Team Data Group Final Report

By Garrett Albracht and Cole Clark

Table of Contents

Introduction

The goal of our project is to enhance data processing and analytics capabilities, especially in handling large-scale datasets. We aim to develop a scalable architecture that efficiently manages data ingestion, querying, and analysis, providing insights from complex data sets. The importance of this project lies in its ability to handle growing data volumes and complexities, a critical need in today's data-driven world. On a smaller scale, our problem focuses on the analysis of the International Consortium of Investigative Journalists (ICIJ) Offshore Leaks Database, which specifically addresses the United States' involvement in the Panama Papers. The Panama Papers consist of a large database of leaked documents that exposed a global network of offshore companies used for tax evasion, fraud, and other illegal activities.

The majority of this problem set is to transform the Panama Papers into a structured network of entities, including persons, companies, and their interrelationships, all anchored in specific timeframes. This network is managed using Neo4j, a graph database that organizes data into nodes (representing entities such as individuals and companies, along with intermediaries and officers) and relationships (the connections between these nodes).

Gaining insight into the Panama Papers is pivotal for this project as it represents an example of handling complex, large-scale datasets to uncover significant real-world problems. The Panama Papers, a massive collection of leaked documents, revealed global patterns of financial secrecy and corruption. Analyzing a dataset like this one demands robust data processing and analytics capabilities. This project's focus on scalable data management and efficient analysis directly aligns with the challenges posed by the Panama Papers, emphasizing the importance of advanced data processing tools in revealing critical insights from in-depth datasets.

One of the important things relating to the Panama Papers is the use of tax haven countries to pay less taxes on assets. Tax haven countries are countries that have lower taxes relating to investments such as capital gains taxes and income taxes. These countries attract wealthy foreign investors who want to pay less in taxes (Investopedia). We wanted to know which tax haven countries were most popular with wealthy U.S. investors in the Panama Papers.

Related Work

We got our data from the International Consortium of Investigative Journalists (*Offshore leaks database).* This data included information from the Panama Papers leak along with four other leaks. We reduced this data by removing all information that was not in the Panama Papers leak. This reduced the data from 617 MB to 98 MB. After this, we decided to only focus on United States nodes and nodes connected to a United States Node. To do this, we added the addresses of each entity, officer, and intermediary to their tuples by using the relationship and address files. After this, we removed the address file and had four files remaining (officers, intermediaries, entities, relationships). Next, we made a subset of the remaining entities, officers, and intermediaries that only included the ones in the United States. We used these subsets to determine which relationships contained a U.S. node and removed the ones that didn’t. To find the data we wanted, we made new subsets of the officers, entities, and intermediaries that were in the relationships file. We ended up with all nodes in the United States along with all foreign nodes they were directly connected to. This reduced our data to about 3.13 MB.

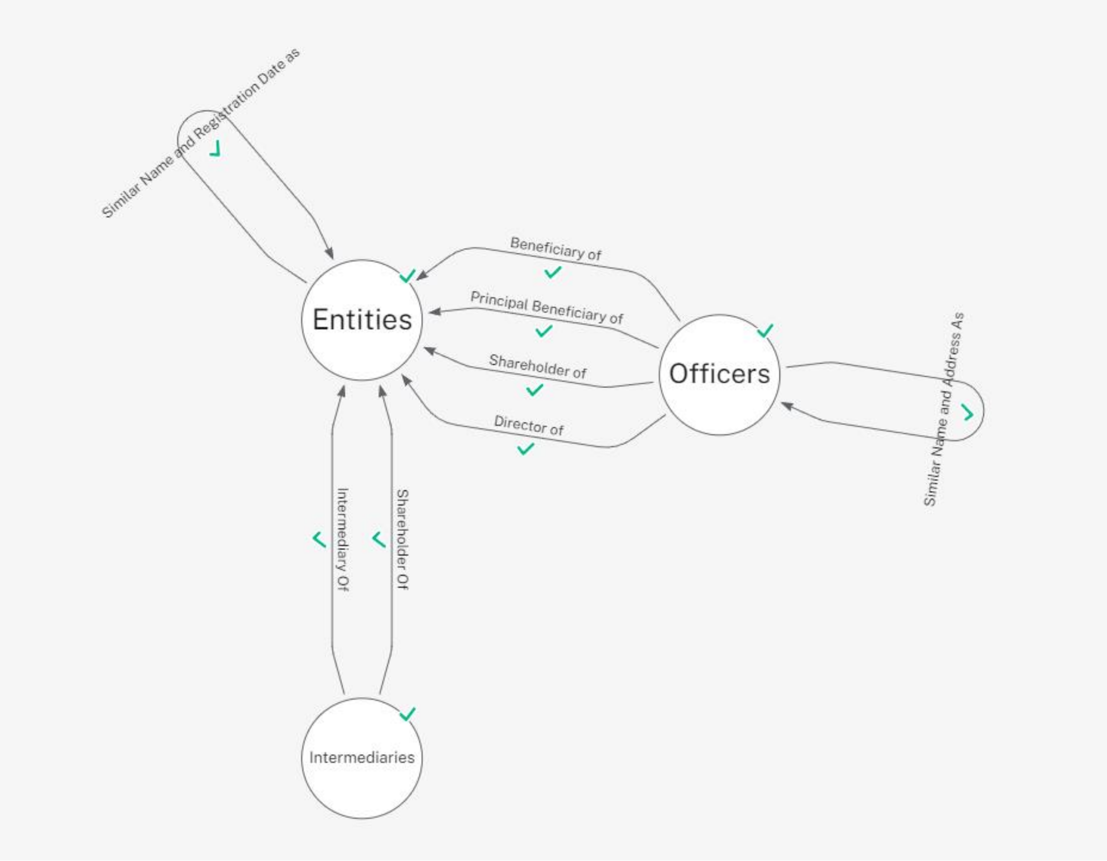
Our database differs from the ICIJ database since it is only focused on U.S. nodes and their direct connections in the Panama Papers rather than all data from the other leaks and nodes without a direct connection to the United States.

Architecture / Design

Overview

We've developed a distributed architecture that uses Apache Spark for efficient data processing and Neo4j for effective data management. We also used pyPlot and Pandas for efficient conversion to plot our data, making the code easy to read and further algorithm additions simple.

Data Model / Database Schema



The image above shows the schema for our database. In this schema, an entity is a business, shell company, or another form of business entity. Officers could be shareholders, directors, or beneficiaries of entities. Intermediaries could be shareholders or intermediaries of entities.

Our data model is designed to maximize efficiency in storage and retrieval, with a schema optimized for the types of queries and analyses our system supports.

Algorithms / Queries

We've developed custom algorithms that are optimized for distributed processing. These include data ingestion routines and complex data queries that can handle large datasets effectively.

getSparkDF(query): This function takes a Cypher query as input and returns a Spark DataFrame. It establishes a connection to the Neo4j database using Spark's read function, allowing for complex graph queries to be executed and their results to be manipulated using Spark's DataFrame API. It is used as a helper function for many of our other functions to get data from the database. The time it takes for this function is heavily influenced by the query being run.

new shortest path (startNode, endNode): This function computes the shortest path between two nodes in the graph, identified by their names. It is particularly useful for tracing the connection chain between two entities or individuals within the Panama Papers network. This function uses the cypher shortest path function. When tested this function took less than 0.4 seconds to run.

Countries(type): This function takes an input one, two, or three and displays a graph of the number of connections from each country. If one is entered in this function a graph displaying the number of connections each country's entities have to U.S. nodes will be displayed. If a two is entered, a similar graph will be displayed with the number of connections between each country’s Intermediaries and the United States. If a three is entered, a graph will display the same thing for Officers. We ran into an issue with the Intermediaries graph. Some countries were listed as “United States;” followed by the actual country. To solve this we used a user-defined function to remove this string if it appears. for all three calls of this function, it took less than 1.7 seconds to run.

sparkConnected(nodeName): This function retrieves all nodes directly connected to a given node (specified by nodeName). It uses a Cypher query to match and return the connected nodes and the type of connection, which is useful for understanding direct relationships in the Panama Papers data. When tested this function took less than a second to run.

timeActive(): This function calculates the active time of entities in the Panama Papers. It takes parameters to order the results and to remove null values. It uses Spark's date functions to compute the duration between incorporation and inactivation dates, providing insights into the longevity of entities. The function then outputs an ordered graph of the data described above. The parameters of this function can be used to order the result in ascending order or choose to keep nodes without an end date. When no parameters are entered, the data is in descending order and all entities with no end date are removed. We ran into an issue when creating this function. Dates for start and end dates were stored as two digits so when we transformed the dates from strings into date values every year was assumed to be after 2000. This caused some issues with ordering entities by time active so we created a user-defined function to add the first two digits to every year. If a year was less than 20, we assumed it was after 2000; if it was over 20, we assumed it was between 1900 and 2000. When tested, this function took about 12 seconds to run.

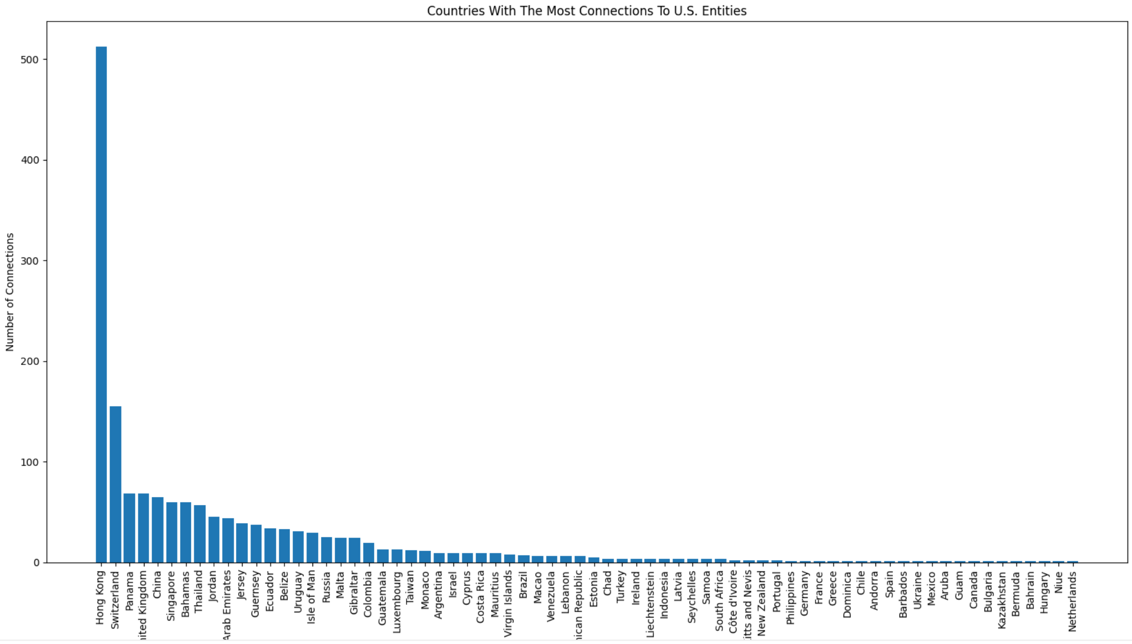
User Interface / User Interaction

The user interface is designed for ease of use using Tkinter, allowing users to interact with the system intuitively. It includes data visualization tools at the click of a button that helps in interpreting the results of data queries and analyses.

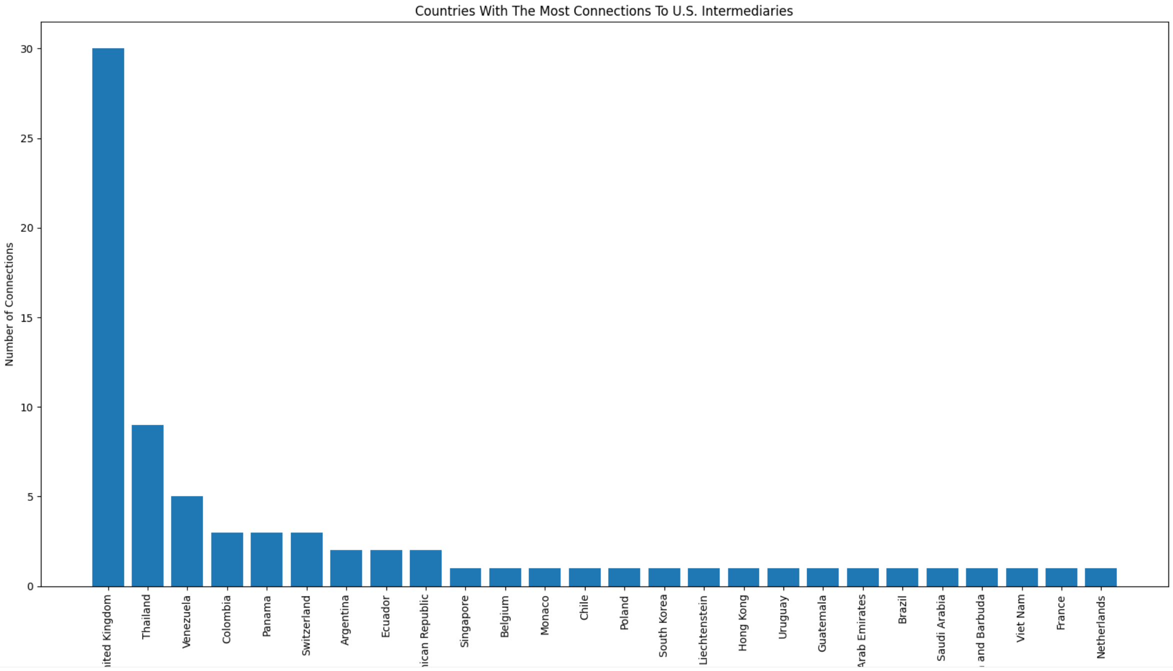
We incorporated Pandas and pyPlot into our algorithms for a smooth inclusion of graph data to go along with our UI. These pyPlot graphs show a colorful and easy-to-read format of the data that is processed in our various functions, making it easy for the user to understand the data being shown and what exactly is represented.

Results and Findings

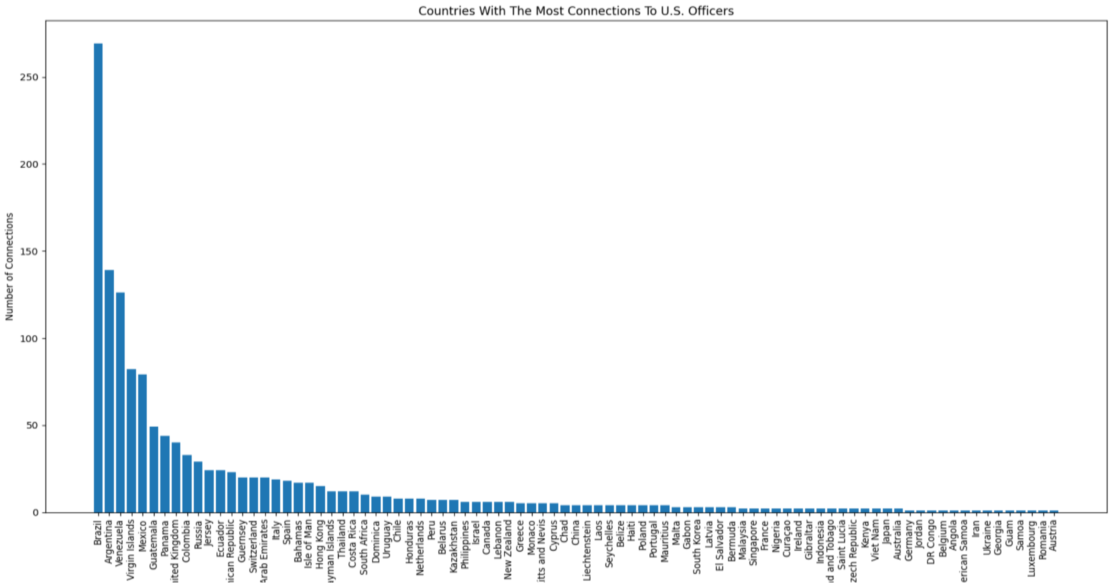
We learned a lot from the results of our algorithms. The countries() function had some particularly interesting results. This function can show the number of entities, intermediaries, or officers from a country that has a connection to a node in the United States. The results for entities shown below indicate that Hong Kong had by far the most entities with a connection to a United States node with over 500. This is interesting because Hong Kong is considered one of the best tax haven countries second only to Bahrain (Goetz). The corporate tax rate in Hong Kong is between 8.25% and 16.5% depending on the amount earned in Hong Kong (Goetz). This is significantly lower than the U.S. corporate tax rate of 21% with the possibility of additional taxes depending on the state (Watson).



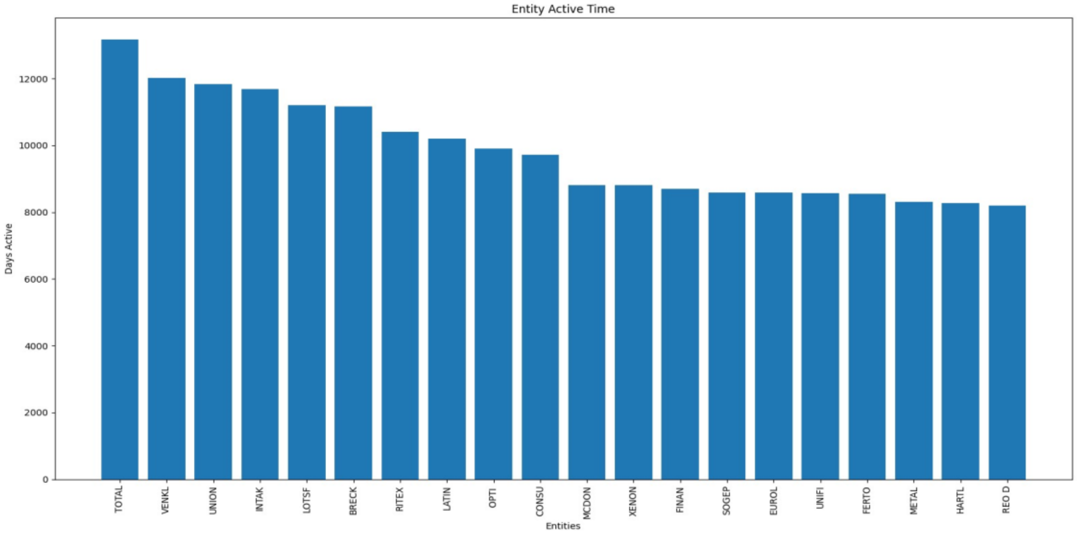
Another interesting result from the countries() function comes from entering intermediaries. As shown below, the most common country for an intermediary with a U.S. connection to be from is the U.K. This makes sense because England is one of the best tax havens in Europe with low capital gains taxes on foreign investments and low income taxes (Parietti). Capital gains tax is a tax on the gain made by holding an asset like a stock. Since these taxes are low in England it is a great place to open a trust with stocks and other assets (Parietti).



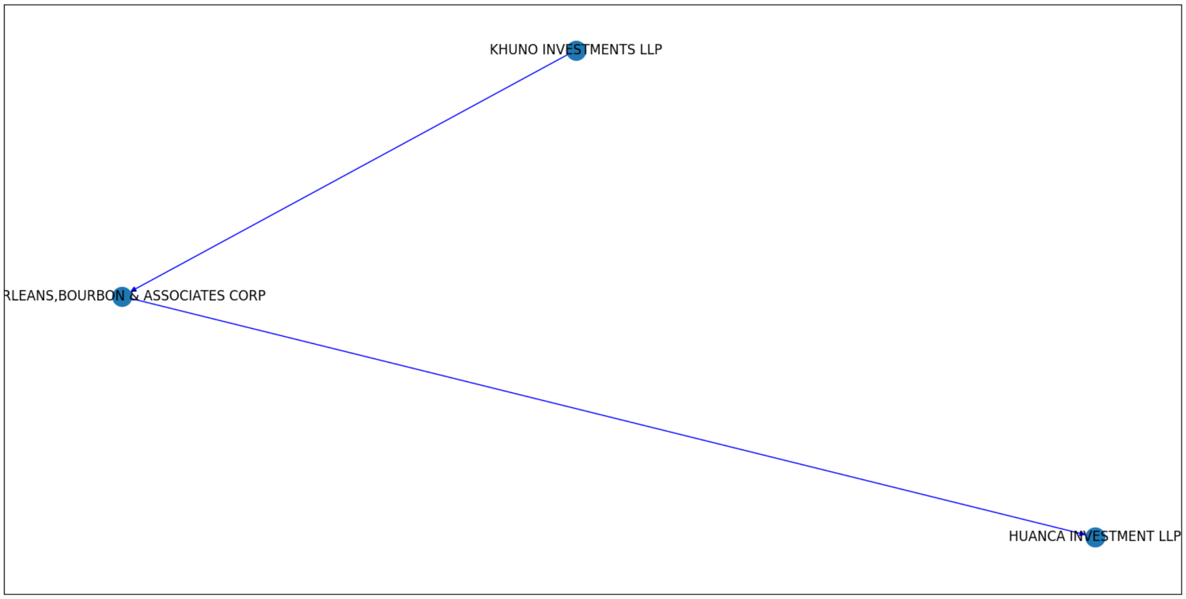
The most interesting discovery comes from our officer graph. This graph is shown below with Brazil as the country with the most officers that have a connection to the United States. Brazil is not a tax haven and has special tax laws that discourage the use of tax haven countries (Barnett). This suggests that people in Brazil are investing in U.S. companies partly because corporate income taxes are 21% in the United States which is slightly above the cutoff for increased taxes from Brazil (Watson). Brazil increases taxes on investments that are in countries with less than a 20% income tax rate (Barnett). This makes businesses in the United States a good option for Brazilian investment.



The following bar graph is associated with our time active function. As we can see, after the first 10 columns, the times active were all very similar. This may be attributed to entities open and closing at similar times from the timeline of the Panama Papers.



This plot is associated with the function shortest path. While there aren't many insights to gain from this, it would be very useful in the analysis of how close 2 entities are connected, or if they have no connections at all.



References

(NNK), Naveen. "Convert Pyspark DataFrame to Pandas." *Spark By {Examples}*, 31 Oct. 2023, sparkbyexamples.com/python/convert-pyspark-dataframe-to-pandas/.

"Matplotlib.Pyplot.Figure#." *Matplotlib.Pyplot.Figure - Matplotlib 3.8.2 Documentation*, matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.figure.html. Accessed 11 Dec. 2023.

"Pandas.Dataframe.Hist#." *Pandas.DataFrame.Hist - Pandas 2.1.4 Documentation*, pandas.pydata.org/docs/reference/api/pandas.DataFrame.hist.html. Accessed 11 Dec. 2023.

"Python GUI Programming with Tkinter." *Real Python*, Real Python, 30 Jan. 2023, realpython.com/python-gui-tkinter/.

Barnett, G. (2022, November 14). *Tax Havens pay a hefty price in Brazil*. Brazil Counsel. https://www.brazilcounsel.com/blog/tax-havens-pay-a-hefty-price-in-brazil#:~:text=Currently%2C%20Brazil's%20list%20of%20tax,state%20is%20on%20the%20list.

Goetz, L. (2022, March 29). *Why is Hong Kong considered a tax haven?*. Investopedia. https://www.investopedia.com/ask/answers/060916/why-hong-kong-considered-tax-haven.asp#:~:text=Tax%20havens%20are%20countries%20with,wealthy%20foreign%20residents%20and%20corporations.

Investopedia. (2023, December 4). *Tax haven: Definition, examples, advantages, and legality*. Investopedia. https://www.investopedia.com/terms/t/taxhaven.asp#:~:text=Tax%20havens%20are%20jurisdictions%20that%20attract%20foreign%20investors%20with%20comparatively,attractive%20destinations%20for%20foreign%20capital.

Offshore leaks database. How to download this database | ICIJ Offshore Leaks Database. (n.d.). https://offshoreleaks.icij.org/pages/database

Parietti, M. (n.d.). The top 10 European tax havens. Investopedia. https://www.investopedia.com/articles/wealth-management/121515/top-10-european-tax-havens.asp#:~:text=Key%20Takeaways,tax%20havens%20on%20the%20continent.

Watson, G. (2023, July 25). *Combined state and federal corporate income tax rates in 2022*. Tax Foundation. https://taxfoundation.org/data/all/state/combined-federal-state-corporate-tax-rates-2022/#:~:text=Combined%20Federal%20and%20State%20Corporate%20Income%20Tax%20Rates%20in%202022,-September%2027%2C%202022&text=Corporations%20in%20the%20United%20States,at%20a%2021%20percent%20rate.