VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Digital Assignment - I

Programme	:	**B.Tech (CSE)**	Semester	:	**WIN:2020-2021**
Course	:	**Mining Massive Dataset**	Code	:	**CSE6017**
Faculty	:	**Dr. Ramesh Ragala**	Slot	:	**F1**
Class Number	:	**CH2020215001372**		:	

Word Count - Hadoop Map Reduce Example

Word count is a typical example where Hadoop map reduce developers start their hands on with. This sample map reduce is intended to count the number of occurrences of each word in the provided input files.

How it works

The word count operation takes place in two stages a mapper phase and a reducer phase. In mapper phase first the test is tokenized into words then we form a key value pair with these words where the key being the word itself and value '1'.

The input text file consists of the following data (Assumption only)
Apple Orange Mango
Orange Grapes Plum
Apple Plum Mango
Apple Apple Plum

The dataset contains four lines of data, then it is divided logically two parts and given to mappers. Each mapper will take single from the corresponding dataset part and process further to generate key-value pair. (This assumption is to understand the working procedure in simple way)

In map phase the sentence would be split as words and form the initial key value pair as
<Apple,1>
<Orange,1>
<Mango,1>
<Orange,1>
<Grapes,1>
<Plum,1>
<Apple,1>
<Plum,1>
<Mango,1>
<Apple,1>
<Apple,1>
<Plum,1>

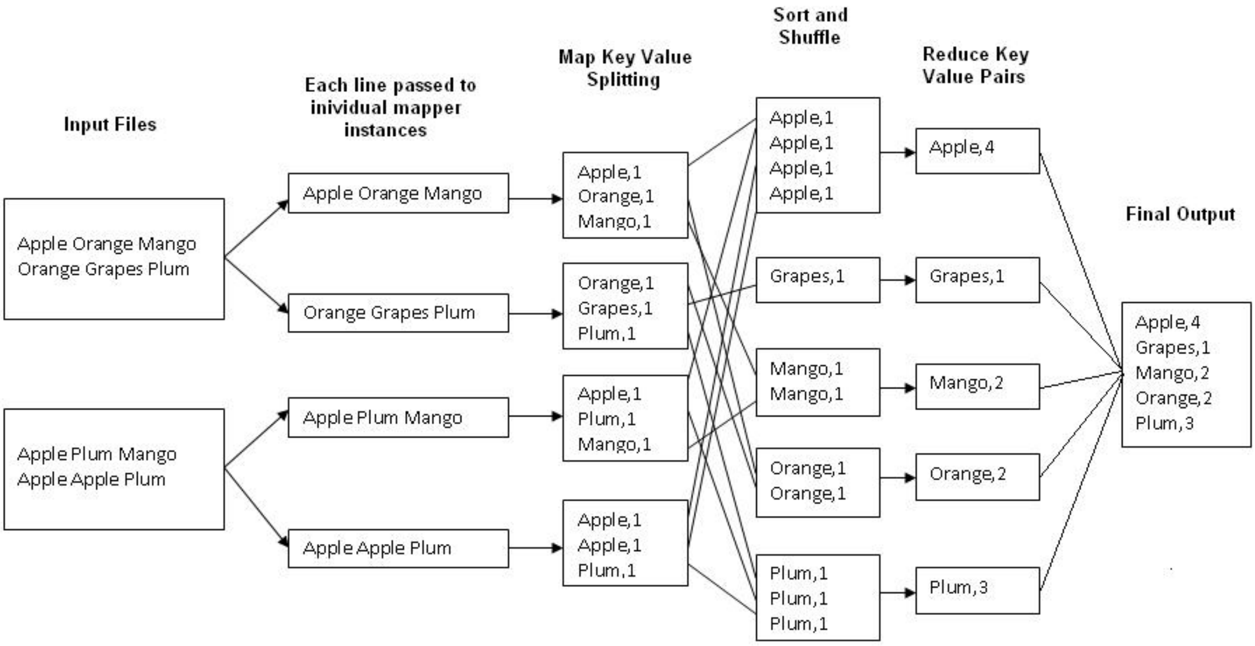
These set of key-value pairs will pass to Sort and Shuffle phase produce the output interm of key-value pair as shown below. <Apple,1>
<Apple,1>
<Apple,1>
<Apple,1>
<Grapes,1>
<Mango,1>
<Mango,1>
<Orange,1>
<Orange,1>
<Plum,1>
<Plum,1>
<Plum,1>

In the reduce phase, the values for similar keys are added. So here there are only one pair of similar keys 'Apple' the values for these keys would be added so the out put key value pairs would be <Apple,4>
<Grapes,1>
<Mango,2>
<Orange,2>
<Plum,3>

This would give the number of occurrence of each word in the input. Thus reduce forms an aggregation phase for keys.

The point to be noted here is that first the mapper class executes completely on the entire data set splitting the words and forming the initial key value pairs. Only after this entire process is completed the reducer starts. Say if we have a total of 10 lines in our input files combined together, first the 10 lines are tokenized and key value pairs are formed in parallel, only after this the aggregation/ reducer would start its operation.

MapReduce flow diagram in Mapper and Reducer Phases:



The solutions to the following questions should be given in the above discussion manner

Question-1

1. Draw a MapReduce flow diagram in Mapper and Reducer phases for K-means clustering.
2. Draw a MapReduce flow diagram in Mapper and Reducer phases for K-means++ clustering.
3. Draw a MapReduce flow diagram in Mapper and Reducer phases for Decision Tree algorithm.

You can use your own dataset for the above problems. And also describe the detailed procedure used in each phases with respect to the stated problems.

Question - 2

Illustrate the mechanisam used in Scalable K-means++ algorithm.