# Top 50 Apache Spark Interview Questions and Answers

Following are frequently asked Apache Spark questions for freshers as well as experienced Data Science professionals.

**1) What is Apache Spark?**

Apache Spark is easy to use and flexible data processing framework. Spark can round on Hadoop, standalone, or in the cloud. It is capable of assessing diverse data source, which includes HDFS, Cassandra, and others.

**2) Explain Dsstream with reference to Apache Spark**

Dstream is a sequence of resilient distributed database which represent a stream of data. You can create Dstream from various source like HDFS, Apache Flume, Apache Kafka, etc.

**3) Name three data source available in SparkSQL**

There data source available in SparkSQL are:

- JSON Datasets
- Hive tables
- Parquet file

**4) Name some internal daemons used in spark?**

Important daemon used in spark are Blockmanager, Memestore, DAGscheduler, Driver, Worker, Executor, Tasks,etc.

**5) Define the term 'Sparse Vector.'**

Sparse vector is a vector which has two parallel arrays, one for indices, one for values, use for storing non-zero entities to save space.


**6) Name the language supported by Apache Spark for developing big data applications**

Important language use for developing big data application are:

- Java
- Python
- R
- Clojure
- Scala

**7) What is the method to create a Data frame?**

In Apache Spark, a Data frame can be created using Tables in Hive and Structured data files.

**8) Explain SchemaRDD**

An RDD which consists of row object with schema information about the type of data in each column is called SchemaRDD.

**9) What are accumulators?**

Accumulators are the write-only variables. They are initialized once and sent to the workers. These workers will update based on the logic written, which will send back to the driver.

**10) What are the components of Spark Ecosystem?**

An important component of Spark are:

- Spark Core: It is a base engine for large-scale parallel and distributed data processing

- Spark Streaming: This component used for real-time data streaming.
- Spark SQL: Integrates relational processing by using Spark's functional programming API
- GraphX: Allows graphs and graph-parallel computation
- MLlib: Allows you to perform machine learning in Apache Spark

**11) Name three features of using Apache Spark**

Three most important feature of using Apache Spark are:

1. Support for Sophisticated Analytics
2. Helps you to Integrate with Hadoop and Existing Hadoop Data
3. It allows you to run an application in Hadoop cluster, up to 100 times faster in memory, and ten times faster on disk.

**12) Explain the default level of parallelism in Apache Spark**

If the user isn't able to specify, then the number of partitions are considered as default level of parallelism in Apache Spark.

**13) Name three companies which is used Spark Streaming services**

Three known companies using Spark Streaming services are:

- Uber
- Netflix
- Pinterest

**14) What is Spark SQL?**

Spark SQL is a module for structured data processing where we take advantage of SQL queries running on that database.

**15) Explain Parquet file**

Paraquet is a columnar format file support by many other data processing systems. Spark SQL allows you to performs both read and write operations with Parquet file.

**16) Explain Spark Driver?**

Spark Driver is the program which runs on the master node of the machine and declares transformations and actions on data RDDs.

**17) How can you store the data in spark?**

Spark is a processing engine which doesn't have any storage engine. It can retrieve data from

another storage engine like HDFS, S3.

**18) Explain the use of File system API in Apache Spark**

File system API allows you to read data from various storage devices like HDFS, S3 or local Fileyste.

**19) What is the task of Spark Engine**

Spark Engine is helpful for scheduling, distributing and monitoring the data application across the cluster.

**20) What is the user of sparkContext?**
SparkContent is the entry point to spark. SparkContext allows you to create RDDs which provided various way of churning data.

**21) How can you implement machine learning in Spark?**

MLif is a versatile machine learning library given by Spark.

**22) Can you do real-time processing with Spark SQL?**

Real-time data processing is not possible directly. However, it is possible by registering existing RDD as a SQL table and trigger the SQL queries on priority.

**23) What are the important differences between Apache and Hadoop**

| Parameter | Apache Spark | Hadoop |
|---|---|---|
| Speed | 100 times faster compares to Hadoop. | It has moderate speed. |
| Processing | Real-time batch processing functionality. | It offers batch processing only. |
| Learning curve | Easy | Hard |
| Interactivity | It has interactive modes | Apart from Pig and Hive, it has not an interactive way. |

**24) can you run Apache Spark On Apache Mesos?**

Yes, you can run Apache Spark on the hardware clusters managed by Mesos.

**25) Explain partitions**

Partition is a smaller and logical division of data. It is the method for deriving logical units of data to speed up the processing process.

**26) Define the term ‘Lazy Evolution’ with reference to Apache Spark**

Apache Spark delays its evaluation until it is needed. For the transformations, Spark adds them to a DAG of computation and only when derive request some data.

**27) Explain the use of broadcast variables**

The most common use of broadcast variables are:

- Broadcast variables help programmer to keep a read-only variable cached on each machine instead of shipping a copy of it with tasks.
- You can also use them to give every node a copy of a large input dataset in an efficient manner.
- Broadcast algorithms also help you to reduce communication cost

**28) How you can use Akka with Spark?**

Spark uses Akka use for scheduling. It also uses Akka for messaging between the workers and masters.

**29) Which the fundamental data structure of Spark**

Data frame is fundamental is the fundamental data structure of Spark.

**30) Can you use Spark for ETL process?**

Yes, you can use spark for the ETL process.

**31) What is the use of map transformation?**

Map transformation on an RDD produces another RDD by translating each element. It helps you to translates every element by executing the function provided by the user.

**32) What are the disadvantages of using Spark?**

The following are some of the disadvantages of using Spark:

- Spark consume a huge amount of data compared with Hadoop.
- You can’t run everything on a single node as work must be distrusted over multiple clusters.
- Developers needs extra care while running their application in Spark.
- Spark streaming does not provide support for record-based window criteria.

**33) What are common uses of Apache Spark?**

- Apache Spark is used for:
- Interactive machine learning
- Stream processing

- Data analytics and processing
- Sensor data processing

**34) State the difference between persist() and cache() functions.**

Persist() function allows the user to specify the storage level whereas cache() use the default storage level.

**35) Name the Spark Library which allows reliable file sharing at memory speed across different cluster frameworks.**

Tachyon is a spark library which allows reliable file sharing at memory speed across various cluster frameworks.

**36) Apache Spark is a good fit for which type of machine learning techniques?**

Apache Spark is ideal for simple machine learning algorithms like clustering, regression, and classification.

**37) How you can remove the element with a critical present in any other Rdd is Apache spark?**

In order to remove the elements with a key present in any other rdd, you need to use substractkey() function.

**38) What is the use of checkpoints in spark?**

Checkpoints allow the program to run all around the clock. Moreover, it helps to make it resilient towards failure irrespective to application logic.

**39) Explain lineage graph**

Lineage graph information computer each RDD on demand. Therefore, whenever a part of persistent RDD is lost. In that situation, you can recover this data using lineage graph information.

**40) What are the file formats supported by spark?**

Spark supports file format json, tsv, snappy, orc, rc, etc.

**41) What are Actions?**

Action helps you to bring back the data from RDD to the local machine. Its execution is the result of all previously created transformations.

**42) What is Yarn?**

Yarn is one of the most important features of Apache Spark.  Running spark on Yarn makes binary distribution of spark as it is built on Yarn support.

**43) Explain Spark Executor**

An executor is a Spark process which runs computations and stores the data on the worker node. The final tasks by SparkContent are transferred to the executor for their execution.

**44) is it necessary to install Spark on all nodes while running Spark application on Yarn?**

No, you don't necessarily need to install spark on all nodes as spark runs on top of Yarn.

**45) What is a worker node in Apache Spark?**

A worker node is any node which can run the application code in a cluster.

**46) How can you launch Spark jobs inside Hadoop MapReduce?**

Spark in MapReduce allows users to run all kind of spark job inside MapReduce without need to obtain admin rights of that application.

**47) Explain the process to trigger automatic clean-up in Spark to manage accumulated metadata.**

You can trigger automatic clean-ups by seeing the parameter 'spark.cleaner.ttf or by separating the long-running jobs into various batches and writing the intermediate results to the disk.

**48) Explain the use of Blinkdb**

BlinkDB is a query engine tool which allows you to execute SQL queries on huge volumes of data and renders query results in the meaningful error bars.

**49) Does Hoe Spark handle monitoring and logging in Standalone mode?**

Yes, a spark can handle monitoring and logging in standalone mode as it has a web-based user interface.

**50) How can you identify whether a given operation is Transformation or Action?**

You can identify the operation based on the return type. If the return type is not RDD, then the operation is an action. However, if the return type is the same as the RDD, then the operation is transformation.

**51) Can You Use Apache Spark To Analyze and Access Data Stored In Cassandra Databases?**

Yes, you can use Spark Cassandra Connector which allows you to access and analyze data stored in Cassandra Database.

**52) State the difference between Spark SQL and Hql**

SparkSQL is an essential component on the spark Core engine. It supports SQL and Hive Query Language without altering its syntax.