ADTA 5240 | Harvesting, Storing and Retrieving Data

Industry
Finance

Use Case
A Credit Card Company identifying precise customer for their product

Team
Indigo

Members
Sumana Sree Nandala
Mani Chandra Navuluri
Varun Kuman Chennuri
Rachana Bairi
Venu Gopalan Krishnagiri Tuppal

Instructor
Tony Fantasia

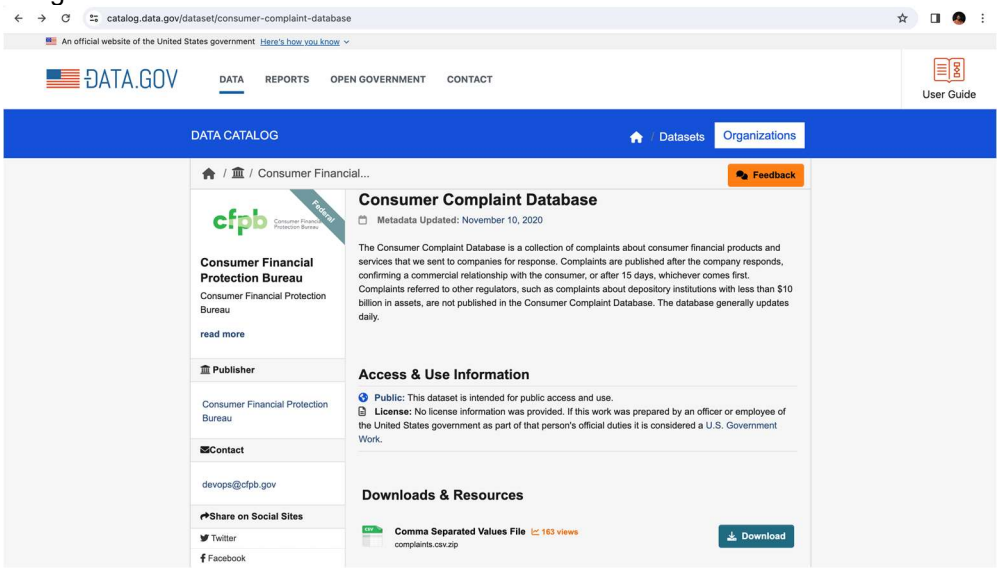
Introduction

With time, there is increase in people using Banking and Finance service. In particular, Credit Card usage has become one of the key services that plays a major role for any finance company. Thus, it is also needed to understand the areas that a company must focus when trying to implement a new product or upgrade an existing one to provide better consumer service and meet the requirements of the customers as efficiently as possible.

Data Source

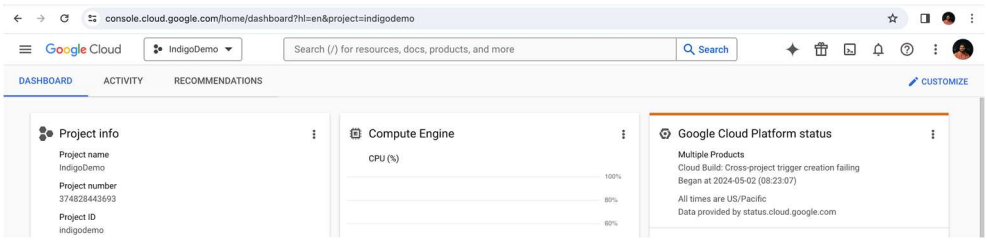
To achieve this, it is crucial that the service provider understands the issues and challenges in the existing products and services. For this, **Consumer Complaint Database** is selected which is collection of complaints and company responses to the customer. This data is published by Consumer Financial Protection Bureau for public access and use which is available in DATA.GOV(https://data.gov/). The data is available in multiple formats and for project Indigo, the CSV File **complaints.csv** is used, which consists of 18 Variables and 4998604 Observations covering various sectors and products of Banking and Finance.

URL for Dataset is as follows:
Consumer Complaint Database: https://catalog.data.gov/dataset/consumer-complaint-database
Reference Image:



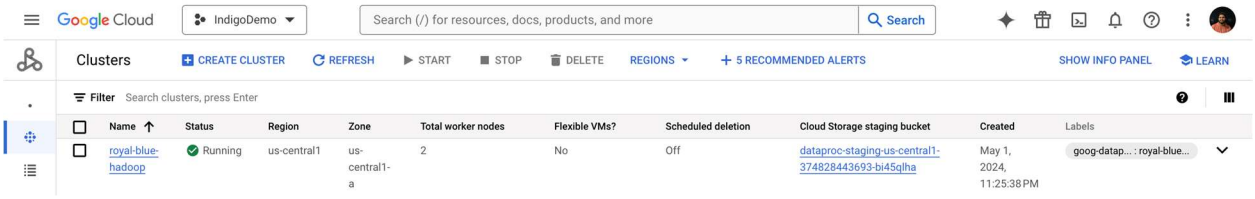
Google Cloud Platform (GCP) Project

For the study and implementation of the above use-case and to understand the complaints and issues of consumers using Credit Card Services, a new project **IndigoDemo** is created in GCP. Actions including Processing, Storage, Management are done in this project. Following screenshot is homepage for Project **IndigoDemo**

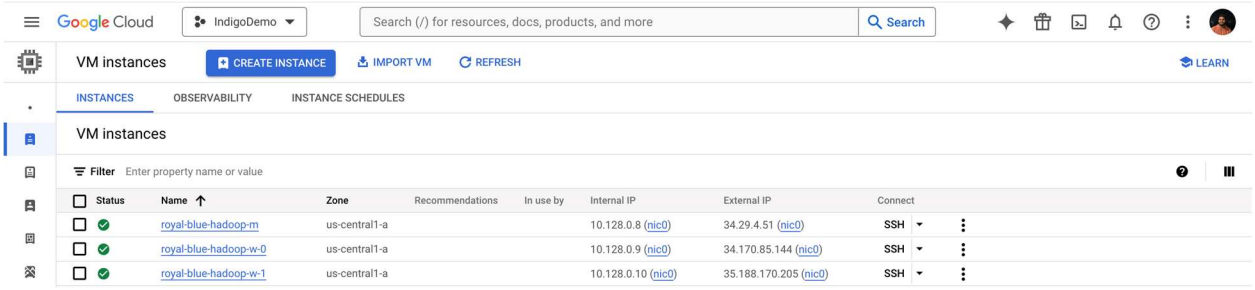


Hadoop Ecosystem Setup

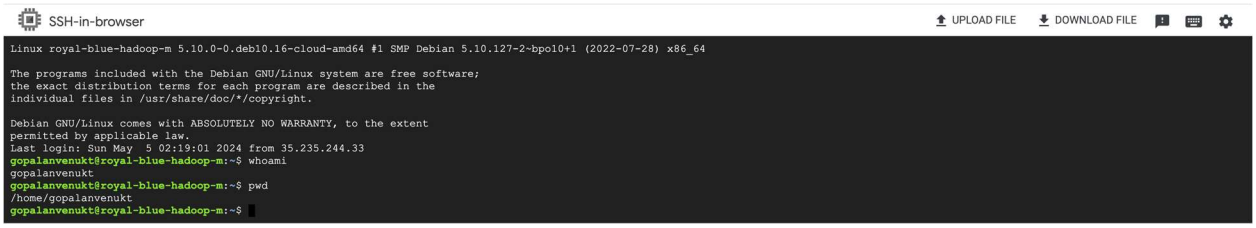
To Implement Hadoop Ecosystem, a Dataproc Cluster **royal-blue-hadoop** is created with 1-Master and 2-Worker Nodes (with same properties mentioned in Course material). Following Screenshot shows **royal-blue-hadoop** cluster which is running.



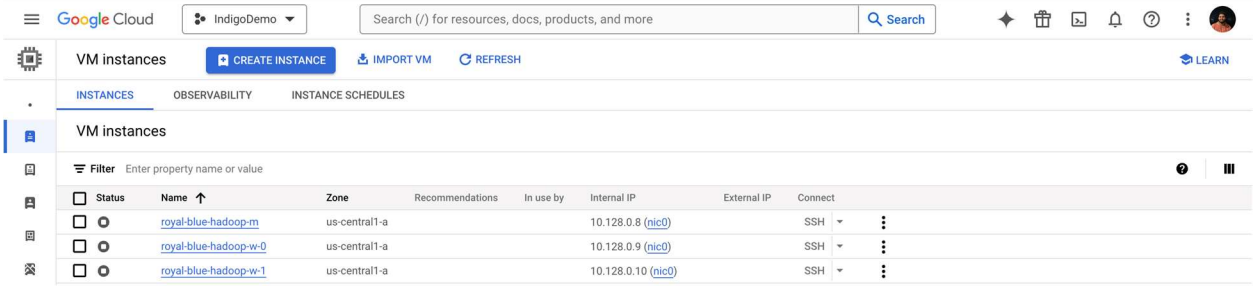
After the cluster is created, Compute Engine API is enabled (was already enabled in this case) and Virtual Machine (VM) Instances are checked to see if HDFS System is accessible or not (shown below)



A sample command is executed in the Linux Machine connected through SSH of Master Node to verify the working of nodes (shown below)

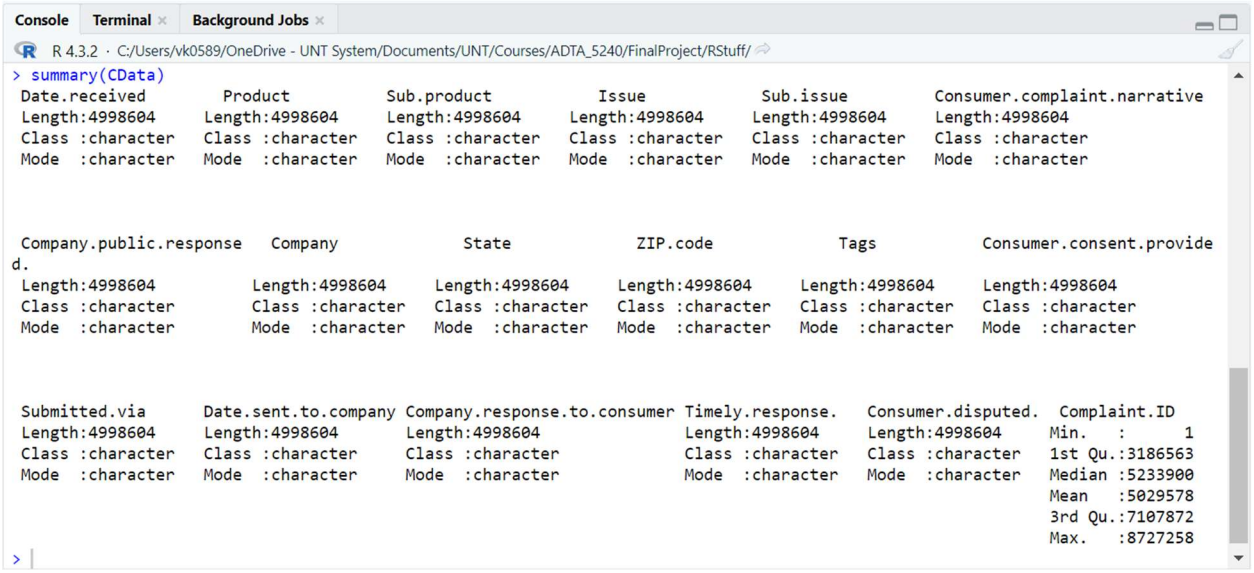


After verification, the Cluster/VM Instances is stopped as not further action is taken now (shown below)



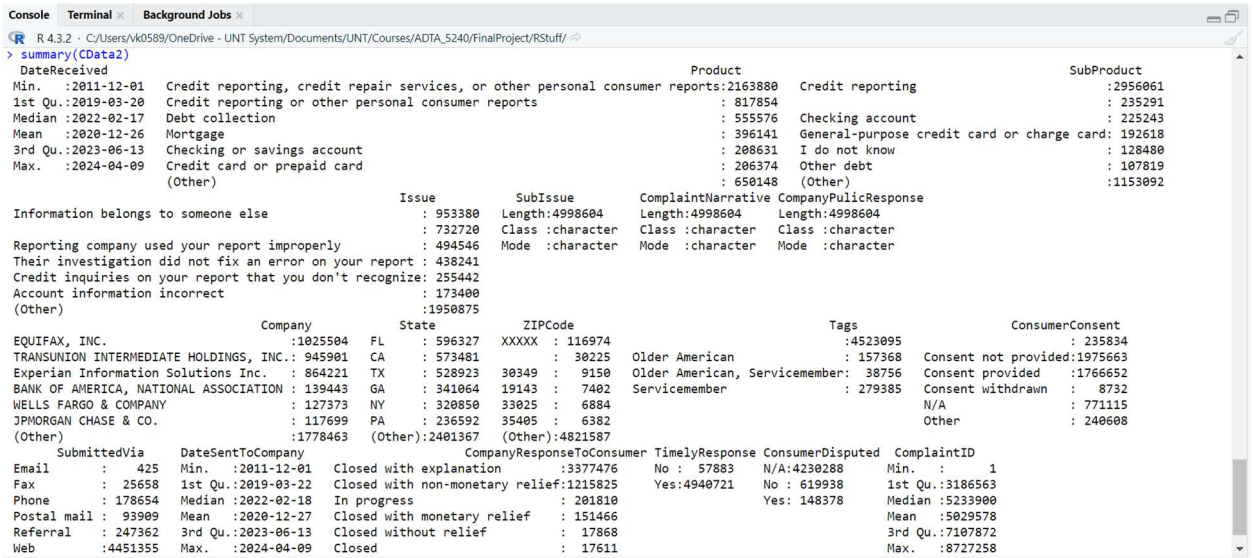
Data Processing

Firstly, data **complaints.csv** is loaded into RStudio (with R Programming) and saved as a dataframe **CData**. Following is the glimpse of CData:



```
Console  Terminal  Background Jobs
R 4.3.2 · C:/Users/vk0589/OneDrive - UNT System/Documents/UNT/Courses/ADTA_5240/FinalProject/RStuff/
> summary(CData)
 Date.received        Product           Sub.product           Issue            Sub.issue         Consumer.complaint.narrative
 Length:4998604     Length:4998604     Length:4998604     Length:4998604     Length:4998604     Length:4998604
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character


 Company.public.response   Company             State              ZIP.code           Tags           Consumer.consent.provide
d.
 Length:4998604          Length:4998604     Length:4998604     Length:4998604     Length:4998604     Length:4998604
 Class :character        Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character        Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character


 Submitted.via      Date.sent.to.company Company.response.to.consumer Timely.response.   Consumer.disputed.  Complaint.ID
 Length:4998604     Length:4998604       Length:4998604               Length:4998604     Length:4998604     Min.   :      1
 Class :character   Class :character     Class :character             Class :character   Class :character   1st Qu.:3186563
 Mode  :character   Mode  :character     Mode  :character             Mode  :character   Mode  :character   Median :5233900
                                                                                                            Mean   :5029578
                                                                                                            3rd Qu.:7107872
                                                                                                            Max.   :8727258
>
```

Copy of CData is taken as CData2 and Column Names are changed. Datatype Conversion (wrangling) is performed on CData2 to convert columns into categorical variables and following summary is obtained:



```
Console  Terminal  Background Jobs
R 4.3.2 · C:/Users/vk0589/OneDrive - UNT System/Documents/UNT/Courses/ADTA_5240/FinalProject/RStuff/
> summary(CData2)
  DateReceived                                                                  Product                                                  SubProduct
 Min.   :2011-12-01   Credit reporting, credit repair services, or other personal consumer reports:2163880   Credit reporting                        :2956061
 1st Qu.:2019-03-20   Credit reporting or other personal consumer reports                        : 817854                                    : 235291
 Median :2022-02-17   Debt collection                                                            : 555576   Checking account                 : 225243
 Mean   :2020-12-26   Mortgage                                                                   : 396141   General-purpose credit card or charge card: 192618
 3rd Qu.:2023-06-13   Checking or savings account                                                : 208631   I do not know                    : 128480
 Max.   :2024-04-09   Credit card or prepaid card                                                : 206374   Other debt                       : 107819
                      (Other)                                                                    : 650148   (Other)                          :1153092
                                                  Issue          SubIssue        ComplaintNarrative CompanyPulicResponse
 Information belongs to someone else                 : 953380   Length:4998604     Length:4998604     Length:4998604
                                                     : 732720   Class :character   Class :character   Class :character
 Reporting company used your report improperly       : 494546   Mode  :character   Mode  :character   Mode  :character
 Their investigation did not fix an error on your report : 438241
 Credit inquiries on your report that you don't recognize: 255442
 Account information incorrect                       : 173400
 (Other)                                             :1950875
                                Company           State           ZIPCode                                Tags                    ConsumerConsent
 EQUIFAX, INC.                     :1025504   FL     : 596327   XXXXX  : 116974                          :4523095                  : 235834
 TRANSUNION INTERMEDIATE HOLDINGS, INC.: 945901   CA     : 573481          :  30225   Older American              : 157368   Consent not provided:1975663
 Experian Information Solutions Inc.   : 864221   TX     : 528923   30349  :   9150   Older American, Servicemember:  38756   Consent provided    :1766652
 BANK OF AMERICA, NATIONAL ASSOCIATION : 139443   GA     : 341064   19143  :   7402   Servicemember               : 279385   Consent withdrawn   :   8732
 WELLS FARGO & COMPANY                 : 127373   NY     : 320850   33025  :   6884                                          N/A                 : 771115
 JPMORGAN CHASE & CO.                  : 117699   PA     : 236592   35405  :   6382                                          Other               : 240608
 (Other)                               :1778463   (Other):2401367   (Other):4821587
      SubmittedVia     DateSentToCompany                  CompanyResponseToConsumer TimelyResponse ConsumerDisputed  ComplaintID
 Email       :    425   Min.   :2011-12-01   Closed with explanation       :3377476     No :  57883    N/A:4230288      Min.   :      1
 Fax         :  25658   1st Qu.:2019-03-22   Closed with non-monetary relief:1215825     Yes:4940721    No : 619938      1st Qu.:3186563
 Phone       : 178654   Median :2022-02-18   In progress                   : 201810                    Yes: 148378      Median :5233900
 Postal mail :  93909   Mean   :2020-12-27   Closed with monetary relief   : 151466                                     Mean   :5029578
 Referral    : 247362   3rd Qu.:2023-06-13   Closed without relief         :  17868                                     3rd Qu.:7107872
 Web         :4451355   Max.   :2024-04-09   Closed                        :  17611                                     Max.   :8727258
```

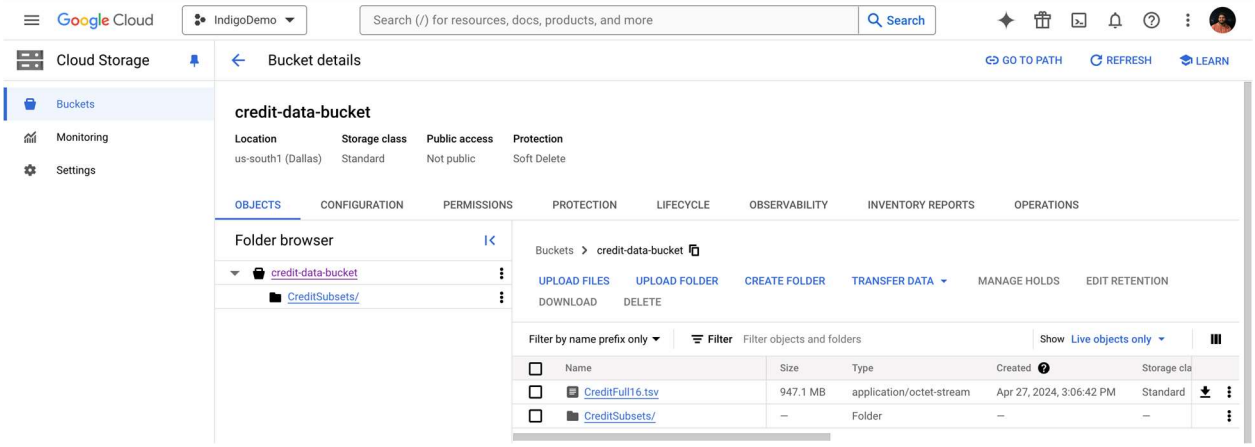
Only the necessary Columns and Product Types is selected from the whole data (Product has 21 Categories from which 5 are related to Credit Card). This new dataframe is then exported as a Tab-Separated Value (.tsv) File which is uploaded to Bucket in Google Cloud Platform (GCP).

(R Script File contains the commands for exporting the excel file which is attached as additional documents which includes the outputs of newly created dataframes)

Google Cloud Platform (GCP) Operations

Cloud Storage and Management

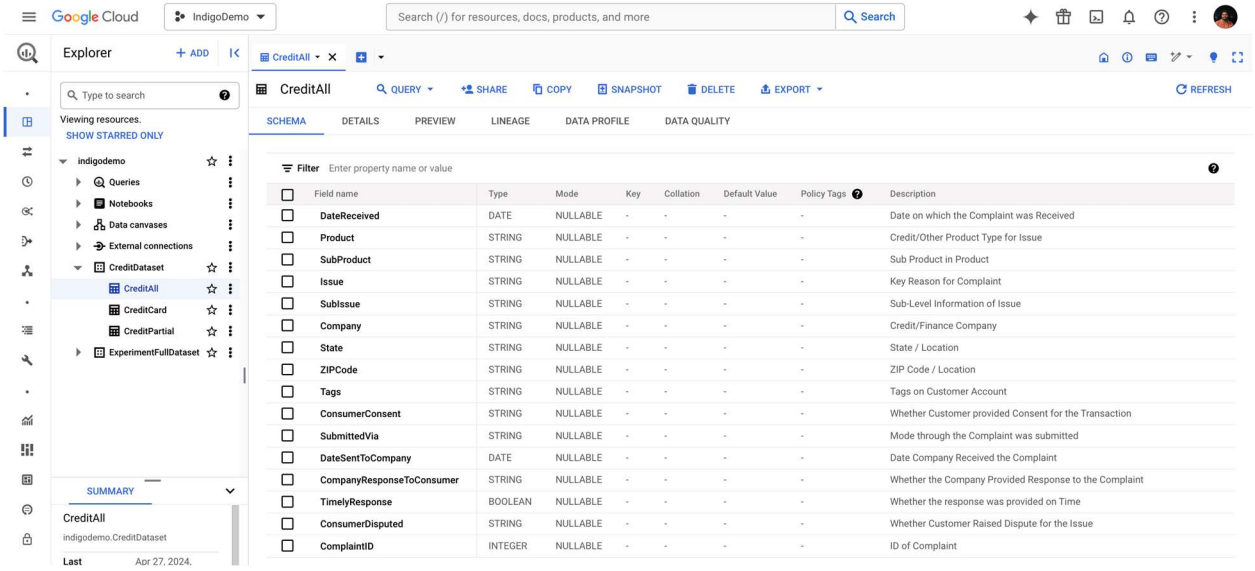
Following the instructions mentioned in the Course Material, a new Bucket – **credit-data-bucket** is created under project **IndigoDemo** for storing and managing data**,** which contains a folder **CreditSubsets** and **CreditFull16.tsv** file. **CreditFull16.tsv** is output file exported from RStudio that contains the required 16 Columns and all 5 Product Categories for Credit Card. Following Screenshot is the image for **credit-data-bucket** with CreditFull16.tsv File



BigQuery

BigQuery Data Loading

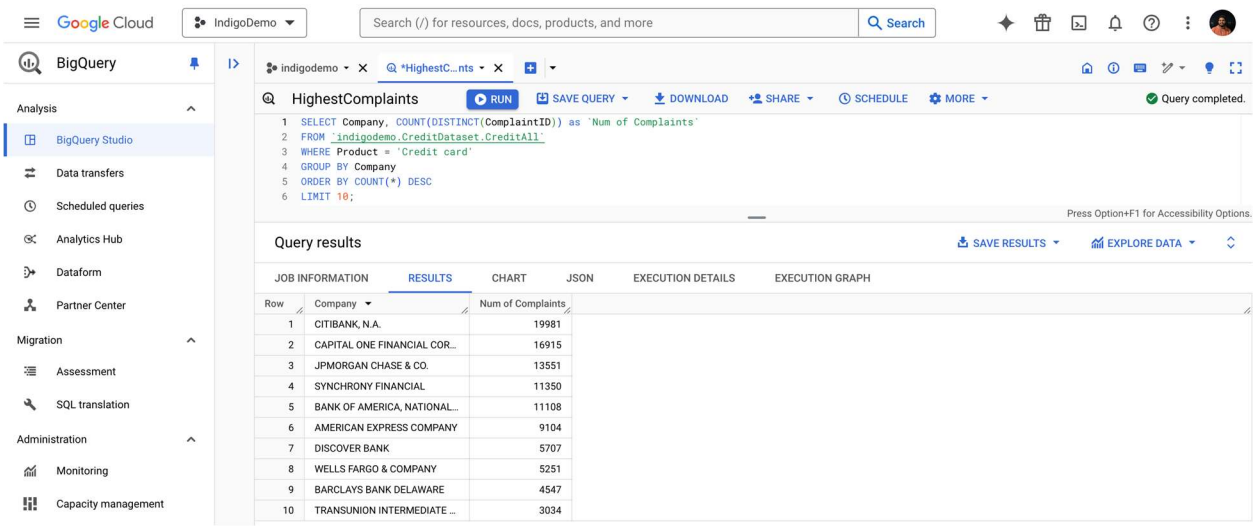
The CreditFull16.tsv is transferred to BigQuery Studio and stored as a Table. For this, a dataset **CreditDataset** is created and CreditFull16.tsv is copied to BigQuery as **CreditAll** Table. Following Screenshot shows the Dataset and Tables created in IndigoDemo Project along with the schema of **CreditAll** Table:



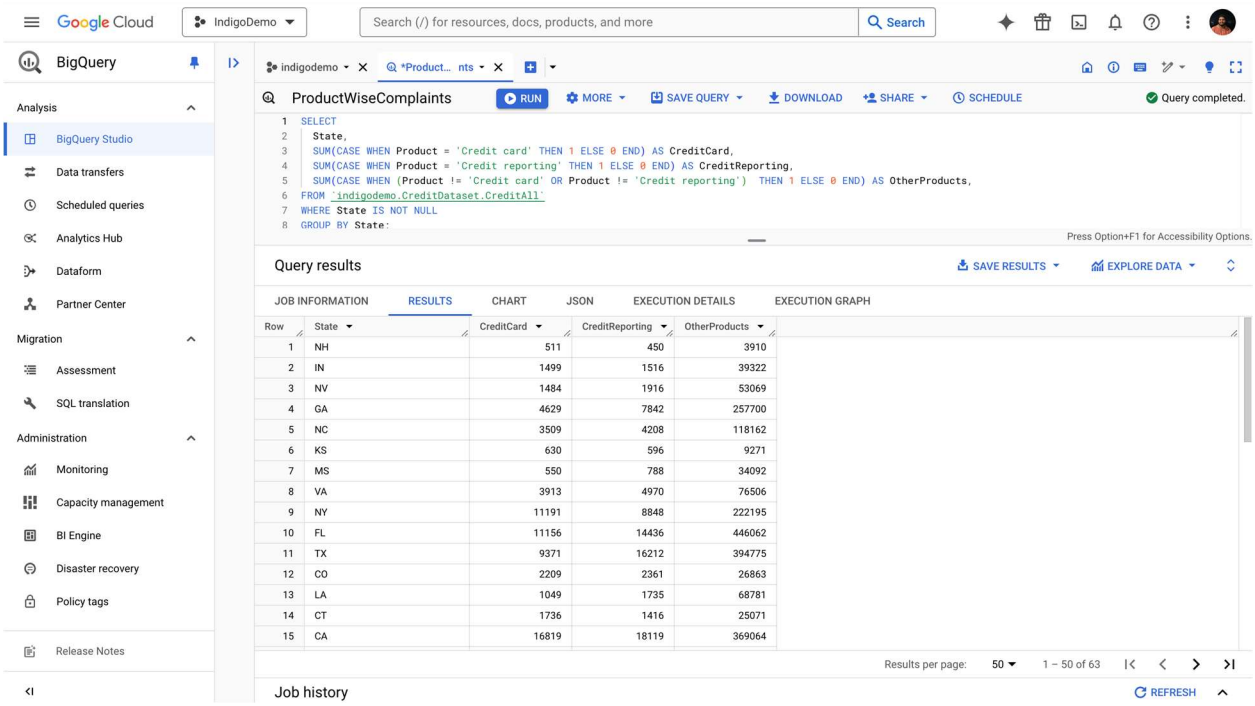
BigQuery Implementation

To explore the data and understand Credit Data further, following Queries were executed to get information and results for Credit Data

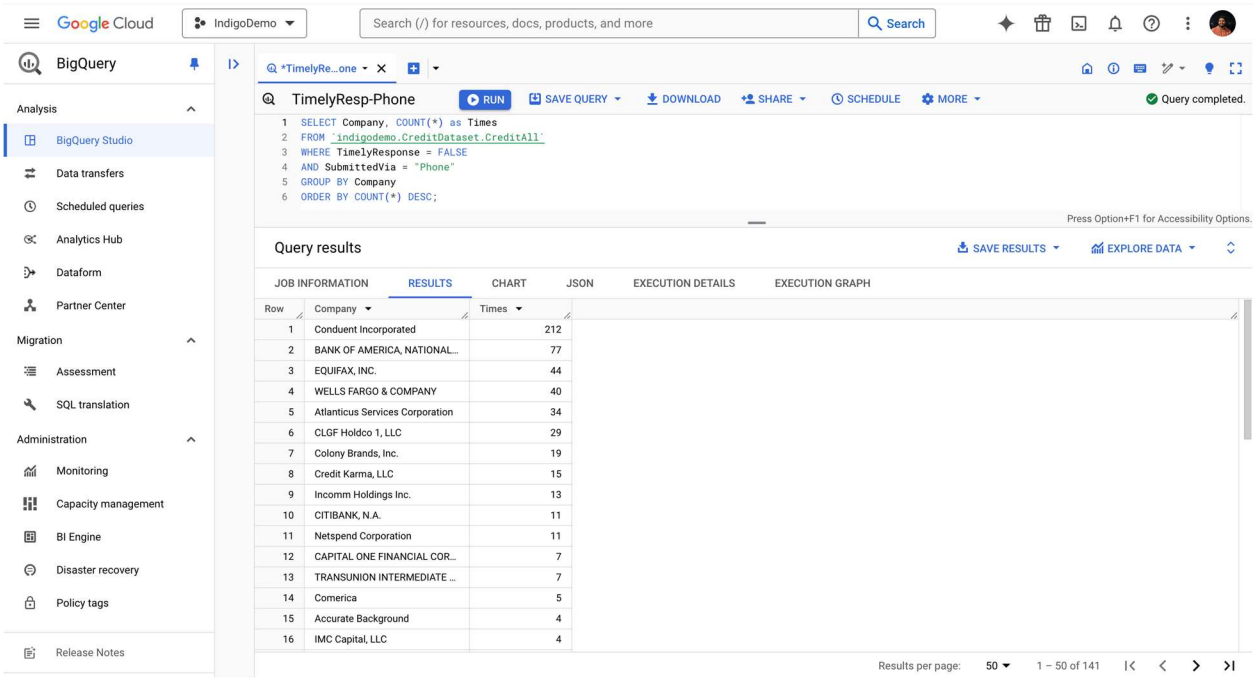
- Query to find the Highest Number of Complaints for ‘Credit card’ Product based on each Company. Below Screenshot gives the query and result for Top 10 Companies with highest complaints



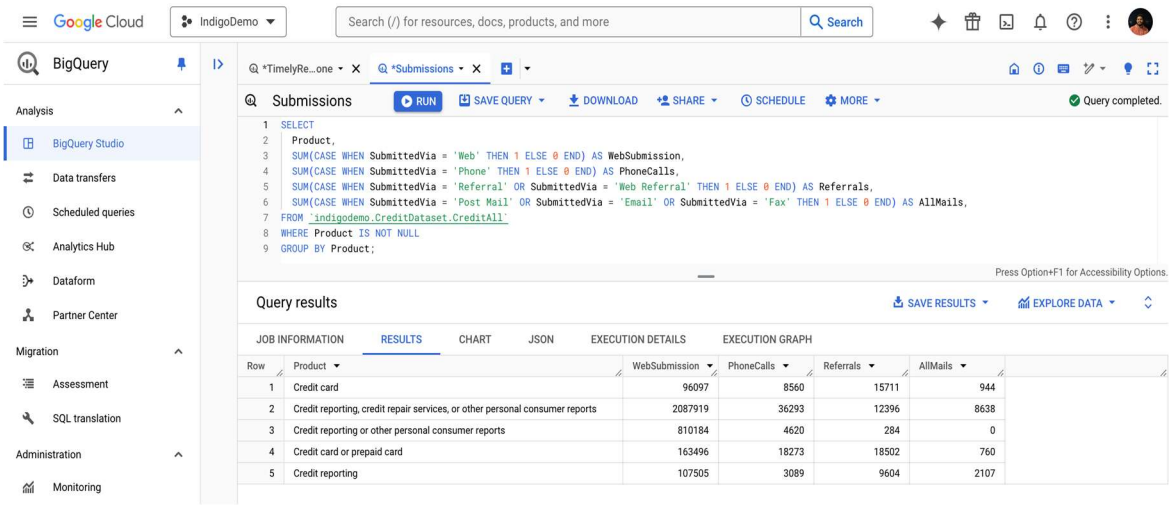
- Query to find the Number of Complaints under each Product Category State-Wise. Below Screenshot includes the query and result for Complaints under ‘Credit card’, ‘Credit Reporting’ and ‘Other Products’ for each State



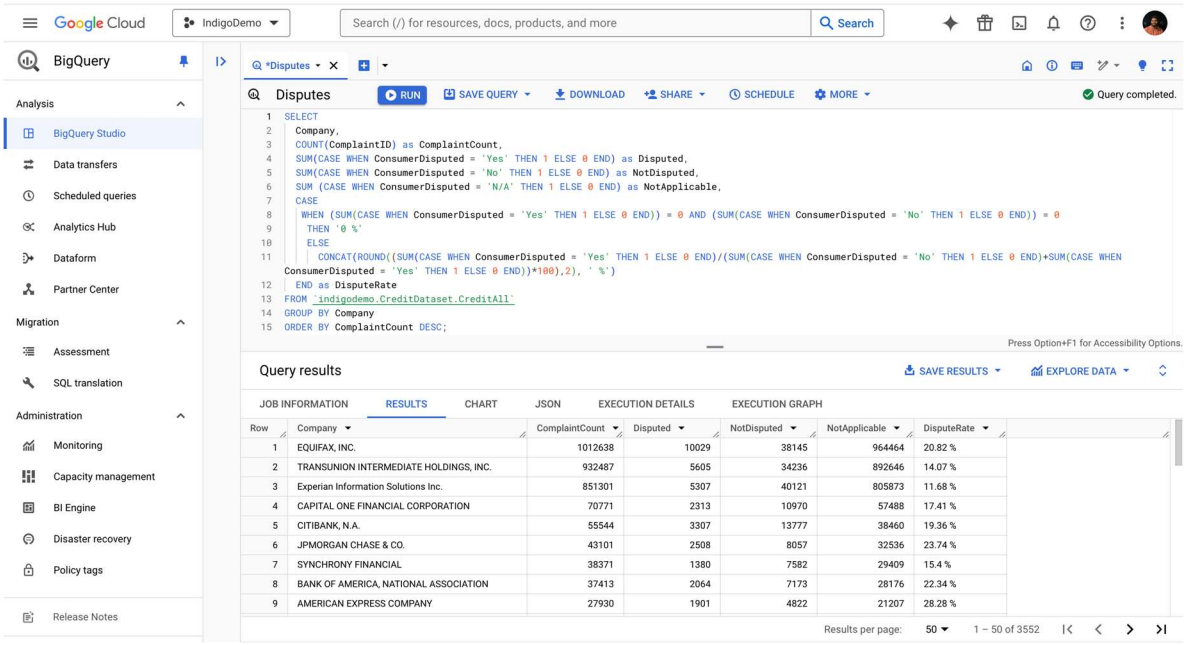
- Query for Case – “Which Company did not Respond on Time when Complaint was submitted through Phone Call?”. For the above use-case question, following Screenshot is the query and result with Complaint Count to match the given conditions



- Query for getting the number of Complaints under each Product and how they were submitted. Below Screenshot gives the query and result for number of Complaints and their mode of Submission under each Product (excluding *NULL* values)

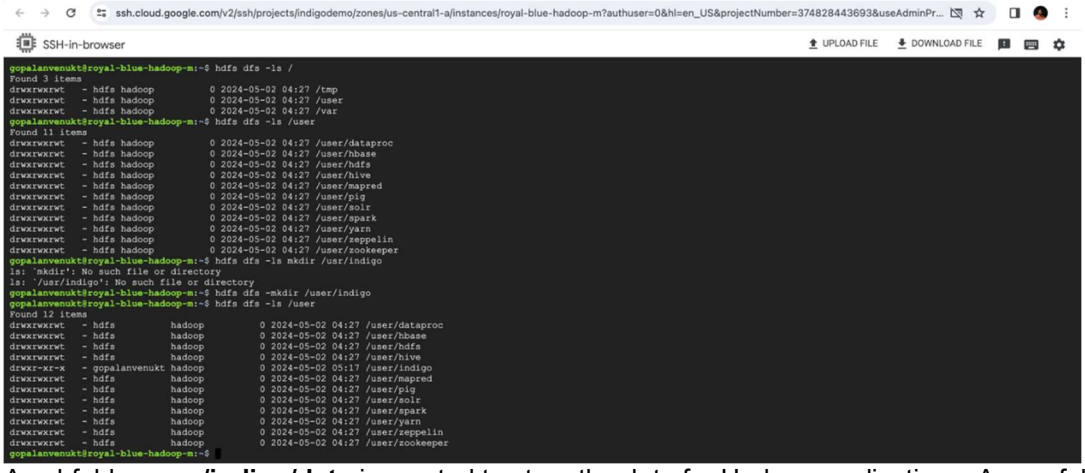


- Query to find the Rate of Disputes for each Company. Below Screenshot includes the query and result for Number of Complaints for which Dispute was raised or not by Consumer, Complaints for which Dispute is 'Not Applicable' and Dispute Rate (Disputed / Dispute-Applicable)

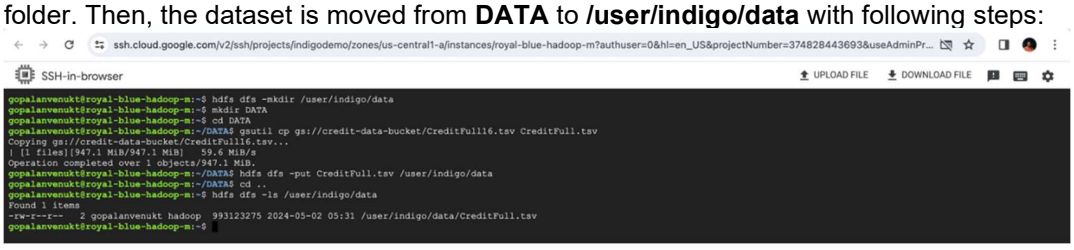


Hadoop File System

Following the instructions mentioned in the Course Material, files present in HDFS is accessed through Master Node's SSH Connection. A new Folder for Indigo Project is created - **user/indigo** (Screenshots for Reference)



A subfolder **user/indigo/data** is created to store the data for Hadoop applications. A new folder **DATA** is created in the Master Node to copy the dataset from Google Cloud Storage **credit-data-bucket** to **DATA** folder. Then, the dataset is moved from **DATA** to **/user/indigo/data** with following steps:



Hive and Spark Implementation

To work with Hadoop Hive and Spark, the above dataset *CreditFull.tsv* that was moved to **/user/indigo/data** is copied by creating a new table *CreditComplaints* using Hive. The Table creation is as follows:

SSH-in-browser

```
gopalanvenukt@royal-blue-hadoop-m:~$ beeline -u jdbc:hive2://localhost:10000
Connecting to jdbc:hive2://localhost:10000
Connected to: Apache Hive (version 3.1.3)
Driver: Hive JDBC (version 3.1.3)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 3.1.3 by Apache Hive
0: jdbc:hive2://localhost:10000> CREATE EXTERNAL TABLE IF NOT EXISTS CreditComplaints
. . . . . .> (
. . . . . .> `DateReceived` DATE,
. . . . . .> `Product` STRING,
. . . . . .> `SubProduct` STRING,
. . . . . .> `Issue` STRING,
. . . . . .> `SubIssue` STRING,
. . . . . .> `Company` STRING,
. . . . . .> `State` STRING,
. . . . . .> `ZIPCode` STRING,
. . . . . .> `Tags` STRING,
. . . . . .> `ConsumerConsent` STRING,
. . . . . .> `SubmittedVia` STRING,
. . . . . .> `DateSentToCompany` DATE,
. . . . . .> `CompanyResponseToConsumer` STRING,
. . . . . .> `TimelyResponse` STRING,
. . . . . .> `ConsumerDisputed` STRING,
. . . . . .> `ComplaintID` INT
. . . . . .> )
. . . . . .> ROW FORMAT DELIMITED
. . . . . .> FIELDS TERMINATED BY '\t'
. . . . . .> STORED AS TEXTFILE
. . . . . .> LOCATION '/user/indigo/data/';
INFO  : Compiling command(queryId=hive_20240502232859_52de6dca-48b0-415b-8589-cb7a82c52d30): CREATE EXTERNAL TABLE IF NOT EXISTS CreditComplaints
(
`DateReceived` DATE,
`Product` STRING,
`SubProduct` STRING,
`Issue` STRING,
`SubIssue` STRING,
`Company` STRING,
`State` STRING,
`ZIPCode` STRING,
`Tags` STRING,
`ConsumerConsent` STRING,
`SubmittedVia` STRING,
`DateSentToCompany` DATE,
`CompanyResponseToConsumer` STRING,
`TimelyResponse` STRING,
`ConsumerDisputed` STRING,
`ComplaintID` INT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE
LOCATION '/user/indigo/data/'
INFO  : Concurrency mode is disabled, not creating a lock manager
```

After creating the table *CreditComplaints*, the creation confirmation is done with Hive/SQL Command as follows:

ssh.cloud.google.com/v2/ssh/projects/indigodemo/zones/us-central1-a/instances/royal-blue-hadoop-m?authuser=0&hl=en_US&projectNumber=374828443693&useAdminPr...

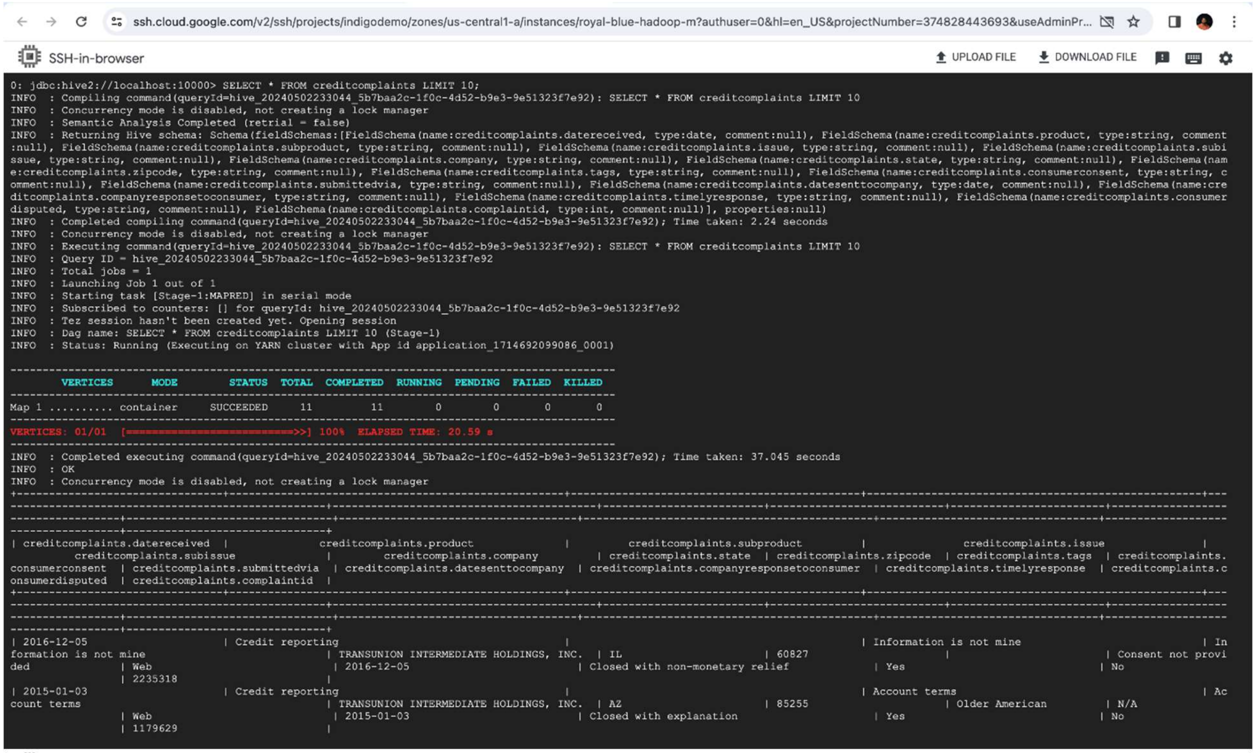
SSH-in-browser UPLOAD FILE DOWNLOAD FILE

```
0: jdbc:hive2://localhost:10000> show tables;
INFO  : Compiling command(queryId=hive_20240502232950_49f83327-d86c-47d6-a87d-838a8d9e0ad1): show tables
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20240502232950_49f83327-d86c-47d6-a87d-838a8d9e0ad1); Time taken: 0.144 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240502232950_49f83327-d86c-47d6-a87d-838a8d9e0ad1): show tables
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20240502232950_49f83327-d86c-47d6-a87d-838a8d9e0ad1); Time taken: 0.033 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------------------+
|     tab_name      |
+-------------------+
| creditcomplaints  |
+-------------------+
1 row selected (0.387 seconds)
```

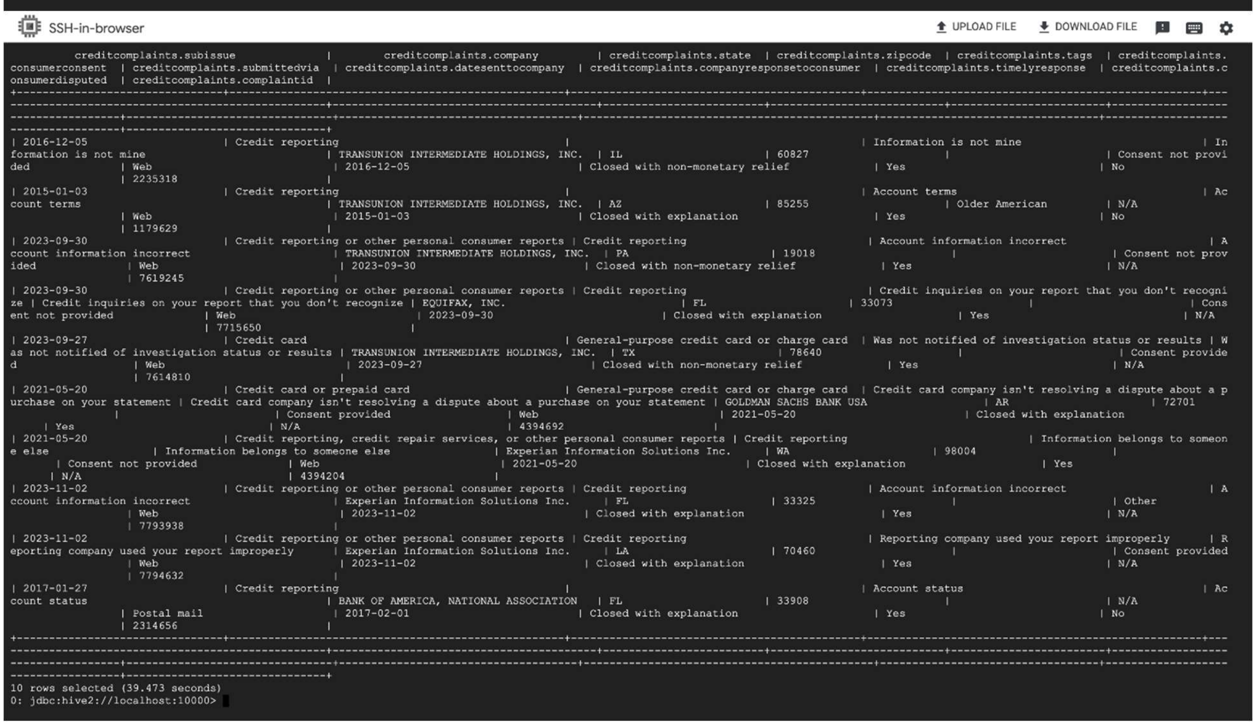
Hive

After table Creation, Queries are executed to verify the Performance and meet the requests which are as follows:

- Query to check the contents of the *CreditComplaints* Table (gives Top 10 Rows of *CreditComplaints* Table):

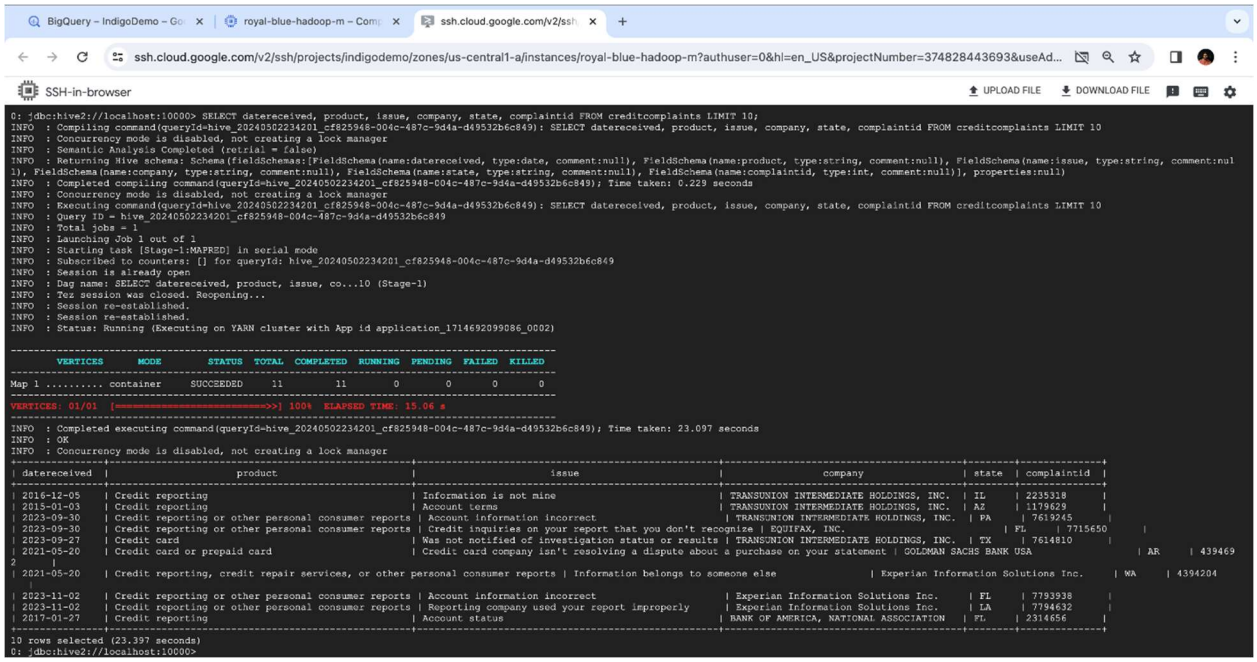


```
ssh.cloud.google.com/v2/ssh/projects/indigodemo/zones/us-central1-a/instances/royal-blue-hadoop-m?authuser=0&hl=en_US&projectNumber=374828443693&useAdminPr...
SSH-in-browser                                                                    UPLOAD FILE   DOWNLOAD FILE

0: jdbc:hive2://localhost:10000> SELECT * FROM creditcomplaints LIMIT 10;
INFO  : Compiling command(queryId=hive_20240502233044_5b7baa2c-1f0c-4d52-b9e3-9e51323f7e92): SELECT * FROM creditcomplaints LIMIT 10
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:creditcomplaints.datereceived, type:date, comment:null), FieldSchema(name:creditcomplaints.product, type:string, comment
:null), FieldSchema(name:creditcomplaints.subproduct, type:string, comment:null), FieldSchema(name:creditcomplaints.issue, type:string, comment:null), FieldSchema(name:creditcomplaints.subi
ssue, type:string, comment:null), FieldSchema(name:creditcomplaints.company, type:string, comment:null), FieldSchema(name:creditcomplaints.state, type:string, comment:null), FieldSchema(nam
e:creditcomplaints.zipcode, type:string, comment:null), FieldSchema(name:creditcomplaints.tags, type:string, comment:null), FieldSchema(name:creditcomplaints.consumerconsent, type:string, c
omment:null), FieldSchema(name:creditcomplaints.submittedvia, type:string, comment:null), FieldSchema(name:creditcomplaints.datesenttocompany, type:date, comment:null), FieldSchema(name:cre
ditcomplaints.companyresponsetoconsumer, type:string, comment:null), FieldSchema(name:creditcomplaints.timelyresponse, type:string, comment:null), FieldSchema(name:creditcomplaints.consumer
disputed, type:string, comment:null), FieldSchema(name:creditcomplaints.complaintid, type:int, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20240502233044_5b7baa2c-1f0c-4d52-b9e3-9e51323f7e92); Time taken: 2.24 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240502233044_5b7baa2c-1f0c-4d52-b9e3-9e51323f7e92): SELECT * FROM creditcomplaints LIMIT 10
INFO  : Query ID = hive_20240502233044_5b7baa2c-1f0c-4d52-b9e3-9e51323f7e92
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20240502233044_5b7baa2c-1f0c-4d52-b9e3-9e51323f7e92
INFO  : Tez session hasn't been created yet. Opening session
INFO  : Dag name: SELECT * FROM creditcomplaints LIMIT 10 (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1714692099086_0001)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     11         11        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 20.59 s
----------------------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20240502233044_5b7baa2c-1f0c-4d52-b9e3-9e51323f7e92); Time taken: 37.045 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------------------------------+---------------------------------------------------+---------------------------------------------+--------------------------------------------------+---
------------------------------------------------+------------------------------------------+-------------------------+---------------------------+-----------------------+------------------
----------------+-------------------------------+-----------------------------------+------------------------------------------------+------------------------------+------------------
----------------+-------------------------------+
| creditcomplaints.datereceived  |              creditcomplaints.product              |          creditcomplaints.subproduct          |               creditcomplaints.issue               |
        creditcomplaints.subissue                 |         creditcomplaints.company          | creditcomplaints.state  | creditcomplaints.zipcode  | creditcomplaints.tags  | creditcomplaints.
consumerconsent  | creditcomplaints.submittedvia  | creditcomplaints.datesenttocompany  | creditcomplaints.companyresponsetoconsumer  | creditcomplaints.timelyresponse  | creditcomplaints.c
onsumerdisputed  | creditcomplaints.complaintid  |
+-------------------------------+---------------------------------------------------+---------------------------------------------+--------------------------------------------------+---
------------------------------------------------+------------------------------------------+-------------------------+---------------------------+-----------------------+------------------
----------------+-------------------------------+-----------------------------------+------------------------------------------------+------------------------------+------------------
----------------+-------------------------------+
| 2016-12-05                     | Credit reporting                                   |                                               | Information is not mine                            | In
formation is not mine                             | TRANSUNION INTERMEDIATE HOLDINGS, INC.  | IL                      | 60827                     |                        | Consent not provi
ded              | Web                            | 2016-12-05                          | Closed with non-monetary relief             | Yes                              | No
                 | 2235318                        |
| 2015-01-03                     | Credit reporting                                   |                                               | Account terms                                      | Ac
count terms                                       | TRANSUNION INTERMEDIATE HOLDINGS, INC.  | AZ                      | 85255                     | Older American         | N/A
                 | Web                            | 2015-01-03                          | Closed with explanation                     | Yes                              | No
                 | 1179629                        |
```



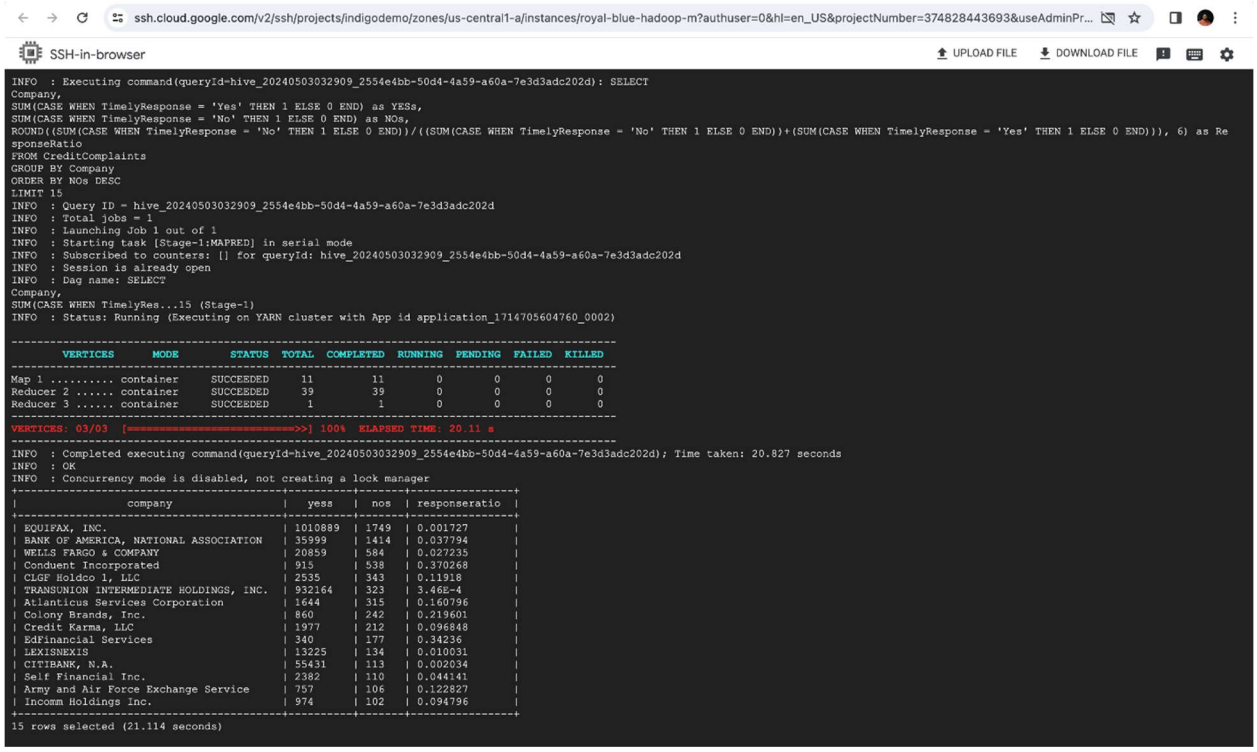
```
SSH-in-browser                                                                    UPLOAD FILE   DOWNLOAD FILE

        creditcomplaints.subissue                 |         creditcomplaints.company          | creditcomplaints.state  | creditcomplaints.zipcode  | creditcomplaints.tags  | creditcomplaints.
consumerconsent  | creditcomplaints.submittedvia  | creditcomplaints.datesenttocompany  | creditcomplaints.companyresponsetoconsumer  | creditcomplaints.timelyresponse  | creditcomplaints.c
onsumerdisputed  | creditcomplaints.complaintid  |
+-------------------------------+---------------------------------------------------+---------------------------------------------+--------------------------------------------------+---
------------------------------------------------+------------------------------------------+-------------------------+---------------------------+-----------------------+------------------
----------------+-------------------------------+-----------------------------------+------------------------------------------------+------------------------------+------------------
----------------+-------------------------------+
| 2016-12-05                     | Credit reporting                                   |                                               | Information is not mine                            | In
formation is not mine                             | TRANSUNION INTERMEDIATE HOLDINGS, INC.  | IL                      | 60827                     |                        | Consent not provi
ded              | Web                            | 2016-12-05                          | Closed with non-monetary relief             | Yes                              | No
                 | 2235318                        |
| 2015-01-03                     | Credit reporting                                   |                                               | Account terms                                      | Ac
count terms                                       | TRANSUNION INTERMEDIATE HOLDINGS, INC.  | AZ                      | 85255                     | Older American         | N/A
                 | Web                            | 2015-01-03                          | Closed with explanation                     | Yes                              | No
                 | 1179629                        |
| 2023-09-30                     | Credit reporting or other personal consumer reports | Credit reporting                             | Account information incorrect                      | A
ccount information incorrect                       | TRANSUNION INTERMEDIATE HOLDINGS, INC.  | PA                      | 19018                     |                        | Consent not prov
ided              | Web                            | 2023-09-30                          | Closed with non-monetary relief             | Yes                              | N/A
                  | 7619245                        |
| 2023-09-30                     | Credit reporting or other personal consumer reports | Credit reporting                             | Credit inquiries on your report that you don't recogni
ze | Credit inquiries on your report that you don't recognize | EQUIFAX, INC.                  | FL                      | 33073                     |                        | Cons
ent not provided                     | Web                            | 2023-09-30                          | Closed with explanation                     | Yes                              | N/A
                                     | 7715650                        |
| 2023-09-27                     | Credit card                                        | General-purpose credit card or charge card  | Was not notified of investigation status or results | W
as not notified of investigation status or results | TRANSUNION INTERMEDIATE HOLDINGS, INC.  | TX                      | 78640                     |                        | Consent provide
d                  | Web                            | 2023-09-27                          | Closed with non-monetary relief             | Yes                              | N/A
                   | 7614810                        |
| 2021-05-20                     | Credit card or prepaid card                        | General-purpose credit card or charge card  | Credit card company isn't resolving a dispute about a p
urchase on your statement | Credit card company isn't resolving a dispute about a purchase on your statement | GOLDMAN SACHS BANK USA                  | AR                      | 72701
                 |                        | Consent provided                     | Web                            | 2021-05-20                          | Closed with explanation
     | Yes                              | N/A                              | 4394692                        |
| 2021-05-20                     | Credit reporting, credit repair services, or other personal consumer reports | Credit reporting                             | Information belongs to someon
e else                 | Information belongs to someone else                | Experian Information Solutions Inc.     | WA                      | 98004                     |
        | Consent not provided                 | Web                            | 2021-05-20                          | Closed with explanation                     | Yes
       | N/A                                   | 4394204                        |
| 2023-11-02                     | Credit reporting or other personal consumer reports | Credit reporting                             | Account information incorrect                      | A
ccount information incorrect                       | Experian Information Solutions Inc.     | FL                      | 33325                     |                        | Other
                  | Web                            | 2023-11-02                          | Closed with explanation                     | Yes                              | N/A
                  | 7793938                        |
| 2023-11-02                     | Credit reporting or other personal consumer reports | Credit reporting                             | Reporting company used your report improperly      | R
eporting company used your report improperly       | Experian Information Solutions Inc.     | LA                      | 70460                     |                        | Consent provided
                  | Web                            | 2023-11-02                          | Closed with explanation                     | Yes                              | N/A
                  | 7794632                        |
| 2017-01-27                     | Credit reporting                                   |                                               | Account status                                     | Ac
count status                                      | BANK OF AMERICA, NATIONAL ASSOCIATION   | FL                      | 33908                     |                        | N/A
                 | Postal mail                    | 2017-02-01                          | Closed with explanation                     | Yes                              | No
                 | 2314656                        |
+-------------------------------+---------------------------------------------------+---------------------------------------------+--------------------------------------------------+---
------------------------------------------------+------------------------------------------+-------------------------+---------------------------+-----------------------+------------------
----------------+-------------------------------+-----------------------------------+------------------------------------------------+------------------------------+------------------
----------------+-------------------------------+
10 rows selected (39.473 seconds)
0: jdbc:hive2://localhost:10000>
```

- Query to access *CreditComplaints* Table with selected columns:



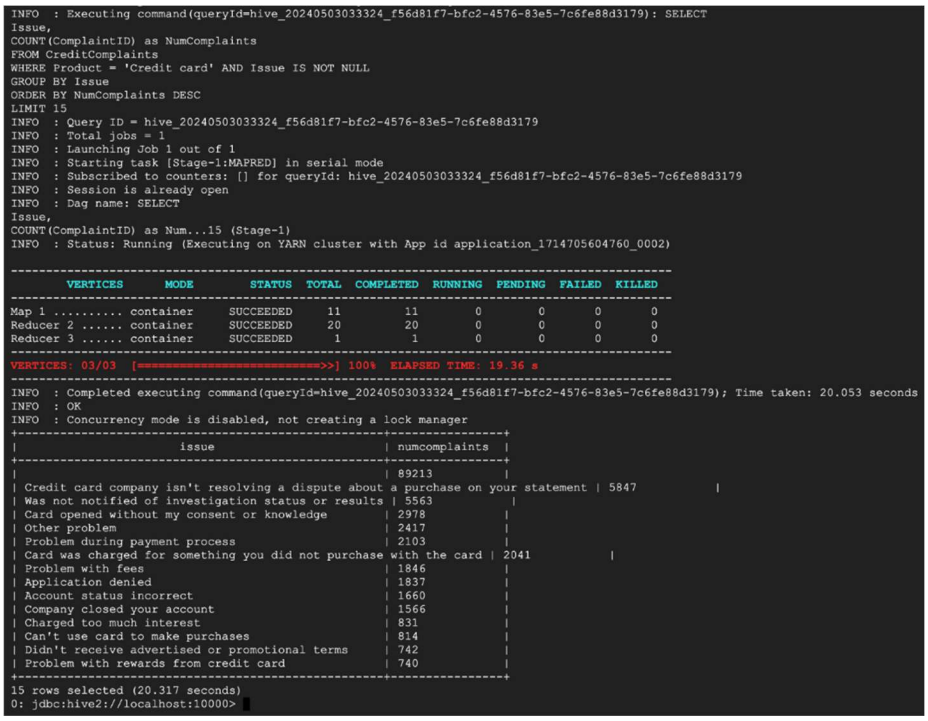
```
0: jdbc:hive2://localhost:10000> SELECT datereceived, product, issue, company, state, complaintid FROM creditcomplaints LIMIT 10;
INFO  : Compiling command(queryId=hive_20240502234201_cf825948-004c-487c-9d4a-d49532b6c849): SELECT datereceived, product, issue, company, state, complaintid FROM creditcomplaints LIMIT 10
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:datereceived, type:date, comment:null), FieldSchema(name:product, type:string, comment:null), FieldSchema(name:issue, type:string, comment:nul
l), FieldSchema(name:company, type:string, comment:null), FieldSchema(name:state, type:string, comment:null), FieldSchema(name:complaintid, type:int, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20240502234201_cf825948-004c-487c-9d4a-d49532b6c849); Time taken: 0.229 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240502234201_cf825948-004c-487c-9d4a-d49532b6c849): SELECT datereceived, product, issue, company, state, complaintid FROM creditcomplaints LIMIT 10
INFO  : Query ID = hive_20240502234201_cf825948-004c-487c-9d4a-d49532b6c849
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20240502234201_cf825948-004c-487c-9d4a-d49532b6c849
INFO  : Session is already open
INFO  : Dag name: SELECT datereceived, product, issue, co...10 (Stage-1)
INFO  : Tez session was closed. Reopening...
INFO  : Session re-established.
INFO  : Session re-established.
INFO  : Status: Running (Executing on YARN cluster with App id application_1714692099086_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ......... container     SUCCEEDED     11         11        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 15.06 s
--------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20240502234201_cf825948-004c-487c-9d4a-d49532b6c849); Time taken: 23.097 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+--------------+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------+--------+--------------+
| datereceived |                       product                       |                        issue                        |              company              | state  | complaintid  |
+--------------+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------+--------+--------------+
| 2016-12-05   | Credit reporting                                    | Information is not mine                             | TRANSUNION INTERMEDIATE HOLDINGS, INC.  | IL     | 2235318      |
| 2015-01-03   | Credit reporting                                    | Account terms                                       | TRANSUNION INTERMEDIATE HOLDINGS, INC.  | AZ     | 1179629      |
| 2023-09-30   | Credit reporting or other personal consumer reports | Account information incorrect                       | TRANSUNION INTERMEDIATE HOLDINGS, INC.  | PA     | 7619245      |
| 2023-09-30   | Credit reporting or other personal consumer reports | Credit inquiries on your report that you don't recognize | EQUIFAX, INC.                      | FL     | 7715650      |
| 2023-09-27   | Credit card                                         | Was not notified of investigation status or results | TRANSUNION INTERMEDIATE HOLDINGS, INC.  | TX     | 7614810      |
| 2021-05-20   | Credit card or prepaid card                         | Credit card company isn't resolving a dispute about a purchase on your statement | GOLDMAN SACHS BANK USA              | AR     | 4394692      |
| 2021-05-20   | Credit reporting, credit repair services, or other personal consumer reports | Information belongs to someone else | Experian Information Solutions Inc.  | WA     | 4394204      |
| 2023-11-02   | Credit reporting or other personal consumer reports | Account information incorrect                       | Experian Information Solutions Inc.  | FL     | 7793938      |
| 2023-11-02   | Credit reporting or other personal consumer reports | Reporting company used your report improperly       | Experian Information Solutions Inc.  | LA     | 7794632      |
| 2017-01-27   | Credit reporting                                    | Account status                                      | BANK OF AMERICA, NATIONAL ASSOCIATION | FL     | 2314656      |
+--------------+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------+--------+--------------+
10 rows selected (23.397 seconds)
0: jdbc:hive2://localhost:10000>
```

- Query to find Ratio of Timely Response and vice-versa for each Company



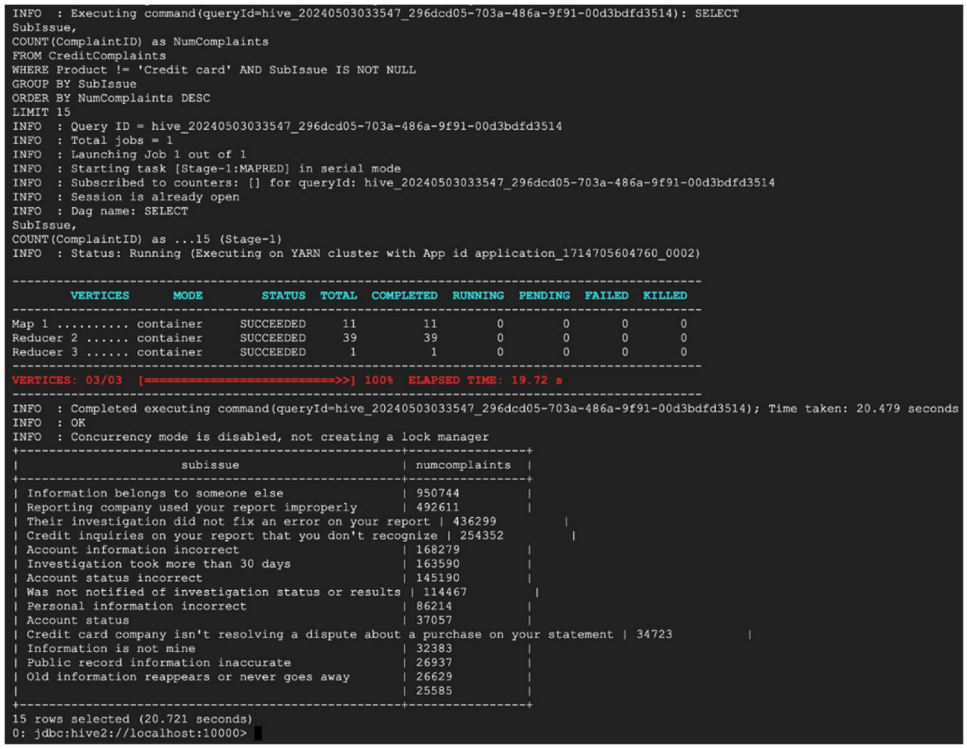
```
INFO  : Executing command(queryId=hive_20240503032909_2554e4bb-50d4-4a59-a60a-7e3d3adc202d): SELECT
Company,
SUM(CASE WHEN TimelyResponse = 'Yes' THEN 1 ELSE 0 END) as YESs,
SUM(CASE WHEN TimelyResponse = 'No' THEN 1 ELSE 0 END) as NOs,
ROUND((SUM(CASE WHEN TimelyResponse = 'No' THEN 1 ELSE 0 END))/((SUM(CASE WHEN TimelyResponse = 'No' THEN 1 ELSE 0 END))+(SUM(CASE WHEN TimelyResponse = 'Yes' THEN 1 ELSE 0 END))), 6) as Re
sponseRatio
FROM CreditComplaints
GROUP BY Company
ORDER BY NOs DESC
LIMIT 15
INFO  : Query ID = hive_20240503032909_2554e4bb-50d4-4a59-a60a-7e3d3adc202d
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20240503032909_2554e4bb-50d4-4a59-a60a-7e3d3adc202d
INFO  : Session is already open
INFO  : Dag name: SELECT
Company,
SUM(CASE WHEN TimelyRes...15 (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1714705604760_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ......... container     SUCCEEDED     11         11        0        0       0       0
Reducer 2 ..... container     SUCCEEDED     39         39        0        0       0       0
Reducer 3 ..... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 20.11 s
--------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20240503032909_2554e4bb-50d4-4a59-a60a-7e3d3adc202d); Time taken: 20.827 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+----------------------------------------+---------+-------+---------------+
|                company                 |   yess  |  nos  | responseratio |
+----------------------------------------+---------+-------+---------------+
| EQUIFAX, INC.                          | 1010889 | 1749  | 0.001727      |
| BANK OF AMERICA, NATIONAL ASSOCIATION  | 35999   | 1414  | 0.037794      |
| WELLS FARGO & COMPANY                  | 20859   | 584   | 0.027235      |
| Conduent Incorporated                  | 915     | 538   | 0.370268      |
| CLGF Holdco 1, LLC                     | 2535    | 343   | 0.11918       |
| TRANSUNION INTERMEDIATE HOLDINGS, INC. | 932164  | 323   | 3.46E-4       |
| Atlanticus Services Corporation        | 1644    | 315   | 0.160796      |
| Colony Brands, Inc.                    | 860     | 242   | 0.219601      |
| Credit Karma, LLC                      | 1977    | 212   | 0.096848      |
| EdFinancial Services                   | 340     | 177   | 0.34236       |
| LEXISNEXIS                             | 13225   | 134   | 0.010031      |
| CITIBANK, N.A.                         | 55431   | 113   | 0.002034      |
| Self Financial Inc.                    | 2382    | 110   | 0.044141      |
| Army and Air Force Exchange Service    | 757     | 106   | 0.122827      |
| Incomm Holdings Inc.                   | 974     | 102   | 0.094796      |
+----------------------------------------+---------+-------+---------------+
15 rows selected (21.114 seconds)
```

- Query to find the Most Common Type of Issue for ‘Credit card’ as Product:



```
INFO  : Executing command(queryId=hive_20240503033324_f56d81f7-bfc2-4576-83e5-7c6fe88d3179): SELECT
Issue,
COUNT(ComplaintID) as NumComplaints
FROM CreditComplaints
WHERE Product = 'Credit card' AND Issue IS NOT NULL
GROUP BY Issue
ORDER BY NumComplaints DESC
LIMIT 15
INFO  : Query ID = hive_20240503033324_f56d81f7-bfc2-4576-83e5-7c6fe88d3179
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20240503033324_f56d81f7-bfc2-4576-83e5-7c6fe88d3179
INFO  : Session is already open
INFO  : Dag name: SELECT
Issue,
COUNT(ComplaintID) as Num...15 (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1714705604760_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     11         11        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     20         20        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 19.36 s
--------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20240503033324_f56d81f7-bfc2-4576-83e5-7c6fe88d3179); Time taken: 20.053 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-----------------------------------------------------+----------------+
|                        issue                        | numcomplaints  |
+-----------------------------------------------------+----------------+
|                                                     | 89213          |
| Credit card company isn't resolving a dispute about a purchase on your statement | 5847          |
| Was not notified of investigation status or results | 5563           |
| Card opened without my consent or knowledge         | 2978           |
| Other problem                                       | 2417           |
| Problem during payment process                      | 2103           |
| Card was charged for something you did not purchase with the card | 2041          |
| Problem with fees                                   | 1846           |
| Application denied                                  | 1837           |
| Account status incorrect                            | 1660           |
| Company closed your account                         | 1566           |
| Charged too much interest                           | 831            |
| Can't use card to make purchases                    | 814            |
| Didn't receive advertised or promotional terms      | 742            |
| Problem with rewards from credit card               | 740            |
+-----------------------------------------------------+----------------+
15 rows selected (20.317 seconds)
0: jdbc:hive2://localhost:10000>
```

- Query to find the most common Sub-Issue faced by Consumers for Products other than ‘Credit card’.

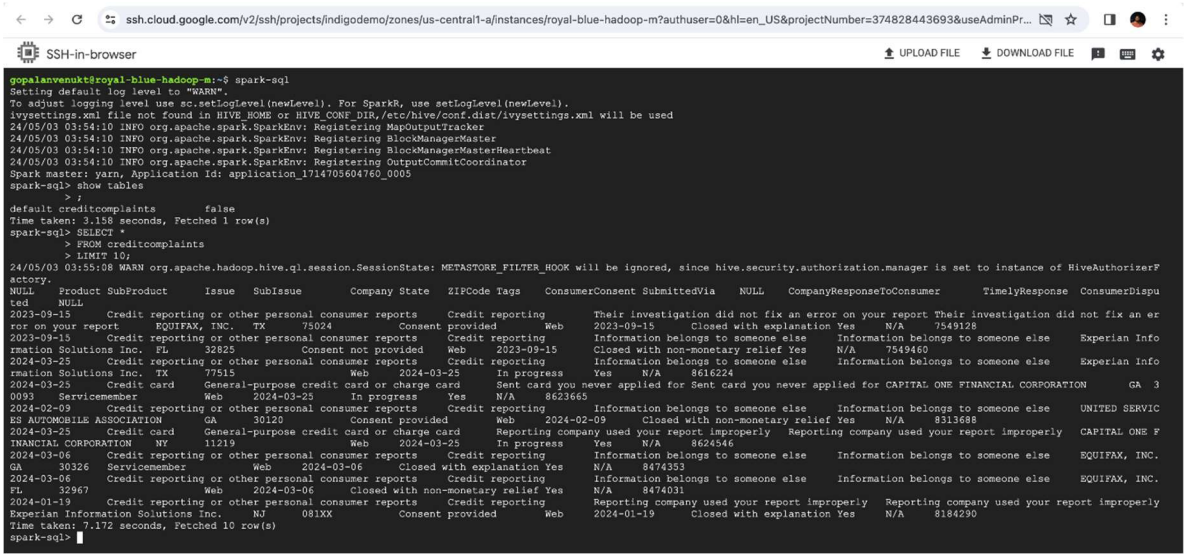


```
INFO  : Executing command(queryId=hive_20240503033547_296dcd05-703a-486a-9f91-00d3bdfd3514): SELECT
SubIssue,
COUNT(ComplaintID) as NumComplaints
FROM CreditComplaints
WHERE Product != 'Credit card' AND SubIssue IS NOT NULL
GROUP BY SubIssue
ORDER BY NumComplaints DESC
LIMIT 15
INFO  : Query ID = hive_20240503033547_296dcd05-703a-486a-9f91-00d3bdfd3514
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20240503033547_296dcd05-703a-486a-9f91-00d3bdfd3514
INFO  : Session is already open
INFO  : Dag name: SELECT
SubIssue,
COUNT(ComplaintID) as ...15 (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1714705604760_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     11         11        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     39         39        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 19.72 s
--------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20240503033547_296dcd05-703a-486a-9f91-00d3bdfd3514); Time taken: 20.479 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-----------------------------------------------------+----------------+
|                      subissue                       | numcomplaints  |
+-----------------------------------------------------+----------------+
| Information belongs to someone else                 | 950744         |
| Reporting company used your report improperly       | 492611         |
| Their investigation did not fix an error on your report | 436299         |
| Credit inquiries on your report that you don't recognize | 254352         |
| Account information incorrect                       | 168279         |
| Investigation took more than 30 days                | 163590         |
| Account status incorrect                            | 145190         |
| Was not notified of investigation status or results | 114467         |
| Personal information incorrect                      | 86214          |
| Account status                                      | 37057          |
| Credit card company isn't resolving a dispute about a purchase on your statement | 34723          |
| Information is not mine                             | 32383          |
| Public record information inaccurate                | 26937          |
| Old information reappears or never goes away        | 26629          |
|                                                     | 25585          |
+-----------------------------------------------------+----------------+
15 rows selected (20.721 seconds)
0: jdbc:hive2://localhost:10000>
```

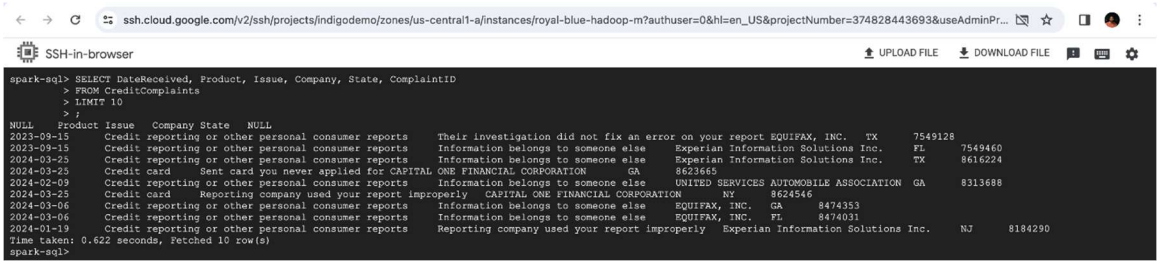
Spark

The same above queries are executed to check the results and efficiency of Spark

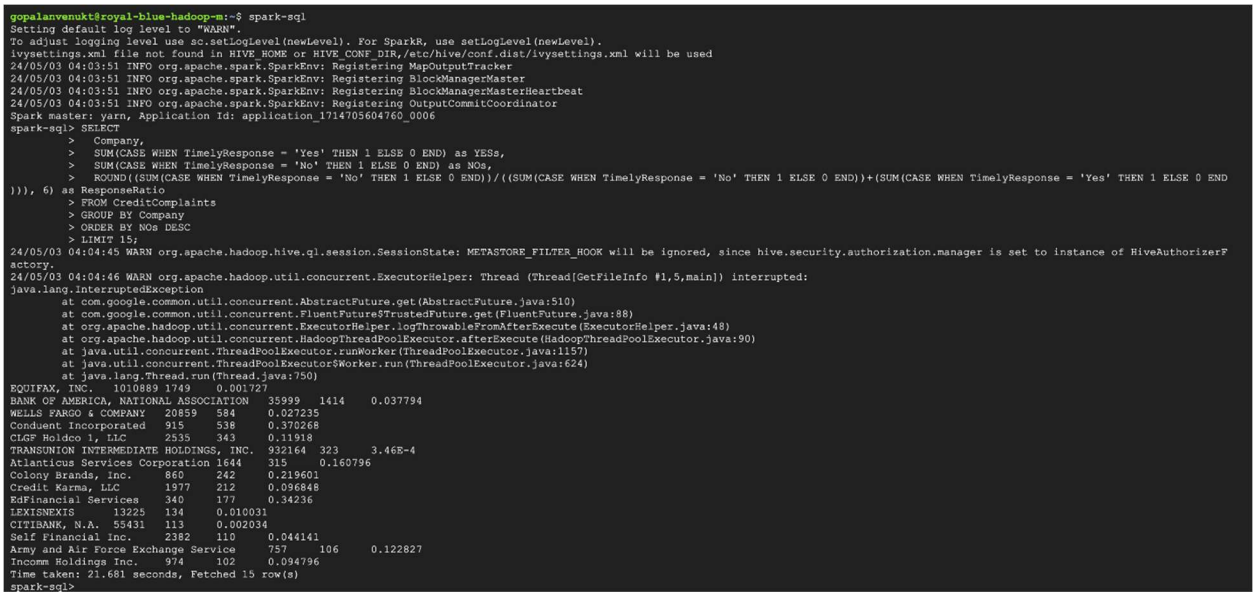
- Query to check the contents of the *CreditComplaints* Table (gives Top 10 Rows of *CreditComplaints* Table):



- Query to access *CreditComplaints* Table with selected columns



- Query to find Ratio of Timely Response and vice-versa for each Company



- Query to find the Most Common Type of Issue for ‘Credit card’ as Product:

```
gopalanvenukt@royal-blue-hadoop-m:~$ spark-sql
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
ivysettings.xml file not found in HIVE_HOME or HIVE_CONF_DIR,/etc/hive/conf.dist/ivysettings.xml will be used
24/05/03 04:38:01 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
24/05/03 04:38:01 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
24/05/03 04:38:01 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
24/05/03 04:38:01 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
Spark master: yarn, Application Id: application_1714710761918_0001
spark-sql> SELECT
         >   Issue,
         >   COUNT(ComplaintID) as NumComplaints
         > FROM CreditComplaints
         > WHERE Product = 'Credit card' AND Issue IS NOT NULL
         > GROUP BY Issue
         > ORDER BY NumComplaints DESC
         > LIMIT 15;
24/05/03 04:38:49 WARN org.apache.hadoop.hive.ql.session.SessionState: METASTORE_FILTER_HOOK will be ignored, since hive.security.authorization.manager is set to instance of HiveAuthorizerF
actory.
24/05/03 04:38:50 WARN org.apache.hadoop.util.concurrent.ExecutorHelper: Thread (Thread[GetFileInfo #1,5,main]) interrupted:
java.lang.InterruptedException
        at com.google.common.util.concurrent.AbstractFuture.get(AbstractFuture.java:510)
        at com.google.common.util.concurrent.FluentFuture$TrustedFuture.get(FluentFuture.java:88)
        at org.apache.hadoop.util.concurrent.ExecutorHelper.logThrowableFromAfterExecute(ExecutorHelper.java:48)
        at org.apache.hadoop.util.concurrent.HadoopThreadPoolExecutor.afterExecute(HadoopThreadPoolExecutor.java:90)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1157)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
        at java.lang.Thread.run(Thread.java:750)
        89213
Credit card company isn't resolving a dispute about a purchase on your statement       5847
Was not notified of investigation status or results     5563
Card opened without my consent or knowledge     2978
Other problem   2417
Problem during payment process  2103
Card was charged for something you did not purchase with the card       2041
Problem with fees       1846
Application denied      1837
Account status incorrect        1660
Company closed your account     1566
Charged too much interest       831
Can't use card to make purchases        814
Didn't receive advertised or promotional terms  742
Problem with rewards from credit card   740
Time taken: 22.611 seconds, Fetched 15 row(s)
spark-sql>
```

- Query to find the most common Sub-Issue faced by Consumers for Products other than ‘Credit card’

```
spark-sql> SELECT
         >   SubIssue,
         >   COUNT(ComplaintID) as NumComplaints
         > FROM CreditComplaints
         > WHERE Product != 'Credit card' AND SubIssue IS NOT NULL
         > GROUP BY SubIssue
         > ORDER BY NumComplaints DESC
         > LIMIT 15;
Information belongs to someone else     950744
Reporting company used your report improperly   492611
Their investigation did not fix an error on your report 436299
Credit inquiries on your report that you don't recognize        254352
Account information incorrect   168279
Investigation took more than 30 days    163590
Account status incorrect        145190
Was not notified of investigation status or results     114467
Personal information incorrect  86214
Account status  37057
Credit card company isn't resolving a dispute about a purchase on your statement        34723
Information is not mine 32383
Public record information inaccurate    26937
Old information reappears or never goes away    26629
        25585
Time taken: 14.464 seconds, Fetched 15 row(s)
spark-sql>
```

Hive v/s Spark

Following are the Time Records of Hive and Spark performance for the above queries:

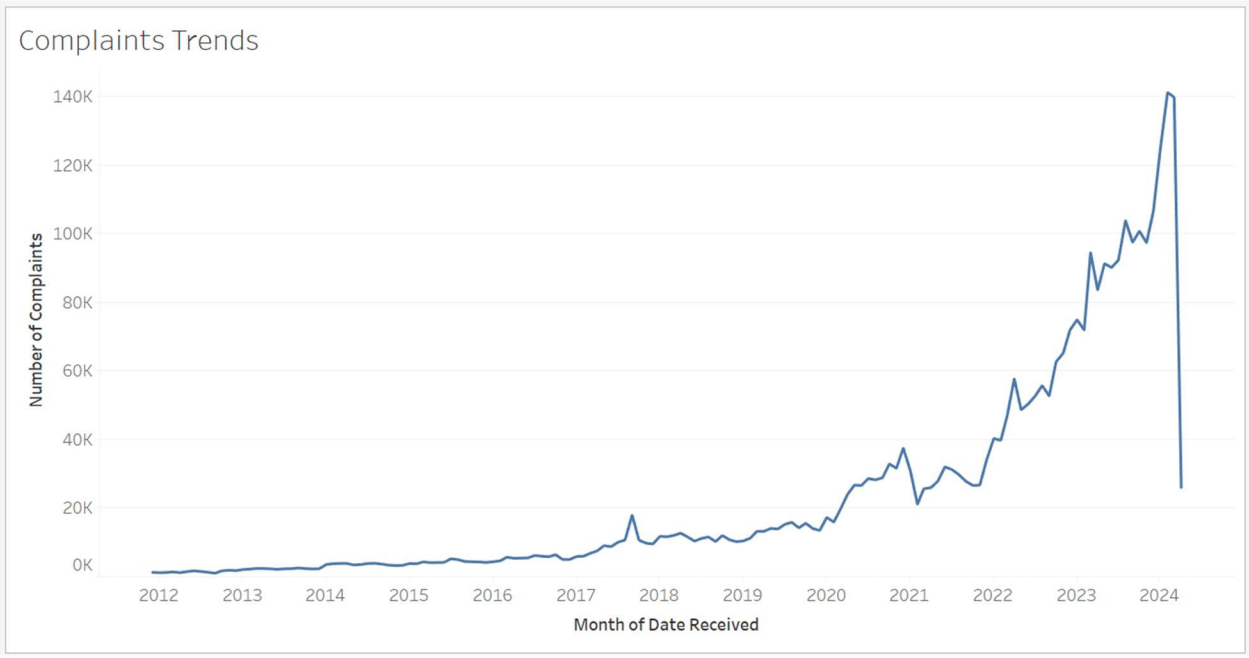
Query	Hive	Spark
Top 10 Rows	39.473	7.172
Top 10 Rows + Selected Columns	23.397	0.622
Response Ratio	21.114	21.681
Most Common Issues in Credit Cards	20.317	22.611
Most Common Sub-Issue in Other Products	20.721	14.464

Except for the 2 cases where Spark has Java-Error/Warning, Spark had quick response when compared to Hive for Query Execution.

Data Visualization

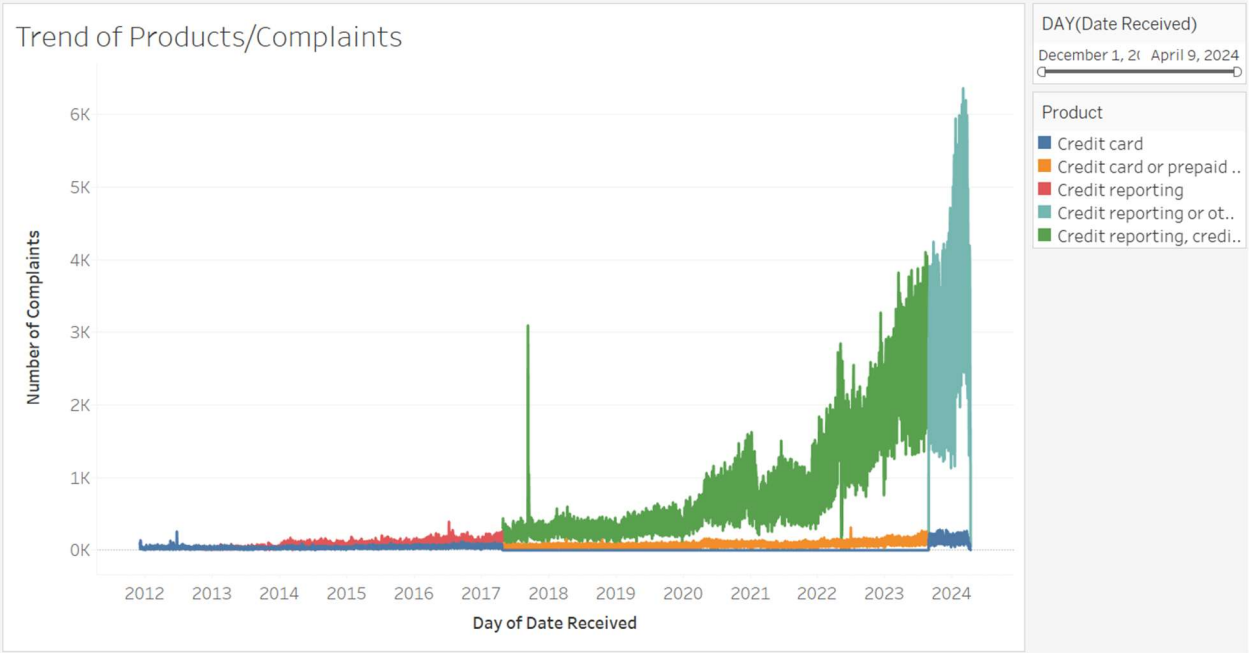
Following Graphs are visualized to further understand the Credit Dataset and find any patterns or other insights,

Graph 1:
Number of Complaints received over Time, plotted with Count of Complaints over Date Received for **CreditFull16.tsv**



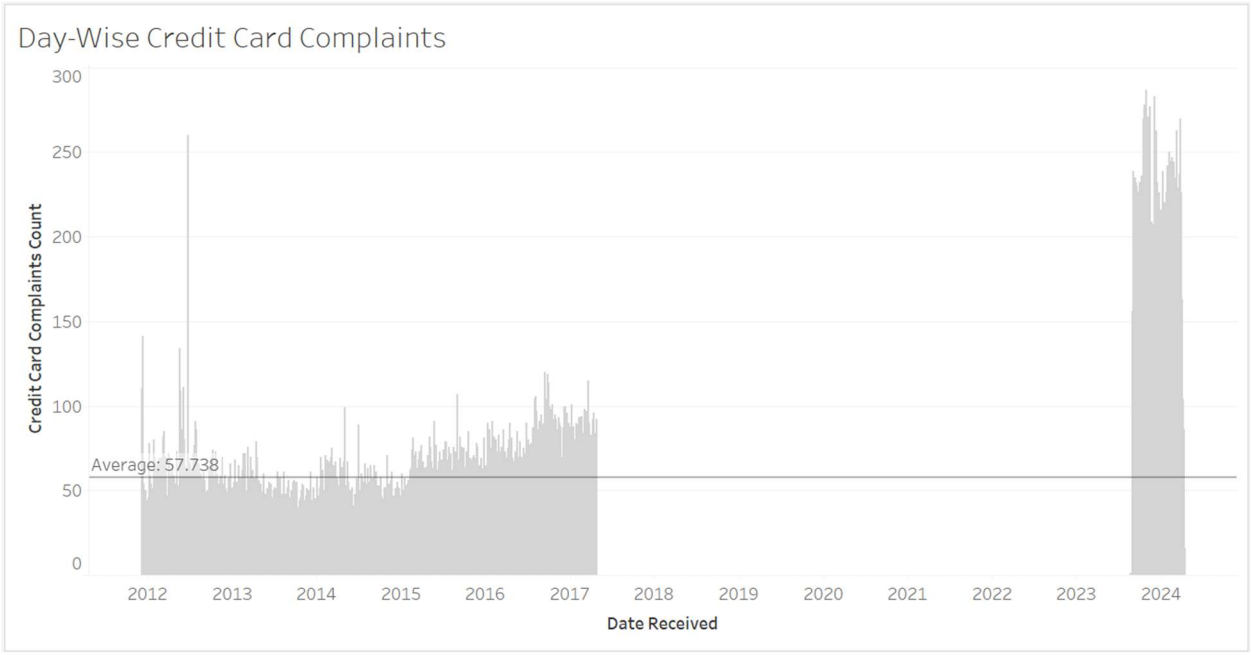
It is observed that there is a spike from 2017 and number of Complaints received grew exponentially.

Graph2:
Number of Complaints under each Product over Time, plotted with Count of Complaints over Date Received and Product in Color to differentiate.



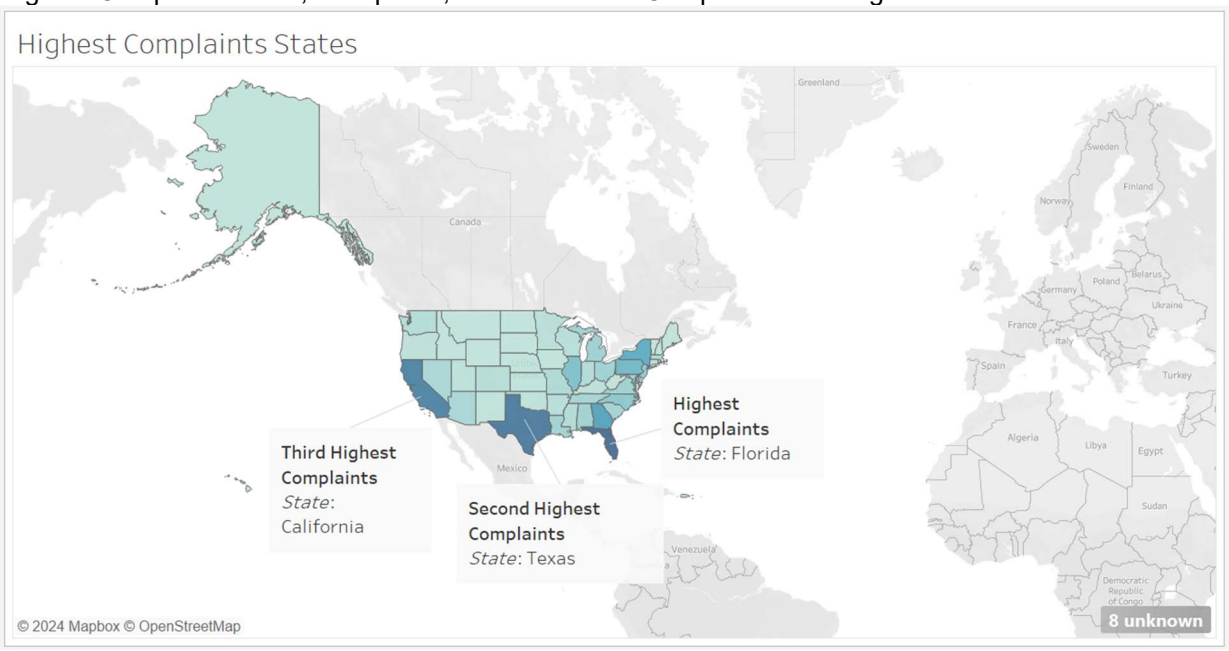
After plotting the Graph for multiple products, it is observed that the reason for spike after 2017 can be because of other complaints (not Credit Card) being shared along with Credit Card Products.

Graph 3:
Day-Wise Credit Card Complaints, plotted with Date Received and Count of Complaints for Product – 'Credit card'.



It was observed that an Average of 57.7 complaints were received daily. Also, from 2017 till end of 2023, there were no complaints for Credit card, and the number of complaints received in 2024 increased a lot.

Graph 4:
Highest Complaint States, a Map Plot, with Number of Complaints as Range.



It is observed that Number of Complaints increased in the Eastern part when looking at the complete map.

References

8 steps in the data life Cycle | HBS Online. (2021a, February 2). Business Insights Blog. https://online.hbs.edu/blog/post/data-life-cycle

Google Cloud Documentation. (n.d.). Google Cloud. https://cloud.google.com/docs

Dataset URL: Consumer Financial Protection Bureau - Consumer Complaint Database https://catalog.data.gov/dataset/consumer-complaint-database

UNT Course Material (Module 6, 11 and 12)

A grammar of data manipulation. (n.d.). https://dplyr.tidyverse.org/ (For Processing Data)