Projet

# *Data Integration and Analysis Using a Data Warehouse and NoSQL*

### I - Dataset Selection and Domain Analysis:

We chose the “Top Global 2000 Largest Companies 2024” dataset from Kaggle due to its relevance in financial and economic analysis. This dataset consolidates key financial information on the world's 2000 largest companies, including their country of operation, revenue, profit, and assets. By leveraging this dataset, we aim to explore large-scale economic dynamics and examine interconnections between companies across different regions.

One of the key objectives of our analysis is to conduct a comparative study of companies based on their financial performance and geographic distribution. Additionally, we seek to explore intercompany relationships using graph database modeling with Neo4j, allowing us to identify regional influences, competitive links, and sectoral trends.

However, before conducting these analyses, we had to address several challenges related to data integration and cleansing. The dataset presents inconsistencies in financial values, where revenue and profit are formatted differently, sometimes using symbols and varying scales such as billions or millions. To ensure uniformity in analysis, we standardized these values into a single unit. Another limitation is the absence of industry classification, which restricts sector-based comparisons. To overcome this, we considered using external datasets or inferencing industries based on company names. Additionally, the dataset lacks a time dimension, making it difficult to track financial performance trends over multiple years.

By integrating structured SQL-based data with graph-based NoSQL modeling, we enable a deeper understanding of economic trends, corporate interconnections, and competitive landscapes. Our data preprocessing efforts ensured consistency, allowing for meaningful analytical insights.

### II - Logical Design of the Data Warehouse:

To efficiently structure our data warehouse, we implemented a star schema, which provides a simple yet powerful approach to organizing data for analytical processing.

### 

The fact table, Fact\_Companies, stores key financial metrics, including sales, profit, assets, and market value. Two supporting dimension tables, Dim\_Company and Dim\_Country, provide contextual information related to companies and their respective countries of operation.

The Dim\_Company table contains unique company identifiers and company names, allowing for efficient mapping of company-related financial data. Similarly, the Dim\_Country table captures country-specific information, linking each company to its respective country. The Dim\_Country table follows a hierarchical structure, organizing locations at different levels, such as country and economic region, to facilitate multi-level aggregations and comparative analyses across broader geographic areas.

The relationships among these tables facilitate insightful analyses. By linking the Fact\_Companies table to the Dim\_Company and Dim\_Country tables, we enable structured queries that provide valuable insights into financial trends. Additionally, the hierarchical structure of Dim\_Country allows for analyses that can be aggregated at different levels, such as total revenue by country or by economic region. This flexibility enhances our ability to generate more granular or broader insights as needed.

We opted for a star schema due to its efficiency in OLAP queries. The structure minimizes the complexity of joins, ensuring faster execution and improved performance. The simplicity of the model also supports scalability, allowing for easy extension of the data warehouse as new requirements emerge.

### III - Physical Implementation:

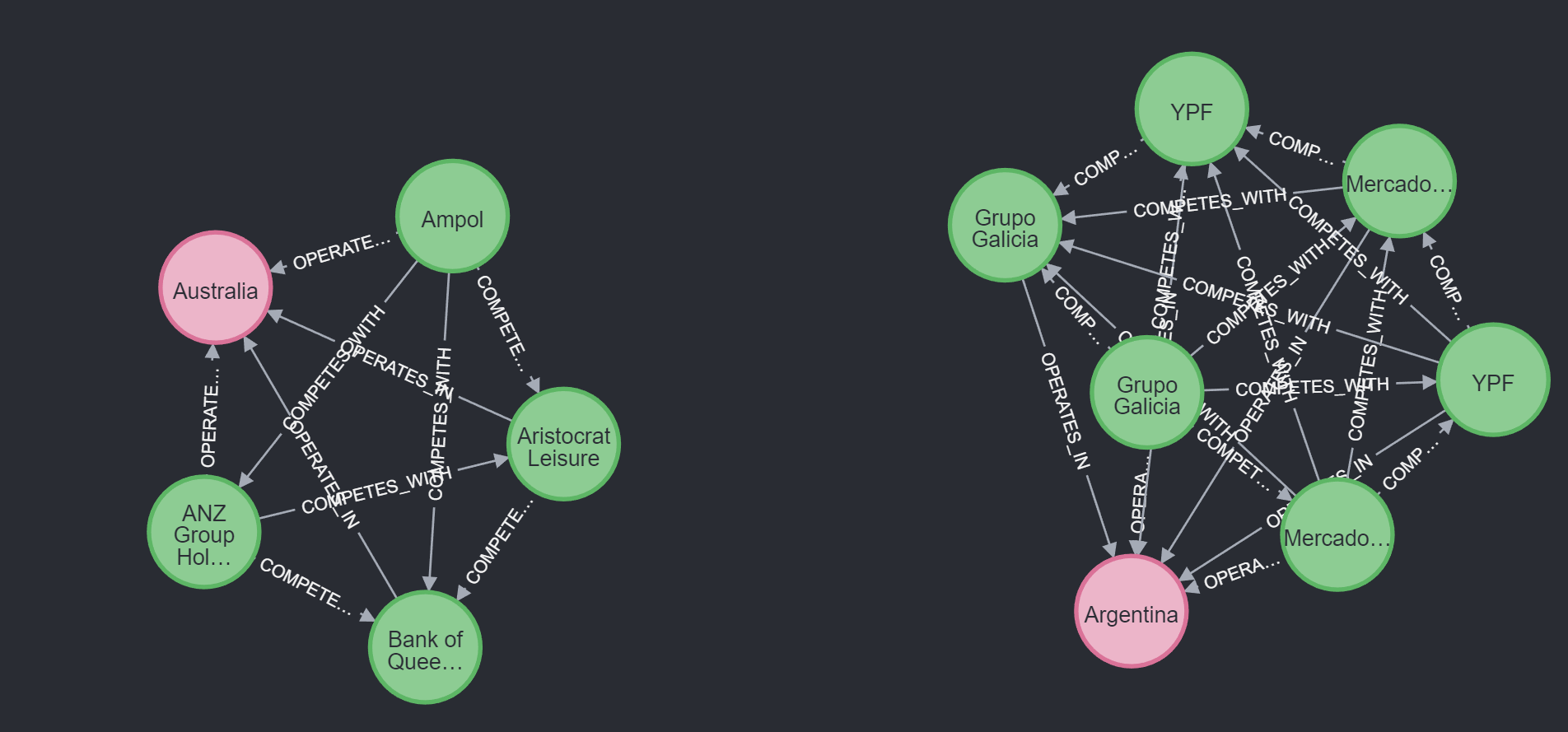
To populate the data warehouse, we first extracted the dataset from Kaggle and converted it into a structured CSV format. During the transformation phase, financial values were standardized into millions of dollars, and country names were normalized for consistency (thanks to the python script named “data\_processing.py”). The cleaned and structured data (in the file “processed\_data”) was then loaded into a MySQL-based WAMP server to facilitate relational storage and querying.

For the NoSQL implementation, we used Neo4j to model and analyze intercompany relationships. Companies were represented as nodes, while relationships such as “OPERATES\_IN” (linking a company to its country) and “COMPETES\_WITH” (indicating competition between companies) were established. The following Cypher query was used to integrate the data into Neo4j:

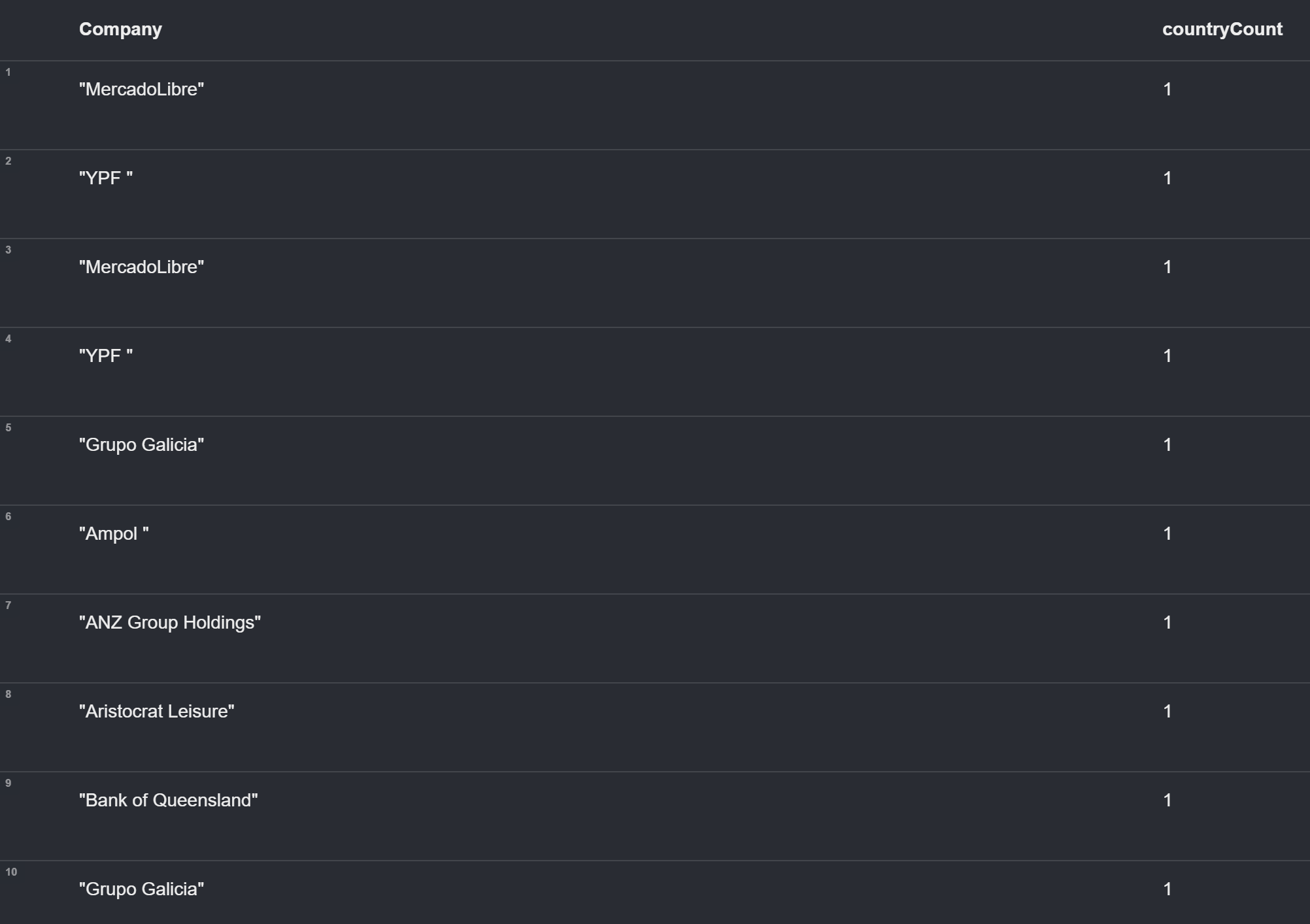
| LOAD CSV WITH HEADERS FROM 'file:///processed\_data.csv' AS **row** MERGE (company:Company {  **name**: row.Name,   sales: toFloat(row.Sales),  profit: toFloat(row.Profit),  assets: toFloat(row.Assets),  marketValue: toFloat(row.`Market Value`) }) MERGE (country:Country {**name**: row.Country})  MERGE (company)-[:OPERATES\_IN]->(**country**);  MATCH (c1:Company)-[:OPERATES\_IN]->(**p**:Country)<-[:OPERATES\_IN]-(c2:Company) WHERE id(c1) < id(c2)  MERGE (c1)-[:COMPETES\_WITH]->(c2); |
| --- |

We have made some queries that allow us to quickly visualize some interesting information.

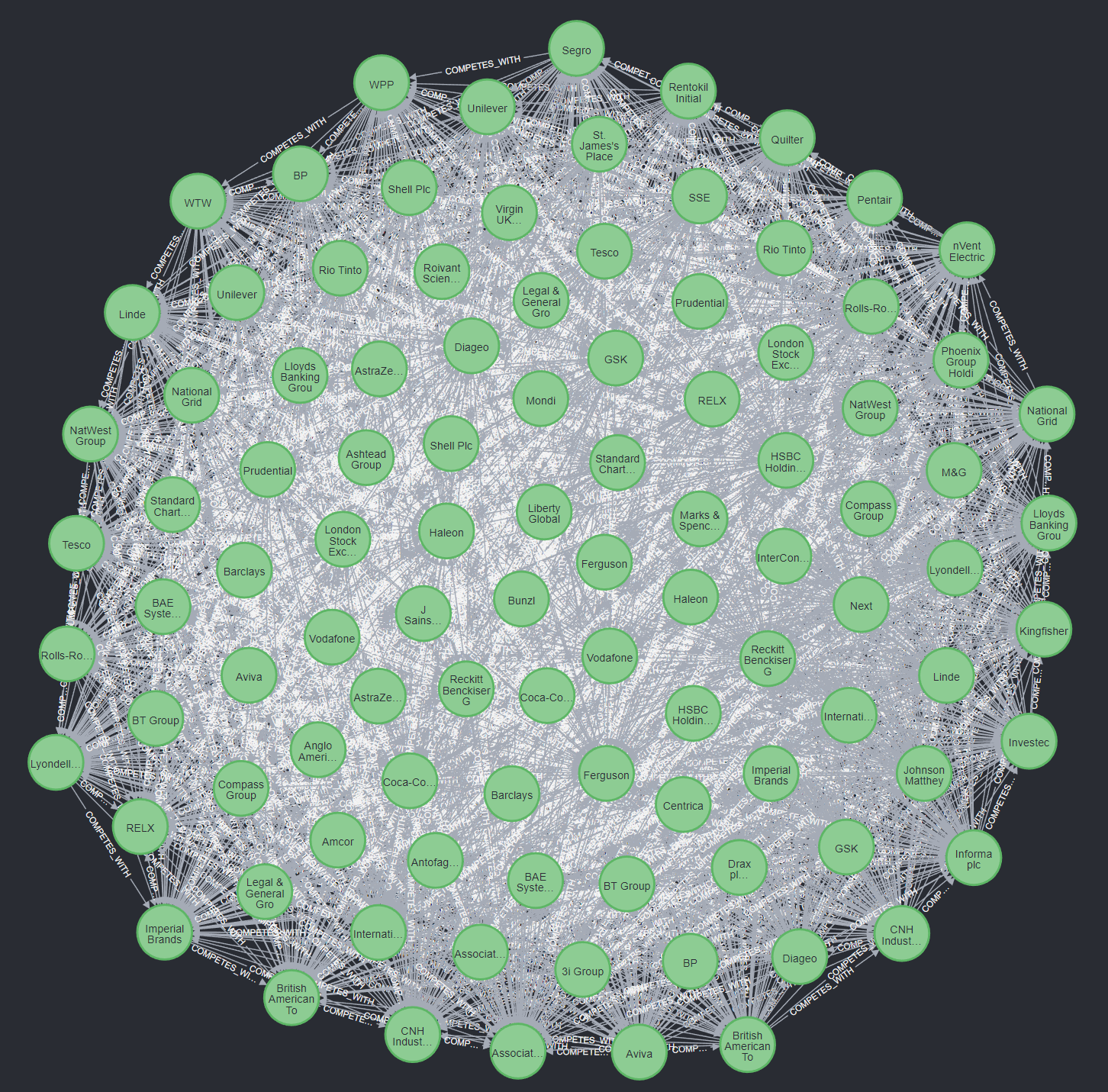
Visualization of 10 connected companies and countries:



### List of companies with the most international connections:



Visualization of competitive relationships in the form of a graph limited to 100 companies:

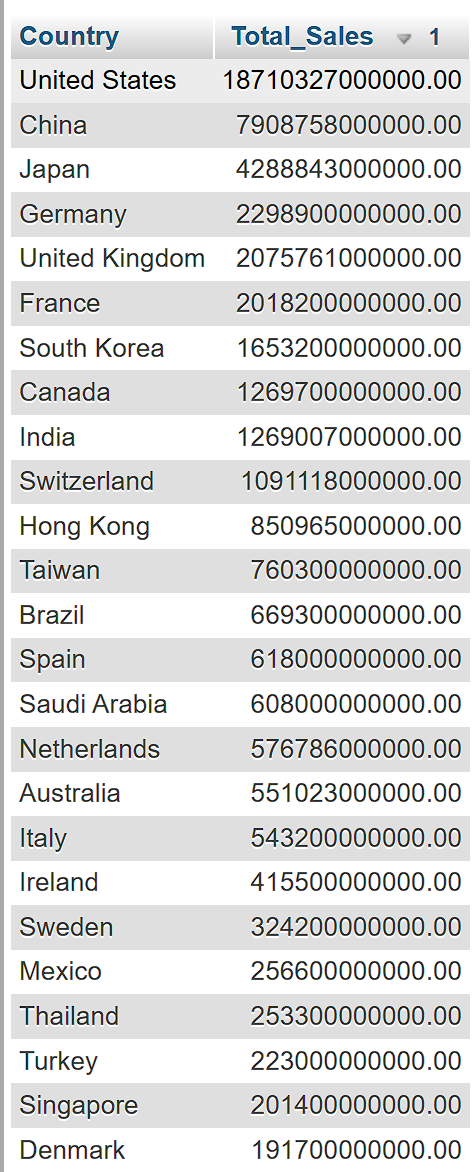


This approach enables the exploration of company interconnections, allowing us to gain insights into economic influences and competitive landscapes.

### IV - Querying the Data Warehouse and NoSQL Database:

With the structured data warehouse and NoSQL database in place, we performed several analytical queries to extract insights. In SQL, we aggregated sales and profit by country to identify key economic regions. The following query retrieves the total sales per country:

| SELECT Country, SUM(Sales) AS Total\_Sales **FROM** Companies GROUP BY Country ORDER BY Total\_Sales DESC; |
| --- |



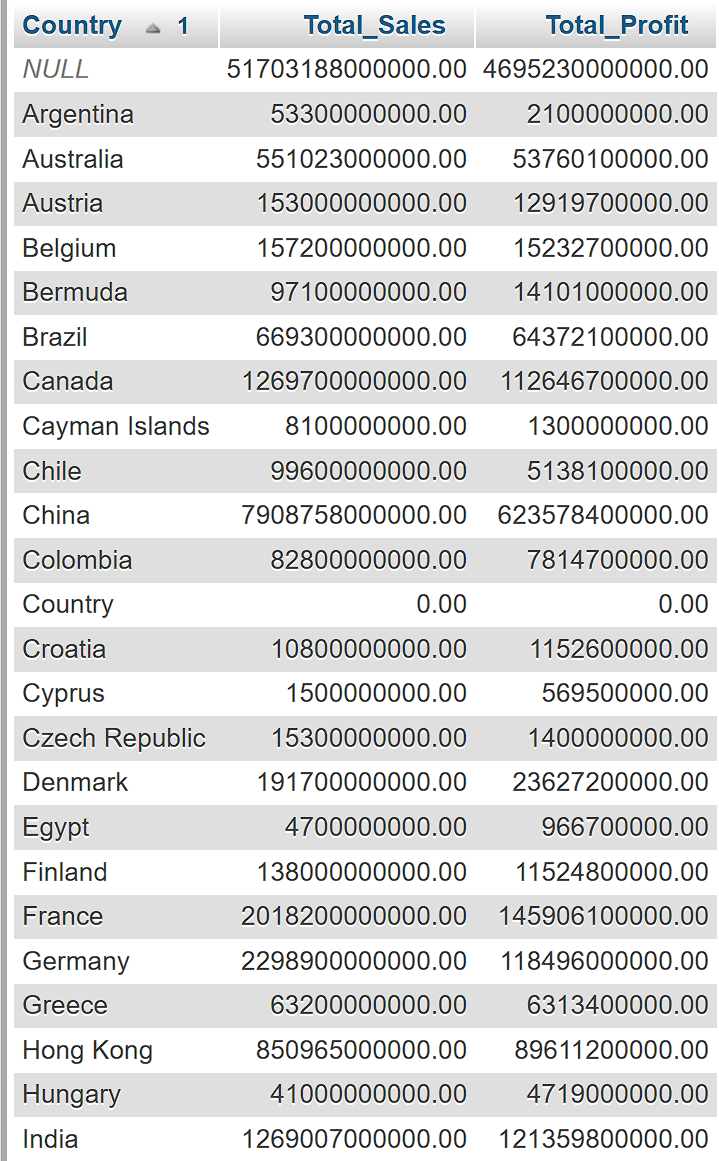
Similarly, we conducted a profit analysis per country:

| SELECT Country, SUM(Profit) AS Total\_Profit **FROM** Companies GROUP BY Country ORDER BY Total\_Profit DESC; |
| --- |



We did a ROLLUP to analyze total sales and profits at multiple levels:

| **SELECT Country, SUM(Sales) AS Total\_Sales, SUM(Profit) AS Total\_Profit FROM Companies GROUP BY Country WITH ROLLUP;** |
| --- |



Since WAMP servers doesn’t handle CUBE operations, we simulated these by doing the following request :

* equivalent of CUBE to get some comprehensive aggregation across dimensions:

| **SELECT Country, SUM(Sales) AS Total\_Sales, SUM(Profit) AS Total\_Profit FROM Companies GROUP BY Country UNION ALL SELECT NULL AS Country, SUM(Sales) AS Total\_Sales, SUM(Profit) AS Total\_Profit FROM Companies;** |
| --- |



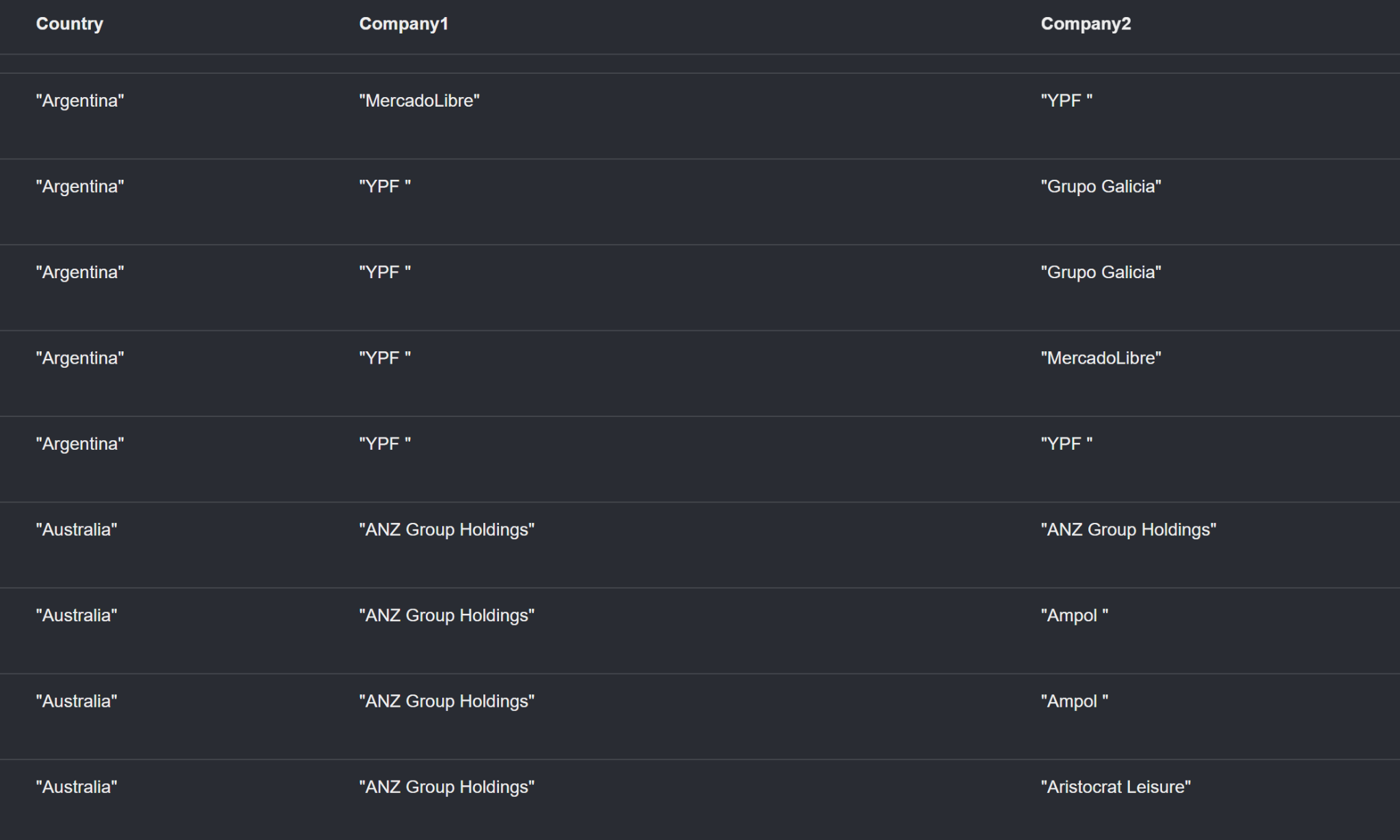
On the Neo4j side, we explored relationships between companies and their geographic presence. The following Cypher query lists companies operating within the same country:

| **MATCH** (**company**:Company)-[:OPERATES\_IN]**-**>(**country**:Country) **RETURN** **country**.name **AS** **Country**, **collect**(**company**.name) **AS** **Companies** **ORDER** **BY** **country**.name; |
| --- |



Additionally, we examined direct competition by identifying companies that operate in the same country:

| **MATCH** (**company1**:Company)-[:OPERATES\_IN]**-**>(**country**:Country)<**-**[:OPERATES\_IN]**-**(**company2**:Company) **WHERE** **id**(**company1**) < **id**(**company2**) **RETURN** **country**.name **AS** **Country**, **company1**.name **AS** **Company1**, **company2**.name **AS** **Company2** **ORDER** **BY** **Country**, **Company1**, **Company2**; |
| --- |



By integrating relational database insights with graph-based analysis, we tackled key challenges in understanding global corporate dynamics.

On the SQL side, we performed financial aggregation queries to identify top-performing companies based on total sales, profit, and market value at both the country level and globally. Using ROLLUP and simulated CUBE operations, we extracted hierarchical financial insights, allowing for a structured comparison of economic strength across regions.

On the NoSQL side (Neo4j), we analyzed corporate relationships to understand market influence and competition. Using OPERATES\_IN relationships, we mapped company presence across different countries. We further established COMPETES\_WITH relationships to visualize competition among businesses within the same market.

By cross-referencing SQL-based financial strength with Neo4j-based market influence, we provided a multi-dimensional analysis that not only highlighted the most financially successful companies but also those with the strongest competitive positioning. This hybrid approach enriched our understanding of how financial power correlates with market dominance, effectively addressing the need for a more comprehensive economic and competitive analysis.

### V - Big Data Integration and Processing:

As data volumes continue to grow, integrating Big Data technologies enhances the ability to process and analyze vast amounts of structured and unstructured data efficiently. Scalability and performance are key challenges in managing large datasets, and tools such as Apache Hadoop and Apache Spark provide solutions tailored to these needs.

Apache Hadoop offers a distributed storage system through HDFS (Hadoop Distributed File System), enabling the management of large datasets across multiple nodes. This architecture ensures high availability, fault tolerance, and horizontal scalability, making it ideal for handling massive amounts of structured and unstructured data. By leveraging Hadoop-based ecosystems, organizations can ingest, store, and retrieve large datasets efficiently, overcoming the storage limitations of traditional relational databases.

While Hadoop facilitates scalable storage, Apache Spark significantly improves data processing performance by utilizing in-memory computation. Compared to disk-based processing methods, Spark accelerates data transformations, filtering, and aggregations, making it particularly suited for large-scale analytical queries. Additionally, Spark SQL provides an optimized way to query massive datasets, and Spark GraphX enables efficient graph processing, which can extend Neo4j-based competitive network analysis.

To analyze large volumes of structured data efficiently, Apache Hive and Impala enable SQL-like querying over Big Data storage solutions. These tools bridge the gap between relational OLAP queries and distributed computing, allowing for efficient aggregation, trend analysis, and pattern discovery in large financial and economic datasets.

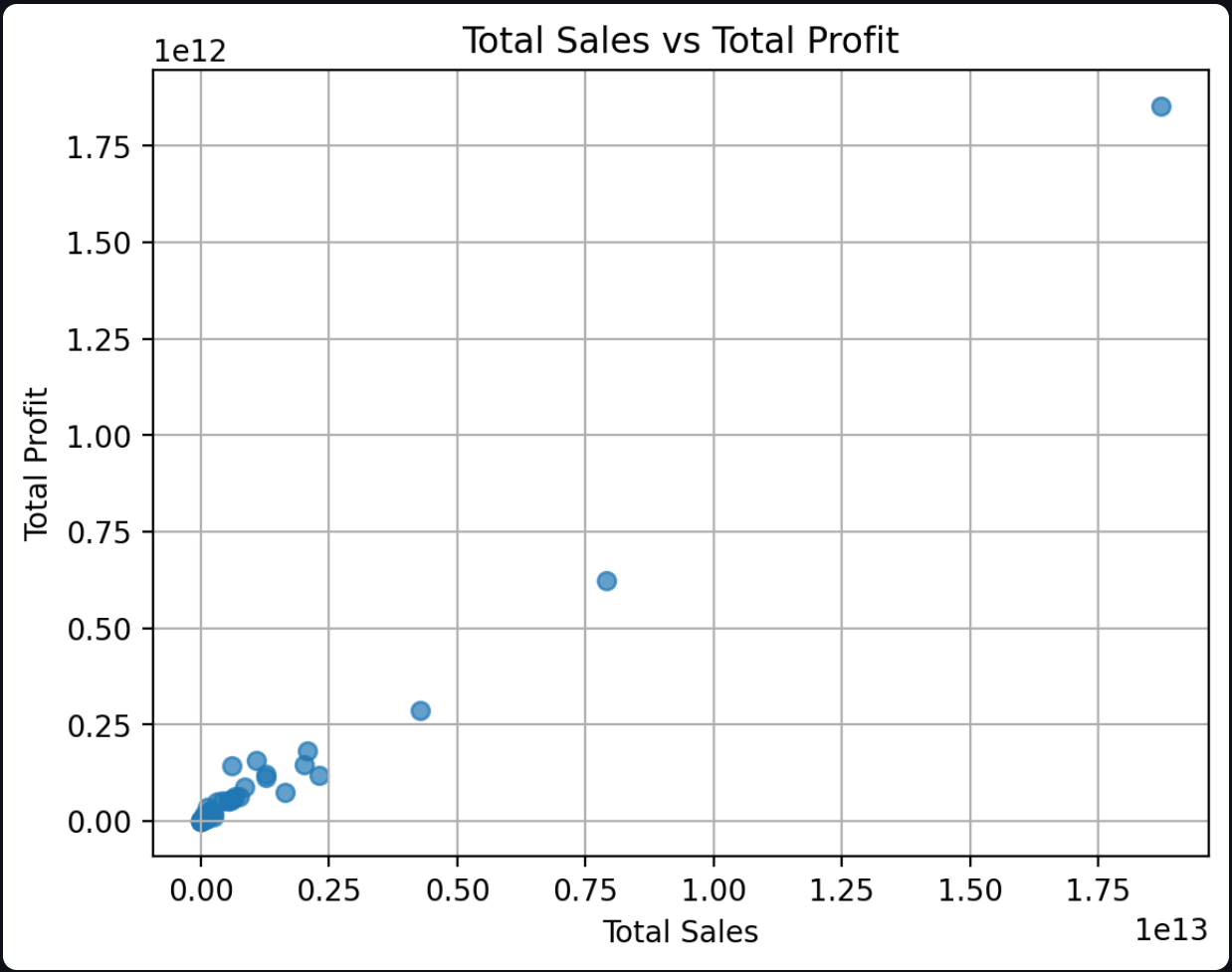
Optimizing query performance in Big Data environments often requires partitioning and indexing strategies. Partitioning data by region, industry, or time period reduces query execution time and enhances accessibility for large-scale financial trend analysis and competitive insights.

Big Data analytics is most effective when insights are visually represented. Tableau and Power BI allow dynamic visualization of SQL and NoSQL analytics, enabling clear presentations of financial trends, sales distributions, and market competition. Additionally, Neo4j’s graph visualization tools, such as Neo4j Bloom, facilitate the exploration of corporate interconnections, competition networks, and market influence rankings using algorithms like PageRank.

Integrating Big Data tools into analytical pipelines improves scalability, performance, and analytical depth. The combination of Hadoop for distributed storage, Spark for high-speed processing, Hive for structured querying, and visualization tools like Tableau and Power BI ensures that data analysis remains efficient and adaptable as datasets grow. This approach supports robust corporate financial trend analysis, deeper competitive insights, and a comprehensive understanding of global economic landscapes.

### VI - Data Analysis and Visualization :

Scatter Plot of Total Sales vs Total Profit:

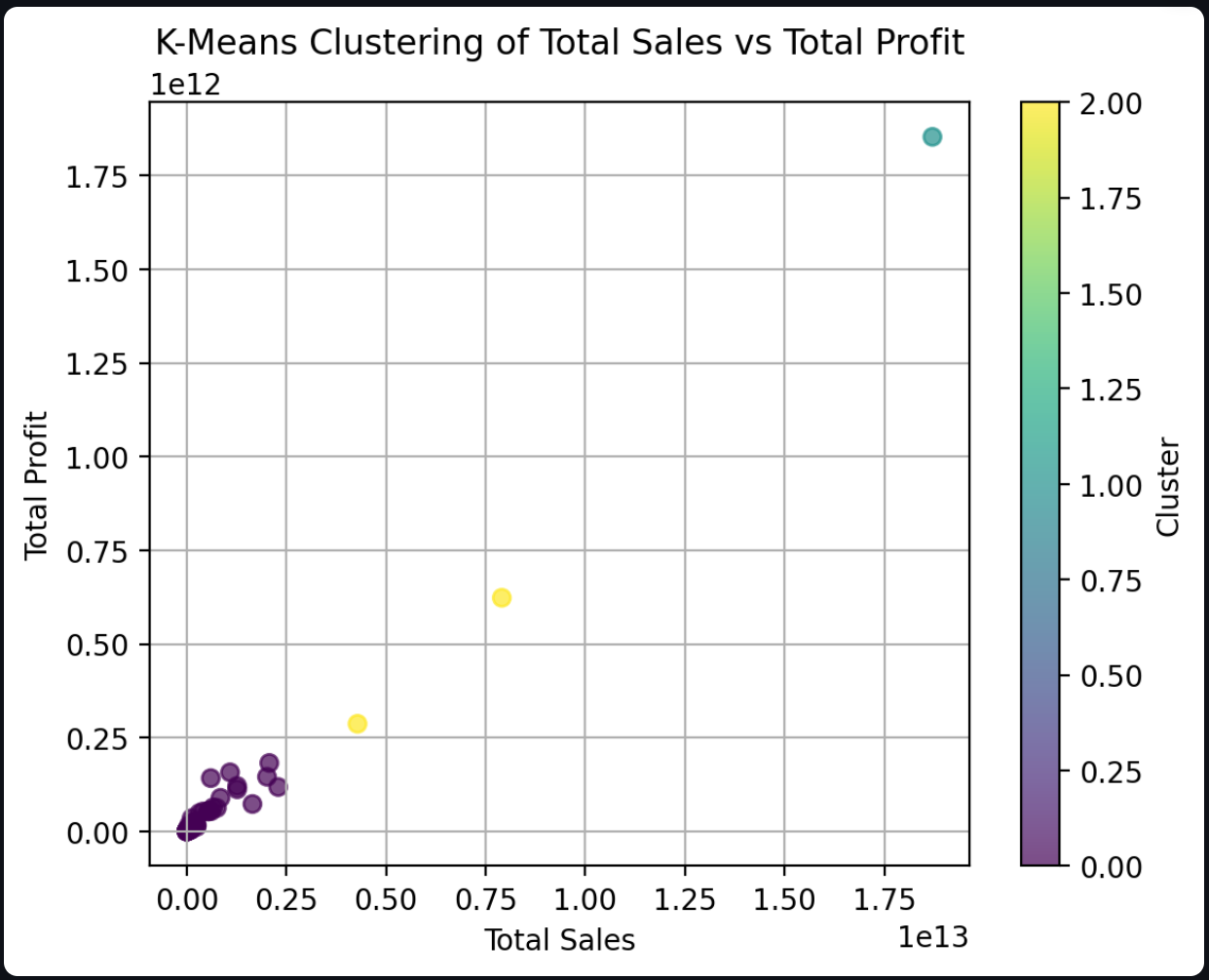


This figure is crucial for observing the raw distribution of data. Each point represents a country, with its position in the graph determined by its total sales and total profit. A general trend can be observed:

* countries with high sales tend to have high profits, indicating a positive correlation.

However, exceptions may exist where certain countries generate high sales but lower profits, which could suggest high costs or low profit margins.

K-Means Clustering Visualization:



After applying the K-Means clustering algorithm, we obtain a similar scatter plot but color-coded based on the clusters assigned to different countries. The goal of this classification is to identify groups of countries with similar economic characteristics. For example:

One cluster may include countries with extremely high sales and profits, representing major economic powerhouses.

Another cluster may represent countries with moderate sales and profits, indicating an intermediate economic performance.

A third cluster may group countries with low sales and low profits, characterizing smaller economies.

This segmentation is useful for understanding how economic performance varies across regions and for adapting business or investment strategies accordingly.

Clustered Data Table:



The final table provides a detailed view of the assigned clusters for each country. It helps to identify specific groupings and compare countries within the same cluster. This table is particularly useful for deeper analysis, such as comparing countries within the same group to identify similar economic development patterns.

General Interpretation:

The use of the K-Means method helps structure the data more effectively and highlights trends that may not be as clear with a simple observation of the raw scatter plot. With this approach, valuable insights can be extracted, such as:

* Which countries dominate the market in terms of sales and profit?
* Which groups of countries exhibit comparable performance?
* How can countries be segmented into categories for more targeted analysis (e.g., investment opportunities, business strategies)?

In summary, these figures provide a comprehensive and analytical view of the economic performance of different countries, helping to better understand the underlying commercial dynamics.