10.

Lab assignment 2
Group A

→ Title : Design a distributed application using Mapreduce.

→ Objective :

a> To explore different big data processing techniques with use cases.

b> To study detailed concepts of map - reduce.

→ Software requirements :

a> Ubuntu 14.10

b> GNU C compiler

c> Hadoop

d> Java

→ Problem statement : Design a distributed application using Map reduce (using Java) which processes a log file of a system. List out the users, who have logged for maximum period on the system. Use simple log file from the internet and process it using a pseudo distribution mode on Hadoop.

→ Theory :

a> Introduction to map reduce -

- Map reduce is a framework using which we can write applications to process huge amount of data in parallel on large datasets of commodity hardware in a reliable manner.
- Map reduce is a processing technique and a program model for distributed computing based on java.
- The map reduce algorithm contains two important tasks, namely map and reduce.

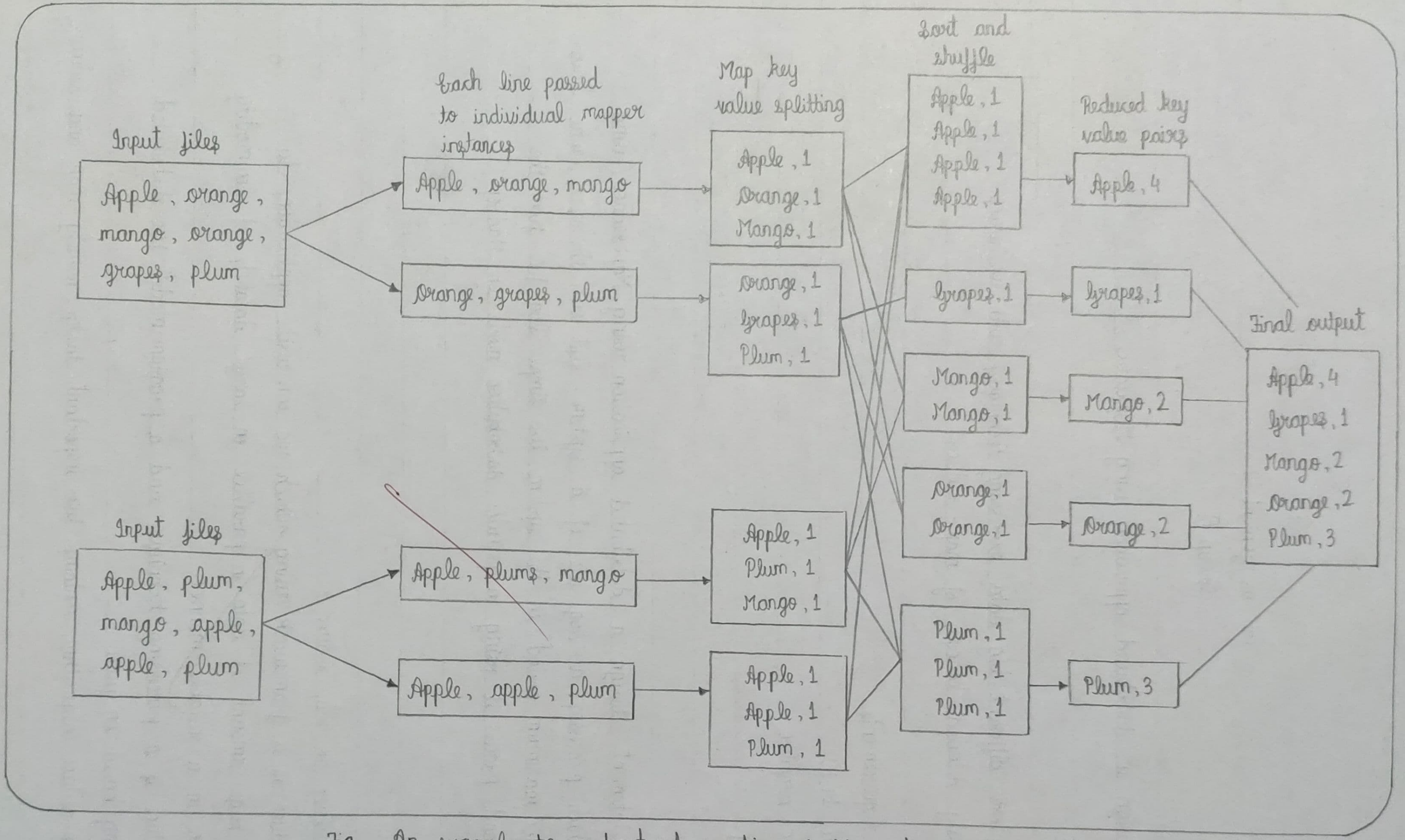

Fig. An example to understand working of Map reduce program

- Map takes a set of data, where individual elements are broken down into tuples and reduce takes output from a map as an input and combines those data tuples into a smaller set of tuples.
- As the sequence of the name Map reduce implies, reduce task is always performed after the map job.
- The main job advantage of Map reduce is that it is easy to scale data processing over multiple computing nodes.

b) Map reduce algorithm -

The map reduce program executes in three stages namely, map stage, shuffle stage and reduce stage.

1. Map stage :
- The map as mapper's job is to process the input data.
- Generally, the input data is in the form of file or directory and is stored in Hadoop file system.
- The input file system is passed to mapper function line by line.
- The mapper processes data and creates several small chunks of data.

2. Reduce stage :
- This stage is combination of shuffle stage and reduce stage.
- The reducer's job is to process the data and create small chunks.
- After processing, it produces a new set of output, which will be stored in HDFS.

c) Inserting data into HDFS -
- The map reduce framework operates on < key, value > pairs that is, the framework views the input to the job as a set of < key, value > pairs and produces a set of < key, value > pairs as the output of the job conceivably of different types.
- The key and the value classes should be in serialized manner by the framework and hence we need to implement the writable comparable interface to facilitate

sorting by the framework.

- Input and output types of map reduce job -

(Input | < k1, v1 > → map → < k2, v2 > → reduce → < k3, v3 > | output)

→ Conclusion : Thus, we have learnt to design a distributed application using map reduce and process log file of system.