211720118030

SPARK FRAMEWORK

HOMEWORK 4

Q 1.Compare Hadoop and Spark.

>Hadoop is designed to handle batch data processing whereas spark is used to handle real-time data efficiently

>Hadoop is a high latency computing framework but spark is a low latency computing framework .

Q 2.What is Apache Spark?

Apache Spark is an open-source, distributed processing system used for big data workloads. It is designed as an interface for large-scale processing. Apache Spark provides the framework of the ETL(Extract Transform Load)

Q 3. Explain the key features of Apache Spark

>Apache Spark is free and open-source system

>It has high fault-tolerance

>It is dynamic in nature.

Q 4.What are the languages supported by Apache Spark and which is the mostpopular one?

>Scala

>Java

>Python

>R

were languages supported by Apache Spark and scala is the most popularly used for Apache spark.

Q 5. What are benefits of Spark over MapReduce?

> Spark processes and retains data in memory for subsequent steps,

whereas MapReduce processes data on disk

> Spark's data processing speeds are up to 100x faster than MapReduce

Q 6.Explain the concept of Resilient Distributed Dataset (RDD).

Resilient Distributed Dataset(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.

Q 7. How do we create RDDs in Spark?

There are two ways to create RDDs,

>RDDs are generally created by parallelized collection.

>created by reading from a text file

>It can be created from dataframe and datasets

Q 8.What is Executor Memory in a Spark application?

Every spark application will have one executor on each worker node. The executor memory is basically a measure on how much memory of the worker node will the application utilize.

Q 9.What do you understand by Transformations in Spark?

> Transformation in spark can create a new dataset from an existing one.

> It takes RDD as input and produces one or more RDD as output .

> The inputs of RDD cannot be changed since RDD are immutable in nature.

Q 10.Define Actions in Spark.

> Action in spark returns a value to the driver program after running a computation on the dataset.

> Action is the spark operations that return raw values.