**Assignment 13 2**

1. **What is RDD?**

*RDD – Resilient Distributed Datasets*

*RDD’s are immutable distributed collection of objects.*

*Each dataset is split into multiple partitions, which may be computed on different nodes of cluster. They are spread across multiple machines.*

1. **Define Partitions.**

*A Partition is a logical chunk of large distributed dataset.*

*🡪Spark manages data using partitions that helps parallelize distributed data processing with minimal network traffic.*

*🡪Worker nodes takes the data for processing that are nearer to them. By doing partitioning network I/O will be reduced so that data can be processed a lot faster.*

1. **What operations does RDD support?**

*RDD performs 2 types of operations:*

* *Transformations🡪 creating new dataset from existing dataset.*
* *Actions🡪returning value to driver program after doing computations on dataset.*

1. **What do you understand by Transformations in Spark?**

*Transformations are functi ons, that take an RDD as input and results in new RDD by performing some computations without changing the original RDD.*

*Transformations are lazy. i.e., data gets loaded only when action is performed.*

*Eg: filter,union*

1. **Define Actions.**

*Actions are operations that gives result. They trigger execution of RDD transformations to return results to the driver program.*

*Eg: collect,count,reduce*