Apache Spark Interview Questions

Spark has become popular among data scientists and big data enthusiasts. If you are looking for the best collection of Apache Spark Interview Questions for your data analyst, big data or machine learning job, you have come to the right place.

In this Spark Tutorial, we shall go through some of the frequently asked Spark Interview Questions.

Entry Level Spark Interview Questions

What is Apache Spark?

Apache Spark is an Open Source Project from the Apache Software Foundation. Apache Spark is a data processing engine and is being used in data processing and data analytics. It has inbuilt libraries for Machine Learning, Graph Processing, and SQL Querying. Spark is horizontally scalable and is very efficient in terms of speed when compared to big data giant Hadoop.

When should you choose Apache Spark?

When the application needs to scale. When the application needs both batch and real-time processing of records. When the application needs to connect to multiple databases like Apache Cassandra, Apache Mahout, Apache HBase, SQL databases, etc. When the application should be able to query structured datasets cumulatively present across different database platforms.

Which builtin libraries does Spark have?

Spark has four builtin libraries. They are :

- SQL and DataFrames
- Spark Streaming
- MLlib (Machine Learning Library)
- GraphX

How fast is Apache Spark when compared to Hadoop ? Give us an example.

Apache Spark is about 100 times faster than Apache Hadoop. For a typical Logistic Regression problem, if Hadoop takes 110ms to complete, then Spark would take around 1ms.

Why is Spark faster than Hadoop?

Spark is so fast because it uses a state-of-the-art DAG scheduler, a query optimizer, and a physical execution engine.

Which programming languages could be used for Spark Application Development?

One can use following programming languages.

- Java
- Scala
- Python
- R
- SQL

On which platforms can Spark run?

When Spark is run in stand-alone cluster mode, it can run on :

- Hadoop YARN
- EC2
- Mesos
- Kubernetes

Which data sources can Spark access?

Spark can access data from hundreds of sources. Some of them are :

- HDFC
- Apache Cassandra
- Apache HBase
- Apache Hive

Can structured data be queried in Spark? If so, how?

Yes, structured data can be queried using :

- SQL
- DataFrame API

What is Spark MLlib?

MLlib is Spark's scalable Machine Learning inbuilt library. The library contains many machine learning algorithms and utilities to transform the data and extract useful information or inference from the data.

How is MLlib scalable?

Spark's DataFrames API realizes scalability.

What kinds of machine learning use cases does MLlib solve?

MLlib contains common learning algorithms that can solve problems like :

- Clustering
- Classification
- Regression
- Recommendation
- Topic Modelling
- Frequent itemsets
- Association rules
- Sequential pattern mining

What is Spark Context?

SparkContext instance sets up internal services for a Spark Application and also establishes a connection to a Spark execution environment. Spark Context should be created by Spark Driver Application.

What is an RDD?

RDD, short for Resilient Distributed Datasets, is a collection of elements. RDD is fault tolerant and can be operated on in parallel. RDD provides the abstraction for distributed computing across nodes in Spark Cluster.

What are RDD transformations?

Transformations are operations which create a new RDD from an existing RDD.

Some of the RDD transformations are :

- map
- filter
- union
- intersection

What are RDD actions?

Actions are operations which consume or run operations on an RDD and produce an output value.

Some of the RDD actions are :

- reduce
- collect
- count
- countByKey

What do you mean by ‘RDD Transformations are lazy’?

Transformations on an RDD are not actually executed until an action is encountered. Hence, RDD Transformations are called lazy.

What do you know about RDD Persistence?

Persistence means Caching. When an RDD is persisted, each node in the cluster stores the partitions (of the RDD) in memory (RAM). When there are multiple transformations or actions on an RDD, persistence helps to cut down the latency by the time required to load the data from file storage to memory.

What is the difference between cache() and persist() for an RDD?

cache() uses default storage level, i.e., MEMORY_ONLY.

persist() can be provided with any of the possible storage levels.

What do you mean by the default storage level: MEMORY_ONLY?

Default Storage Level – MEMORY_ONLY mean store RDD as deserialized Java objects in the JVM. If the RDD does not fit in memory, some partitions will not be cached and will be recomputed on the fly each time they’re needed.

What is Sliding Window Operation?

Medium Level Spark Interview Questions

How does Spark Context in Spark Application pick the value for Spark Master?

That can be done in two ways.

1. Create a new SparkConf object and set the master using its setMaster() method. This Spark Configuration object is passed as an argument while creating the new Spark Context.

```
SparkConf conf = new SparkConf().setAppName("JavaKMeansExample")
        .setMaster("local[2]")
        .set("spark.executor.memory","3g")
        .set("spark.driver.memory", "3g");

JavaSparkContext jsc = new JavaSparkContext(conf);
```

2. **<apache-installation-directory>/conf/spark-env.sh** file, located locally on the machine, contains information regarding Spark Environment configuration. Spark Master is one the parameters that could be provided in the configuration file.

How do you configure Spark Application?

Spark Application could be configured using properties that could be set directly on a SparkConf object that is passed during SparkContext initialization.

Following are the properties that could be configured for a Spark Application.

- Spark Application Name
- Number of Spark Driver Cores
- Spark Driver's Maximum Result Size
- Spark Driver's Memory
- Spark Executors' Memory
- Spark Extra Listeners
- Spark Local Directory
- Log Spark Configuration
- Spark Master
- Deploy Mode of Spark Driver
- Log Application Information
- Spark Driver Supervise Action

Reference : Configure Spark Application

What is the use of Spark Environment Parameters? How do you configure those?

Spark Environment Parameters affect the behavior, working and memory usage of nodes in a cluster.

These parameters could be configured using the local config file spark-env.sh located at **<apache-installation-directory>/conf/spark-env.sh**.

Reference: Configure Spark Ecosystem

How do you establish a connection from Apache Mesos to Apache Spark?

A connection to Mesos could be established in two ways.

- Spark Driver program has to be configured to establish a connection to Mesos. Also, Spark binaries location should be accessible to the Apache Mesos program.
- The other way to connect to Mesos is that installing Spark in the same location as that of Mesos, and configure the property, `spark.mesos.executor.home` to point to the location of Mesos installation.

How do you minimize data transfers between nodes in a cluster?

Minimizing shuffles can minimize the data transfers between nodes, and thus making the process fast. There are some practices that can reduce the usage of shuffling operations. Use broadcast variables to join small and large RDDs; and accumulators to update the variable's values.

To run Spark Applications, should we install Spark on all the nodes of a YARN cluster?

Spark programs can be executed on top of YARN. So, there is no need to install Spark on the nodes a YARN cluster, to run spark applications.

To run Spark Applications, should we install Spark on all the nodes of a Mesos cluster?

Spark programs can be executed on top of Mesos. So, there is no need to install Spark on the nodes of a Mesos cluster, to run spark applications.

What is the usage of GraphX module in Spark?

GraphX is a graph processing library. It can be used to build and transform interactive graphs. Many algorithms are available with GraphX library. PageRank is one of them.

How does Spark handle distributed processing?

Spark provides an abstraction to the distributed processing through Spark RDD API. A general user does not need to worry about how data is processed in a distributed cluster. There are some exceptions though. When you optimize an application for performance, you should understand about operations and actions which require the data transfer between nodes.

Advanced Level Spark Interview Questions

Learn Apache Spark

- Apache Spark Tutorial
- Install Spark on Ubuntu
- Install Spark on Mac OS
- Scala Spark Shell - Example
- Python Spark Shell - PySpark
- Setup Java Project with Spark
- Spark Scala Application - WordCount Example
- Spark Python Application
- Spark DAG & Physical Execution Plan
- Setup Spark Cluster
- Configure Spark Ecosystem
- Configure Spark Application
- Spark Cluster Managers

Spark RDD

- Spark RDD
- Spark RDD - Print Contents of RDD
- Spark RDD - foreach
- Spark RDD - Create RDD
- Spark Parallelize
- Spark RDD - Read Text File to RDD

- Spark RDD - Read Multiple Text Files to Single RDD
- Spark RDD - Read JSON File to RDD
- Spark RDD - Containing Custom Class Objects
- Spark RDD - Map
- Spark RDD - FlatMap
- Spark RDD - Filter
- Spark RDD - Distinct
- Spark RDD - Reduce

Spark Dataseet

- Spark - Read JSON file to Dataset
- Spark - Write Dataset to JSON file
- Spark - Add new Column to Dataset
- Spark - Concatenate Datasets

Spark MLlib (Machine Learning Library)

- Spark MLlib Tutorial
- KMeans Clustering & Classification
- Decision Tree Classification
- Random Forest Classification
- Naive Bayes Classification
- Logistic Regression Classification
- Topic Modelling

Spark SQL

- Spark SQL Tutorial
- Spark SQL - Load JSON file and execute SQL Query

Spark Others

- **Spark Interview Questions**