UNIT-V

1 Mark Questions

1. What is Apache Spark?

**Apache Spark is a powerful open source engine. Since, it offers real-**

**time stream processing, interactive processing, graph processing, in-**

**memory processing as well as batch processing.**

1. What is the use of Standalone Scheduler?

Is at cluster manger level and deal with resuorces for multiple spark applicaptions.

1. What is Machine Learning?

Machine learning is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead.

1. Define RDD?

• **Resilient:** Fault tolerant and is capable of rebuilding data on failure

• **Distributed:** Distributed data among the multiple nodes in a cluster

• **Dataset: Collection** of partitioned data with values

1. What are the basic functionalities of Spark?

Task scheduling ,memory management ,fault recovery ,intercting storage systems

1. Define Spark SQL?

Spark SQL is Spark’s package for working with structured data. It allows querying data via SQL as well as the Apache Hive variant of SQL—called the Hive Query Language (HQL)—and it supports many sources of data, including Hive tables, Parquet, and JSON.

1. What is the use of Spark Streaming?

Spark Streaming is a Spark component that enables processing of live streams of data.Eg. logfiles generated by production web servers.

1. What is meant by degree of fault tolerance?

To what extent does the system in work properly without any failure.

1. Define Mlib?

Spark comes with a library containing common machine learning (ML) functionality, called MLlib.

1. What is GraphX?

GraphX is a library for manipulating graphs (e.g., a social network’s friend graph) and performing graph-parallel computations.

1. Explain the difference between Spark SQL and HIVE?
2. What are the variety of Spark Cluster Managers?

**Spark** Standalone. Apache Mesos. **Hadoop YARN**.

1. What are the components of Distributed Spark?

Apache **Spark** Ecosystem which empower to Apache **Spark**- **Spark** Core, **Spark** SQL, **Spark** Streaming, **Spark** MLlib, **Spark** GraphX, and SparkR

1. Define Spark Context?

**SparkContext** (aka **Spark context**) is the entry point to the services of Apache **Spark** (execution engine) and so the heart of a **Spark** application.

1. Explain the role of Worker Node in Spark?
2. What are the operations performed on RDDs?

Transormations, actions and shuffles

1. What is the use of map() and filter().
2. What are the actions performed on RDD?

countByKeys,collectAsMap,lookup

1. What is meant by Lazy Evaluation?

Spark will not begin to execute until it sees an action.

1. What are the languages supported by Apache Spark for developing big data applications? Python scala
2. Explain about the major libraries that constitute the Spark Ecosystem?
3. What are the common mistakes developers make when running Spark applications?

Deciding on the number of executors, cores and memory. ...

**Mistake** #2. **Application** failureWhy?No **Spark** shuffle block can be greater than 2GB. ...

**Mistake** #3. Slow jobs on Join/ShuffleYour dataset takes 20 seconds to **run** over a with a map job, but takes 4 hours when joined or shuffled

1. What are the various data sources available in SparkSQL?

Generic Load/Save Functions.

Parquet Files.

ORC Files.

JSON Files.

Hive Tables.

JDBC To Other Databases.

Avro Files.

1. How Spark uses Hadoop?

Speed multiple lanuguages and advanced analytics

1. What do you understand by Pair RDD?
2. Explain about the different types of transformations?

reduceBykey,groupbykey

Subtractbykey,join

1. Explain about the popular use cases of Apache Spark?

Easy to use ,easy of deployment real time streaming,advanced analytics

1. How can you achieve high availability in Apache Spark?

By using Zookeeper stand by masters.

1. What do you understand by SchemaRDD?

**SchemaRDDs are** composed of Row objects, along with a schema that describes the data types of each column in the row. A **SchemaRDD** is similar to a table in a traditional relational database.

1. Is it necessary to install spark on all the nodes of a YARN cluster  while running Apache Spark on YARN ?

No, it is not **necessary to install Spark on all** the 3 **nodes**. Since **spark runs** on top of **Yarn**, it utilizes **yarn** for the execution of its commands over the **cluster's nodes**. So, you just have to **install Spark** on one **node**

Essay Questions.

1. a) What is spark? Write its features.
2. Write about RDDS and Implement wordcount program using spark.
3. What is RDD. Write about the transformations and actions on rdds with example.
4. What is pair RDD. Write about the transformations and actions on pair rdds with example.
5. Explain about Spark Stack.
6. What is spark and write its features.
7. Implement wordcount program in Spark.
8. Write about Lineage Graph.
9. Distinguish RDDS and Paired RDDS.
10. Explain about Accumulators with example.
11. Explain about the working of Broadcast variables with example.