# Spark Framework

## Homework

Sujatha.k

211720104148

3rd year CSE C

1. Compare Hadoop and Spark.

| HADOOP | SPARK |
| --- | --- |
| *mostly associated with MapReduce | *A general-purpose processing |
| Batch applications,through it supports | engine for batch jobs,machine |
| Other types of workloads. | Learning and other uses. |
| *It was primarily developed to store | *Spark is primarily developed to |
| and processing big data. | Offer faster processing. |

2. What is Apache Spark?

Apache spark is a multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters.

3.Explain the key features of Apache Spark.

*Fault tolerance.

*Lazy Evaluation.

*Real-Time Stream Processing.

*Speed.

*Reusability.

*Advanced Analytics.

*In Memory Computing.

4.What are the languages supported by Apache Spark and which is the most popular one?

* Scala, Java, python, R

5.What are benefits of Spark over MapReduce?

*Spark is relatively easier to program and requires a lot less of actual coding than MapReduce.

* Spark uses a data abstraction, RDD, to make the features more productive, whereas MapReduce does not have any concept.

*Spark executes batch processing jobs almost 10X to 100X times faster than MapReduce.

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

RDD was the primary user-facing API in Spark since its inception. At the core, an RDD is an immutable distributed collection of elements of your data, partitioned across nodes in your cluster that can be operated in parallel with a low-level API that offers transformations and actions.

7.How do we create RDDs in Spark?

There are three ways to create an RDD in Spark.

*Parallelizing already existing collection in driver program.

*Referencing a dataset in an external storage system (e.g. HDFS, Hbase, shared file system).

*Creating RDD from already existing RDDs.

8.What is Executor Memory in a Spark application?

An executor is a process that is launched for a Spark application on a worker node. Each executor memory is the sum of yarn overhead memory and JVM heap memory

9.What do you understand by Transformations in Spark?

Spark Transformation is a function that produces new RDD from the existing RDDs.. It takes RDD as input and produces one or more RDD as output. Each time it creates new RDD when we apply any transformation. Thus, the so input RDDs, cannot be changed since RDD are immutable in nature.

10.Define Actions in Spark.

Actions are RDD's operation, that value returns back to the spar driver programs, which kick off a job to execute on a cluster