ATLYS DATA ENGINEERING TASK

Stock Market Data Pipeline

Shikhar Sheoran

Files:

1. ***get_historical_data.py*** - This is the script for one time historical dump
2. ***get_daily_data.py*** - This is the script that will run daily and fetch data for D-1.
3. ***queries*** - This folder contains queries for the tasks mentioned.

Running the code:

1. Download this repository.
2. Activate the virtual env: **source my-venv/bin/activate**
3. Install dependencies: **python3 -m pip install -r requirements.txt**
4. Run the scripts: **python3 <script_name>**

Querying the data:

1. Make sure you have mysql shell installed : link
2. You can connect to the database by running the command: **mysqlsh -h stock-database.ctuyy6isws2v.ap-south-1.rds.amazonaws.com -u admin -p**
3. This will then ask for the password, which is: **atlys-data**
4. Then run the command: **use central_stock_db;**
5. Then you can run any query you want on either of the two tables: **historical_stock_data** or **daily_stock_data**
6. To run the query files in the queries folder, run the command: **source <your_path>/stock-market-de/queries/<query_file_name>.sql** in the mysql shell.

Code Overview:

1. We fetch data from the **AlphaVantage** API.
2. We fetch data for each company and then write it to our db.

3. I've used an **AWS RDS** instance for our database, so that you can also connect to it and query the data.
4. For the historical dump, the dates can be specified in the script itself.
5. For the daily script, it just calculates the date for yesterday and filters data based on that.
6. You might find the script is running a bit slow, that is due to the 12 second timeout added for the API rate limits.

Notes:

1. Indexing of the tables has been done on columns **Company** and **Date** to speed up the queries. You can check this by running the command: **show indexes from <table_name>;**
2. For median daily variation, **variation in closing price** has been used as the metric. This query is run on the **historical_stock_data** table, and you can specify the date range in the query for which the median needs to be found. For now, I have chosen it as the current date and 70 days before the current date.
3. For the **daily_price_variation** and **daily_volume_variation**, these queries are run on the **daily_stock_data** table, which for now only contains data for 2 days - 4th and 5th June. This query gives the variation in the metrics for all the days data is present in the table.
4. The data from the scripts is also stored in csv files in the data folder for quick validation purposes.
5. If you run the historical data script, you will get a key error as data is already stored in the db. You can then run the command: **drop table historical_stock_data;** , and then run the script and it will run fine.