1. What is RDD?

RDD (Resilient Distributed Dataset) is main logical data unit in Spark. An RDD is distributed collection of objects. Distributed means, each RDD is divided into multiple partitions. Each of these partitions can reside in memory or stored on disk of different machines in a cluster. RDDs are immutable (Read Only) data structure. You can’t change original RDD, but you can always transform it into different RDD with all changes you want.

2. Define Partitions.

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?

* Transformation
* Action

4. What do you understand by Transformations in Spark?

A transformation generates a new RDD from an existing one, this are the functions applied on RDD. The map () method is a transformation. It does not execute until an action occurs. if the return type is RDD, then it’s a transformation.

5. Define Actions.

Spark provides two categories of operations on RDDs, that is transformations and actions. action triggers a computation on an RDD and does something with the results—either returning them to the user, or saving them to external storage. Actions have an immediate effect, but transformations do not—they are lazy, in the sense that they don’t perform any work until an action is performed on the transformed RDD.