Jay Kwon | Metis Data Science and Engineering (flex program) | Module 3 - Data Eng (Feb, 2022)

**Historical Earnings per Share (EPS) EDGAR Data Pipeline**

**Abstract**

The main goal of this project was to create a data pipeline that takes raw historical financial data from the Security and Exchange Commission’s (SEC) EDGAR database and make the diluted Earnings per Share (EPS) data easily accessible on a PostgreSQL database hosted by AWS. Of the many company data available, the data companies that comprise the Standard and Poor’s 500 (SP500) index was chosen to be entered into the database, resulting in table with 100,368 rows. A Streamlit frontend was created to query this database for a company of interest, filter the data to quarterly periods, and showcase a timeseries graph visualizing the historical EPS for that company.

**Design**

Although historical price and trade volume of stocks are readily available for free, financial statement data for companies are not available for free. This data pipeline makes such data readily available in an easily accessible manner. Anyone interested in financial statement data of public US companies will benefit from this data pipeline and the insights gained from this project. This includes stock investors, financial media, financial services, and rating agencies. Although the project chose to process EPS data, the pipeline may be adjusted to obtain the many other financial metrics on the income statement, balance sheet, and cash flow statements for companies of interest. Importantly, this serial, historical data may serve as important features for time-series machine learning models to predict targets such as weekly stock returns.

**Data**

Historical financial statement data were downloaded in JSON format from the SEC’s EDGAR database. The raw data included metrics from the balance sheet, income statement, and cash flow statements from SEC filings (quarterly 10-Q, and yearly 10-K) for 15,556 companies. EPS data for the SP500 companies were inserted into a PostgresSQL database hosted on AWS. In total the database table included 100,368 rows, each representing a SEC financial statement filing for a particular SP500 company. In addition to the EPS itself, there were 9 other columns, including 4 categorical data that described where the particular EPS was from (filing date, start/end date, filing type, company). For most companies the data spanned the period from 2009 to 2022.

**Algorithm**

\* this project did not involve and machine learning

**Tools**

Psycopg2 and SQL to insert and query data

AWS to host the PostgreSQL server from the cloud

Pandas to filter and manipulate data

Matplotlib for time series visualization

Streamlit to create a frontend

**Communication**

- 5 minute presentation with Slides

- GitHub repo: <https://github.com/jaykwon2/EDGAR-dilutedEPS>

- Streamlit frontend: <https://share.streamlit.io/jaykwon2/edgar-dilutedeps/main/EDGAR_streamlit_app.py>