What is RDD?

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

1. **Parallelized Collections :** The existing RDD’s running parallel with one another.
2. **Hadoop datasets :** perform function on each file record in HDFS or other storage system

Define Partitions.

Ans:-Partition is a smaller and logical division of data  similar to ‘split’ in MapReduce. Partitioning is the process to derive logical units of data to speed up the processing process. Everything in Spark is a partitioned RDD.

What operations does RDD support?

Ans:-

* Transformations.
* Actions

What do you understand by Transformations in Spark?

Ans:-

Transformations are functions applied on RDD, resulting into another RDD. It does not execute until an action occurs. map() and filer() are examples of transformations, where the former applies the function passed to it on each element of RDD and results into another RDD. The filter() creates a new RDD by selecting elements form current RDD that pass function argument.

Define Actions.

Ans:-

An action helps in bringing back the data from RDD to the local machine. An action’s execution is the result of all previously created transformations. reduce() is an action that implements the function passed again and again until one value if left. take() action takes all the values from RDD to local node.