By OnlineInterviewQuestions.com

Apache Spark Interview Questions

Reportedly a fact has been introduced that Apache Spark's market share is been around 4.9 to 5 %. So there is a lot in the market for everyone to gain. Many international firms have opened the doors for research and development with Apache Spark. With the absolute preparation through **Apache Spark interview questions** and practicing advanced technology, you can definitely achieve your dream job as an Apache Spark developer.

Apart from the fundamental proficiency, a candidate might also be asked about skills in Hadoop, Hive, Sqoop, and many more. You will get a perfect combination of Apache spark interview questions for fresher as well as experienced candidates here. In the most specific segment like Spark SQL programming, there are enough job opportunities. If you have given a thought to it then keep yourself assured with your skills and below listed **Apache Spark interview questions.**

Q1. What is Apache Spark?

Apache Spark is basically a processing framework which is extremely fast and convenient to use. With an advanced execution engine supporting it offers the cyclic data flow and in-memory computation. Apache spark can also run on Hadoop, cloud or standalone. It is capable to access the diverse data including Cassandra, HDFS, HBase.

Q2. Can you mention some features of spark?

On a general note, the most essential features of Apache Spark are-

- It allows the integration with Hadoop and files including HDFS.
- It also has an interactive language shell.
- It consists of RDD (Resilient Distributed Datasheets) it can be cached across multiple computing nodes in a cluster.
- It also supports analytical tools which are utilized for interactive analysis and real-time analysis.

Q3. Can you define RDD?

The acronym for Resale in Distributed Datasheets is RDD. It is a fault-tolerant collection for all of the elements that run parallel. The sorted data in RDD is immutable and primarily of two types –

- Parallelized collections
- Hadoop datasets

Q4. Do you know the comparative differences between Apache Spark and Hadoop?

Yes there are several segments on which they can be differentiated. Few of them are-

Feature	**Apache spark**	**Hadoop**
Speed	It is almost 100 times faster than Hadoop	It has moderate speed
Processing	Offers real time and batch processing functionality	It offers batch processing only
Difficulty	It has high level modules hence it is easy	It is tough to learn
Recovery	It allows the partition recovery	MapReduce
Interactivity	It has interactive modes	Other than Pig and Hive, it has no interactive mode

Q5. Name the languages which are supported by Apache Spark and which one is most popular?

Apache Spark supports the languages Java, Python, Scala and R. among them Scala and Python have interactive shares for Apache Spark and Scala shell can be easily accessed through the ./bin/spark-shell and Python can be accessed through ./bin/pyspark. Among them, Scala is the most popular because Apache Spark is written in Scala.

Q6. List few benefits of Apache spark over map reduce?

The benefits are –

- With the in-memory processing, Spark implements it around 100 times faster than the head of MapReduce.
- Spark provides inbuilt libraries for most of the multidimensional task as compared to map reduce.
- Spark is independent of the disk storage on the other hand Hadoop is highly dependent.
- Spark can perform multiple computations within the same datasheet.

Q7. What do you understand about yarn?

The YARN is a key feature in Spark which provides a central resource management for most of the operational deliveries across a cluster. It is also a container manager like Mesos. Spark can easily run on YARN which eventually emphasizes a binary distribution of Apache Spark built on its support.

Q8. If map reduce is inferior to Spark then is there any benefit of learning it?

Apache Spark is far better than MapReduce but still learning MapReduce is essential. MapReduce is a paradigm which is even used by Spark as big data tools. When the data is large and grows bigger, in that case, MapReduce is much relevant. Data tools like pig and hive convert their message queries into MapReduce in order to optimize them properly.

Q9. How can we create RDDs in Apache spark?

In order to create RDD there are basically two methods –

- In a driver program parallelizes a collection. This will use the sparkcontext's “Parallelize”.
- From the external storage, an external data sheet can be loaded into the file system.

Q10. What do you understand by the partitions in spark?

Partitions are done in order to simplify the data as they are the logical distribution of entire data. It is similar to the split in MapReduce. In order to enhance the processing speed, this logical distribution is carried out. Each and every association in Apache Spark is a partitioned RDD.

Q11. Name the operations supported by RDD?

As the major logical data units in Apache Spark, RDD possesses a distributed collection of data. It is a read-only data structure and you cannot change the original format but it can always be transformed into a different form with the changes. The two operations which are supported by RDD are -

- Transformation - It creates a new RDD from the former one. They are executed only on demand.
- Actions - The final outcomes of the RDD computations are returned by actions.

Q12. Can you explain about the cluster manager of Apache spark?

There are three different cluster manager in Spark which are as-

- YARN
- Apache Mesos - Spark along with several other applications can easily be scheduled and run over it. Its priority is to scale down the allocations between several commands in order to provide interfaces when several users run their shells.
- Standalone deployments - For the easy setup and convenience, for the new deployments.

Q13. What are accumulators in Apache spark?

The write only variables which are initially executed once and send to the workers are accumulators. On the basis of the logic written, these workers will be updated and sent back to the driver which will process it on the basis of logic. A driver has the potential to exercise accumulator’s value.

Q14. What is Spark SQL?

Spark SQL is basically a module which is formulated to provide structured data processing. The advantage of SQL message queries running on the datasheets can be taken from it.

Q15. What is a Hive on Apache spark?

For the Apache Spark processing, Hive contains the support. High execution is formulated to spark.

Q16. Name few companies that are the uses of Apache spark?

The companies that are using Apache Spark with the production are Pinterest, Shopify, Open Table, ComViva.

Please Visit OnlineInterviewquestions.com to download more pdfs