1. What is RDD?

Resilient Distributed Datasets (RDD) is a fundamental data structure of Spark. It is an immutable distributed collection of objects. Each dataset in RDD is divided into logical partitions, which may be computed on different nodes of the cluster. RDDs can contain any type of Python, Java, or Scala objects, including user-defined classes.

In Spark data is stored in partitions of the RDD.

2. Define Partitions

Partition is basic unit of parallelism in a RDD, and each partition is one logical division of data which is immutable and created through some transformations on existing partitions. Immutability helps to achieve consistency in computations.

3. What operations does RDD support?

Transformation Operations: Transformations are kind of operations which transform RDD data from one form to another and when this operation is applied on any RDD, a new RDD with transformed data is formed. Operations like map, filter, and flatMap are transformations.

Action Operations: This kind of operation also gives another RDD but this operation will trigger all the lined up transformation on the base RDD (or in the DAG) and then execute the action operation on the last RDD. Operations like collect, count, first, saveAsTextFile are actions.

4. What do you understand by transformations in spark?

Transformations are functions applied on RDD, resulting in a new RDD. It does not execute until an action occurs. Map and filter are examples of transformations, where map applies the function assigned to it on each element of the RDD and results in another RDD. The filter creates a new RDD by selecting elements from the current RDD.

5. Define actions

Actions trigger all the lined up transformation on the base RDD and return final results of RDD computations. It triggers execution using lineage graph to load the data into original RDD, carry out all intermediate transformations and return final results to Driver program or write it out to file system. Few are collect, count, first, saveAsTextFile are actions.