Assignment – Hive – 01

This assignment contains 2 parts.

Part 1 – Simulating the data

Part 2 – Creating the Hive tables, Loading the data created in part 1 and doing the exploratory analysis on the tables.

**Submission Deadline**:

28-05-2018 – 6:00 PM EST

Part 1:

The data should be simulated with help of programming language of your choice (if no programming language is known then use excel to simulate the data). Create 3 files and load them into hive.

Submission details:

The program should be uploaded in LMS, Hive load screen shot and sample 10 records screen should be provided in word document.

1. Create Account\_master table with below columns.

   Account number

   Account Opening Balance

   Ledger Balance

   Clearing Hold

   Account status code

   Country Code

   Account Currency

Account Creation TimeStamp

   Simulate the data for 1000 Account numbers.

   Account number  - 10 digit number

   Account Opening Balance  - random number between 100 to 1000000

   Ledger Balance           - random number between 100 to 1000000

   Clearing Hold            - random number between 100 to 1000

   Account status code      - Allowable values 1/8/5

   Country Code             - 3 Char

   Account Currency         - USD/INR/CAD

Account Creation TimeStamp - Timestamp

Account status code values significance      - 1 – Active / 8 – Opened /5 - Dormant

2. Create account\_name table with below columns

   Account Number

   First name

   Middle Name

   Last name

   Date of birth

   Joint account or not

   Joint Account Number

   Similar to above.

Note: The account number should be from one of the values from account numbers which were created in step 1.

3. Create transaction\_details table with below columns

   Account numnber

   Transaction Posting date

   Transaction capture date

   Transaction amout

   Debit or Credit

   Transaction code

    Transaction Curreny (USD/INR/CAD/HKD)

   Simulate the data for 3000 records.

   Account number    - Should be valid account number from account numnber list

   Transaction Posting date  - The date can be anything from 01-July-2017 to current date

   Transaction capture date  - The date can be anything from 01-July-2017 to current date

   Transaction amount         - 100 to 1500 any random number

   Debit or Credit           - D for debit or C for Credit

   Transaction code          - 10001 for debit or 2001 for debit

Transaction Amount - Should be able to store the amount fields with the precision of 2

**Step 2**:

**Requirement:**

Perform exploratory data analysis on the data which was created and loaded into Hive table.

What are the social and business values of those insights?

**Skill Set to be used:**

SQL skills to join multiple table, aggregations and other operations.

**What should be submitted:**

A word document with all the queries used and the insights with explanation.

**Sample Queries:**

1. How many foreign currency transactions were done?

    Join the account master table and transaction table if the account currency from account master table is not equal to transaction currency from transaction table.

2. How many accounts are there with each account status and what is the sum of ledger balance with respect to account status codes.

Account status code      - Allowable values 1/8/5

3. How many joint account holders are there and which are the countries they belong? i.e., prepare a country level report for joint account holders.

Please note that above are sample queries only. Please come up different scenarios and write the queries accordingly.