**Scenario 1**

Attached are files received from two different transaction systems in a bank.



The data from both the systems come in slightly different format.

It provides, for each account number, transaction amounts as per transaction dates and transaction types. Name of account holder is masked.

**Problem Statement 1**

1. Ingest both the files to HDFS and write a map reduce program to find average transaction amount for all the credit and debit transactions. Note that a single program should execute to find the details. You may use multiple map classes as per design decision.

(Output required – map reduce program, snapshot of execution, snapshot of result)

1. Create a hive external table on the above ingested file and load the data to a managed table with partition on year. Using hive queries, find the average transaction amount per year for credit and debit transaction each.

(Output required – hive queries, snapshot of execution, and snapshot of result)

1. Merge both the files in PIG. In file “trnx\_data\_branch1.csv”, where channel is missing, add data value as “Missing” in the merged file. Using PIG or hive, mine the data to find account\_id whose balance amount is highest after all the transactions in both the files

(Output required – PIG & Hive script, snapshot of execution, snapshot of result)

**Scenario 2**

Install MYSQL database in your environment and create a table in MYSQL with a structure like:



**Problem Statement 2**

1. Load the merged HDFS file created in problem statement 1 (3) to this MySQL table using SQOOP

(Output required – snapshot of MYSQL table)

1. Import this data back to HBASE table (transaction data for year 2014 only) using SQOOP.

RowKey needs to be Account-id

Write a Java program using HBASE APIs to which will accept account-id as input and provides last 3 transaction details. This should be a program that runs on command line and ends only after use inputs account id as “000000”.

(Output required – Java program & snapshot of execution)