

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:

The screenshot shows the DATA.GOV website interface. At the top, there's a navigation bar with links for DATA, REPORTS, OPEN GOVERNMENT, and CONTACT. On the right side of the header, there are icons for a user profile and a 'User Guide'. Below the header, the main content area has a blue header bar with 'DATA CATALOG' and navigation links for 'Datasets' and 'Organizations'. The main content area displays the 'Consumer Complaint Database' entry. It includes a sidebar with the CFPB logo and contact information (read more, Publisher, Contact, devops@cfpb.gov, Share on Social Sites). The main content section has a heading 'Consumer Complaint Database' with a 'Metadata Updated: November 10, 2020' link. It describes the database as a collection of complaints about consumer financial products and services. Below this is a section titled 'Access & Use Information' with details about public access and license. At the bottom, there's a 'Downloads & Resources' section featuring a 'Comma Separated Values File' (complaints.csv.zip) with 163 views and a 'Download' button.

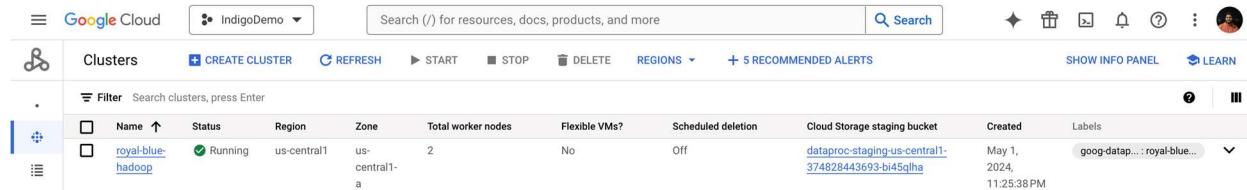
Google Cloud Platform (GCP) Project

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**

The screenshot shows the Google Cloud Platform dashboard for the 'IndigoDemo' project. The top navigation bar includes links for Google Cloud, IndigoDemo, and a search bar. Below the navigation is a 'DASHBOARD' tab and 'ACTIVITY' and 'RECOMMENDATIONS' buttons. The main content area is divided into three cards: 'Project info' (Project name: IndigoDemo, Project number: 374828443693, Project ID: indigodemo), 'Compute Engine' (CPU (%)), and 'Google Cloud Platform status' (Multiple Products: Cloud Build: Cross-project trigger creation failing, Begun at 2024-05-02 08:23:07, All times are US/Pacific, Data provided by status.cloud.google.com).

Hadoop Ecosystem Setup

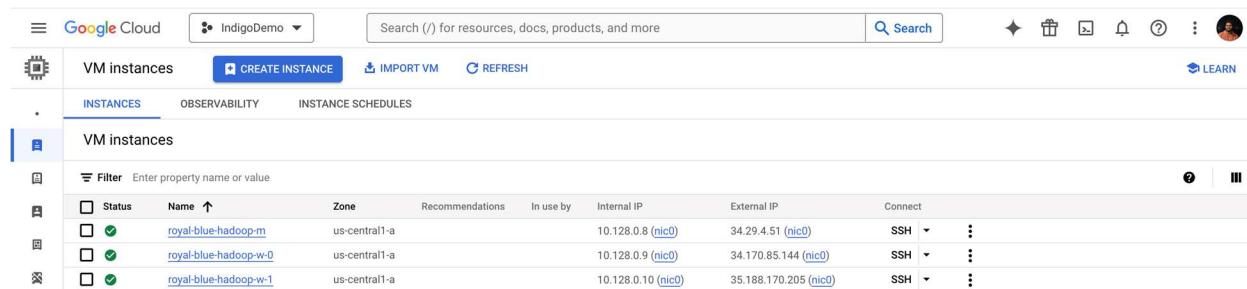
To Implement Hadoop Ecosystem, a Dataproc Cluster **royal-blue-hadoop** is created with 1-Master and 2-Worker Nodes (with same properties mentioned in Course material). Following Screenshot shows **royal-blue-hadoop** cluster which is running.



The screenshot shows the Google Cloud Platform interface for the 'IndigoDemo' project. The 'Clusters' tab is selected. A table lists one cluster: 'royal-blue-hadoop'. The cluster details are as follows:

Name	Status	Region	Zone	Total worker nodes	Flexible VMs?	Scheduled deletion	Cloud Storage staging bucket	Created	Labels
royal-blue-hadoop	Running	us-central1	us-central1-a	2	No	Off	dataproc-staging-us-central1-374828443693-bi45qlha	May 1, 2024, 11:25:38 PM	goog-datap... : royal-blue...

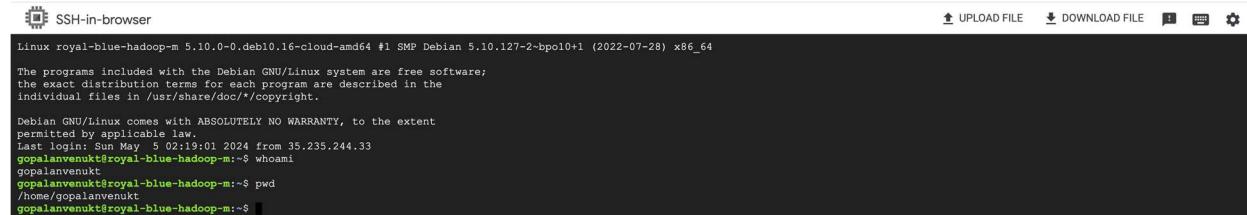
After the cluster is created, Compute Engine API is enabled (was already enabled in this case) and Virtual Machine (VM) Instances are checked to see if HDFS System is accessible or not (shown below)



The screenshot shows the Google Cloud Platform interface for the 'IndigoDemo' project. The 'VM instances' tab is selected. A table lists three instances:

Status	Name	Zone	Recommendations	In use by	Internal IP	External IP	Connect
Running	royal-blue-hadoop-m	us-central1-a			10.128.0.8 (nic0)	34.29.4.51 (nic0)	SSH
Running	royal-blue-hadoop-w-0	us-central1-a			10.128.0.9 (nic0)	34.170.85.144 (nic0)	SSH
Running	royal-blue-hadoop-w-1	us-central1-a			10.128.0.10 (nic0)	35.188.170.205 (nic0)	SSH

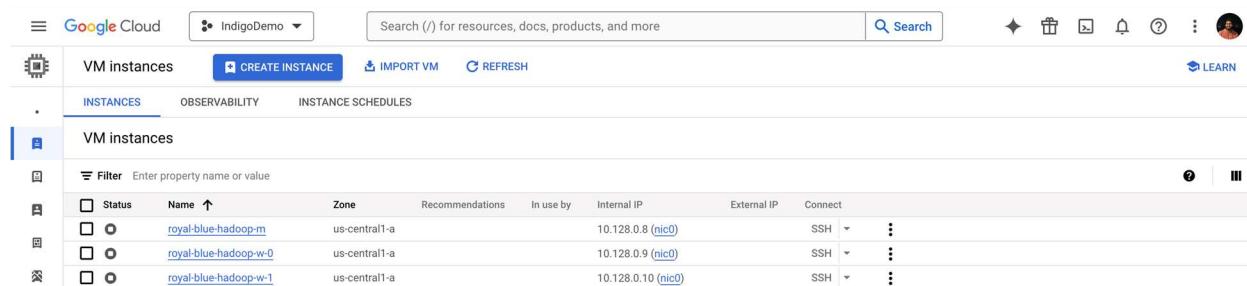
A sample command is executed in the Linux Machine connected through SSH of Master Node to verify the working of nodes (shown below)



```
Linux royal-blue-hadoop-m 5.10.0-0-deb10.16-cloud-amd64 #1 SMP Debian 5.10.127-2-bpo10+1 (2022-07-28) x86_64
The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Sun May  5 02:19:01 2024 from 35.235.244.33
gopalvenukut@royal-blue-hadoop-m:~$ whoami
gopalvenukut
gopalvenukut@royal-blue-hadoop-m:~$ pwd
/home/gopalvenukut
gopalvenukut@royal-blue-hadoop-m:~$
```

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



The screenshot shows the Google Cloud Platform interface for the 'IndigoDemo' project. The 'VM instances' tab is selected. A table lists three instances, all of which are currently stopped:

Status	Name	Zone	Recommendations	In use by	Internal IP	External IP	Connect
Stopped	royal-blue-hadoop-m	us-central1-a			10.128.0.8 (nic0)		SSH
Stopped	royal-blue-hadoop-w-0	us-central1-a			10.128.0.9 (nic0)		SSH
Stopped	royal-blue-hadoop-w-1	us-central1-a			10.128.0.10 (nic0)		SSH

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)
Product      SubProduct
Credit reporting :2163880  Credit reporting :2956061
:817854          :235291
Checking account :555576   Checking account :225243
:396141          :192618
General-purpose credit card or charge card :192618
:I do not know :128480
Other debt       :206374   Other debt       :107819
:(Other)          :650148   :(Other)          :1153092

Issue      SubIssue      ComplaintNarrative      CompanyPublicResponse
953380    Length:4998604  Length:4998604    Length:4998604
:732720    Class :character  Class :character   Class :character
:494546    Mode :character  Mode :character    Mode :character

Information belongs to someone else
Reporting company used your report improperly
Their investigation did not fix an error on your report
Credit inquiries on your report that you don't recognize
Account information incorrect
:(Other)

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 : 7482  Servicemember : 279385  Consent withdrawn : 8732
WELLS FARGO & COMPANY :127373  NY      :320850  33025 : 6884  N/A : 771115
JPMORGAN CHASE & CO. :117699  PA      :236592  35495 : 6382  Other : 240608
:(Other)          :(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
Postal mail : 93909  Mean :2020-12-27  Closed with monetary relief : 151466
Referral    : 247362  3rd Qu.:2023-06-13  Closed without relief : 17868
Web         :4451355  Max. :2024-04-09  Closed : 17611
```

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

The screenshot shows the Google Cloud Storage interface. In the left sidebar, 'Buckets' is selected. A single bucket named 'credit-data-bucket' is listed. It has the following details:

Location	Storage class	Public access	Protection
us-south1 (Dallas)	Standard	Not public	Soft Delete

The 'OBJECTS' tab is selected, showing a 'Folder browser' view. Inside the 'credit-data-bucket' folder, there are two entries: 'CreditFull16.tsv' (application/octet-stream, 947.1 MB) and 'CreditSubsets/' (Folder). There are also buttons for UPLOAD FILES, UPLOAD FOLDER, CREATE FOLDER, TRANSFER DATA, DOWNLOAD, and DELETE.

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:

The screenshot shows the Google BigQuery Studio interface. On the left, the 'Explorer' sidebar lists datasets like 'indigodemo', 'CreditDataset', and 'ExperimentFullDataset'. The 'CreditDataset' dataset is selected, and its 'CreditAll' table is shown in the main pane. The table schema is as follows:

Field name	Type	Mode	Key	Collation	Default Value	Policy Tags	Description
DateReceived	DATE	NULLABLE	-	-	-	-	Date on which the Complaint was Received
Product	STRING	NULLABLE	-	-	-	-	Credit/Other Product Type for Issue
SubProduct	STRING	NULLABLE	-	-	-	-	Sub Product in Product
Issue	STRING	NULLABLE	-	-	-	-	Key Reason for Complaint
SubIssue	STRING	NULLABLE	-	-	-	-	Sub-Level Information of Issue
Company	STRING	NULLABLE	-	-	-	-	Credit/Finance Company
State	STRING	NULLABLE	-	-	-	-	State / Location
ZIPCode	STRING	NULLABLE	-	-	-	-	ZIP Code / Location
Tags	STRING	NULLABLE	-	-	-	-	Tags on Customer Account
ConsumerConsent	STRING	NULLABLE	-	-	-	-	Whether Customer provided Consent for the Transaction
SubmittedVia	STRING	NULLABLE	-	-	-	-	Mode through the Complaint was submitted
DateSentToCompany	DATE	NULLABLE	-	-	-	-	Date Company Received the Complaint
CompanyResponseToConsumer	STRING	NULLABLE	-	-	-	-	Whether the Company Provided Response to the Complaint
TimelyResponse	BOOLEAN	NULLABLE	-	-	-	-	Whether the response was provided on Time
ConsumerDisputed	STRING	NULLABLE	-	-	-	-	Whether Customer Raised Dispute for the Issue
ComplaintID	INTEGER	NULLABLE	-	-	-	-	ID of Complaint

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

The screenshot shows the Google Cloud BigQuery interface. The left sidebar includes sections for Analysis, Migration, and Administration. The main area displays a query titled "HighestComplaints" which selects the company name and the count of distinct complaint IDs from the "CreditAll" table where the product is 'Credit card'. The results table shows the top 10 companies with their respective counts.

Row	Company	Num of Complaints
1	CITIBANK, N.A.	19981
2	CAPITAL ONE FINANCIAL COR...	16915
3	JPMORGAN CHASE & CO.	13551
4	SYNCHRONY FINANCIAL	11350
5	BANK OF AMERICA, NATIONAL...	11108
6	AMERICAN EXPRESS COMPANY	9104
7	DISCOVER BANK	5707
8	WELLS FARGO & COMPANY	5251
9	BARCLAYS BANK DELAWARE	4547
10	TRANSUNION INTERMEDIATE ...	3034

- 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

The screenshot shows the Google Cloud BigQuery interface. The left sidebar includes sections for Analysis, Migration, and Administration. The main area displays a query titled "ProductWiseComplaints" which uses a CASE WHEN statement to categorize complaints into 'CreditCard', 'CreditReporting', and 'OtherProducts' based on the product type. The results table shows the number of complaints for each state across these three categories.

Row	State	CreditCard	CreditReporting	OtherProducts
1	NH	511	450	3910
2	IN	1499	1516	39322
3	NV	1484	1916	53069
4	GA	4629	7842	257700
5	NC	3509	4208	118162
6	KS	630	596	9271
7	MS	550	788	34092
8	VA	3913	4970	76506
9	NY	11191	8848	222195
10	FL	11156	14436	446062
11	TX	9371	16212	394775
12	CO	2209	2361	26863
13	LA	1049	1735	68781
14	CT	1736	1416	25071
15	CA	16819	18119	369064

- 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

```

SELECT Company, COUNT(*) as Times
FROM `indigodemo.CreditDataset.CreditAll`
WHERE TimelyResponse = FALSE
AND SubmittedVia = "Phone"
GROUP BY Company
ORDER BY COUNT(*) DESC;
    
```

Query results

Company	Times
Conduent Incorporated	212
BANK OF AMERICA, NATIONAL...	77
EQUIFAX, INC.	44
WELLS FARGO & COMPANY	40
Atlanticus Services Corporation	34
CLGF Holdco 1, LLC	29
Colony Brands, Inc.	19
Credit Karma, LLC	15
Incomm Holdings Inc.	13
CITIBANK, N.A.	11
Netspend Corporation	11
CAPITAL ONE FINANCIAL COR...	7
TRANSUNION INTERMEDIATE ...	7
Comerica	5
Accurate Background	4
IMC Capital, LLC	4

- 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)

```

SELECT
Product,
SUM(CASE WHEN SubmittedVia = 'Web' THEN 1 ELSE 0 END) AS WebSubmission,
SUM(CASE WHEN SubmittedVia = 'Phone' THEN 1 ELSE 0 END) AS PhoneCalls,
SUM(CASE WHEN SubmittedVia = 'Referral' OR SubmittedVia = 'Web Referral' THEN 1 ELSE 0 END) AS Referrals,
SUM(CASE WHEN SubmittedVia = 'Post Mail' OR SubmittedVia = 'Email' OR SubmittedVia = 'Fax' THEN 1 ELSE 0 END) AS AllMails,
FROM `indigodemo.CreditDataset.CreditAll`
WHERE Product IS NOT NULL
GROUP BY Product;
    
```

Query results

Product	WebSubmission	PhoneCalls	Referrals	AllMails
Credit card	96097	8560	15711	944
Credit reporting, credit repair services, or other personal consumer reports	2087919	36293	12396	8638
Credit reporting or other personal consumer reports	810184	4620	284	0
Credit card or prepaid card	163496	18273	18502	760
Credit reporting	107505	3089	9604	2107

- 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)

Google Cloud IndigoDemo Search (/) for resources, docs, products, and more Search

BigQuery Disputes + RUN SAVE QUERY DOWNLOAD SHARE SCHEDULE MORE Query completed.

Analysis +

BigQuery Studio

Data transfers

Scheduled queries

Analytics Hub

Dataform

Partner Center

Migration +

Assessment

SQL translation

Administration +

Monitoring

Capacity management

BI Engine

Disaster recovery

Policy tags

Release Notes

1 SELECT
2 Company,
3 COUNT(ComplaintID) as ComplaintCount,
4 SUM(CASE WHEN ConsumerDisputed = 'Yes' THEN 1 ELSE 0 END) as Disputed,
5 SUM(CASE WHEN ConsumerDisputed = 'No' THEN 1 ELSE 0 END) as NotDisputed,
6 SUM(CASE WHEN ConsumerDisputed = 'NA' THEN 1 ELSE 0 END) as NotApplicable,
7 CASE
8 WHEN (SUM(CASE WHEN ConsumerDisputed = 'Yes' THEN 1 ELSE 0 END)) = 0 AND (SUM(CASE WHEN ConsumerDisputed = 'No' THEN 1 ELSE 0 END)) = 0
9 THEN '0 %'
10 ELSE
11 CONCAT(ROUND((SUM(CASE WHEN ConsumerDisputed = 'Yes' THEN 1 ELSE 0 END)/(SUM(CASE WHEN ConsumerDisputed = 'No' THEN 1 ELSE 0 END)+SUM(CASE WHEN ConsumerDisputed = 'Yes' THEN 1 ELSE 0 END))*100),2), ' %')
12 END as DisputeRate
13 FROM `indigodemo.CreditDataset.CreditAll`
14 GROUP BY Company
15 ORDER BY ComplaintCount DESC;

Press Option+F1 for Accessibility Options

Query results SAVE RESULTS EXPLORE DATA

Row	Company	ComplaintCount	Disputed	NotDisputed	NotApplicable	DisputeRate
1	EQUIFAX, INC.	1012638	10029	38145	964464	20.82 %
2	TRANSUNION INTERMEDIATE HOLDINGS, INC.	932487	5605	34236	892646	14.07 %
3	Experian Information Solutions Inc.	851301	5307	40121	805873	11.68 %
4	CAPITAL ONE FINANCIAL CORPORATION	70771	2313	10970	57488	17.41 %
5	CITIBANK, N.A.	55544	3307	13777	38460	19.36 %
6	JPMORGAN CHASE & CO.	43101	2508	8057	32536	23.74 %
7	SYNCHRONY FINANCIAL	38371	1380	7582	29409	15.4 %
8	BANK OF AMERICA, NATIONAL ASSOCIATION	37413	2064	7173	28176	22.34 %
9	AMERICAN EXPRESS COMPANY	27930	1901	4822	21207	28.28 %

Results per page: 50 ▶ 1 – 50 of 3552

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)

```
gopalanvenkut@royal-blue-hadoop-m:~$ hdfs dfs -ls /
Found 3 items
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /tmp
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /user
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /var
gopalanvenkut@royal-blue-hadoop-m:~$ hdfs dfs -ls /user
Found 11 items
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /user/dataproc
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /user/hbase
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /user/hdfs
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /user/hive
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /user/mapred
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /user/pig
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /user/solr
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /user/spark
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /user/yarn
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /user/zeppelin
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /user/zookeeper
gopalanvenkut@royal-blue-hadoop-m:~$ hdfs dfs -mkdir /user/indigo
ls: '/user/indigo': No such file or directory
gopalanvenkut@royal-blue-hadoop-m:~$ hdfs dfs -mkdir /user/indigo
gopalanvenkut@royal-blue-hadoop-m:~$ hdfs dfs -ls /user
Found 12 items
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /user/dataproc
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /user/hbase
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /user/hdfs
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /user/hive
drwxrwxrwt - hdfs hadoop 0 2024-05-02 05:17 /user/indigo
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /user/mapred
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /user/pig
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /user/solr
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /user/spark
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /user/yarn
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /user/zeppelin
drwxrwxrwt - hdfs hadoop 0 2024-05-02 04:27 /user/zookeeper
gopalanvenkut@royal-blue-hadoop-m:~$
```

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:

```
gopalvenkut@royal-blue-hadoop-m:~$ hdfs dfs -mkdir /user/Indigo/data
gopalvenkut@royal-blue-hadoop-m:~$ hadoop DATA
gopalvenkut@royal-blue-hadoop-m:~$ hdfs dfs -put DATA
gopalvenkut@royal-blue-hadoop-m:~$ /GATX gutil cp gs://credit-data-bucket/CreditFull116.tsv CreditFull.tsv
Copying gs://credit-data-bucket/CreditFull116.tsv...
| [files](947.1 MB) 59.6 MB/s
Operation completed over 1 objects/947.1 MB.
gopalvenkut@royal-blue-hadoop-m:~$ /GATX hdfs dfs -put CreditFull.tsv /user/indigo/data
gopalvenkut@royal-blue-hadoop-m:~$ /GATX cd ..
gopalvenkut@royal-blue-hadoop-m:~$ hdfs dfs -ls /user/indigo/data
Found 1 items
-rw-r--r-- 2 gopalvenkut hadoop 993123275 2024-05-02 05:31 /user/indigo/data/CreditFull.tsv
gopalvenkut@royal-blue-hadoop-m:~$
```

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]
gopalanvenkut@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-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 (retryal - false)
INFO : Returning Hive schema: Schema([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:DQL] 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):

```
← → C 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
FILE 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 (retrying = false)
INFO : Returning Hive schema: Schema[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.state, type:string, comment:null), FieldSchema(name:creditcomplaints.zipcode, type:string, comment:null), FieldSchema(name:creditcomplaints.tags, type:string, comment:null), FieldSchema(name:creditcomplaints.consumerconsent, type:string, comment:null), FieldSchema(name:creditcomplaints.consumerdisputed, type:string, comment:null), FieldSchema(name:creditcomplaints.company, type:string, comment:null), FieldSchema(name:creditcomplaints.datesenttocompany, type:string, comment:null), FieldSchema(name:creditcomplaints.timelyresponse, type:string, comment:null), FieldSchema(name:creditcomplaints.consumerconsent, type:string, comment:null), FieldSchema(name:creditcomplaints.consumerdisputed, type:string, comment:null), FieldSchema(name:creditcomplaints.complaintid, type:int, comment:null), FieldSchema(name:creditcomplaints.timelyresponse, type:string, comment:null), FieldSchema(name:creditcomplaints.consumerconsent, type:string, comment:null), FieldSchema(name:creditcomplaints.consumerdisputed, type:string, comment:null), FieldSchema(name:creditcomplaints.company, type:string, comment:null), FieldSchema(name:creditcomplaints.datesenttocompany, type:string, comment:null), FieldSchema(name:creditcomplaints.timelyresponse, type:string, comment:null), FieldSchema(name:creditcomplaints.consumerconsent, type:string, comment:null), FieldSchema(name:creditcomplaints.consumerdisputed, 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] 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 SUCCEEDEDED 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.state | creditcomplaints.zipcode | creditcomplaints.tags | creditcomplaints.consumerconsent | creditcomplaints.consumerdisputed | creditcomplaints.submittedvia | creditcomplaints.datesenttocompany | creditcomplaints.company | creditcomplaints.datesenttocompany | creditcomplaints.timelyresponse | creditcomplaints.consumerconsent | creditcomplaints.consumerdisputed | creditcomplaints.complaintid |
+-----+
| 2016-12-05 | Credit reporting | TRANSUNION INTERMEDIATE HOLDINGS, INC. | IL | 60827 | Information is not mine | Yes | Consent not provided |
formation is not mine ded | Web | 2016-12-05 | Closed with non-monetary relief | Yes | No |
| 2235318 | |
+-----+
| 2015-01-03 | Credit reporting | TRANSUNION INTERMEDIATE HOLDINGS, INC. | AZ | 85255 | Account terms | Older American | N/A | Ac |
count terms | Web | 2015-01-03 | Closed with explanation | Yes | No |
| 1179629 | |
+-----+
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
Map 1 ..... container SUCCEEDEDED 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.subissue | creditcomplaints.company | creditcomplaints.state | creditcomplaints.zipcode | creditcomplaints.tags | creditcomplaints.consumerconsent | creditcomplaints.submittedvia | creditcomplaints.datesenttocompany | creditcomplaints.companyresponsetoconsumer | creditcomplaints.timelyresponse | creditcomplaints.consumerdisputed | creditcomplaints.complaintid |
+-----+
| 2016-12-05 | Credit reporting | TRANSUNION INTERMEDIATE HOLDINGS, INC. | IL | 60827 | Information is not mine | Yes | Consent not provided |
formation is not mine ded | Web | 2016-12-05 | Closed with non-monetary relief | Yes | No |
| 2235318 | |
+-----+
| 2015-01-03 | Credit reporting | TRANSUNION INTERMEDIATE HOLDINGS, INC. | AZ | 85255 | Account terms | Older American | N/A | Ac |
count terms | Web | 2015-01-03 | Closed with explanation | Yes | No |
| 1179629 | |
+-----+
| 2023-09-30 | Credit reporting or other personal consumer reports | Credit reporting | TRANSUNION INTERMEDIATE HOLDINGS, INC. | PA | 19018 | Account information incorrect | Yes | Consent not provided |
account information incorrect | Web | 2023-09-30 | Closed with non-monetary relief | Yes | No |
| 7619425 | |
+-----+
| 2023-09-30 | Credit reporting or other personal consumer reports | Credit reporting | TRANSUNION INTERMEDIATE HOLDINGS, INC. | FL | 33073 | Credit inquiries on your report that you don't recognize | Yes | Consent not provided |
ze | Credit inquiries on your report that you don't recognize | EQUIFAX, INC. | FL | 33073 | |
| Not provided | Web | 2023-09-30 | Closed with explanation | Yes | No |
| 7715650 | |
+-----+
| 2023-09-27 | Credit card | General-purpose credit card or charge card | TRANSUNION INTERMEDIATE HOLDINGS, INC. | TX | 78640 | Was not notified of investigation status or results | Yes | Consent provided |
as not notified of investigation status or results | Web | 2023-09-27 | Closed with non-monetary relief | Yes | No |
| 76194810 | |
+-----+
| 2021-05-20 | Credit card or prepaid card | General-purpose credit card or charge card | GOLDMAN SACHS BANK USA | AR | 72701 | Credit card company isn't resolving a dispute about a purchase on your statement | Yes |
purchase on your statement | Credit card company isn't resolving a dispute about a purchase on your statement | Goldman Sachs Bank USA | AR | 72701 |
| Yes | N/A | 4394692 | 2021-05-20 | Closed with explanation |
+-----+
| 2021-05-20 | Credit reporting, credit repair services, or other personal consumer reports | Credit reporting | Experian Information Solutions Inc. | WA | 98004 | Information belongs to someone else | Yes |
e else | Web | 2021-05-20 | Closed with explanation | Yes | No |
| Consent not provided | 4394204 | |
+-----+
| 2023-11-02 | Credit reporting or other personal consumer reports | Credit reporting | Experian Information Solutions Inc. | FL | 33325 | Account information incorrect | Yes | Other |
count information incorrect | Web | 2023-11-02 | Closed with explanation | Yes | N/A |
| 7793938 | |
+-----+
| 2023-11-02 | Credit reporting or other personal consumer reports | Credit reporting | Experian Information Solutions Inc. | LA | 70460 | Reporting company used your report improperly | Yes | Consent provided |
reporting company used your report improperly | Web | 2023-11-02 | Closed with explanation | Yes | N/A |
| 7794632 | |
+-----+
| 2017-01-27 | Credit reporting | BANK OF AMERICA, NATIONAL ASSOCIATION | FL | 33908 | Account status | Yes | N/A |
count status | 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:

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

```
← → G ssh.cloud.google.com/v2/ssh/projects/indigodemo/instances/us-central1-a/instances/royal-blue-hadoop-m?authuser=0&hl=en_US&projectNumber=374828443693&useAdminPr... └☆ └□ └○ └●

SSH-in-browser
UPLOAD FILE DOWNLOAD FILE
INFO : Executing command(queryId=hive_20240503032909_2554e4bb-50d4-4a59-a60a-7e3d3adc202d): SELECT
Company,
SUM(CASE WHEN TimelyResponse = 'Yes' THEN 1 ELSE 0 END) as YES,
SUM(CASE WHEN TimelyResponse = 'No' THEN 1 ELSE 0 END) as NO,
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 ResponseRatio
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 : Day name: SELECT
Company,
SUM(CASE WHEN TimelyResponse=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: 03/01 [----->>>] 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 | yes | nos | responseratio |  

+---  

| EQUIFAX, INC. | 101089 | 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 |  

| TRANSLATION INTERMEDIATE HOLDINGS, INC. | 932164 | 323 | 3.46E-4 |  

| Atlanticus Services Corporation | 1644 | 315 | 0.04936 |  

| First Data, Inc. | 860 | 942 | 0.213601 |  

| Credit Karma, LLC | 1977 | 212 | 0.09648 |  

| 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 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_296dc0d05-703a-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_296dc0d05-703a-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_296dc0d05-703a-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_296dc0d05-703a-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):

```
gopalvenkut@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 03:54:10 INFO org.apache.spark.SparkKvNs: Registering MapOutputTracker
24/05/03 03:54:10 INFO org.apache.spark.SparkKvNs: Registering BlockManagerMaster
24/05/03 03:54:10 INFO org.apache.spark.SparkKvNs: Registering BlockManagerMasterHeartbeat
24/05/03 03:54:10 INFO org.apache.spark.SparkKvNs: Registering OutputCommitCoordinator
Spark master: yarn, Application Id: application_1714705604760_0005
spark-sql> show tables
+-----+
| creditcomplaints |
+-----+
1 row(s) fetched. Time taken: 3.159 seconds.
spark-sql> SELECT *
> FROM creditcomplaints
> LIMIT 10;
24/05/03 03:55:08 WARN org.apache.hadoop.hive.session.SessionState: METASTORE_FILTER_HOOK will be ignored, since hive.security.authorization.manager is set to instance of HiveAuthorizerFactory.
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| NULL | Product | SubProduct | Issue | SubIssue | Company | State | ZIPCode | Tags | ConsumerConsent | SubmittedVia | NULL | CompanyResponseToConsumer | TimelyResponse | ConsumerDisputed |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
2023-09-15 | Credit reporting or other personal consumer reports | Credit reporting | Their investigation did not fix an error on your report | Their investigation did not fix an error on your report | N/A | 7549128
ror on your report | EQUIFAX, INC. | TX | 75024 | Consent provided | Web | 2023-09-15 | Closed with explanation Yes | N/A | 7549128
2023-10-18 | Credit reporting or other personal consumer reports | Credit reporting | Information belongs to someone else | Information belongs to someone else | Experian Info
rmation Solutions Inc. | FL | 30120 | Consent provided | Web | 2023-10-18 | Closed with non-monetary relief Yes | N/A | 7549460
2024-03-25 | Credit reporting or other personal consumer reports | Credit reporting | Information belongs to someone else | Information belongs to someone else | Experian Info
rmation Solutions Inc. | TX | 77515 | Web | 2024-03-25 | In progress Yes | N/A | 8616224
2024-03-25 | Credit card | General-purpose credit card or charge card | Sent card you never applied for | Sent card you never applied for | CAPITAL ONE FINANCIAL CORPORATION | GA | 3
0093 | Servicemember | Web | 2024-03-25 | In progress Yes | N/A | 8623665
2024-09-09 | Credit reporting or other personal consumer reports | Credit reporting | Information belongs to someone else | Information belongs to someone else | UNITED SERVICE
AUTOMOBILE ASSOCIATION | NY | 11219 | Web | 2024-09-09 | Closed with non-monetary relief Yes | N/A | 8313688
2024-03-25 | Credit card | General-purpose credit card or charge card | Reporting company used your report improperly | Reporting company used your report improperly | CAPITAL ONE F
INANCIAL CORPORATION | NY | 11219 | Web | 2024-03-25 | In progress Yes | N/A | 8624546
2024-03-06 | Credit reporting or other personal consumer reports | Credit reporting | Information belongs to someone else | Information belongs to someone else | EQUIFAX, INC.
GA | 30326 | Servicemember | Web | 2024-03-06 | Closed with explanation Yes | N/A | 8474353
2024-03-06 | Credit reporting or other personal consumer reports | Credit reporting | Information belongs to someone else | Information belongs to someone else | EQUIFAX, INC.
DE | 30367 | Web | 2024-03-06 | Closed with non-monetary relief Yes | N/A | 8474031
2024-01-19 | Credit reporting or other personal consumer reports | Credit reporting | Reporting company used your report improperly | Reporting company used your report improperly
Experian Information Solutions Inc. | NJ | 081XX | Consent provided | Web | 2024-01-19 | Closed with explanation Yes | N/A | 8184290
Time taken: 7.172 seconds, Fetched 10 row(s)
spark-sql>
```

- Query to access *CreditComplaints* Table with selected columns

```
gopalvenkut@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
spark-sql> SELECT DateReceived, Product, Issue, Company, State, ComplaintID
> FROM CreditComplaints
> LIMIT 10
> ;
+-----+-----+-----+-----+-----+-----+
| NULL | Product | Issue | Company | State | ComplaintID |
+-----+-----+-----+-----+-----+-----+
2023-09-15 | Credit reporting or other personal consumer reports | Their investigation did not fix an error on your report | EQUIFAX, INC. | TX | 7549128
2023-09-15 | Credit reporting or other personal consumer reports | Information belongs to someone else | Experian Information Solutions Inc. | FL | 7549460
2024-03-25 | Credit reporting or other personal consumer reports | Information belongs to someone else | Experian Information Solutions Inc. | TX | 8616224
2024-03-25 | Credit card | Sent card you never applied for | CAPITAL ONE FINANCIAL CORPORATION | GA | 8623665
2024-02-09 | Credit reporting or other personal consumer reports | Information belongs to someone else | UNITED SERVICES AUTOMOBILE ASSOCIATION | GA | 8313688
2024-03-25 | Credit card | Reporting company used your report improperly | CAPITAL ONE FINANCIAL CORPORATION | NY | 8624546
2024-03-06 | Credit reporting or other personal consumer reports | Information belongs to someone else | EQUIFAX, INC. | GA | 8474353
2024-03-06 | Credit reporting or other personal consumer reports | Information belongs to someone else | EQUIFAX, INC. | FL | 8474031
2024-01-19 | Credit reporting or other personal consumer reports | Reporting company used your report improperly | Experian Information Solutions Inc. | NJ | 8184290
Time taken: 0.622 seconds, Fetched 10 row(s)
spark-sql>
```

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

```
gopalvenkut@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 04:03:51 INFO org.apache.spark.SparkKvNs: Registering MapOutputTracker
24/05/03 04:03:51 INFO org.apache.spark.SparkKvNs: Registering BlockManagerMaster
24/05/03 04:03:51 INFO org.apache.spark.SparkKvNs: Registering BlockManagerMasterHeartbeat
24/05/03 04:03:51 INFO org.apache.spark.SparkKvNs: Registering OutputCommitCoordinator
Spark master: yarn, Application Id: application_1714705604760_0006
spark-sql> 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 ResponseRatio
FROM CreditComplaints
GROUP BY Company
ORDER BY NOs DESC
LIMIT 15;
24/05/03 04:04:45 WARN org.apache.hadoop.hive.session.SessionState: METASTORE_FILTER_HOOK will be ignored, since hive.security.authorization.manager is set to instance of HiveAuthorizerFactory.
24/05/03 04:04:46 WARN org.apache.hadoop.util.concurrent.ExecutorHelper: Thread (Thread[GetFileInfo@1,main]) interrupted:
java.lang.InterruptedExecutionException
at com.google.common.util.concurrent.AbstractFuture.get(AbstractFuture.java:510)
at com.google.common.util.concurrent.FluentFuture.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)
EQUIFAX, INC. 1010898 1749 0.001727
BANK OF AMERICA, NATIONAL ASSOCIATION 35959 1414 0.037794
MERRILL LYNCH COMPANY 20659 810 0.027235
Conduit Incorporated 915 338 0.370268
CLIF Holdings LLC 2535 343 0.119168
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 3225 154 0.010039
OICU, INC. 5331 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
Time taken: 21.681 seconds, Fetched 15 row(s)
spark-sql>
```

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

```
palanumekut@royal-blue-hadoop-m:~$ spark-sql
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel).
hivesettings.xml file not found in HIVE_HOME or HIVE_CONF_DIR, /etc/hive/conf.dist/hivesettings.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 HiveAuthorizerFactory.
24/05/03 04:38:50 WARN org.apache.hadoop.util.concurrent.ExecutorHelper: Thread (Thread[GetFileInfo #1,5,main]) interrupted;
java.lang.InterruptedException
at java.util.concurrent.AbstractFuture.get(AbstractFuture.java:510)
at com.google.common.util.concurrent.FluentFuture.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)
...
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
...
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

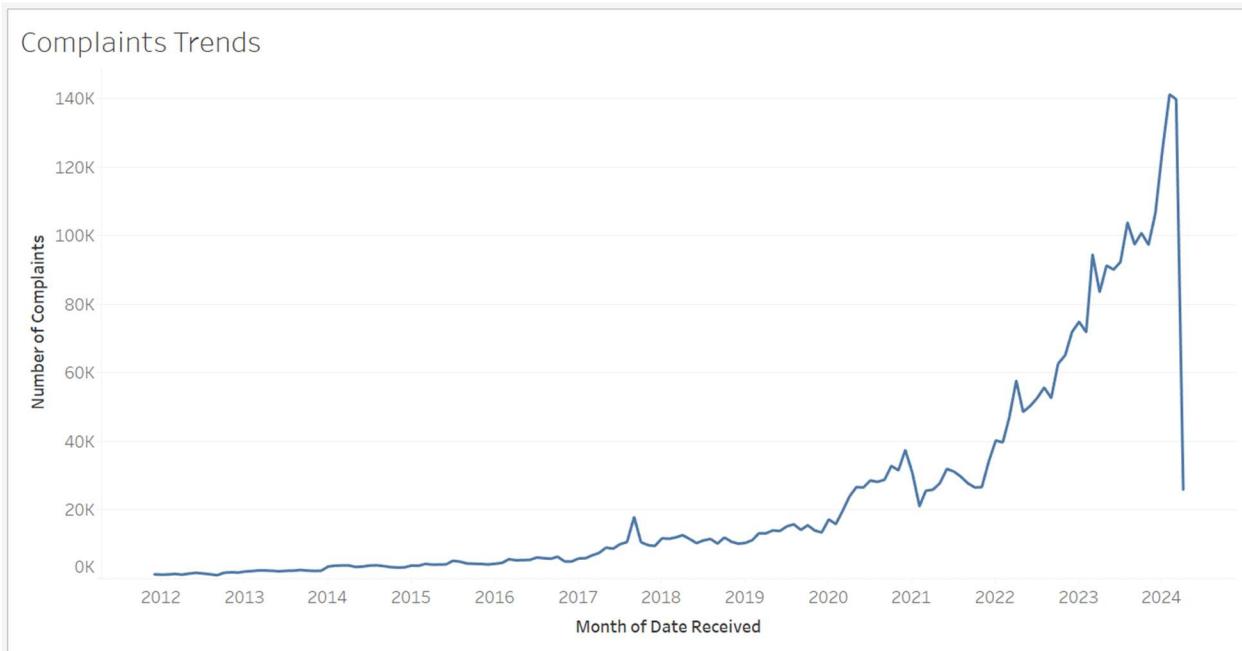
Except for the 2 cases where Spark has Java-Error/Warning, Spark had quick response when compared to Hive for Query Execution.

Data Visualization

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

Graph 1:

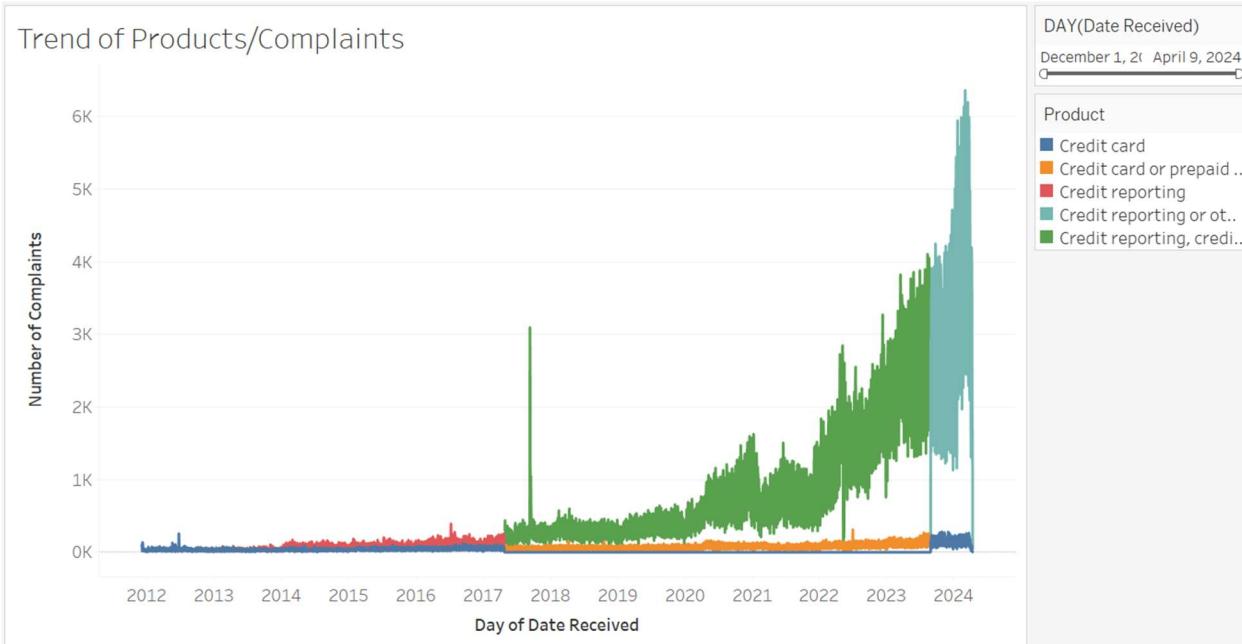
Number of Complaints received over Time, plotted with Count of Complaints over Date Received for **CreditFull16.tsv**



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

Graph2:

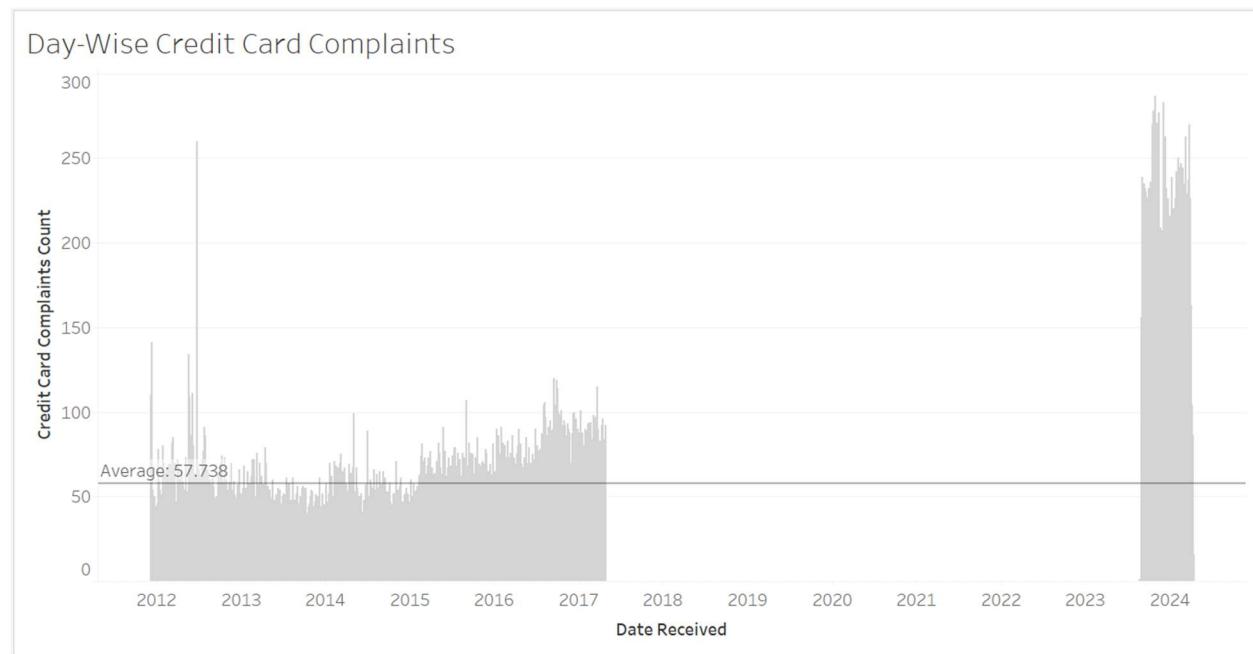
Number of Complaints under each Product over Time, plotted with Count of Complaints over Date Received and Product in Color to differentiate.



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:

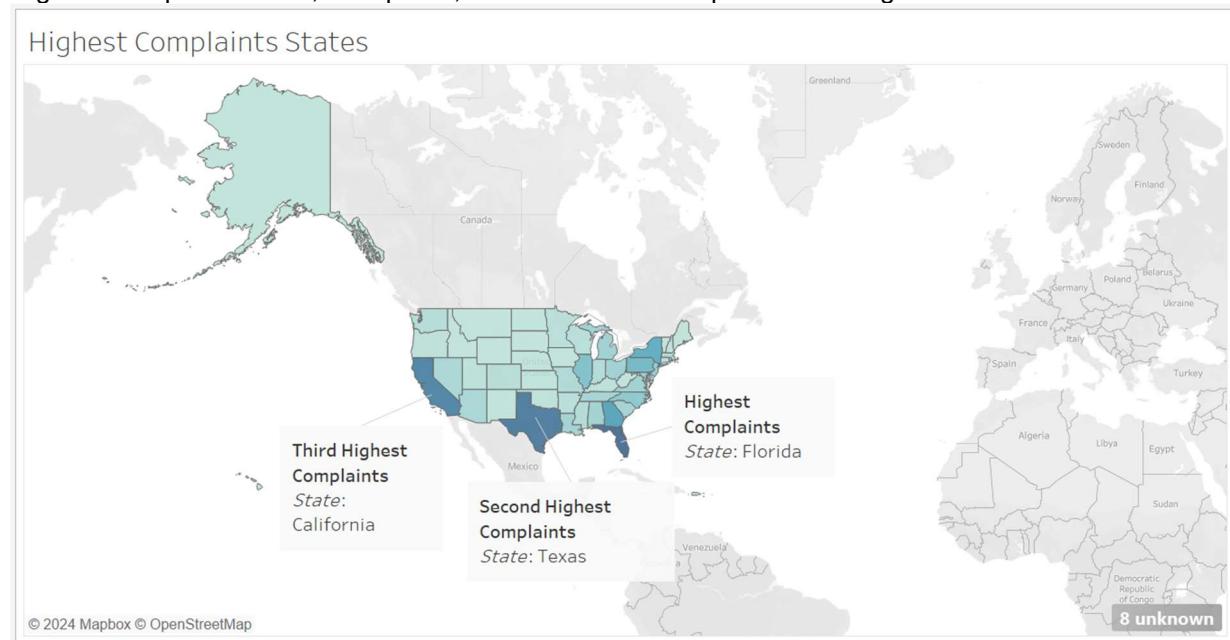
Day-Wise Credit Card Complaints, plotted with Date Received and Count of Complaints for Product – ‘Credit card’.



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)