HISTORICAL EARNINGS PER SHARE (EPS) EDGAR DATA PIPELINE

Jay Kwon

Metis Data Science and Engineering (flex program)

Module 3 – Data Engineering

Feb 23, 2022

<u>Goal</u>:

- **Create a Data Pipeline for historical financial statement data (EPS) for companies in the S&P500 index.**

  1. obtain raw data from SEC's EDGAR database, curate EPS data, make available on PostgreSQL server hosted on AWS
  2. create interactive frontend for end users access to data/visualization

<u>Background</u>:

- historical price and trade volume of stocks are easily available for free
- historical **financial statement data** for companies are difficult to track, curated data **not available for free**
  - without this data ML regression models to *<u>predict returns</u>* proved to be ineffective:

    R^2 of -0.055 (worse than simple average)

Motivation:

- Serial, historical data may serve as important features for time-series ML models (e.g. predicting stock returns)
- insights gained would be beneficial to anyone interested in company fundamentals including:
  - **investors** (individuals, wealth management firms, hedge funds, pension funds)
  - **financial news** (Bloomberg, CNBC, Wall Street Journal)
  - **ratings agencies** (S&P, Moody's, Fitch)
  - **financial services**

DATASET

- Raw JSON files obtained from SEC's EDGAR database: https://www.sec.gov/edgar/sec-api-documentation
- 498 JSON files each representing a company in the SP500 index
- PostgreSQL database had 1 table for EPS with 100,368 rows and 10 columns (4 categorical, 6 quantitative)
- time period: 2009 to 2022

- **Why EPS data?**

  - represents a company's profit per share

  - great candidate as a feature to predict stock returns

Tools used:

- Psycopg2, PostgreSQL, AWS, Pandas, Matplotlib, Streamlit

HISTORIC TREND OF STOCK MARKET

* Standard and Poor's 500 Index (source: Google)

period examined
(Feb 2010 – Jan 2020)

<u>Period of our Data</u>: Feb 2010 to Feb 2020

- want to avoid extreme events that are hard to predict
- after 2008 financial crisis, before 2020 Covid-19 pandemic
- relatively stable bull market

Historical Financial Statement Data Pipeline

DATA INGESTION

Download JSON files for companies from EDGAR website

TESTING/ROBUSTNESS

check that new data was written by accessing new rows, check that values make sense (reasonable range, negative values)

DATA STORAGE

Parse JSON files and store diluted Earnings Per Share (EPS) data for SP500 companies on a PostgresSQL server hosted on AWS using Psycopg2 and SQL commands

PROCESSING

Clean/filter data into a usable format using Pandas;

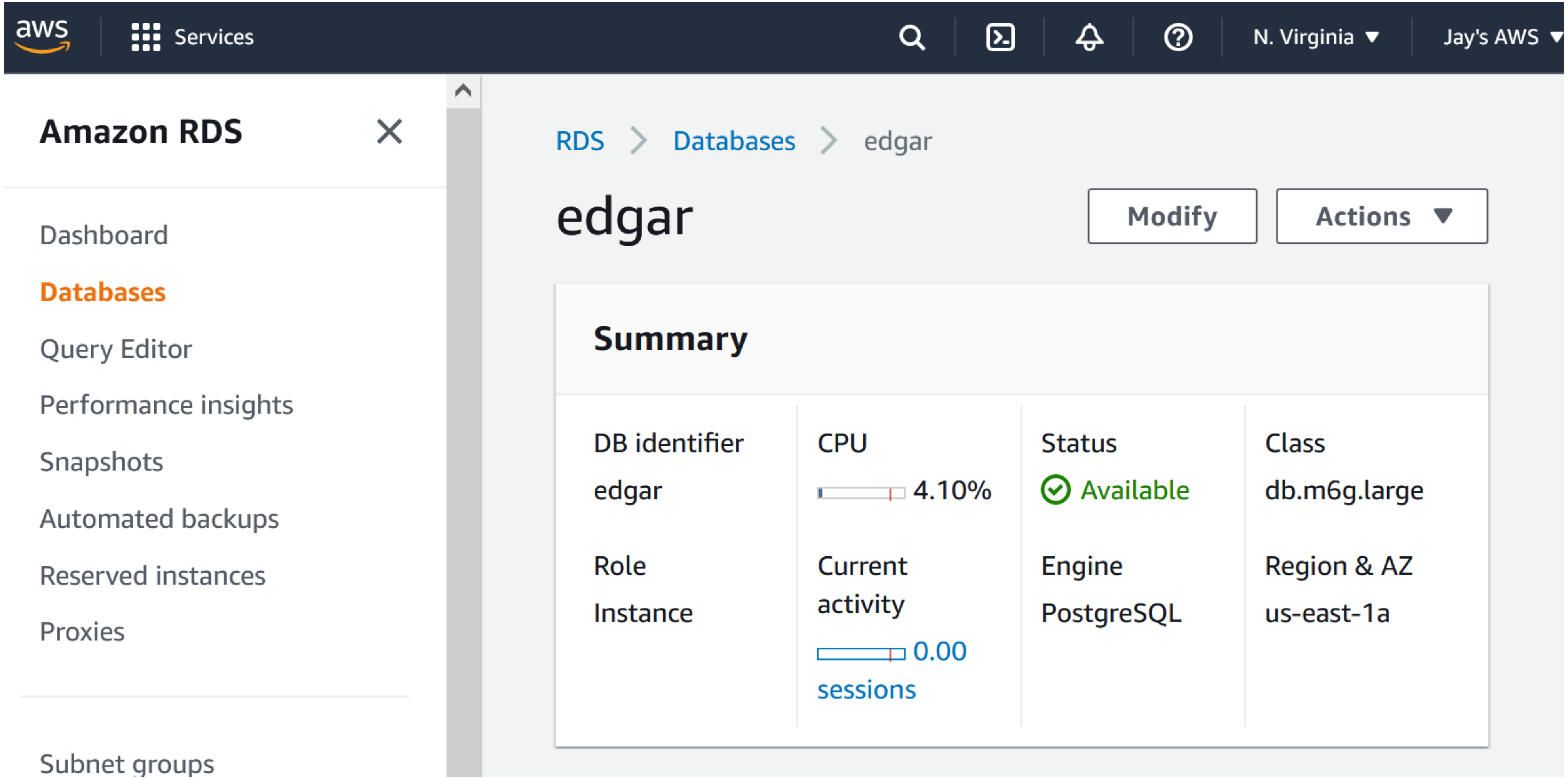
Engineer features of interest

DEPLOYMENT

Frontend hosted on Streamlit with table and visualization (Matplotlib) of historical EPS data

Allow access to AWS database; Jupyter notebook and python scripts access on GitHub

AWS



AWS → RDS → Databases → Instance → databases → tables

PostgreSQL

Schema:

Table Name	Object ID	Owner	Tablespace	Row Count Estimate
earningspersharediluted	16,468	postgres	pg_default	100,368

EPS Table:

earningspersharediluted Enter a SQL expression to filter results (use Ctrl+Space)

Grid Text

	id	companyname	startdate	enddate	val	accn	fy	fp	form	filed
1	1	3M COMPANY	2007-01-01	2007-12-31	5.6	0001104659-10-029054	2,009	FY	8-K	2010-05-17
2	2	3M COMPANY	2007-01-01	2007-12-31	5.6	0001104659-10-007295	2,009	FY	10-K	2010-02-16
3	3	3M COMPANY	2008-01-01	2008-06-30	2.7	0001104659-09-046329	2,009	Q2	10-Q	2009-07-31
4	4	3M COMPANY	2008-04-01	2008-06-30	1.33	0001104659-09-046329	2,009	Q2	10-Q	2009-07-31
5	5	3M COMPANY	2008-01-01	2008-09-30	4.11	0001104659-09-061503	2,009	Q3	10-Q	2009-10-30
6	6	3M COMPANY	2008-07-01	2008-09-30	1.41	0001104659-09-061503	2,009	Q3	10-Q	2009-10-30
7	7	3M COMPANY	2008-01-01	2008-12-31	4.89	0001104659-11-031722	2,010	FY	8-K	2011-05-26
8	8	3M COMPANY	2008-01-01	2008-12-31	4.89	0001104659-11-007845	2,010	FY	10-K	2011-02-16
9	9	3M COMPANY	2008-01-01	2008-12-31	4.89	0001104659-10-029054	2,009	FY	8-K	2010-05-17
10	10	3M COMPANY	2008-01-01	2008-12-31	4.89	0001104659-10-007295	2,009	FY	10-K	2010-02-16

Frontend: Streamlit

Quarterly diluted Earnings per Share

Select an SP500 Company from the list below:

3M COMPANY

You selected: 3M COMPANY

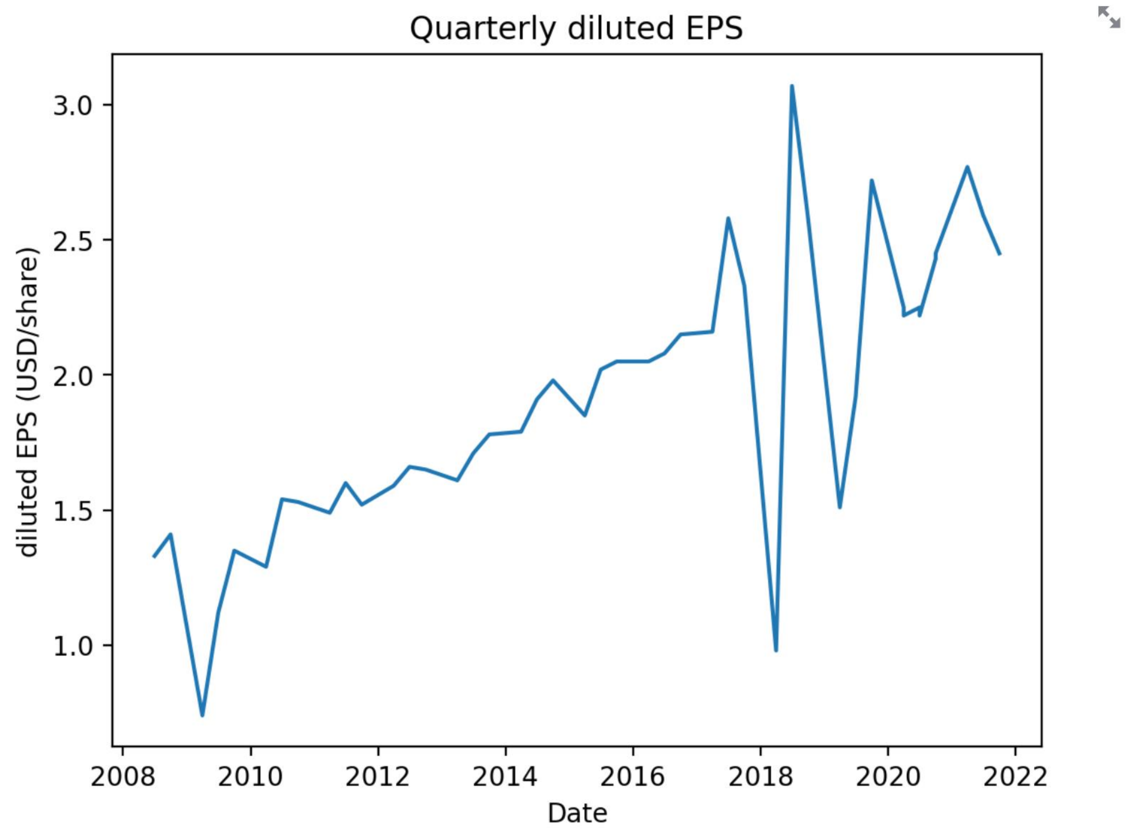
3M COMPANY - quarterly diluted Earnings per Share (EPS)

'val' column is the diluted EPS (in USD/share)

	enddate	val	accn	fy	fp	form	filed	period
1	2008-06-30	1.3300	0001104659-09-046329	2009	Q2	10-Q	2009-07-31	90
3	2008-09-30	1.4100	0001104659-09-061503	2009	Q3	10-Q	2009-10-30	91
4	2009-03-31	0.7400	0001104659-10-025776	2010	Q1	10-Q	2010-05-05	89
7	2009-06-30	1.1200	0001104659-10-041912	2010	Q2	10-Q	2010-08-04	90

Time Series graph of quarterly diluted EPS

- [x] Show Graph



- User chooses one of the companies
- shows EPS data and time-series graph
- https://share.streamlit.io/jaykwon2/edgar-dilutedeps/main/EDGAR_streamlit_app.py

Conclusions and Future Projects

<u>Conclusions</u>:

- It is possible to create one's own data pipeline for historic financial statement data
- handling financial data on local machine even only for 500 companies is inefficient
- using AWS proves useful for handling big data

<u>Future Projects</u>:

- Expand to other financial metrics: debt-to-equity, quick ratio, working capital ratio
- Apply ML models on data to predict targets such as stock returns
- Cron job on EC2 instance to regularly update database using API