipelines
* Pyspark has machine learning and graph libraries
* It allows to process data efficiently in a distributed fashion.

**PySpark RDD**

**Resilient Distributed Datasets**

Resilient Distributed Datasets (RDD) is a fundamental data structure of Spark. It is an immutable distributed collection of objects. Each dataset in RDD is divided into logical partitions, which may be computed on different nodes of the cluster.

RDD operations

On PySpark RDD, you can perform two kinds of operations.

**RDD transformations –** Transformations are lazy operations. When you run a transformation(for example update), instead of updating a current RDD, these operations return another RDD.

**RDD actions** – operations that trigger computation and return RDD values to the driver.

**Pyspark Architecture**

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

Worker Node

Executer cache

task

Cluster Manager

Worker Node

Executer cache task

Spark Context

Driver program