1. What is RDD?

RDD stands for Resilient Distributed Dataset.

* RDD’s in Spark are immutable distributed collection of objects.
* Each RDD is split into multiple partitions, which may be computed on different nodes of the cluster.
* The RDDs can contain any type of Python, Java or Scala objects, including user-defined classes.
* They are not actual data, but they are Objects, which contain information about data residing on the cluster.
* The RDDs try to solve these problems by enabling fault tolerant, distributed In-memory computations
* Data is stored in partitions of the RDDs and store in worker nodes (datanodes) which are computed in parallel across all the nodes.
* RDDs load the data for us and are resilient, which means they can be recomputed.
* RDDs perform two types of operations: Transformations, which creates a new dataset from the previous RDD or new source and Actions, which return a value to the driver program after performing the computation on the dataset.
* RDDs keeps track of Transformations and check them periodically. If a node fails, it can rebuild the lost RDD partition on the other nodes, in parallel

2. Define Partitions.

Partitions is way to control grouping/distribution of data across different nodes in cluster which leads to parallel computing and reduced network traffic by ensuring data locality. All these result in better speed to process big data sets.

Each RDD is split into multiple partitions, which may be computed on different nodes of the cluster. RDDs are stored in partitions and operated in parallel. In other words RDDs in Apache Spark are collection of partitions.

Types of Partitioning in Apache Spark

* Hash Partitioning
* Range Partitioning
* Custom Partitioning

3. What operations does RDD support?

RDDs support two types of operations:

* Transformations, which creates a new dataset from the previous RDD or new source (file, database, external dataset)
* Actions, which return a value to the driver program after performing the computation on the dataset.

4. What do you understand by Transformations in Spark?

Transformations create a new data set either from the previous RDD or using new RDD creation operations (like parallelize, textfile, etc) based upon the various transformation operation like the following. Transformations are lazy which means they only get evaluated when an action is performed.

|  |  |  |
| --- | --- | --- |
| map() | intersection() | cartesian() |
| flatmap() | distinction() | coalesce() |
| filter() | groupbyKey() | repartition() |
| mapPartitions() | reducebyKey() | partitionby() |
| mapPartitionsWithIndex() | sortbyKey() | join() |
| sample() | union() | cogroup() |

5. Define Actions.

Actions return a value to the driver program after performing the computation on the dataset. When an action is called the previously lined-up transformations are evaluated and the final result is sent back to the driver program.