# Top 40 Apache Spark Interview Questions and Answers in 2021

By Chandanp 9.8 K Views 21 min read Updated on December 13, 2021

In this list of the top most-asked Apache Spark interview questions and answers, you will find all you need to clear your Spark job interview. Here, you will learn what Apache Spark key features are, what an RDD is, what a Spark engine does, Spark transformations, Spark Driver, Hive on Spark, the functions of Spark SQL, and so on. Learn Apache Spark from Intellipaat's Apache Spark Course and fast-track your career!

## Top Answers to Apache Spark Interview Questions

As a professional in the field of Big Data, it is important for you to know all the terms and technologies related to this field, including Apache Spark, which is among the most popular and in-demand technologies in Big Data. Go through these Apache Spark interview questions to prepare for job interviews to get a head start in your career in Big Data:



These Apache Spark interview questions and answers are majorly classified into the following categories:



## Basic Interview Questions

### 1. What is Apache Spark?

Spark is a fast, easy-to-use, and flexible data processing framework. It is an open-source analytics engine that was developed by using Scala, Python, Java, and R. It has an advanced execution engine supporting acyclic data flow and in-memory computing. It uses in-memory caching and optimized execution of queries for faster query analytics of data of any size. Apache Spark can run standalone, on Hadoop, or in the cloud and is capable of accessing diverse data sources including HDFS, HBase, and Cassandra, among others.

### 2. Explain the key features of Spark.

- Apache Spark allows integrating with Hadoop.
- It has an interactive language shell, Scala (the language in which Spark is written).
- Spark consists of RDDs (Resilient Distributed Datasets), which can be cached across the computing nodes in a cluster.
- Apache Spark supports multiple analytic tools that are used for interactive query analysis, real-time analysis, and graph processing
- Apache Spark supports stream processing in real-time.
- Spark helps in achieving a very high processing speed of data, which it achieves by reducing the read or write operations to disk.
- Apache Spark codes can be reused for data streaming, running ad-hoc queries, batch processing, etc.
- Spark is considered a better cost-efficient solution when compared to Hadoop.

***Learn more key features of Apache Spark in this Apache Spark Tutorial!***

### 3. What is MapReduce?

It is a software framework and programming model which is used for processing huge datasets. MapReduce is basically split into two parts, Map and Reduce. Map handles data splitting and data mapping, meanwhile, Reduce handles shuffle and reduction in data.

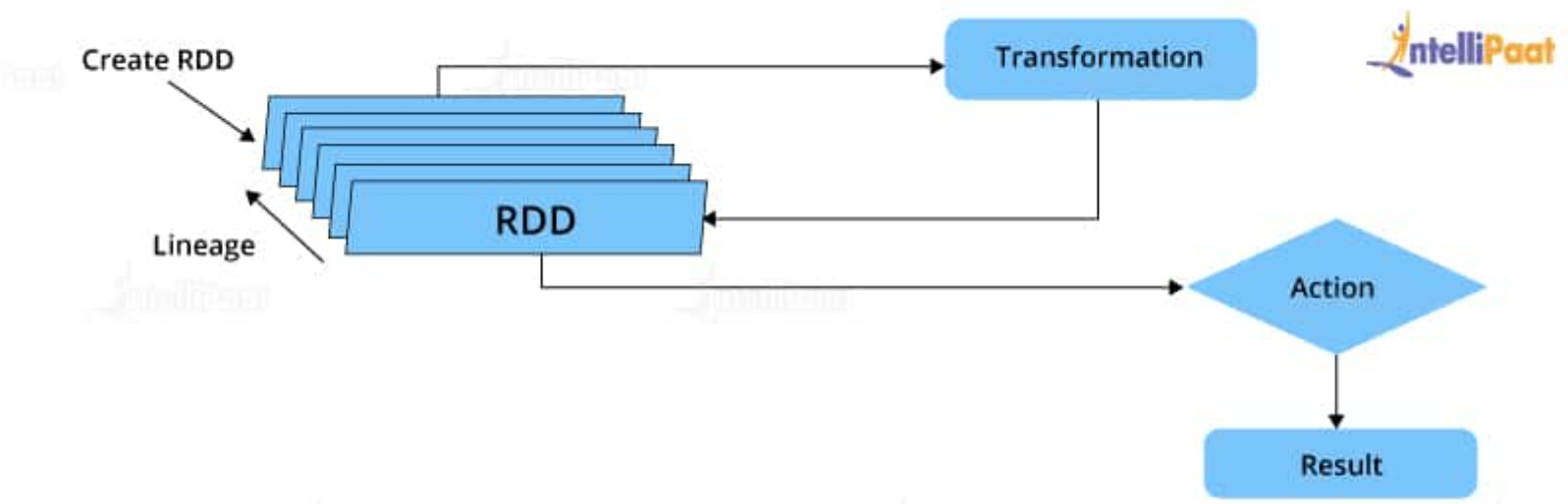
### 4. Compare MapReduce with Spark.

| Criteria | MapReduce | Spark |
|---|---|---|
| Processing speed | Good | Excellent (up to 100 times faster) |
| Data caching | Hard disk | In-memory |

| | | |
|---|---|---|
| Performing iterative jobs | Average | Excellent |
| Dependency on Hadoop | Yes | No |
| Machine Learning applications | Average | Excellent |

### 5. Define RDD.

RDD is the acronym for Resilient Distribution Datasets—a fault-tolerant collection of operational elements that run in parallel. The partitioned data in an RDD is immutable and distributed. There are primarily two types of RDDs:




RDD in SparkRDD in Spark

- Parallelized collections: The existing RDDs running in parallel with one another
- Hadoop datasets: Those performing a function on each file record in HDFS or any other storage system

### 6. What does a Spark Engine do?

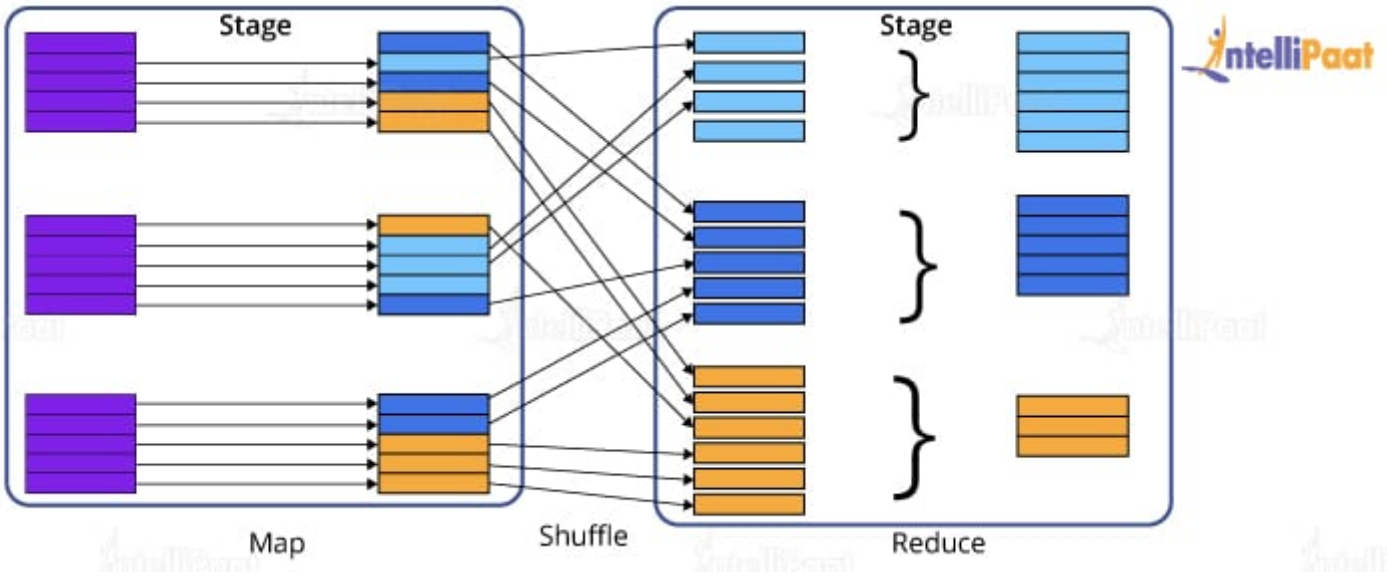
A Spark engine is responsible for scheduling, distributing, and monitoring the data application across the cluster. Spark Engine is used to run mappings in Hadoop clusters. It is suitable for wide-ranging circumstances. It includes SQL batch and ETL jobs in Spark, streaming data from sensors, IoT, ML, etc.

***Read on Spark Engine and more in this Apache Spark Community!***



### 7. Define Partitions.

As the name suggests, a partition is a smaller and logical division of data similar to a 'split' in MapReduce. Partitioning is the process of deriving logical units of data to speed up data processing. Everything in Spark is a partitioned RDD.



### 8. What operations does an RDD support?

- **Transformations:** Transformations produce a new RDD from an existing RDD, every time we apply a transformation to the RDD. Always it takes an RDD as input and ejects one or more RDD as output.
- **Actions:** Actions are used when we wish to use the actual RDD instead of working with a new RDD after we apply transformations. Actions eject out non-RDD values unlike transformations, which only eject RDD values.

### 9. What do you understand about Transformations in Spark?

Transformations are functions applied to RDDs, resulting in another RDD. It does not execute until an action occurs. Functions such as map() and filer() are examples of transformations, where the map() function iterates over every line in the RDD and splits into a new RDD. The filter() function creates a new RDD by selecting elements from the current RDD that passes the function argument.

### 10. Define Actions in Spark.

Actions are operations in Spark; they help in working with the actual data set. They help in transferring data from executor to driver. In Spark, an action helps in bringing back data from an RDD to the local machine. They are RDD operations giving non-RDD values, which is unlike transformations operations, which only eject RDD as output. The reduce() function is an action that is implemented again and again until only one value is left. The take() action takes all the values from an RDD to the local node.

**Check out this insightful video on Spark Tutorial for Beginners:**


### 11. Define the functions of Spark Core.

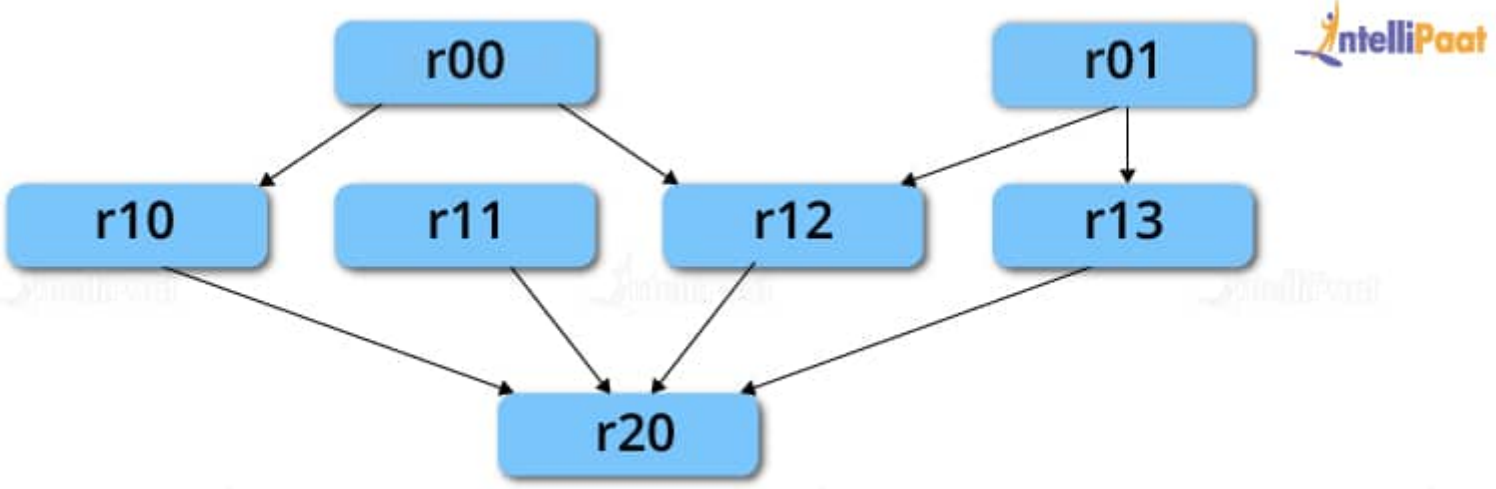
Serving as the base engine, Spark Core performs various important functions like memory management, basic I/O functionalities, monitoring jobs, providing fault-tolerance, job scheduling, interaction with storage systems, distributed task dispatching, and many more. Spark Core is the base of all projects. The above-mentioned functions are Spark Core's primary functions.

***Learn more about Spark from this Spark Training in New York to get ahead in your career!***

## Intermediate Interview Questions

### 12. What is RDD Lineage?

Spark does not support data replication in memory and thus, if any data is lost, it is rebuilt using RDD lineage.




RDD lineage is a process that reconstructs lost data partitions. The best thing about this is that RDDs always remember how to build from other datasets.

## Career Transition

### 13. What is Spark Driver?

Spark Driver is the program that runs on the master node of a machine and declares transformations and actions on data RDDs. In simple terms, a driver in Spark creates SparkContext, connected to a given Spark Master. It also delivers RDD graphs to Master, where the standalone Cluster Manager runs.

### 14. What is Hive on Spark?

Hive contains significant support for Apache Spark, wherein Hive execution is configured to Spark:

```
hive> set spark.home=/location/to/sparkHome;
hive> set hive.execution.engine=spark;
```

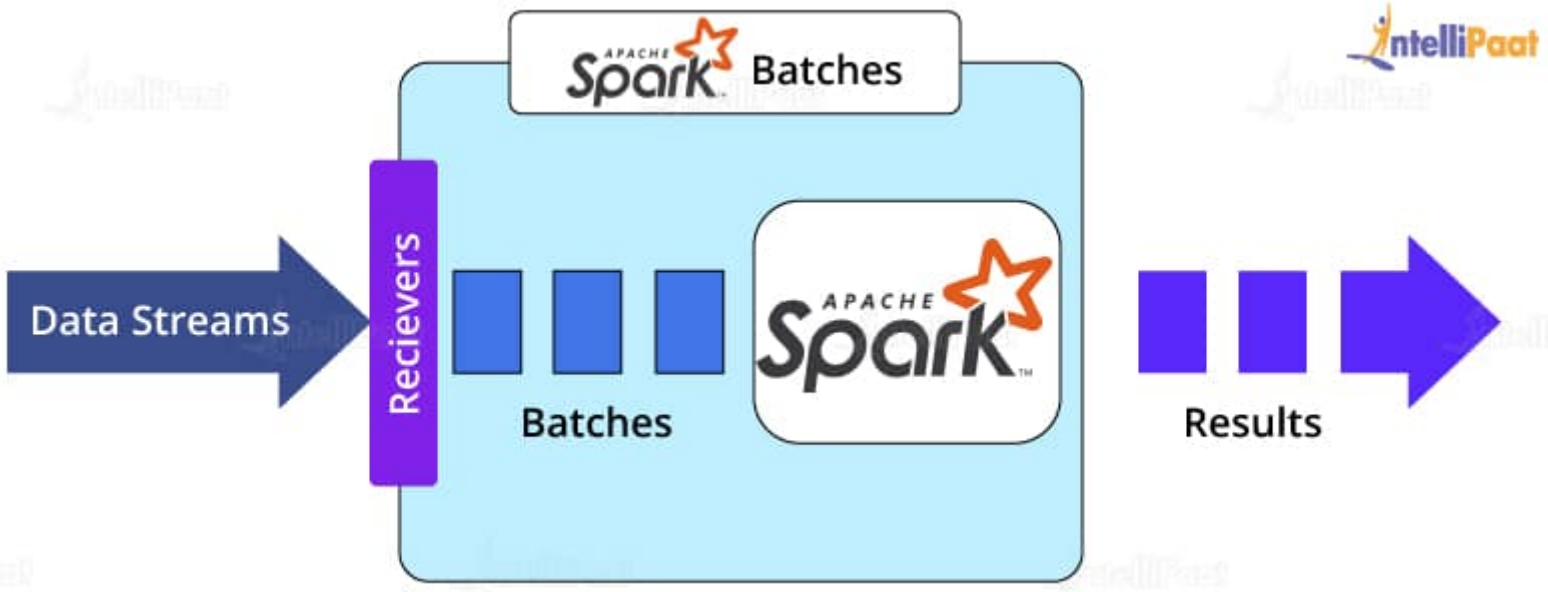
Hive supports Spark on YARN mode by default.

### 15. Name the commonly used Spark Ecosystems.

- Spark SQL (Shark) for developers
- Spark Streaming for processing live data streams
- GraphX for generating and computing graphs
- MLlib (Machine Learning Algorithms)
- SparkR to promote R Programming in the Spark engine

### 16. Define Spark Streaming.

Spark supports stream processing—an extension to the Spark API allowing stream processing of live data streams.




Data from different sources like Kafka, Flume, Kinesis is processed and then pushed to file systems, live dashboards, and databases. It is similar to batch processing in terms of the input data which is here divided into streams like batches in batch processing.

***Learn in detail about the Top Four Apache Spark Use Cases including Spark Streaming!***

### 17. What is GraphX?

Spark uses GraphX for graph processing to build and transform interactive graphs. The GraphX component enables programmers to reason about structured data at scale.

### 18. What does MLlib do?

MLlib is a scalable Machine Learning library provided by Spark. It aims at making Machine Learning easy and scalable with common learning algorithms and use cases like clustering, regression filtering, dimensional reduction, and the like.

### 19. What is Spark SQL?

Spark SQL, better known as Shark, is a novel module introduced in Spark to perform structured data processing. Through this module, Spark executes relational SQL queries on data. The core of this component supports an altogether different RDD called SchemaRDD, composed of row objects and schema objects defining the data type of each column in a row. It is similar to a table in relational databases.

### 20. What is a Parquet file?

Parquet is a columnar format file supported by many other data processing systems. Spark SQL performs both read and write operations with the Parquet file and considers it to be one of the best Big Data Analytics formats so far.



### 21. What file systems does Apache Spark support?

Apache Spark is a powerful distributed data processing engine that processes data coming from multiple data sources. The file systems that Apache Spark supports are:

- Hadoop Distributed File System (HDFS)
- Local file system
- Amazon S3
- HBase
- Cassandra, etc.

### 22. What is Directed Acyclic Graph in Spark?

Directed Acyclic Graph or DAG is an arrangement of edges and vertices. As the name implies the graph is not cyclic. In this graph, the vertices represent RDDs, and the edges represent the operations applied to RDDs. This graph is unidirectional, which means it has only one flow. DAG is a scheduling layer that implements stage-oriented scheduling and converts a plan for logical execution to a physical execution plan.

### 23.What are deploy modes in Apache Spark?

There are only two deploy modes in Apache Spark, client mode and cluster mode. The behavior of Apache Spark jobs depends on the driver component. If the driver component of Apache Spark will run on the machine from which the job is submitted, then it is the client mode. If the driver component of Apache Spark will run on Spark clusters and not on the local machine from which the job is submitted, then it is the cluster mode.

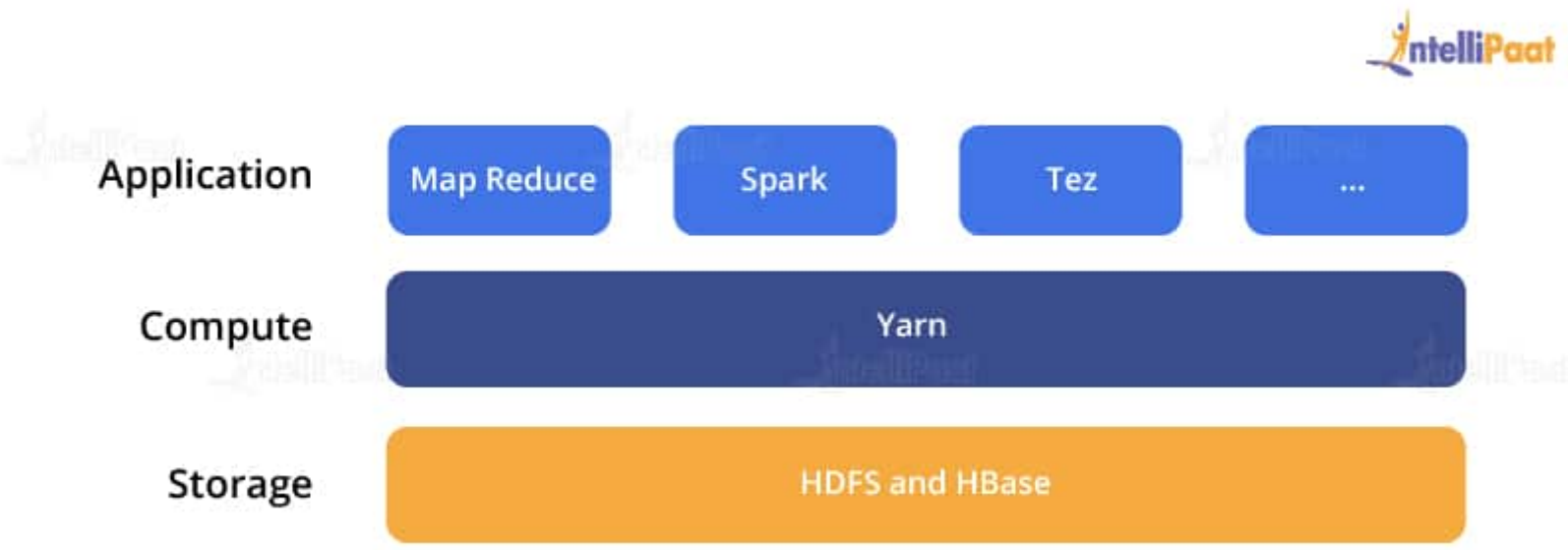
### 24. Roles of receivers in Apache Spark Streaming?

Within Apache Spark Streaming Receivers are special objects whose only goal is to consume data from different data sources and then move it to Spark. You can create receiver objects by streaming contexts as long-running tasks on various executors. There are two types of receivers. They are **Reliable receivers:** This receiver acknowledges data sources when data is received and replicated successfully in Apache Spark Storage. **Unreliable receiver:** These receivers do not acknowledge data sources even when they receive or replicate in Apache Spark Storage.

## Advanced Interview Questions

### 25. What is YARN?

Similar to Hadoop, YARN is one of the key features in Spark, providing a central and resource management platform to deliver scalable operations across the cluster. Running Spark on YARN needs a binary distribution of Spark that is built on YARN support.




*Enroll in Intellipaat's Spark Course in London today to get a clear understanding of Spark!*

### 26. List the functions of Spark SQL.

Spark SQL is capable of:

- Loading data from a variety of structured sources
- Querying data using SQL statements, both inside a Spark program and from external tools that connect to Spark SQL through standard database connectors (JDBC/ODBC), e.g., using Business Intelligence tools like Tableau
- Providing rich integration between SQL and the regular Python/Java/Scala code, including the ability to join RDDs and SQL tables, expose custom functions in SQL, and more.

### 27. What are the benefits of Spark over MapReduce?

- Due to the availability of in-memory processing, Spark implements data processing 10–100x faster than Hadoop MapReduce. MapReduce, on the other hand, makes use of persistence storage for any of the data processing tasks.
- Unlike Hadoop, Spark provides in-built libraries to perform multiple tasks using batch processing, steaming, Machine Learning, and interactive SQL queries. However, Hadoop only supports batch processing.
- Hadoop is highly disk-dependent, whereas Spark promotes caching and in-memory data storage.

- Spark is capable of performing computations multiple times on the same dataset, which is called iterative computation. Whereas, there is no iterative computing implemented by Hadoop.

***For more insights, read on Spark vs MapReduce!***

### 28. Is there any benefit of learning MapReduce?

Yes, MapReduce is a paradigm used by many Big Data tools, including Apache Spark. It becomes extremely relevant to use MapReduce when data grows bigger and bigger. Most tools like Pig and Hive convert their queries into MapReduce phases to optimize them better.

### 29. What is a Spark Executor?

When SparkContext connects to Cluster Manager, it acquires an executor on the nodes in the cluster. Executors are Spark processes that run computations and store data on worker nodes. The final tasks by SparkContext are transferred to executors for their execution.

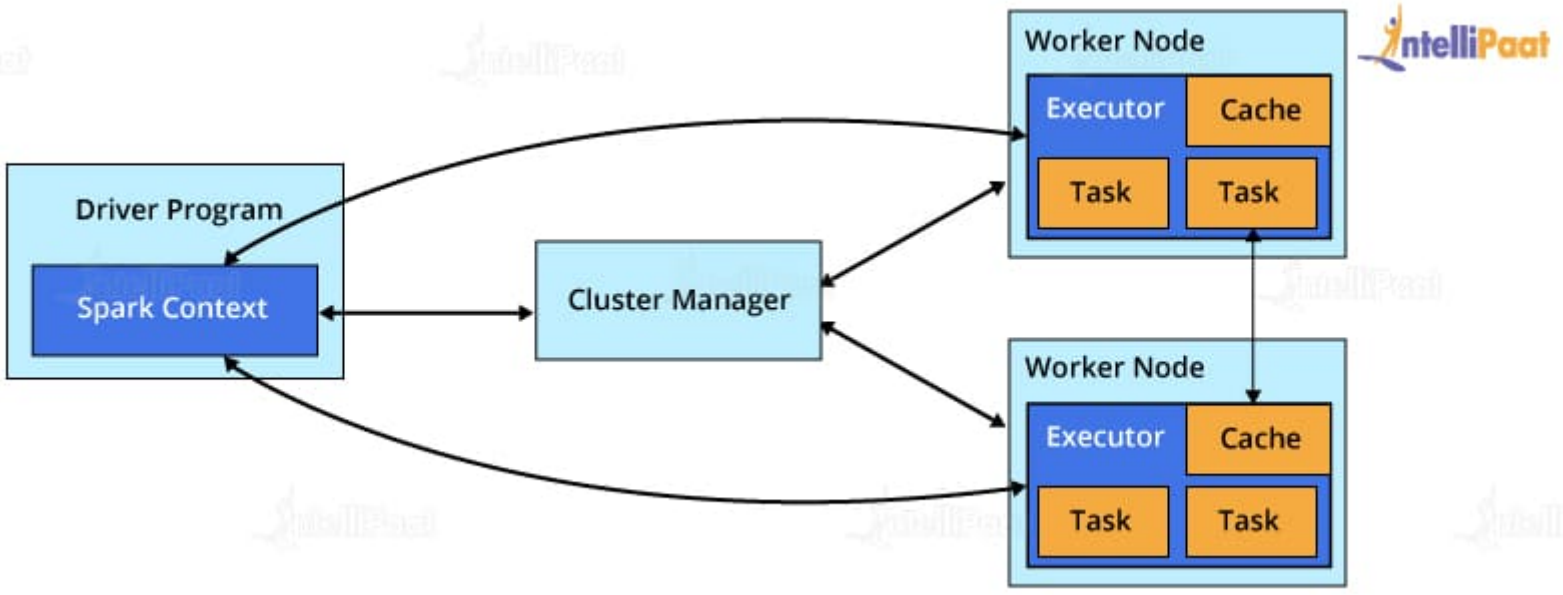
### 30. Name the types of Cluster Managers in Spark.

The Spark framework supports three major types of Cluster Managers.

- **Standalone:** A basic Cluster Manager to set up a cluster
- **Apache Mesos:** A generalized/commonly-used Cluster Manager, running Hadoop MapReduce and other applications
- **YARN:** A Cluster Manager responsible for resource management in Hadoop

### 31. What do you understand by a Worker node?

A worker node refers to any node that can run the application code in a cluster.




### 32. What is PageRank?

A unique feature and algorithm in GraphX, PageRank is the measure of each vertex in a graph. For instance, an edge from *u* to *v* represents an endorsement of *v*'s importance w.r.t. *u*. In simple terms, if a user on Instagram is followed massively, he/she will be ranked high on that platform.

### 33. Do you need to install Spark on all the nodes of the YARN cluster while running Spark on YARN?

No, because Spark runs on top of YARN.



### 34. Illustrate some demerits of using Spark.

Since Spark utilizes more storage space when compared to Hadoop and MapReduce, there might arise certain problems. Developers need to be careful while running their applications of Spark. To resolve the issue, they can think of distributing the workload over multiple clusters, instead of running everything on a single node.

### 35. How to create an RDD?

Spark provides two methods to create an RDD:

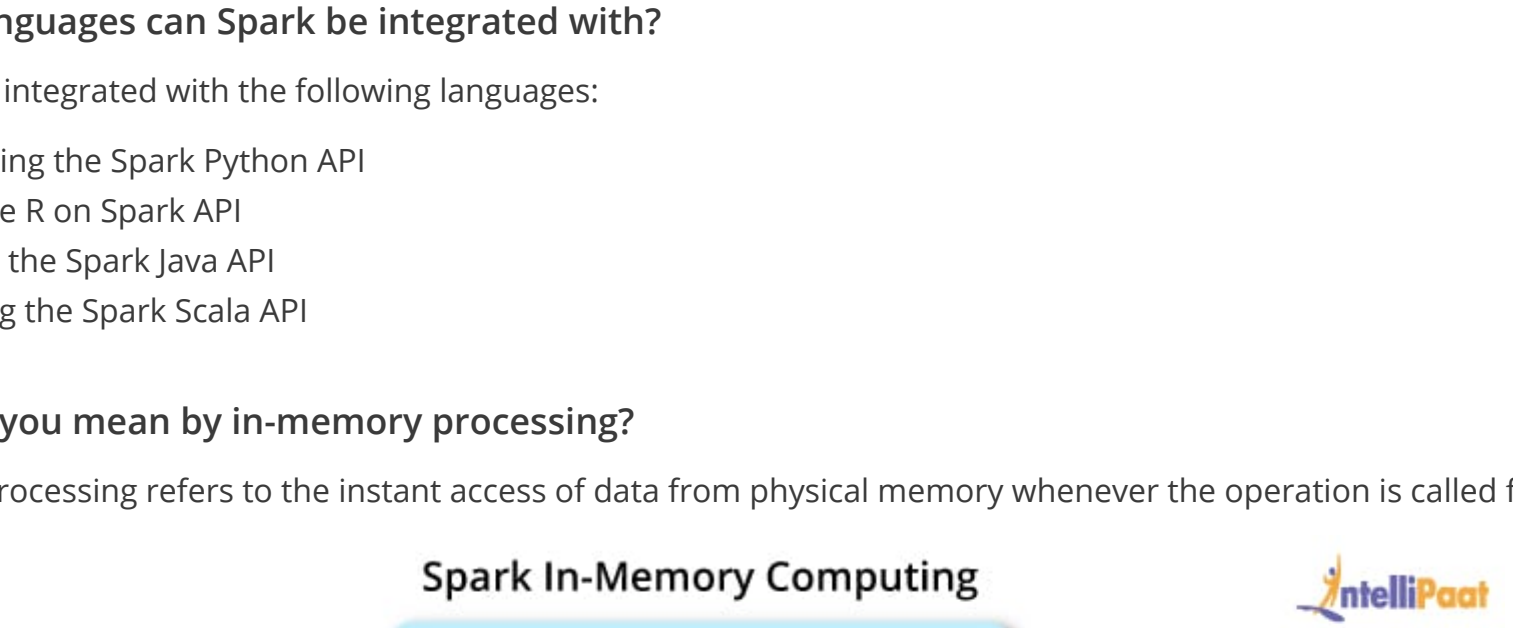
- By parallelizing a collection in the driver program. This makes use of SparkContext's 'parallelize' method **val**

```
IntellipaatData = Array(2,4,6,8,10)
val distIntellipaatData = sc.parallelize(IntellipaatData)
```

By loading an external dataset from external storage like HDFS, the shared file system

### 36. What are Spark DataFrames?

When a dataset is organized into SQL-like columns, it is known as a DataFrame.




This is, in concept, equivalent to a data table in a relational database or a literal 'DataFrame' in R or Python. The only difference is the fact that Spark DataFrames are optimized for Big Data.

### 37. What are Spark Datasets?

Datasets are data structures in Spark (added since Spark 1.6) that provide the JVM object benefits of RDDs (the ability to manipulate data with lambda functions), alongside a Spark SQL-optimized execution engine.

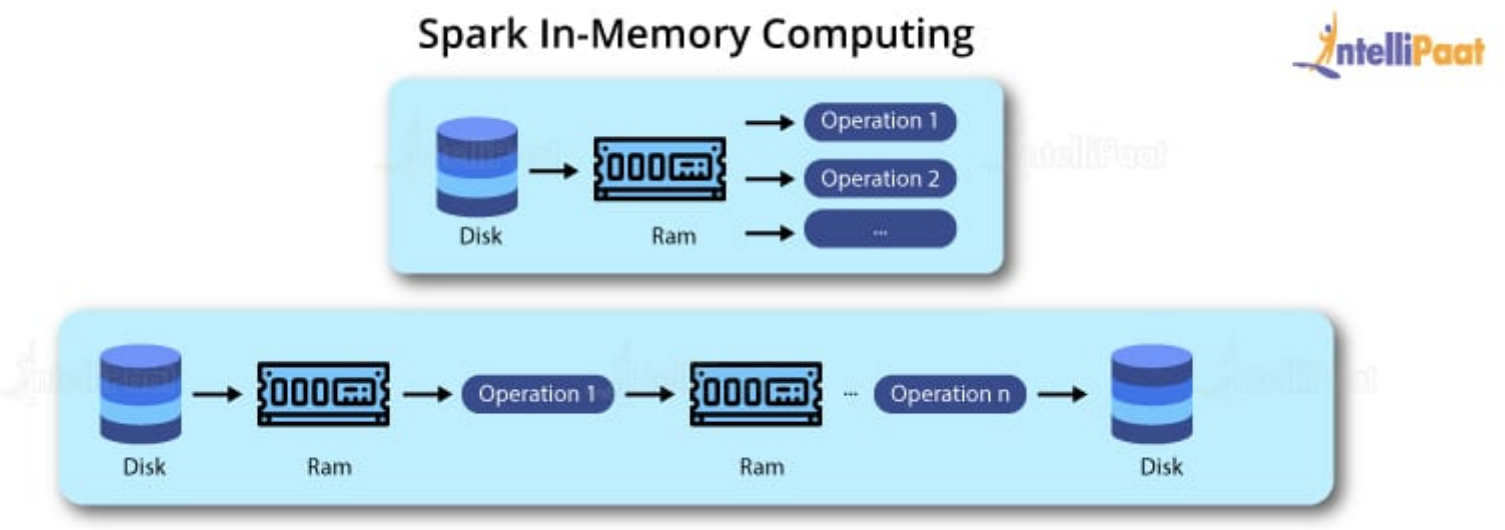
### 38. Which languages can Spark be integrated with?

Spark can be integrated with the following languages:

- Python, using the Spark Python API
- R, using the R on Spark API
- Java, using the Spark Java API
- Scala, using the Spark Scala API

### 39. What do you mean by in-memory processing?

In-memory processing refers to the instant access of data from physical memory whenever the operation is called for.



This methodology significantly reduces the delay caused by the transfer of data. Spark uses this method to access large chunks of data for querying or processing.

### 40. What is lazy evaluation?

Spark implements a functionality, wherein if you create an RDD out of an existing RDD or a data source, the materialization of the RDD will not occur until the RDD needs to be interacted with. This is to ensure the avoidance of unnecessary memory and CPU usage that occurs due to certain mistakes, especially in the case of Big Data Analytics.

***Interested in learning Spark? Take up our Spark Training in Sydney now!***

Next

## Course Schedule

| Name | Date | |
|---|---|---|
| Big Data Course | 2021-12-25 2021-12-26<br>(Sat-Sun) Weekend batch | View Details |
| Big Data Course | 2022-01-01 2022-01-02<br>(Sat-Sun) Weekend batch | View Details |
| Big Data Course | 2022-01-08 2022-01-09<br>(Sat-Sun) Weekend batch | View Details |

### 11 thoughts on "Top 40 Apache Spark Interview Questions and Answers in 2021"

**Vishakha** says:

Good compilation of questions, thank you

DECEMBER 18, 2015 AT 1:35 PM

Reply

**RAFAEL VALVERDE BUSTOS** says:

Simple, accurate, useful; brilliant definitively

MARCH 29, 2016 AT 12:23 AM

Reply

**Eric O. LEBIGOT** says:

Thanks. A question about shuffling would be quite relevant, I find.

APRIL 28, 2016 AT 3:28 AM

Reply

**Monika** says:

Thanks for sharing very useful Interview Q and A.

MAY 2, 2016 AT 6:31 PM

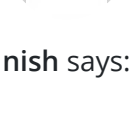
Reply

Excellent Tutorial. It's very helpful for beginner's as well as experienced.

MAY 3, 2016 AT 6:18 PM

Reply

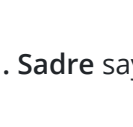
**Joshi** says:



So nice tutorial, very well explained…Thanks to Intellipaat team. MAY 3, 2016 AT 6:19 PM

Reply

**Apporva** says:



Awesome Apache Spark Interview Questions and Answers. It's easy to understand and very informative. MAY 3, 2016 AT 6:22 PM

Reply

**Meena** says:



**Deepak Kumar** says:

Good one JULY 28, 2016 AT 8:50 PM

Reply

**nish** says:

That's great,veary helpful:) JANUARY 8, 2020 AT 12:33 AM

Reply

**Md. Sadre** says:

nice SEPTEMBER 5, 2020 AT 2:43 PM

Reply

**Hr** says:

## Leave a Reply

Your email address will not be published. Required fields are marked *

**Comment**

**Name** *

**Email** *

Post Comment