1. What is RDD?

Ans: 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:

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

2. Define Partitions.

Ans: A Partition is a smaller and logical division of data, that is similar to the split in Map Reduce. Partitioning is the process that helps derive logical units of data in order to speed up data processing.

3. What operations does RDD support?

Ans: RDDs perform two types of operations:

- Transformations
- Actions

4. 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 filter() 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 from current RDD that pass function argument.

5. Define Actions.

Ans: Action produces an output like printing the data and storing the data. This kind of operation will also give you another RDD but this operation will trigger all the lined up transformation on the base RDD (or in the DAG) and than execute the action operation on the last RDD. Operations like collect, count, first, saveAsTextFile are actions.