Project 2: SEC Corporate Filings Insights

Introduction

The Financial Statement Data Set is a comprehensive resource created by the U.S. Securities and Exchange Commission (SEC) to facilitate the analysis and consumption of financial data submitted by public companies. This dataset is derived from the EX-101 attachments to various filings, which utilize the eXtensible Business Reporting Language (XBRL) to present structured financial information.

Key Features of the Dataset

The dataset includes quarterly and annual numeric data from primary financial statements, which encompass:

- Balance Sheet
- Income Statement
- Cash Flow Statement
- Changes in Equity
- Comprehensive Income

Structure

The dataset consists of four key components:

1. **Submission Data Set (SUB):** Contains records for each XBRL submission, providing essential information about the filing entity and the submission itself.
2. **Number Data Set (NUM):** Includes individual numeric values reported in the primary financial statements, detailing every line item from each submission.
3. **Tag Data Set (TAG):** Provides metadata about the tags used in the submissions, including documentation labels, taxonomy versions, and tag attributes.
4. **Presentation Data Set (PRE):** Offers insights into how tags and corresponding numbers were presented in the financial statements, including their sequence and structure.

Update Frequency: The Financial Statement Data Set is updated quarterly, incorporating data from filings submitted after the close of business on the last day of the quarter. This ensures users have access to the most recent financial information available.

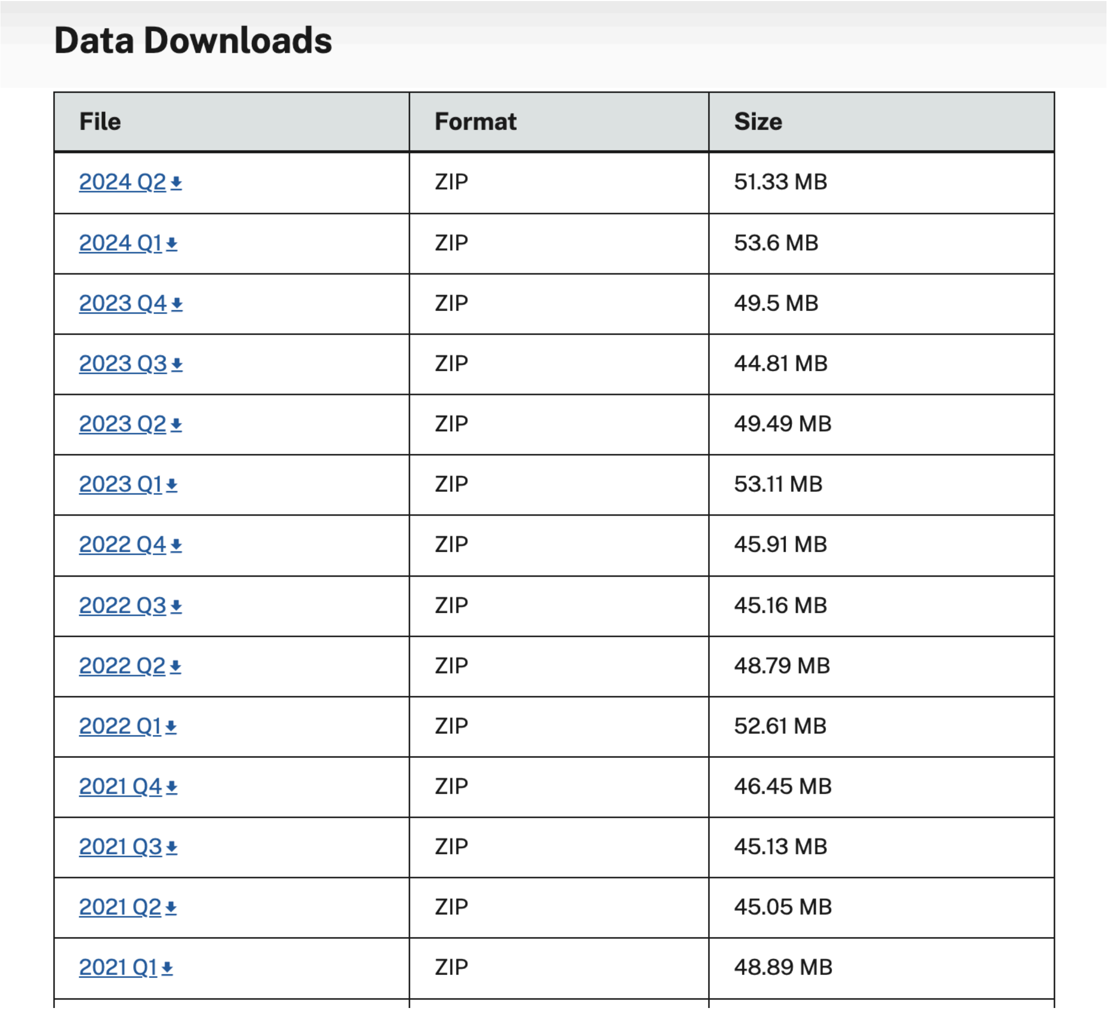
Each zip file has a README with the data format too.

Data Coverage

Start Year: 2009
End Year: Present (the dataset continues to be updated quarterly)
Frequency of Updates: Quarterly

Data Access

Details of the dataset are present on the SEC website.
The dataset is publicly available. The data for each quarter starting 2009 can be downloaded as shown below:



Data Downloads

File	Format	Size
2024 Q2	ZIP	51.33 MB
2024 Q1	ZIP	53.6 MB
2023 Q4	ZIP	49.5 MB
2023 Q3	ZIP	44.81 MB
2023 Q2	ZIP	49.49 MB
2023 Q1	ZIP	53.11 MB
2022 Q4	ZIP	45.91 MB
2022 Q3	ZIP	45.16 MB
2022 Q2	ZIP	48.79 MB
2022 Q1	ZIP	52.61 MB
2021 Q4	ZIP	46.45 MB
2021 Q3	ZIP	45.13 MB
2021 Q2	ZIP	45.05 MB
2021 Q1	ZIP	48.89 MB

This project will be undertaken in groups of two, with a different composition than Project 1.

We will focus on data from 2013 to 2023. Each group will be assigned one year of non-overlapping data; for example, Group 1 will work with data from 2013, Group 2 with 2014, and so on. The dataset allocation will be assigned by the course TA.

Tooling

You are expected to use a combination of Neo4j, pyspark or pandas, graph analytics, and GraphRAG in Google Colabs for this project. My recommendation is to upload your dataset to your GWU Google Drive and then load it into Neo4j.
Of course, you will have to sign up for a free Neo4j Aura account. Note that the free tier has a 200,000 Nodes and 400,000 Relationships limit. Therefore, when loading data into Neo4j limit the number of each node and edge type ingested into the graph based on the analysis.
You can use any Python library, such as `matplotlib`, `yfiles_jupyter_graphs_for_neo4j`, etc. of your choosing to create the plots. In addition, all documentation/comments should be in the form of Markdown within the notebook itself.

Analysis

For each piece of analysis, provide an explanation of the process as well as an interpretation of the results.

1. Analyze financial statements for companies. Financial Statement Comparison allows analysts and investors to compare financial statements of different companies, aiding in benchmarking and performance evaluation.
2. Cluster companies based on financial health. Financial health is a function of revenue and debt, i.e. a health company will have high revenue and low debt.
3. Identify unusual reporting patterns or significant deviations from historical data, which may indicate potential fraud or misrepresentation.
4. Analyze how corporate executives and board members are connected across different companies and perform centrality analysis. Note that this dataset directly does not contain the names of office holders for a company. This data will need to be extracted from other sources such as Form 8-K or Form 10-K, or any third party data source.
5. Financial Query and Report Generation via GraphRAG. The user should be able to query specific financial metrics, trends, or insights from the dataset using plain English.

Due Date

15 October, 2024 - 11:59AM

You can submit the link to your Google Colab via Blackboard in the Assignments section.