Design Documentation for Task #2

The problem:

I have two datasets:

1. Users (id, email, language, location)
2. Transactions (transaction-id, product-id, user-id, purchase-amount, item-description)

Given these datasets, I want to find the number of locations in which each product has been sold. To do that, I need to join the two datasets together.

MapReduce Join

Joining two large dataset can be achieved using MapReduce Join. However, this process involves writing lots of code to perform actual join operation.

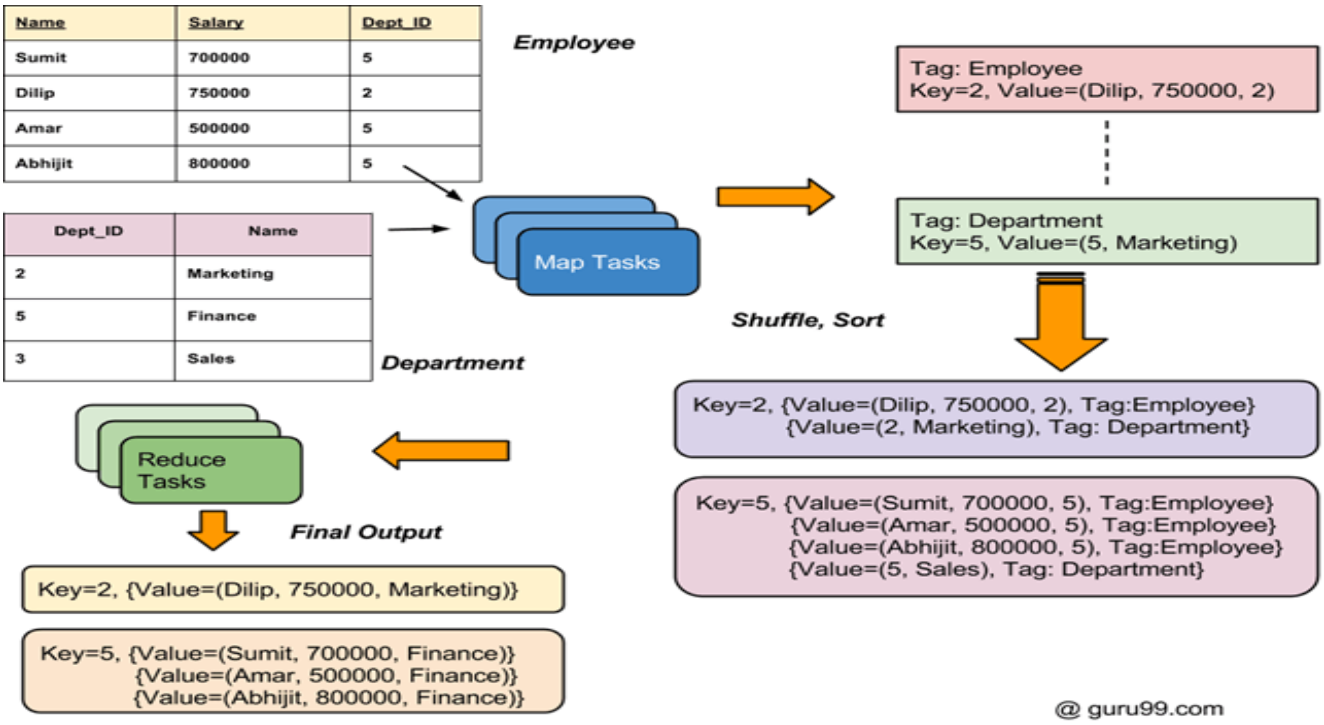
Joining of two datasets begin by comparing size of each dataset. If one dataset is smaller as compared to the other dataset then smaller dataset is distributed to every data node in the cluster. Once it is distributed, either Mapper or Reducer uses smaller dataset to perform lookup for matching records from large dataset and then combine those records to form output records. Depending upon the place where actual join is performed, this join is classified into-

1. Map-side join - When the join is performed by the mapper, it is called as map-side join. In this type, the join is performed before data is actually consumed by the map function. It is mandatory that the input to each map is in the form of a partition and is in sorted order. Also, there must be an equal number of partitions and it must be sorted by the join key.

2. Reduce-side join - When the join is performed by the reducer, it is called as reduce-side join. There is no necessity in this join to have dataset in a structured form (or partitioned).

Here, map side processing emits join key and corresponding tuples of both the tables. As an effect of this processing, all the tuples with same join key fall into the same reducer which then joins the records with same join key.

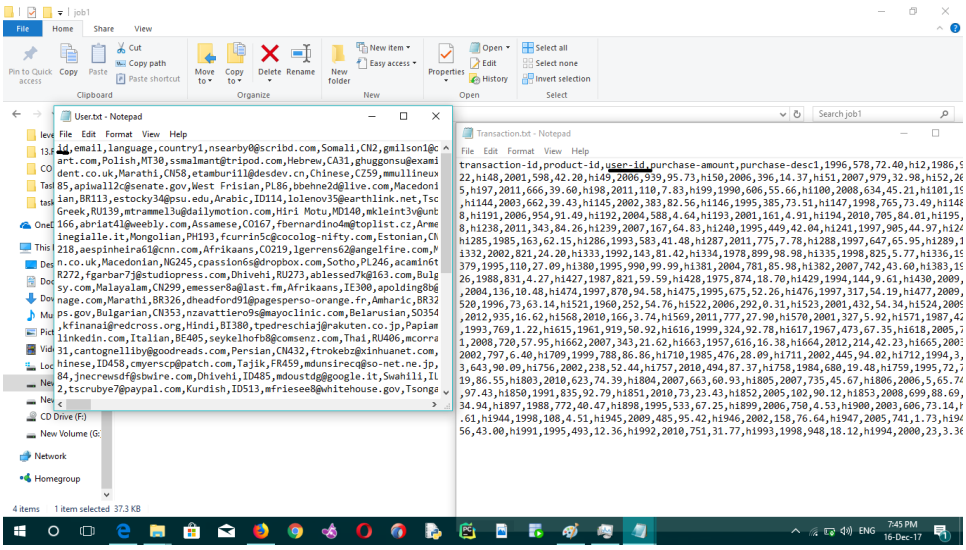
Overall process flow is depicted in below diagram.



MapReduce Hadoop Program To Join Data

Problem Statement:

There are 2 Sets of Data in 2 Different Files



The Key User_ID is common in both files.

The goal is to use MapReduce Join to combine these files

Input: Our input data set is a txt file, **User.txt & Transaction.txt**

MapReduce the second one

This one to get the result which expected which that get the number of locations in which the product has been sold this step will be finished when the output of the first mapreduce will be input to the second mapreduce

Mapper side :

Will do-nothing actually it will print the input to the standard output

Reducer side:

The reducer will take the standard output as standard input then will count the locations that have the same product then will print the product & the number of those locations.