3.2.1 Resilient Distributed Dataset (RDD): it works as a collection of elements which are operated in parallel. In the distributed file system Spark runs on Hadoop cluster, and RDD is created from files in the format of text or sequence files. RDD is used for reading the objects in the collection and when some partition is lost, it can be rebuilt because RDDs are distributed across a set of machines.

Summarized by © lakhasly.com