Large Dataset for Applications
Assignment 3
Abdul Basit

Part 2
Section C

C.1

Basically her each command split function is supposed to run on worker node while in her case it is running in master node. Her print statement is in the split function. Cluster manager actually launches enforcers on worker nodes. Result is aggregated from worker nodes while the output is from master node and as the split function is printing just the worker node hence she can't see.

C.2

It is the case as it is easy and safe process to share data across numerous pathways. Furthermore, it is uncomplicated process to reestablish RDD's. With the help of RDD, the computation of process can be strengthened. As allocation, duplication and accumulation of data becomes very simple.

C.3

All the data from RDD is transferred to master nodes from worker nodes after instructions from the master with run.collect(), which memory of master node then eventually saves. It causes a problem as memory of master node is not big enough to store such large amount of data therefore creating a problem in cluster.

C.4

Resilience here in this sense is defined as when faced with failures in nodes, RDD's can re manage the absent or affected partitions. Lineage graph keeps track of all parent RDD's of a certain RDD and this graph is used by the RDD itself. In case of data loss, RDD uses lineage graphs to regenerate the missing data.

Part D

- **Abstain from using UDF's**
  Python processes which are to be performed are not rapid enough specially when we use user defined python functions. Furthermore, there are extra expenditures for serialization and de-serialization.

- **Apply caching**
  Spark has it's own caching techniques, that can basically be used as .persist() and .cache(). These certain processes if used can rapidly increase performance.
- **Data Serilization Optimization**
  Data sterilizers can help improve performance if used appropriately, such as Kryo serialization as these spark jobs are distributed.
- **Bucketing**
  Bucketing and data partition are some what identical, and can be simultaneously used along with SparkSQL. It mingles the data in advance f our using of specific functions for example table joins.