

CONSUMER COMPLAINTS ANALYSIS

Instructor: Dr. Jongwook Woo

GROUP – 1

Dhwani Vaishnav (Team Lead, Big Data Developer, QA Engineer)

Manimozhi Neethinayagam (Big Data Developer, QA Engineer)

Akanksha Khaire (Big Data Developer)

Mansi Dhoke (Big Data Developer)

Agenda



Introduction



Dataset Details



Cluster Specification



Project Workflow



Hive Tables for Visualizations



Analysis & Visualization



Challenges / Solutions



Conclusion

Introduction

Some Insights on Dataset

- ✓ U.S. government agency dedicated to consumer protection in BANKS and Financial Institutions.
- ✓ Provides data on finance-related issues faced by consumers.

Project Introduction

- ✓ Analyzed consumer complaints in the **USA**.
- ✓ Conducted **year-on-year growth analysis** to identify trends.
- ✓ Analyzed complaints **statistics** specifically for **California**.
- ✓ Performed **sentiment analysis** on consumers' complaints sentiment.
- ✓ Focused on **Ngram Text Processing** for meaningful **insights** from Complaints Narrative column.



Project GitHub link

https://github.com/dvaishna/BigDataDriven_Consumer_Complaints_Analysis

Dataset Details

- ✓ **DATASET NAME:** Consumer Financial Protection Bureau Dataset
- ✓ **DATASET URL:** <https://catalog.data.gov/dataset/consumer-complaint-database>
- ✓ **SIZE:** 3.6 GB
- ✓ **NUMBER OF FILES:** 1
- ✓ **FORMAT:** JSON
- ✓ **COUNTRY CONSIDERED:** USA

- ✓ **DATASET COLUMNS:** “Date received”, “Product”, “Sub-product”, “Issue”, “Sub-issue”, “Consumer complaint narrative”, “Company public response”, “Company”, “ZIP code”, “Tags”, “Consumer consent provided?”, “Submitted via”, “Date sent to company”, “Company response to consumer”, “Timely response?”, “Consumer disputed?”, “Complaint Id”

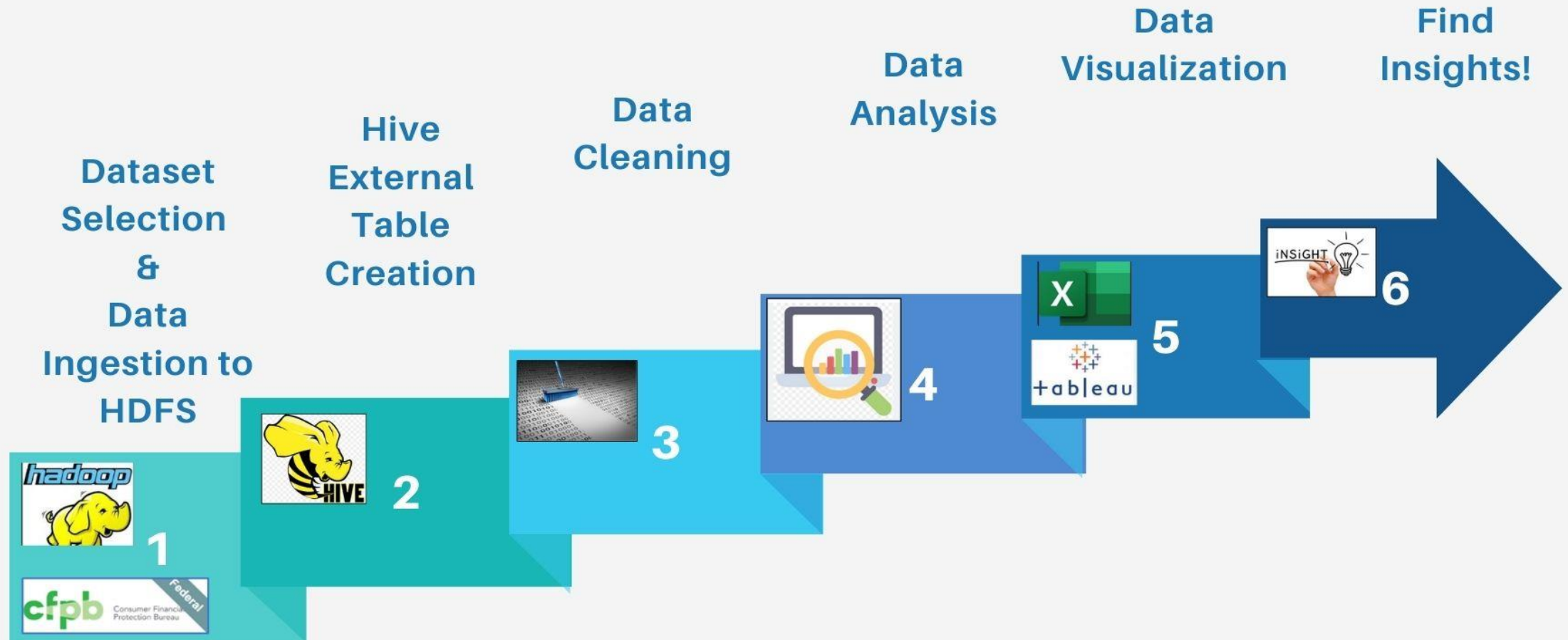
Hadoop Cluster Specification

- ✓ **CLUSTER VERSION:** Hadoop 3.1.2
- ✓ **CLUSTER NODES:** 5 (2 master nodes & 3 data nodes)
- ✓ **MEMORY SIZE:** Memory Used – 367.68 GB, Memory Remaining – 20.96 GB
- ✓ **CPU SPEED:** 1995.312 MHz



Building a
Cluster

6-STEP ANALYSIS PROCESS



Hive Tables for Visualizations

EXTERNAL BASE TABLE

Consumer Complaint
Complaint ID
Date received
Product
Sub-product
Issue
Sub-issue
Consumer complaint narrative
Company public response
Company
State
ZIP code
Tags
Consumer consent provided?
Submitted via
Date sent to company
Company response to consumer
Timely response?
Consumer disputed?

HIVE TEXT PROCESSING

Top Product	Top Companies	Top Issue	Submitted Via
Product	Company	Issue	State
No of Complaints	No of Complaints	No of Complaints	Submitted Via Medium
			No of Complaints

VISUALIZATION 1

Complaints by Date
Date Received
No of Complaints

VISUALIZATION 2

State Table
Complaint ID
Date Received
Product
Company
Issue
State
Zip

VISUALIZATION 3

CA Product	CA Company
State	Company
No of Complaints	No of Complaints
Product	

VISUALIZATION 4

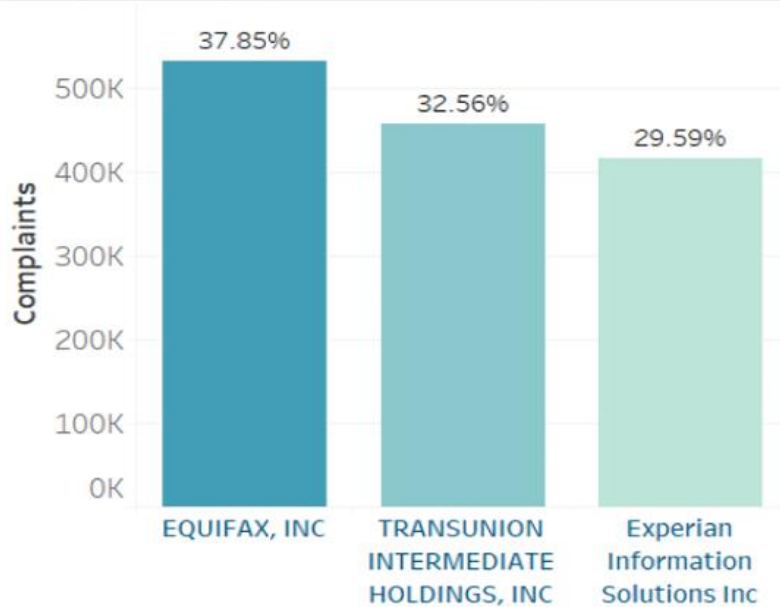
CA Company Public Response
Company
Company Response

Tables For Sentiment Analysis

VISUALIZATION 5

Overall Complaints Statistics

Top Companies w.r.t Highest Complaints



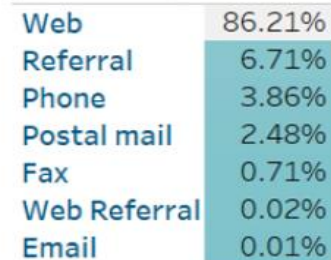
Top Issues w.r.t. Highest Complaints



Top Products w.r.t. Highest Complaints



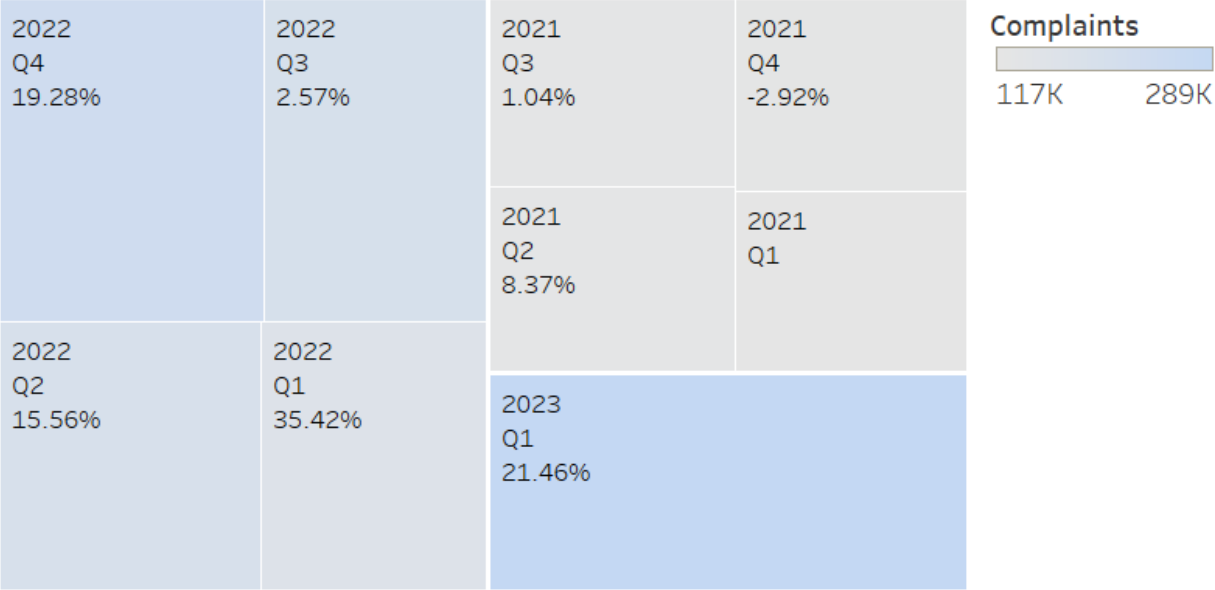
Mediums used for Complaints Submission



- ✓ **Equifax, Inc** received approx. 38% complaints in last 3 years.
- ✓ **Credit reporting** garnered approx. 80% complaints.
- ✓ **Incorrect Information on the report** received approx. 49% complaints.
- ✓ Out of all the complaints, approx. 86% complaints are raised using CFPB **website**.

Year-On-Year Complaints Statistics (2021 to 2023)

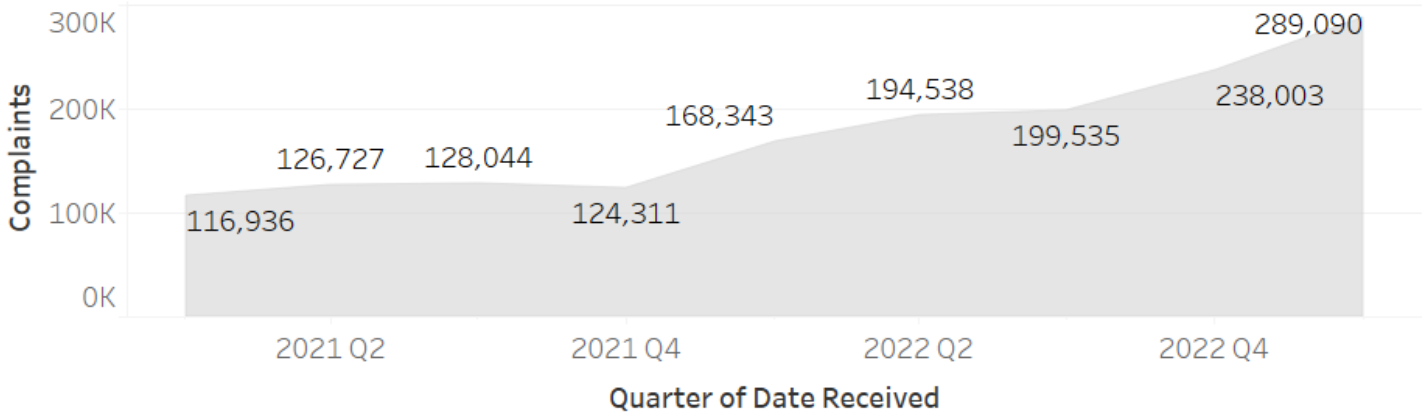
Year on Year Growth Rate



The number of overall complaints increased significantly from **117K** to **289K**

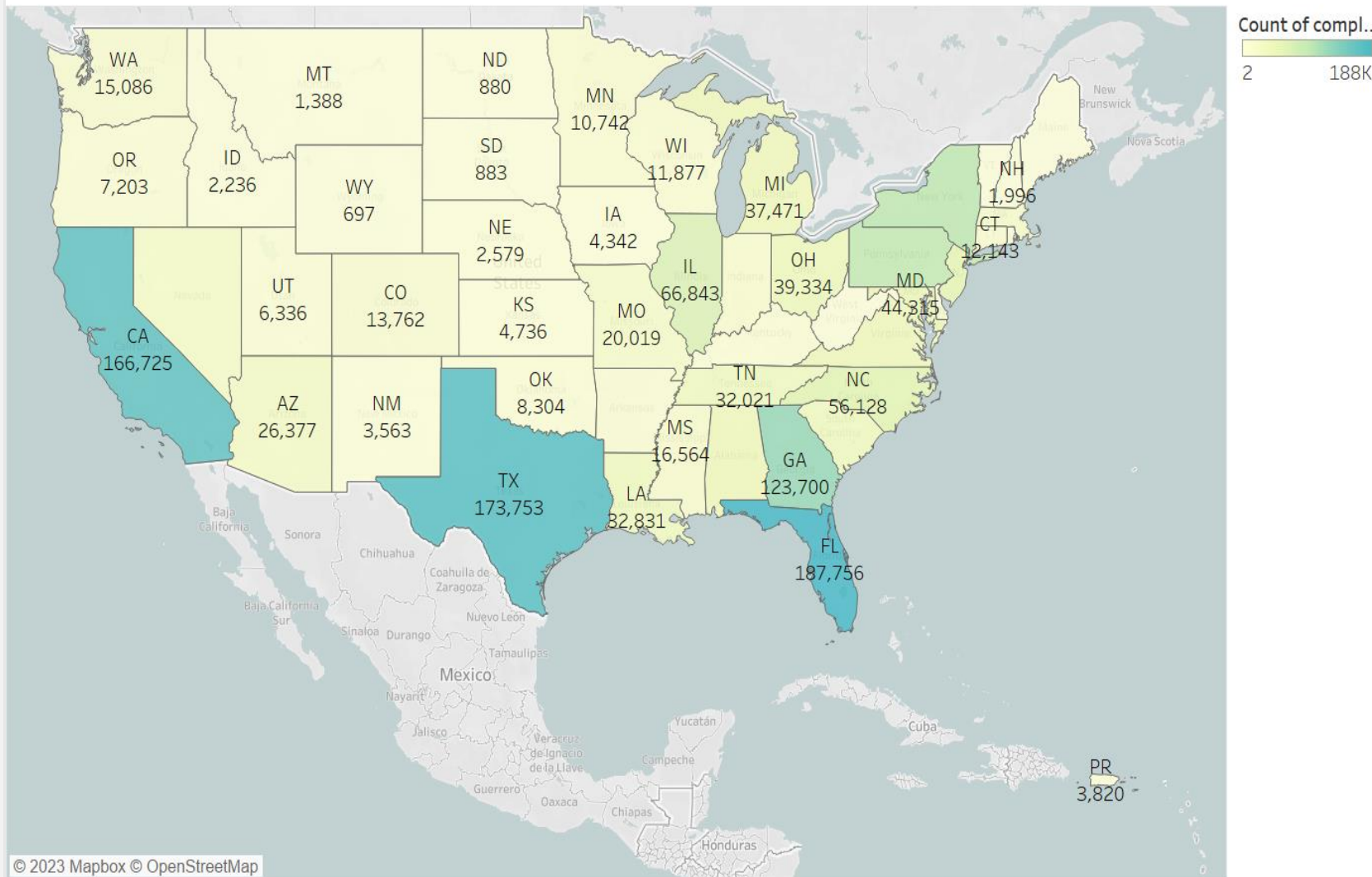
Year	Quarter	Change	% Diff
2021	Q2	↑	8%
2021	Q3	↑	1%
2021	Q4	↓	3%
2022	Q1	↑	35%
2022	Q2	↓	15%
2022	Q3	↓	3%
2022	Q4	↑	19%
2023	Q1	↑	21%

Rate of Complaints



State Based Statistics

Complaints distribution across USA



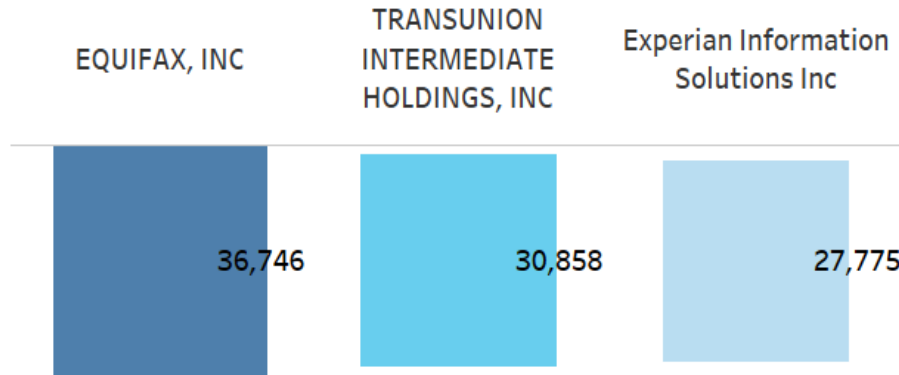
Map based on Longitude (generated) and Latitude (generated). Color shows count of complaint_id. The marks are labeled by state and count of complaint_id. Details are shown for state. The data is filtered on zip, which excludes Null.

- ✓ Consumers in **Florida** raised the highest complaints with approx. **187K** complaints
- ✓ Followed by is **Texas** with approx. **174K** complaints.
- ✓ **California** is the third state with approx. **166K** complaints
- ✓ **Wyoming** is the state with the **lowest complaints** - only **697!**

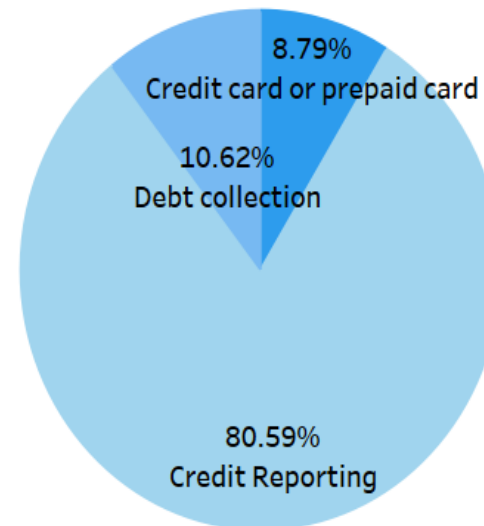
California Complaints Statistics



Companies with Highest Complaints



Products with Highest Complaints



EQUIFAX, INC Public Response	
Closed with explanation	91.68%
In progress	8.12%
Closed with non-monetary relief	0.20%

- ✓ **EQUIFAX, INC** has received the highest complaints for California too!
- ✓ **Credit reporting** has almost 8 times higher complaints rate than the other products/services EQUIFAX Inc offers.
- ✓ Approx. 91% complaints have been **closed with an explanation**
- ✓ Just **0.2%** complaints are **closed without any monetary relief**.

Complaints Narrative - NGram Text Processing

```
+-----+
|                snippet                |
+-----+
| {"ngram":["on","my","credit","report"],"estfrequency":50175.0} |
| {"ngram":["fair","credit","reporting","act"],"estfrequency":29238.0} |
| {"ngram":["the","fair","credit","reporting"],"estfrequency":28003.0} |
| {"ngram":["from","my","credit","report"],"estfrequency":19497.0} |
| {"ngram":["with","the","fair","credit"],"estfrequency":18703.0} |
| {"ngram":["in","accordance","with","the"],"estfrequency":16724.0} |
| {"ngram":["accordance","with","the","fair"],"estfrequency":16129.0} |
| {"ngram":["narrative","in","accordance","with"],"estfrequency":13427.0} |
| {"ngram":["victim","of","identity","theft"],"estfrequency":12536.0} |
| {"ngram":["my","credit","report","i"],"estfrequency":11675.0} |
+-----+
```

- ✓ **"Victim of identity theft"** appears frequently in the list, indicating a significant number of consumers faced this issue.
- ✓ EQUIFAX, INC. needs to address Identity Theft in its credit reporting services.

```
| {"ngram":["narrative","in","accordance","with"],"estfrequency":13427.0} |
| {"ngram":["victim","of","identity","theft"],"estfrequency":12536.0} |
| {"ngram":["my","credit","report","i"],"estfrequency":11675.0} |
```



Complaints Narrative Sentiment Analysis

April 2022

Negative = 0
Neutral = 1
Positive = 2

Layer 1

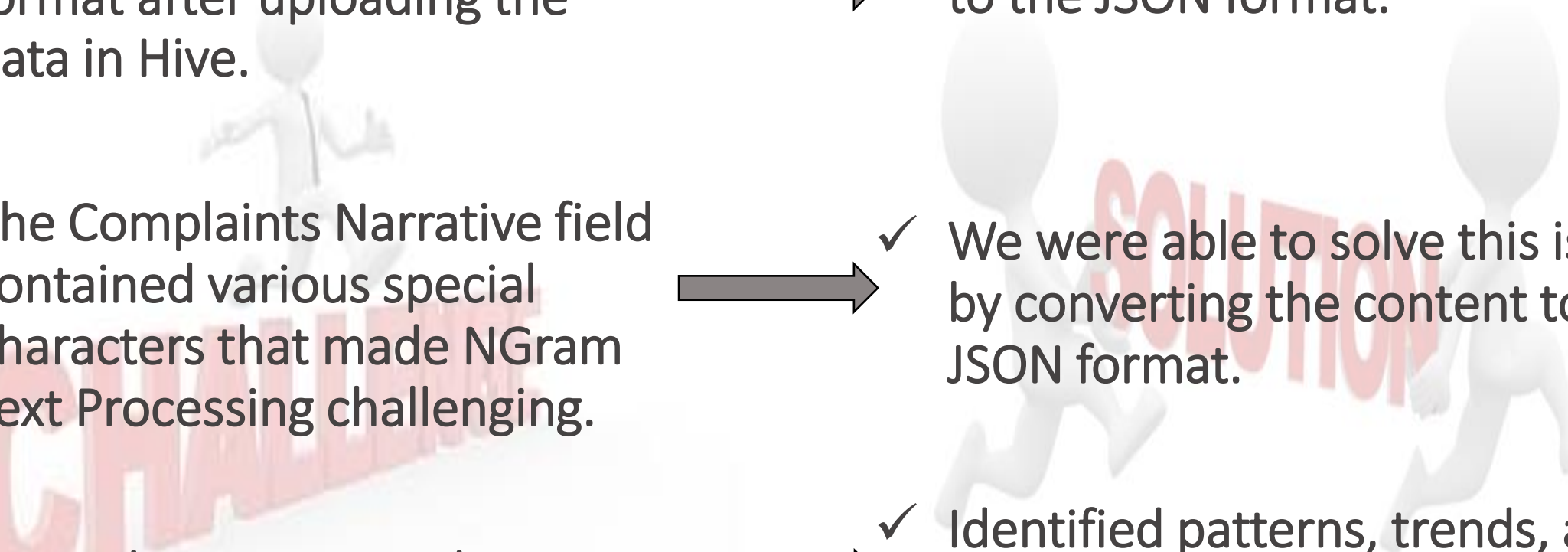
0

1

2



Challenges And Solutions

- 
- ✓ Difficulties in handling CSV file format after uploading the data in Hive. → ✓ To overcome this, we switched to the JSON format.
 - ✓ The Complaints Narrative field contained various special characters that made NGram Text Processing challenging. → ✓ We were able to solve this issue by converting the content to JSON format.
 - ✓ Limited measures in dataset. → ✓ Identified patterns, trends, and relationships using the number of complaints as a measure.



Conclusion

- ✓ The company that received the highest number of complaints related to credit reporting is **EQUIFAX, INC.**
- ✓ The product reported by consumers in the **Credit Reporting** category is **Identity Theft.**
- ✓ **Florida** has the highest number of reported complaints, followed by **Texas & California.**
- ✓ The general sentiment expressed in the complaints is **NEGATIVE.**
- ✓ Most consumers chose the **WEBSITE** as their preferred medium for filing complaints.

THANK
YOU

