Name: K. Praveen Kumar

Pyspark Assignment-2

**Apache Spark:**

* It is an open-source data processing engine to store and process data in real-time across various clusters of computers using single programming constructs.
* Apache spark is a lightning-fast cluster computing technology, designed for fast computation.
* It is based on Hadoop Map reduce.

**Features of Pyspark:**

* Fast Processing
* In-memory computing.
* Flexible
* Fault Tolerance
* Better Analytics

**Resilient Distributed Datasets (RDD):**

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 cluster.

RDD’s can contain any type of python, java or scala objects, including user-defined classes.

Formally, an RDD is a read-only, partitioned collection of records.

**Two ways to create RDD:**

**1.** Parallelizing an existing collection in your driver program.

**2.** Referencing a dataset in an external storage system, such as a shared file system, HDFS, HBase or any data source offering a Hadoop input format.

**Components:**

* **Spark core**
* **Spark SQL**
* **Spark streaming**
* **Spark MLlib**
* **Graphx**

1. **Spark Core:** It is the base engine for large-scale parallel and distributed data processing.

**Responsible for:**

1.Memory management

2.Fault Recovery

3.Scheduling, distributing and monitoring jobs.

4.Interacting with Storage Systems.

This is embedded with a special collection called RDD.

There are 2 operations performed on RDD:

1.Transformation: Function that produces new RDD from the existing RDD’s.

2.Action: In Transformation, RDD’s are created from each other. But when we need to work with the actual dataset, then at that point we use Action.

**2. Spark SQL:** It is a distributed framework which is used for structured and semi-structured data processing.

**Features:**

1.Cost based Optimizer

2.Mid query fault-tolerance

3.Full compatibility with existing hive data

**3. Spark Streaming:** It is an add-on to core spark API that allows developers to perform batch processing and real-time streaming of data.

**3-Phases of spark streaming:**

**1.Gathering** data from basic sources which are available in streaming context API like file systems, Socket Connections and from the advanced sources like Kafka, Flume, Kinesis etc…

**2.** **Processing:** Data is processed by using different algorithms.

**3.** **Data Storage:** Processed data is pushed out to file systems, Databases and live dashboards.

**4.Spark MLlib:** It is a scalable machine learning library that discusses both high-quality algorithm and high speed.

The motive is to make machine learning scalable and easy.

It contains ML algorithms like Clustering, Regression, Classification.

**Spark Architecture:**

It uses a Master-slave Architecture that consists of a driver, that runs on a master node and multiple executors which run across the worker nodes in cluster.

Worker Node

Executor cache

Task Task

Driver Program

Cluster Manager

Spark Context

Worker Node

Executor cache

Task Task

**Cluster Manager Types:**

1.Standalone

2.Apache Mesos

3.Hadoop yarn

4.Kubernetes

**Pyspark Modules and Packages:**

* PySpark RDD (pyspark.RDD)
* PySpark DataFrame and SQL (pyspark.sql)
* PySpark Streaming (pyspark.streaming)
* PySpark MLib (pyspark.ml, pyspark.mllib)
* PySpark GraphFrames (GraphFrames)