Analyzing Insider Trading Data Using Hadoop

Authors: Jeevan Hall, Karen Marcos, Markus Just, Oliver Diaz Department of Information Systems, California State University Los Angeles CIS5200-03 System Analysis and Design

[jhall15@calstatela.edu,](mailto:dmanato@calstatela.edu) [kmarcos@calstatela.edu,](mailto:kmarcos@calstatela.edu,) [mjust@calstatela.edu,](mailto:mjust@calstatela.edu,) [odiaz33@calstatela.edu,](mailto:odiaz33@calstatela.edu,)

**1. Introduction**

This document summarizes the project our group has been working on this semester. We have learned Hadoop and Hive and to work with big data. Our project is about a big data set on insider trading. We downloaded the big data file from Kaggle and uploaded it to Hadoop servers. We then used commands we learned in class this semester to handle big data. Cleaning and summarizing the data directly into Excel or Tableau would not work since the application would crash. This is where the power and advantages of cloud servers come into play. We cleaned and analyzed the data set using Oracle Cloud, and then we extracted sample sizes from our selected rows to visualize data. This project taught us how to deal with big data, use Hadoop and Hive, troubleshoot, and about insider trading.

2. Related Work

A related work to this insider trading data includes the following article entitled Large-scale insider trading analysis: patterns and discoveries. A recent research study looked at how people who work for companies or "insiders" buy and sell stocks within their own company. The researchers used a large dataset from the government with over 12 million transactions made by 370 thousand insiders between 1986 and 2012. Insights from this article include the following: how insiders' trading patterns changed over time because of government rules, company policies, and the economy. In the article some insiders became adept at making a profit based on good timing obviously understanding the insider POV of the trade Additionally, the researchers discovered that insiders often form groups where they share information about trades. The study could help government agencies and law makers create strategies and methods to prevent such illegal activities.

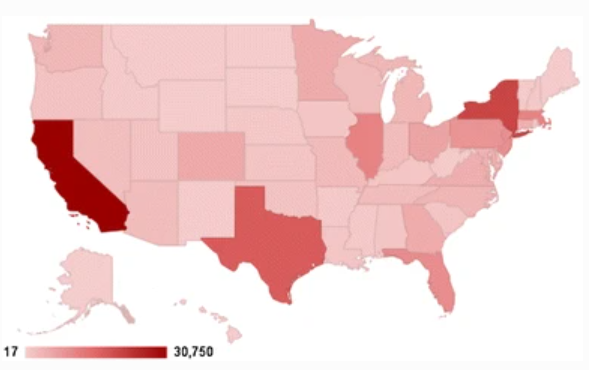


Figure 1: USA Insider Trading Zip Code Analysis

The previous graph is interesting as it shows the number of trades per zip code. The highest is in California even though California has some of the strictest policies in the United States of America. The caption from the image form the article states Geographical distribution of the number of transactions based on the zip codes of the insiders’ companies. Darker color indicates higher number. The highest number of transactions initiates from the state of California

3. Technical Specifications

The dataset is comprised of insider trading activity at publicly traded companies. The SEC (Securities and Exchange Commission) has made these insider trading reports available on its website in a structured format since mid-2003. The dataset is of 32.47GB and covers several years of insider trading data starting from 2003 (05/2003) to 2023 (03/2023). Although the dataset proved to be a big size, we analyzed four files in csv format, adding up to 14.16GB.

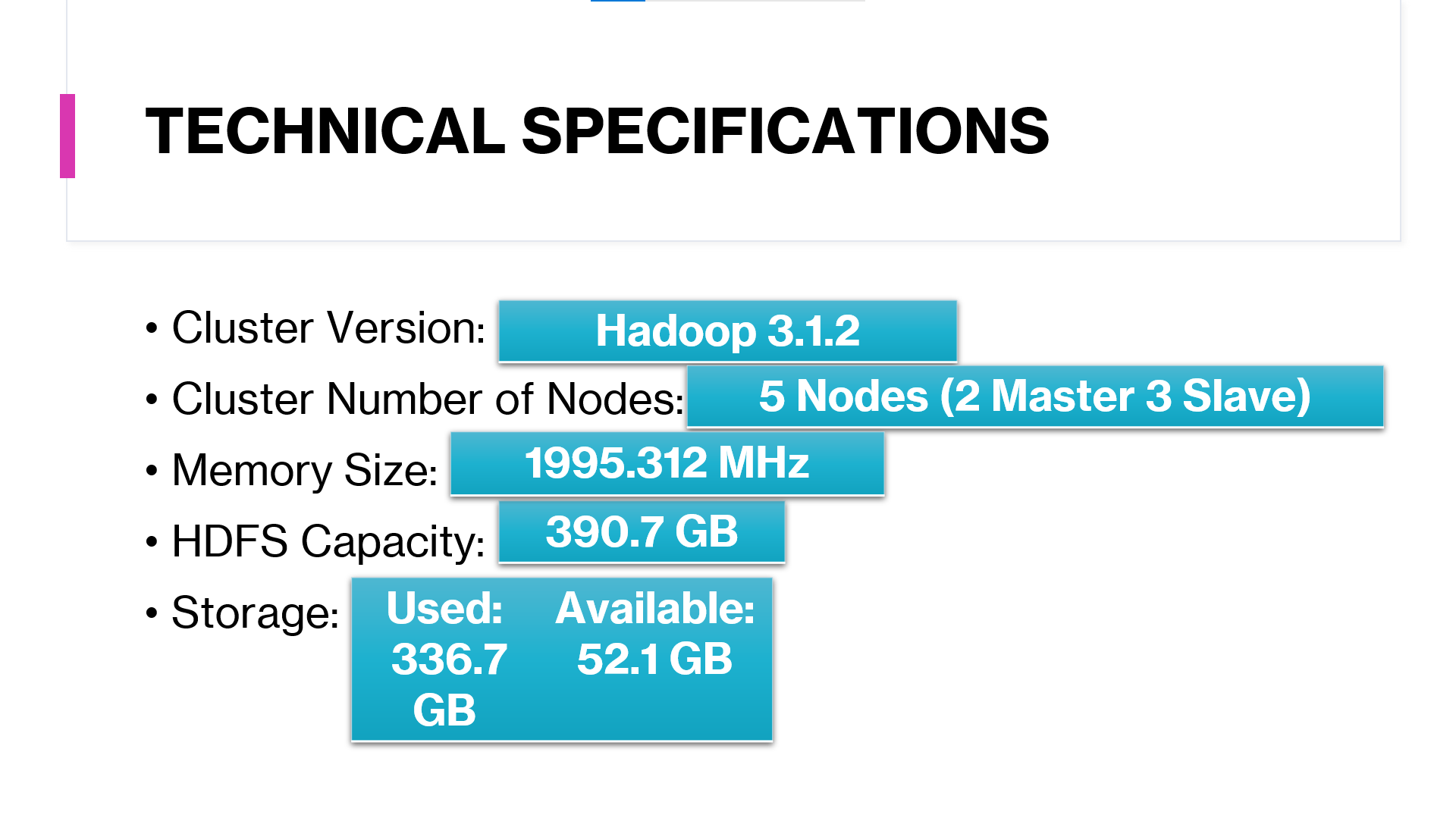


Figure 2: Technical Specifications of Oracle Hadoop Cluster

**3.1 DataSet Information**

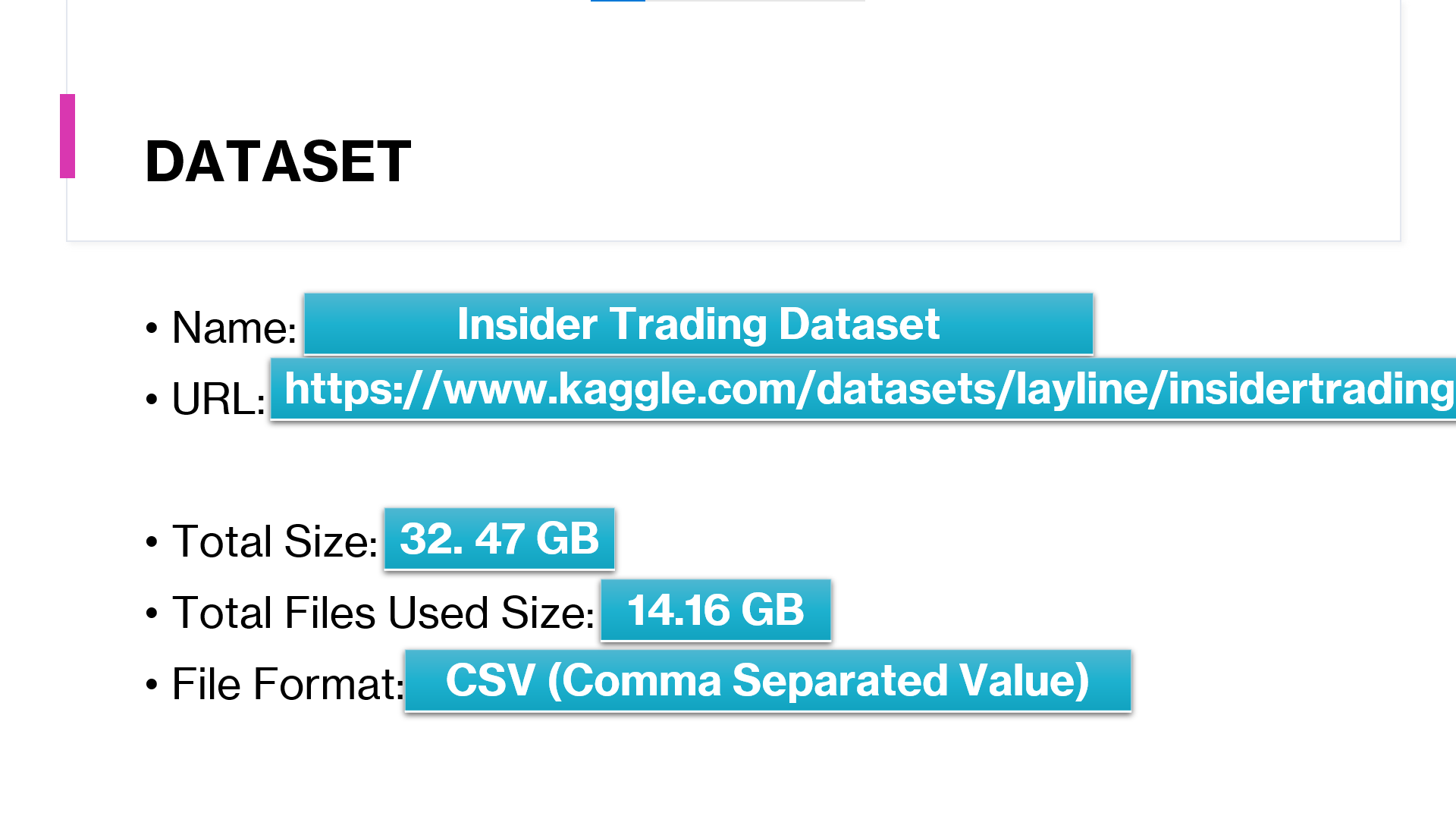


Figure 3: Dataset Technical Details

**Abstract:** After much research, our group decided on the dataset“InsiderTradingDataset”(https://www.kaggle.com/datasets/layline/insidertrading?resource=download). This is an extensive collection of data that focuses on insider trading activity at publicly traded companies. The information was captured by the Securities and Exchange commission that dates back to mid-2003. The dataset that was presented is created from the original regulatory filings in order to overcome limitations because most academic papers use proprietary commercial database that can be quite opaque. This means that this dataset is updated daily and includes all the information that’s reported by insiders without any alteration. This specific dataset had various variables such as their security title, date of transaction, shares, price shares etc. Our group will be analyzing the type of stocks, for example, if it’s an employee stock, director stock, incentive stock, etc. and their activity over time, like the amount of shares and transaction prices.

**4. Flow Chart**

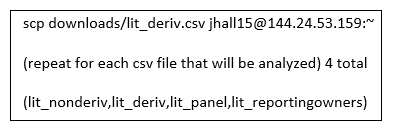
Here you can see our flow chart or data flow. The data source for this project origninated from the populat dataset provider kaggle. Kaggle is the world's largest data science community and also provides resources and forms to help students and professionals grow in the field of data science. The dataset was a relatively clean one, however we did have to clean the data as show in the next section. We had 4 files out of an archive of 8. These files included lit\_deriv,lit\_nonderiv,lit\_panel (merged dataset), and lit\_reporting owners. Most of the group was Mac OS users and therefore utilized the built in terminal in Mac OS however one user utilzed GitBash as a secure shell which in a linux based shell with command line interface. The first step of dealing with the data was downloading 4 csv files from the kaggle website. To download data from kaggle one must create a user profile and be logged in to download the csv file. Once the csv file is downloaded then the process of cleaning the data and organizing it can begin. First we used a secure shell command to the remote server to connect. To get the data inside of HDFS we needed to spc from the downloads folder to the home path of our linux file system. 

Figure 4: Secure Copy Command Local Computer to HDFS

Once the data was inside of HDFS we created a directory for each file which would house each csv.

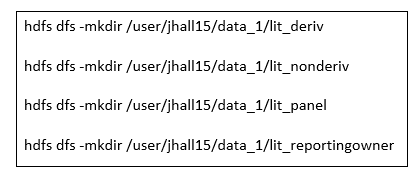


Figure 5: Mkdir Command for eaceh csv

Then we copied each file from the home path to the directories we created.

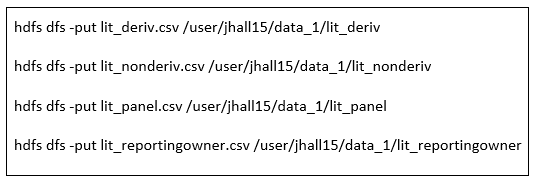


Figure 6: Copy CSV File from Home Path to Directories

Next we connected to HIVE via beeline to start table creation and data cleaning. When connecting to HIVE it is imperative to remind the user to (use db) to create and modify tables in the correct data silo. We created 4 tables, applied random samples of each table and then created queries to take data from a large dataset to a more manageable size. Once we ran the 4 queries we downloaded the csv file from gitbash to the local file system and utlized Tableau as our data visualization tool.

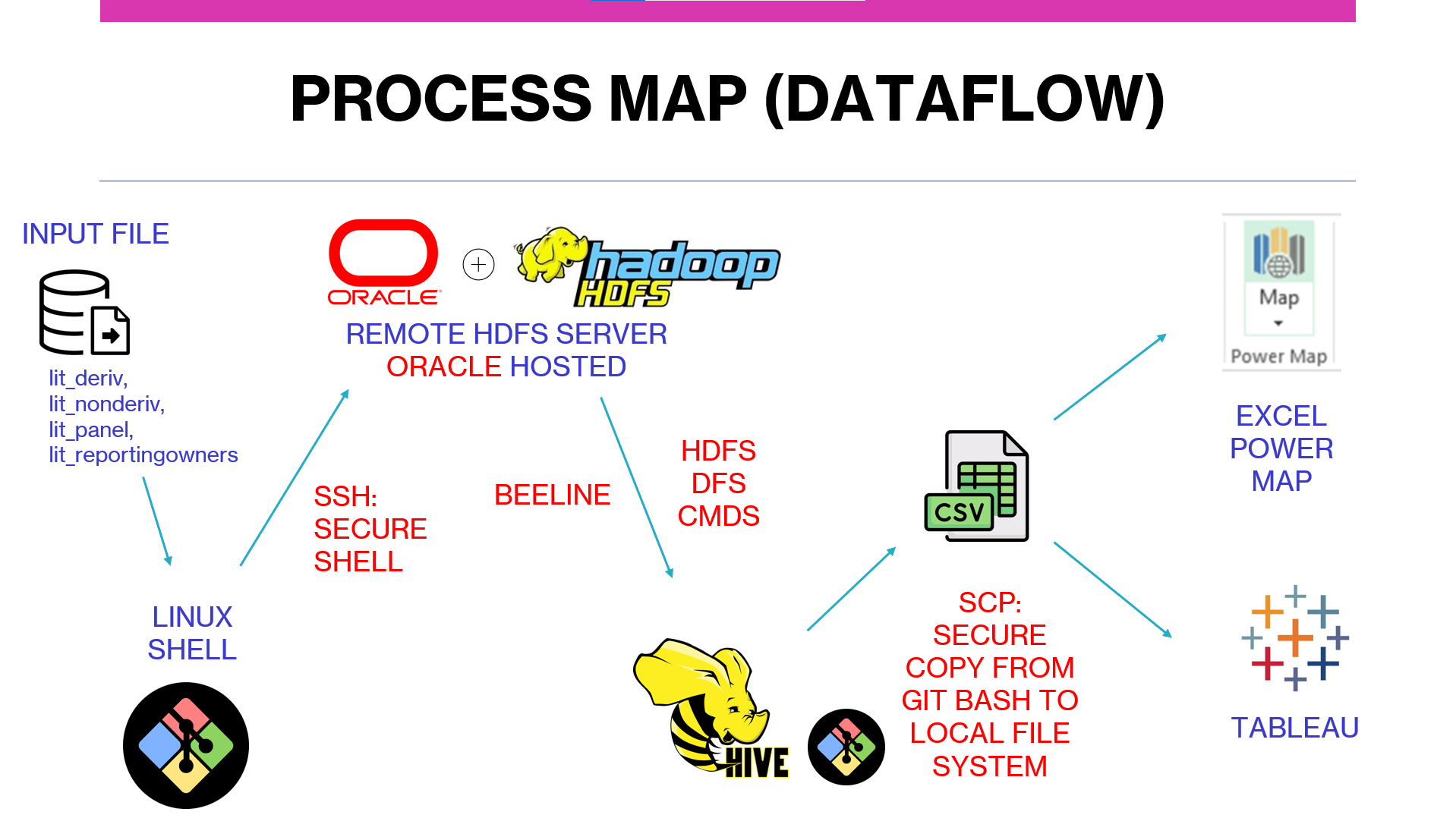


Figure 6 7: DataFlow & Process Map

**5. Data Cleaning**

We utilized various data cleaning methods. To deal with the big dataset we utliized the random function in SQL to gather a much smaller section of data. We are going to walk through the steps we did to clean the data and extract a usable file we could use in Excel and Tableau.

1. First, we downloaded the big data set into Hadoop server. We made a shared directory so every team member couald access it from their own computers.
2. Then, we made a table regarding the dataset we wanted to analyze. The big dataset we had consisted of multiple tables, therefore each group member chose their own data to analyze.
3. After we created a table, we needed to insert the data into it. We did that through the “Serde” command. We then used the “count” function to see if the data got extracted, which it did.
4. Then, we created a table with the table name and then underscore rand (\_rand), because this is the table we wanted to insert random values from the table into.
5. Then, we added random data into the teable through the “INSERT OVERWRITE TABLE” function. Most of the team members chose to pick 100,000 coluns of data to extract, which we thought was a good amount.
6. Then., we created a query table that showed the results from the random table. We populated the table in tmp path of hdfs.
7. Lastly, we secured a copy csv file for analysis from hdfs to local file system.

This made it possible for us to analyze big amounts of data and make visualizations, because we extracted random data that was already in the dataset. We can trust the values and resukts because we extracted a large amount (100,000 clumns), which makes the chances for errors quite slim.

**6. Analysis & Visualizations**

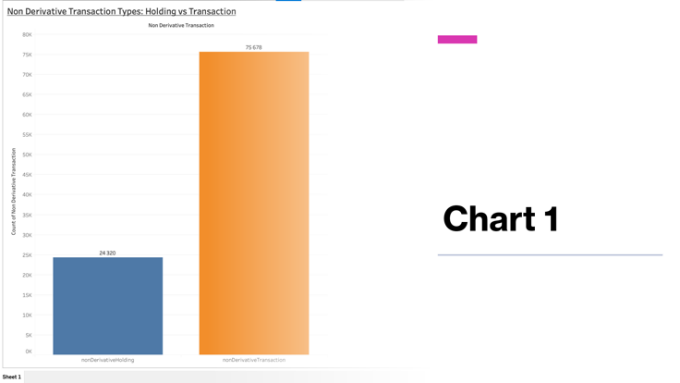


Figure 8: Bar Chart Full

Regarding Chart 1, we visualized the difference of non-derivative transactions and non-derivative holdings. **A non-derivative transaction** refers to a simple purchase, sale, or exchange of a security, such as a stock or bond, without the use of any derivative products like options or futures. This means that the transaction is straightforward and does not involve any added complexity from derivative instruments. While **non-derivative holdings** refers to the ownership of a security without the use of any derivative products. In other words, the investor has a direct ownership interest in the security itself, without any additional exposure to risks or returns that come from derivative products. To summarize, the difference between non-derivative transactions and non-derivative holdings is that transactions involve buying or selling securities without the use of derivative products, while holdings refer to the direct ownership of securities without any added complexity from derivatives. Below is the making of the visualization.

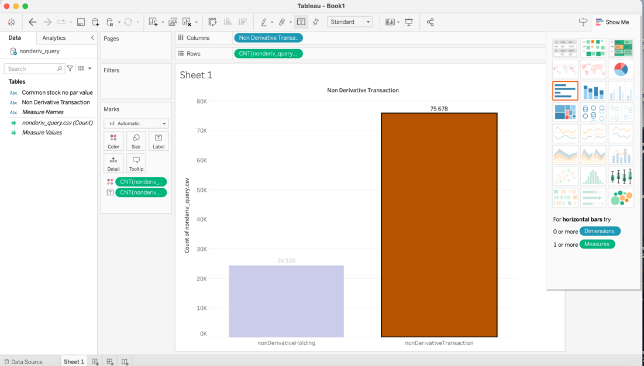


Figure 9: Bar Graph

To make this visualization, we dragged “non-derivative transactions” to columns and dragged the total query we did to rows. As we can see on the visualization, around 24% where non-derivative holding items, where around 75% were non-derivative transaction items.

Before we look at the chart data for chart 2 it is imperative to explain some financial terms which will provide the appropriate context to understand and gain insights from our visualization**. Transaction Share** refers to the percentage of ownership in a company that is represented by a particular transaction.**Direct ownership** refers to a situation where an individual or entity owns an asset or property outright and has legal control and possession over it. **Indirect ownership**, on the other hand, refers to a situation where an individual or entity has a claim or interest in an asset or property, but does not have legal control or possession over it. Indirect ownership can take several forms.The second chart that our group created was from the deriv.csv dataset. This was an interesting dataset as we analyzed not only data by year from 1995 to 2022, but we also looked at direct and indirect ownership. There seems to be a trend of 20% indirect ownership versus 80% direct ownership when we look at the table as a whole.

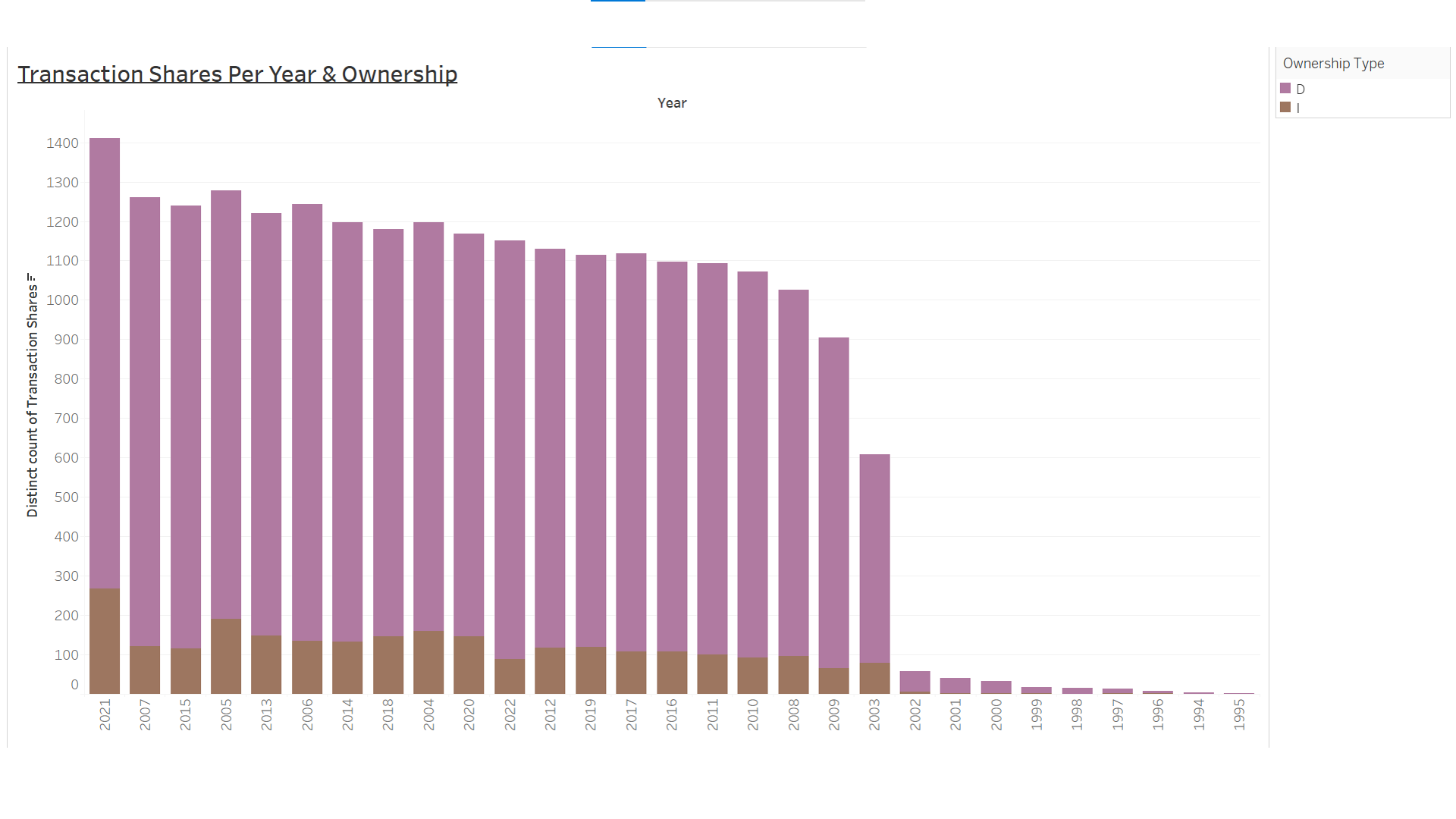
****

Figure 10: Stacked Bar Chart

Here are some quick analysis from our chart. In 2021there was the maximum amount of transaction shares versus in

1995 there was the minimum amount of transaction shares. We had different hypothesis on why there was low reporting in the 1990s and 2000. Some hypothesis included the Y2K dilemna as well as the ENRON scandal of 2001 where ENRON greatly inflated their stock prices. During our class presentation there was some insight from CSULA Wethanie Law which suggests that the governemnt provided bonuses to companies who reported insider trading after 2003 hence the steep increase in the nubmers reported.In this chart Hive helped us segment the data into years and distinguish between direct and indirect ownership. While the graph itself will not completely elimiante insider trading, it does help us track the activity as a whole and possibly understand which companies are at risks for such activities based on the historical numbers and data that we have witnessed for the past decade.

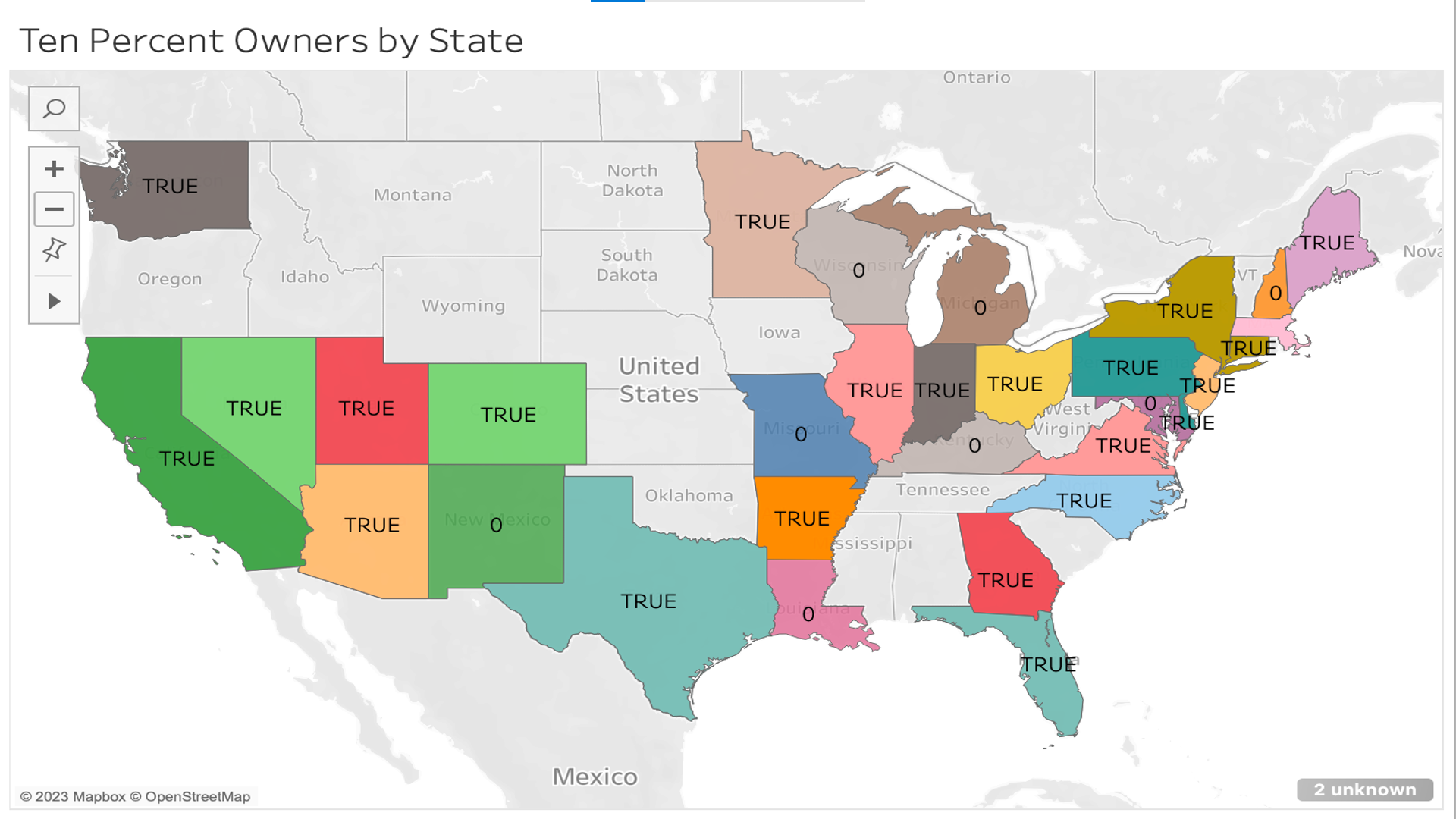
****

Figure 11: Geo-Spatial Map (Top 10% Owners)

Our intital approach to analyze data in insider trading was to create a geo-map to showcase our insights. Before that, we have to understand the data to figure out what it means and why it’s important. 10% Shareholder means a person who owns, directly or indirectly, stock possessing more than 10% of the total combined voting power of all classes of stock of the Company. This is vital in how company is run because those 10% shareholders have certain influence in a company or stock. They are allowed to vote and bake board decisions influencing stocks/shares. From a regulation perspective, this map can provide very valuable insight as to which states hold a 10% majority. Using tableau, I was able to set parameters to show count majority, so if there were more 10% shareholders, I wanted to display “True” on that state. The location of 10% shareholders can assist in regulatory compliance, taxation and reporting, and legal considerations. In other words, this visualization can ensure compliance, enables accurate reporting and holds shareholders responsible while conducting due dilligence.

The final visualization is a box and whisker plot which examines common stocks grouped by month. Examining the dataset and analyzing the relationship between filing dates and security titles, it gains insights into the timing, frequency, and types of securities involved in the insider trading activities. It can help identity patterns, trends or potential correlations between insider trading events and the performance of specific securities. The box in the plot represents the interquartile range (IQR), which contains the middle 50% of the data. The line within the box represents the median. The whiskers extend from the box to the range of the data, excluding outliers. Outliers, if present, are represented by individual points beyond the whiskers. The two values we focused on was the filing date and security title. **Filing date** indicates the date when the insider trading activity was officially reported or filed with the relevant regulatory authorities. By analyzing the filing dates, you can track the timing and frequency of insider trading activities. **Security Title** contains information about the specific security or financial instrument involved in the insider trading transaction. It indicates the name of the security being trade like common stock, preferred stock, options or other financial instruments. It provides insights into the types of securities that insiders are buying or selling. By evaluating the security title and months in the box and whisker plot, I was able to gain insights into the distrubtion of insider trading activities across different securities and time periods. It allowed me to compare the central tendencies,variability, and potential outliers within each category. I was able to obsrve if certain security titles or months have higher or lower levels of insider trading actiity and helps identify any patterns or trends.

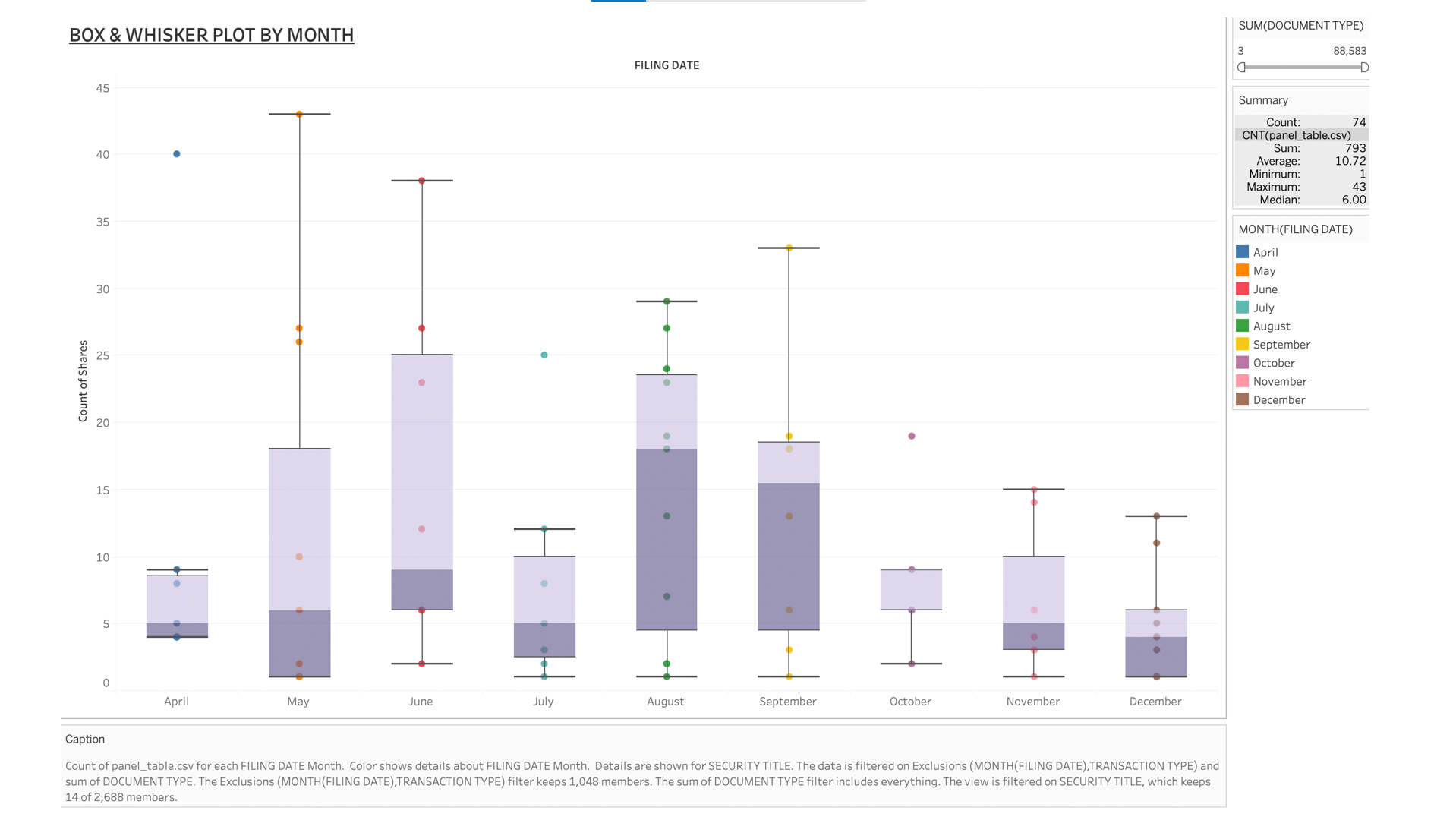
****

Figure 12: Box & Whisker Plot (Share Types by Month)

**7. Challenges**

As in every group project, you will have challenges. Typical challenges most groups meet, including ours was to allocate time, allocate work, and meet deadlines. However, we feel we contributed fairly even, and worked together well. We met often in-person, and sometimes over Microsoft Teams to handle new issues or challenges.

A big challenge we had for weeks was to get the dataset into Hadoop from Kaggle. We met multiple weekends at the library trying to figure it out for 4-5 hours each time. This was a frustrating part of the project, because we feel we were stuck. However, with the help of the professor and us working together, we found out eventually. Even though it was tough at that moment, we understand and appreciate it now since we learned a lot from testing different strategies and troubleshooting.

### References

[1] Tamersoy, A., Khalil, E., Xie, B. et al. Large-scale insider trading analysis: patterns and discoveries. Soc. Netw. Anal. Min. 4, 201 (2014). https://doi.org/10.1007/s13278-014-0201-9

[2]http://www.ohsdba.cn/index.php?m=Article&a=show&id=351

[3] Balogh, A. Layline Insider Trading Dataset. Harvard Dataverse https://doi.org/10.7910/DVN/VH6GVH (2023).

[4] Balogh, A. Insider trading. Scientific Data 10, 237, https://doi.org/10.1038/s41597-023-02147-6 (2023).

[5]