CONSUMER COMPLAINT DATA ANALYSIS

In this Project, We have analyzed consumer complaint database for products,subproducts,company.state,zipcode.To analyze the data follow the steps. You can find the Insurance data in the same repository named as "ConsumerComplaintData.csv". We had run various queries, which are in the file "CODE".

Prerequisites:

Microsoft Azure account Click here for trial account

Cloud Berry for Azure To download click here

Memory Configuration:

working Node Configuration -

- no. of core- 4
- Ram -14 GB
- disks -8
- local SSD -200 GB

Total no. of working nodes: 2

Head Node Configuration -

- core- 4
- Ram -14 GB
- disks-8
- local SSD- 200 GB

Total no. of head nodes: 2

Steps:

1. After creating an account on Microsoft's Azure, create an HDInsight cluster.

- Go to Azure portal after creating an account in Azure.
- Click on New
- In Data+Analytics, Select HDInsight
- A new pop-up will appear, Type the cluster name. Select type as "Hadoop", Operating System a "Windows", Select Total node as 2, Put Login name and password,
- Click on *create*.
- Deployment of cluster will start. It will take few minutes to complete the process.

2. After Creating the cluster, we will upload data into the cluster.

3. To upload data in the cluster we will use Cloudberry for Azure.

- Install the cloudberry.
- Once cloudberry installed ,open cloudberry.
- Select Azure Blob In the dialog box, add the same details mentioned while creating cluster and upload file to the cluster.

4. After uploading the data, Go to the Azure where the cluster is created.

5.Click on URL OR ON HIVE TO open the hive editor.

6.Add details mentioned while creating cluster.

5. In Hive, First step is to create table.

- For, Creating the table we run this query in a query editor:
- CREATE TABLE CONSUMER_COMPLAINT_DATABASE1(Date_Received string,Product string,Sub_Product string,Issue String,Sub_issue string,Consumer_complaint_narrative string,Company_public_response string,Company string,State string,Zip_code int,Submitted_via string,Date_sent_to_company string,Company_response_to_consumer string,Timely_response string,Consumer_disputed string,Complaint_ID string)

  ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

  TBLPROPERTIES("skip.header.line.count=1");

- After the table is created, load data in the table by the following query:
- LOAD DATA INPATH '\HdiSamples\Consumer_Complaints.csv' OVERWRITE INTO TABLE CONSUMER_COMPLAINT_DATABASE1;
- Now data is ready to analyse.

6. Now, Analyse the data using following queries.

- Product which is popularly being complained about:

  select product,count(product) AS INC_OCCURE from CONSUMER_COMPLAINT_DATABASE1 GROUP BY product ORDER BY INC_OCCURE desc;

- Sub-Product which is popularly being complained about:

  select sub_product,count(sub_product) AS INC_OCCURE from CONSUMER_COMPLAINT_DATABASE1 GROUP BY sub_product ORDER BY INC_OCCURE desc;

- States where most issues are arising:

  select state,count(state) AS INC_OCCURE from CONSUMER_COMPLAINT_DATABASE1 GROUP BY state ORDER BY INC_OCCURE desc;

- Zip Code where most issues are arising:

  select zip_code,count(zip_code) AS INC_OCCURE from CONSUMER_COMPLAINT_DATABASE1 GROUP BY zip_code ORDER BY INC_OCCURE desc;

- Number of Complaints Per Company:

  select company,count(company) AS INC_OCCURE from CONSUMER_COMPLAINT_DATABASE1 GROUP BY company ORDER BY INC_OCCURE desc;

- Complaints Submitted via:
  select submitted_via,count(submitted_via) AS INC_OCCURE from CONSUMER_COMPLAINT_DATABASE1 GROUP BY product ORDER BY INC_OCCURE desc;
- Mortgage Complaints According to Year:
  SELECT YEAR,COUNT(YEAR) from DB3 GROUP BY YEAR;

7. After data is analysed in Hive, go to Microsoft excel.

- Select Data tab on top.
- From Other Sources.
- From Microsoft Query.
- Hive
- Add the columns
- Next, and finish.
- In Insert tab, Select power query.
- Project the data.