**MINI PROJECT**

**Project Goal:**

The goal of this project is to showcase how hadoop eco-system components are   
Inter- connected and how they can be leveraged to solve complex business problems.

We have **Health Insurance** related data coming as a csv file and few relational tables coming from mysql database. We have to integrate these and answer specific business questions using the hadoop framework.

The dataset HI.csv contains Health insurance data with working hours of wives.

**The description of the columns of the file is given below:**

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| whrswk | Hours worked per week by wife |
| hhi | Wife covered by husband's HI? |
| whi | Wife has HI thru her job? |
| hhi2 | Husband has HI thru own job? |
| education | A factor with levels Ex. <9years,9-11years;12years;13-15years;16years;>16years |
| race | One of W,L,O |
| hispanic | Hispanic? |
| experience | Years of potential work experience |
| kidslt6 | Number of kids under age of 6 |
| kids618 | Number of kids 6-18 years old |
| husby | Husband's income in thousands of dollars |
| region | One of O,NC,W,S |
| wght | Sampling weight |

Following tasks needs to be performed.

1. **Import Data**

1. Import tables both Race and Region from POC schema of Mysql and create the table hive using sqoop command.

**Note:** a. Make sure mysql services are up and running.

b. In HIVE a database needs to be created with Name “POC”.

1. **Clean Dataset**
2. Clean the dataset using Pig to eliminate the double quotes and discard the last column and store it on HDFS with '|' as delimiter.
3. Also create an external table in hive on this dataset. Create a normalised table by joining the external table and the imported table which can be used for analysis.
4. **Write the hive queries for below questions**

1. Display and count the number of records where both husband and wife do not have insurance. Which region has the maximum records in this case?

2. Display and count the number of records where husband has his HI trough his job and wife has been covered by her job and husband both. What is the min and max salary of husbands for this specific case?

3. At distinct education level, find the average experience and average working hours for only working wives.

4. At distinct education levels and race, find out the minimum salary and maximum salary and store it in an appropriate table.

5. Under each region what is the count of years of education, order the count in descending to show the highest count on top.

6. What are the Average working hours per week for women who have HI through her job by education factor?

7. Rank Women by years of education, meaning show women count by no. of years of education and rank them by count descending.

8. What is the average years of education based on the hours worked by wife in each region? Order it based on the region and descending working hrs

**Note:** Since education is in levels, write a UDF function, to convert the education level to an integer number. For the level with '<', decrease by 1 year, for '>', increase by a year. For education level in a range, take the average.

For example, <9years = 8; 9-11years = 10; >17years = 18

1. **Write the data onto a** **file**

Save this data onto a file where both wife and husband have been covered HI through their jobs and have at least one kid and finally export the data to MySQL.

**Here is the plan of execution on a high level:**

1. **Data Ingestion Phase:** 
   1. We are going to use **SQOOP** to bring relational data into HDFS.
   2. Using hadoop commands we will bring Health Insurance flat file into HDFS.
2. **Data Cleaning Phase:**
   1. In this phase we are going to USE **PIG Latin language** to read and cleanse the Health Insurance file.
3. **Data Integration Phase:**
   1. We are going to use **HIVE** to define the structure and make the Health insurance data available for querying.
   2. We are going to integrate both Health Insurance and Relational Data using tables in HIVE.
4. **Data Analysis Phase:** 
   1. Write analytical queries in **HIVE** to answer specific business questions.
   2. Write custom user defined functions in Java to convert some characters that are found in the Health insurance dataset.
5. **Data Export Phase:**
   1. Using **SQOOP** we are going to transfer data/results from HDFS to relational tables so that this result set can be combined or integrated with traditional data ware house/marts for further analysis.