Hadoop:-

1-Creating Datasets

Write a java program that creates two datasets (two files), Customers and Transactions. Each line in Customers file represents one customer, and each line in Transactions file represents one transaction.

The attributed within each line are comma separated.

The Customers dataset should have the following attributes for each customer:

* ID: unique sequential number (integer) from 1 to 50,000 (that is the file will have 50,000 line)
* Name: random sequence of characters of length between 10 and 20 (do not include commas)
* Age: random number (integer) between 10 to 70
* CountryCode: random number (integer) between 1 and 10
* Salary: random number (float) between 100 and 10000

The Transactions dataset should have the following attributes for each transaction:

* TransID: unique sequential number (integer) from 1 to 5,000,000 (the file has 5M transactions)
* CustID: References one of the customer IDs, i.e., from 1 to 50,000 (on Avg. a customer has 100 trans.)
* TransTotal: random number (float) between 10 and 1000
* TransNumItems: random number (integer) between 1 and 10
* TransDesc: random text of characters of length between 20 and 50 (do not include commas)

Writing MapReduce Jobs:

1. Write a job(s) that reports the customers whose CountryCode between 2 and 6 (inclusive).
2. Write a job(s) that reports for every customer, the number of transactions that customer did and the total sum of these transactions. The output file should have one line for each customer containing: CustomerID, NumTransactions, TotalSum You are required to use a Combiner in this query
3. Write a job(s) that joins the Customers and Transactions datasets (based on the customer ID) and reports for each customer the following info: CustomerID, Name, Salary, NumOf Transactions, TotalSum, MinItems Where NumOfTransactions is the total number of transactions done by the customer, TotalSum is the sum of field “TransTotal” for that customer, and MinItems is the minimum number of items in transactions done by the customer
4. Write a job(s) that reports for every country code, the number of customers having this code as well as the min and max of TransTotal fields for the transactions done by those customers. The output file should have one line for each country code containing: CountryCode, NumberOfCustomers, MinTransTotal, MaxTransTotal.
5. Write a job(s) that reports the customer names whose associated number of transactions is larger than the average number of transactions of all customers. That is, calculate the average number of transactions over all customers (Say value x), you goal is to report the customer names who have more than x transactions