Experiment 7

Shashwat Shah
60004220126
TY BTECH COMPS B2

Aim: Implement RDD using Pyspark.

Theory: Apache spark is an open source distributed processing system used for big data workloads. It utilizes in-memory caching, and optimized query execution for fast analytic queries against data of any size.

Spark not only supports 'Map' and 'Reduce'. It provides developement API's in Java, Scala, Python and R. and supports code reuse across multiple workloads.

RDD is a core abstraction in spark which stands for resilent distributed dataset. It enables partition of large data into smaller data that fit's each machine. So that computational can be done parallely on multiple machines.

RDD supports two types of operations:

Transformations are operations (such as map, filter, join and so on) that are performed on an RDD and which yield a new RDD containing the result.

Actions are operations (such as reduce, count, first and so on). that return a value after running a computation on an RDD.

Conclusion: Thus we have implemented RDD using Pyspark.