**Assignment 16.2**

1) Pen down the limitations of MapReduce.

2) What is RDD? Explain few features of RDD?

3) List down few Spark RDD operations and explain each of them.

***Limitations of Map Reduce:***

* Data processing happens in different parts with help of Mappers and Reducers
* Quite difficult in programming Map Reduce programs
* Data processing take time because there are multiple IO Operations involved
* MR is mainly used in batch processing

***RDD:***

* RDD is resilient distributed datasets, which is the representation of the data that is coming into our system in object format.
* RDD’s allows us to do computations
* Fault tolerant is a main feature in RDD’s whenever a failure happens they use prior information with lineage so that results get recomputed.
* Lazy execution, the execution happens only when an action is triggered
* They are immutable where the state of RDD’s cannot be changed
* Records are portioned where the splits happen in logical manner.
* Writes in RDD’s are coarse grained
* Recovery happens based on the lineage graphs where they regenerate results when failure happens
* Consistency is trivial in RDD’s
* Easy programming

***Spark RDD Operations:***

* Transformations:
  + Transformation is what we do to RDD’s to get other resultant RDD’s
  + Opening a file and creating a RDD, performing functions like filters.
  + Filter, reduceByKey, map, join are few transformation functions
* Actions:
  + Actions are the place where actual computation happens and where the result is derived
  + Performing a count function comes under actions.
  + Only during the action performed the actual execution happens. (Lazy Evaluation)