

Extraction, Transformation, and Load Technical Report

Headlines and Stock Price Trends

TABLE OF CONTENTS

1.	Introduction	3
1.1	Summary	3
1.2	Scope	3
1.3	Technologies and resource contributions	3
1.4	Definitions, Acronyms and Abbreviations	4
2.	ETL Details	4
2.1	Data Import/Extract Sources and Method	4
2.2	Data Acquisition	4
2.3	Data Transform	4
2.4	Data Integrity	4
2.5	Data Refresh Frequency	4
2.6	Data Security	4
2.7	Data Loading and Availability	5
3.	Data Quality	6

1. INTRODUCTION

The purpose of the Extraction, Transformation, and Load (ETL) Technical Report is to capture details that pertain specifically to ETL portion of the data pipeline that is to be used in a data science project. This however does keep in mind the final target objective while performing the ETL.

1.1 Summary

The objective of this project is to track the relationship between news headlines and fluctuations in stock prices. Looking at these two data sets allows us to further analyze the correlation between the success/decline of stock and how the company is reflected in the media.

1.2 Scope

The integrated data sources are comprised of the Alpha Advantage API and the News API. Alpha Advantage provides real-time and historical stock data. News API provides a means to search worldwide news articles and headlines from all over the web in real-time. Determining positive or negative words that influence stock price is outside the scope of this project.

1.3 Technologies and resource contributions

Janita Brock - Assisted in API interface for obtaining data, cleaning data, and providing information for the technical report.

Ola Browne - Created visualizations for report.

Parul Garg - Created a host connection among all team members for data collaboration and DB sharing. Tried to establish a connection off campus but could not troubleshoot beyond GT network security issues. Inserting data in Postgres SQL and running queries.

Reza Gharahgozly - Created an ER diagram for SQL databases and used Postgres SQL to make tables in the database.

Sandy Lake - Worked with Robert on obtaining headlines from News API and converted headlines to CSV

Robert Patterson - Worked with Sandy on obtaining headlines from News API. Munged data in pandas to be consistent with how it would be loaded into the Postgresql database. Helped to construct Second Normal Form data structure with Parul and Reza.

Selvi Ramalingam - Retrieved stock API for 5 companies of choice, and set a template for further retrieval if there is more interest.

Challenges encountered:

While setting up the Postgres Server, we tried to set up one machine as Postgres Server which can be accessed by other systems in the same network. Considering security settings of the GTVisitor network - this kind of setup was taking more effort and time than the time allotted. With the similar setting change, this kind of setup was achieved in a home intranet setup.

Setting change:

> For the system setup as a server, login to root as 'su' and make changes to the 'pg_hba.conf' file

```
sh-3.2# pwd
/Library/PostgreSQL/11/data
sh-3.2# ls
PG_VERSION      pg_commit_ts    pg_multixact    pg_stat          pg_wal           postmaster.pid
base             pg_dynshmem     pg_notify       pg_stat_tmp      pg_xact
current_logfiles pg_hba.conf     pg_replslot     pg_subtrans      postgresql.auto.conf
global          pg_ident.conf   pg_serial       pg_tblspc        postgresql.conf
log             pg_logical      pg_snapshots    pg_twophase      postmaster.opts
sh-3.2#
```

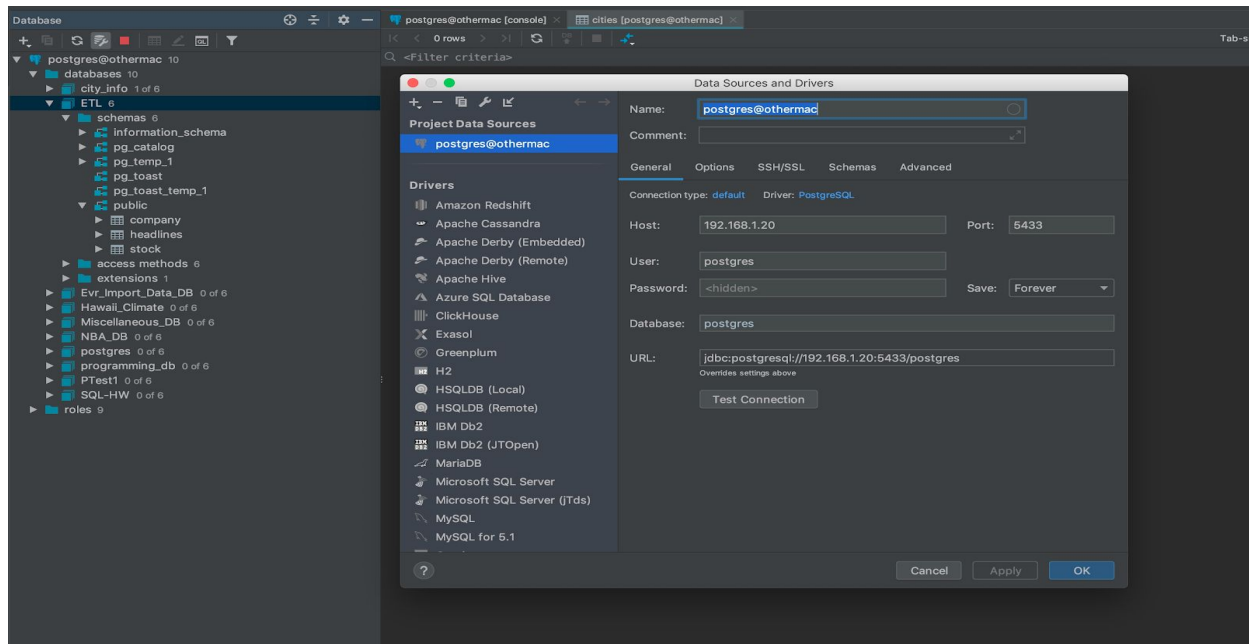
> Add a new line for the IP setting for the network:

```
# "local" is for Unix domain socket connections only
local    all             all                                     md5
host     all             all      192.168.1.0/24             password
host     all             all      107.77.236.230/32         password
host     all             all      10.136.16.8/32            password
host     all             all      10.136.0.0/16             password
```

The highlighted setting above is for the home network where 'client' DB was able to connect to Server system. (Screenshot of client shown below)

1.4 Definitions, Acronyms, and Abbreviations

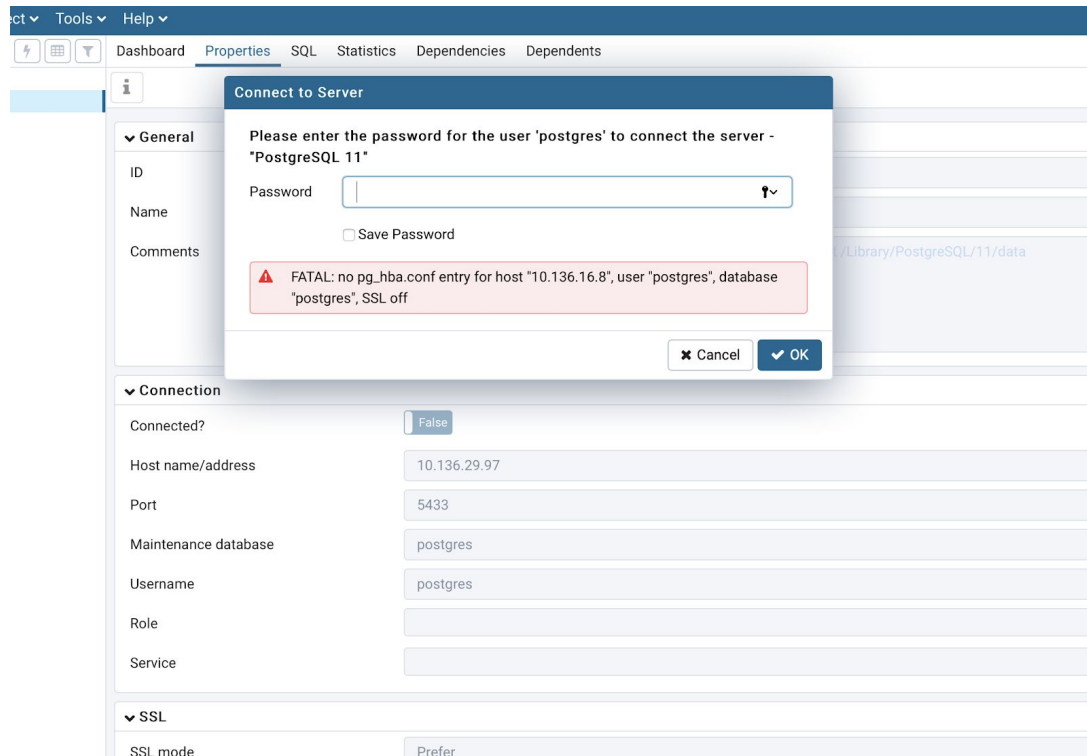
API - Application Program Interface



> Similar settings on GT network did not allow the client to access server DB. (SSL error)

"local" is for Unix domain socket connections only

local	all	all		md5
host	all	all	192.168.1.0/24	password
host	all	all	107.77.236.230/32	password
host	all	all	10.136.16.8/32	password
host	all	all	10.136.0.0/16	password



This was a great learning and challenging experience.

2. ETL DETAILS

2.1 Data Import/Extract Sources and Method

	Source - APIs	Urls
1	Alpha Vantage	https://www.alphavantage.co/documentation
2	News API	https://newsapi.org/

Method included in 2.2 Data Acquisition

2.2 Data Acquisition

We used two APIs to help gather data for this project.

The first used was [Alpha Vantage](https://www.alphavantage.co/documentation) (<https://www.alphavantage.co/documentation>) which is a free API that provides realtime and historical data on stocks. We are able to choose the parameters of the data as described in their documentation which led us to use the 'Time Series Daily Adjusted' API so we were able to receive daily information on the opening and closing prices, the low and high for the day, along with volume. This allowed us to truly send the trends of the stocks during the date range we selected. There were additional options to see the API data in a full or compact output size. The full data would return all current data along with 20+ years of historical data, but the compact would provide the last 100 data points on all stocks collected. As we knew we would be working with two API datasets we chose to utilize the compact output.

Additionally, newsapi.org was used to collect headlines related to the companies we previously gathered stock data for. This news API is also free and the API key allows for 60 requests per minute. Creating the request required a call to newsapi.org with appended search parameters (i.e. to_date, from_date, keyword, etc.) A difficulty we encountered was that the API provided a max return of 20 hits/page, and in order to obtain all hits, we created a for loop which iterates through all of the return pages. This data was munged in a pandas data frame to have clean and concise columns. This data frame was exported straight to a CSV file to be uploaded into the SQL database.

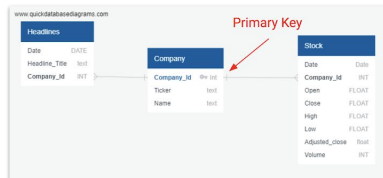
Creating The DB

- Created 3 Tables in SQL
 - Declared Data Types
 - Applied 2NF

Visualizing The DB

- Developed Entity Relationship Diagram

```
1 CREATE TABLE Company(  
2 Company_ID INT PRIMARY KEY,  
3 Ticker text,  
4 Name text,  
5 CEO text,  
6 Location text);  
7  
8 CREATE TABLE Headlines(  
9 Headline_Title text,  
10 Company_ID INT,  
11 Date DATE,  
12 Content text,  
13 FOREIGN KEY(Company_ID) REFERENCES Company(Company_ID));  
14  
15 CREATE TABLE Stock(  
16 Company_ID INT,  
17 Timestamp text,  
18 Open FLOAT,  
19 High FLOAT,  
20 Low FLOAT,  
21 Close FLOAT,  
22 Volume INT,  
23 FOREIGN KEY(Company_ID) REFERENCES Company(Company_ID));
```



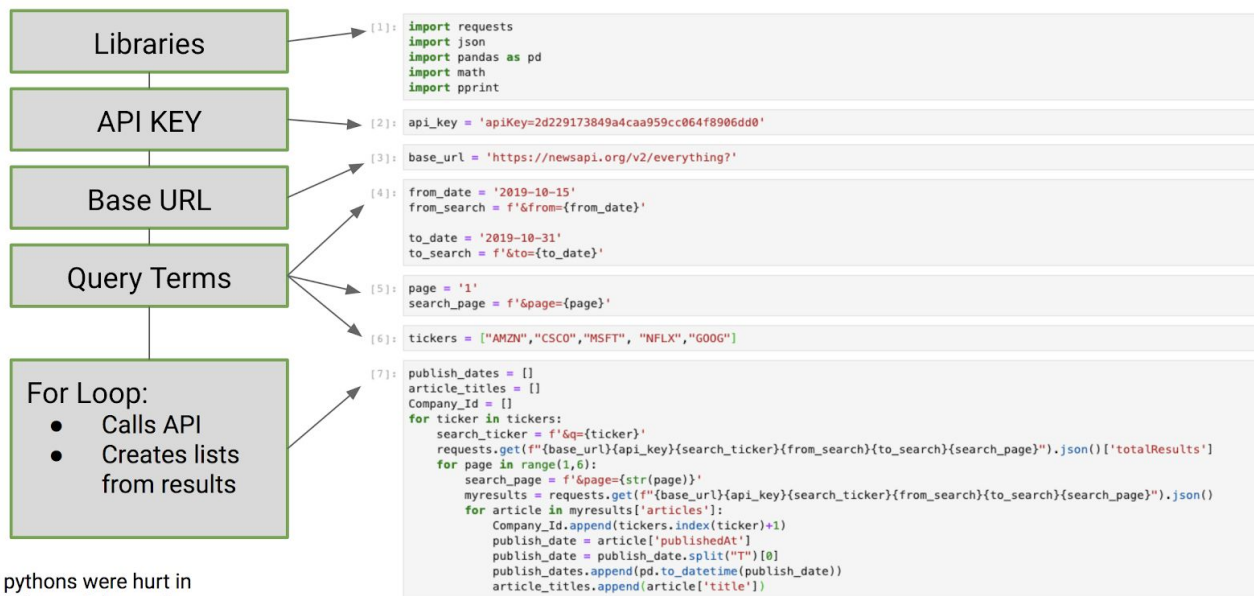
2.3 Data Transform

The data was transformed to allow for Second Normal Form. The primary key company_id was used to normalize data across three different tables. The first table contained information about various companies and is where the primary key is assigned. Each company has a unique company_id. Stock and headline data was transformed to only present information for a two-week period. This was done primarily to reduce gaps in information and to quickly confirm gaps and discrepancies between data tables. In the stock data table, price information and the corresponding date is itemized across rows for each day in the two week period. In headline data table, headlines and dates are itemized across rows for each day in the two week period.

2.4 Data Integrity

The data from both resources had high data integrity as they both provided full data sets on a daily basis. The one gap we did have between APIs was the news API provided information on each day of the week (Sunday - Saturday), while the Alpha Vantage API provided information during the week (Monday - Friday) when the stock exchange is open.

Both APIs are updated daily and the code is written to allow the user to place in desired data points, i.e. reference dates and companies.



No pythons were hurt in
making of this code

2.5 Data Refresh Frequency

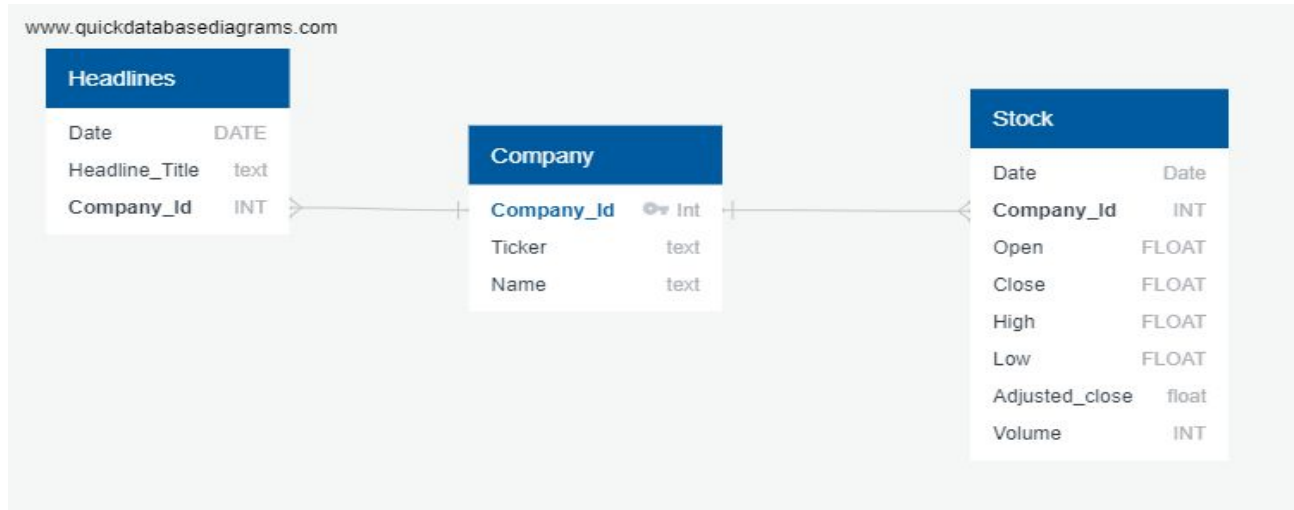
Due to the data automatically refreshing on a daily basis, we have built our code to allow us to input the desired dates and companies.

2.6 Data Security

No data anonymity and security requirements needed to be satisfied in this ETL project.

2.6 Data Loading and Availability

ERD for this project includes three tables - “company” (master table), “headlines” and “stock”. The database for this ETL project has been created in Postgres server.





Company table targets to store the companies which are being considered to be queried and compared during the data search. Currently this database tries to capture stocks and headline data for Amazon, Cisco, Microsoft, netflix and Google.

```
7 select * from company
```

	company_id	ticker	name
	[PK] integer	text	text
1	3	MSFT	Microsft
2	5	GOOGL	Google
3	1	AMZN	Amazon
4	2	CSCO	Cisco
5	4	NFLX	Netflix

“Stock” data is a set of open, close, high low values of each company for a particular date range. This is retrieved using respective exposed API in the form of CSV files which were later imported in the database.

```
30 select * from stock
```

Data Output		Explain	Messages	Notifications				
	date date	 company_id integer	open double precision	close double precision	high double precision	low double precision	adjusted_close double precision	volume integer
1	2019-10-31	3	144.9	143.37	144.93	142.99	143.37	24605100
2	2019-10-30	3	143.52	144.61	145	142.79	144.61	18496600
3	2019-10-29	3	144.08	142.83	144.5	142.65	142.83	20589500
4	2019-10-28	3	144.4	144.19	145.67	143.51	144.19	35280100
5	2019-10-25	3	139.34	140.73	141.14	139.2	140.73	25959700
6	2019-10-24	3	139.39	139.94	140.42	138.67	139.94	37029300

“Headlines” data is data set retrieved using news API for respective dates and requested companies. This is also saved as CSV files and imported into DB

```
16 select * from Headlines
```

Data Output		Explain	Messages	Notifications
	company_id integer		headline_title text	date date
1		1	How the retail industry will top \$5.5 trillion by 2020 (TGT, WMT, AMZN)	2019-10-31
2		1	Over 200 musicians have pledged to boycott Amazon's music festival unless the comp...	2019-10-24
3		1	I spent the last week with Amazon's new Echo Buds, and it showed me a lot about wher...	2019-10-29
4		1	Toy brands are reportedly paying Amazon millions of dollars for the chance to be featur...	2019-10-21
5		1	Trump reportedly tried to stop Amazon from winning a \$10 billion cloud deal, but exper...	2019-10-29

3. DATA QUALITY

Address in this section success criteria for this project. Summarize the parameter KPIs such as Totals and expected counts. What user acceptance testing was performed and what were the outcomes. What is the recommended site acceptance testing that your client can perform to ensure the expected outcomes meets their expectations?

If time permitted, this System would expose an API which will give user the power to make a search for a company (from Company table) stocks data and news headlines regarding the same company in a given date range. The expected result would look like this but response structure would be JSON:

	date date	company_id integer	company name text	open double precision	close double precision	headline_title text
1	2019-10-15	1	AMZN	1742.14	1767.38	Is Amazon Stock Ready to Break Out Again?
2	2019-10-15	1	AMZN	1742.14	1767.38	Sustainable Investing: For ESG investors, the newest challenge is separating fact fro...
3	2019-10-15	1	AMZN	1742.14	1767.38	Momentum investors are now buying shares of Apple, Amazon and Netflix
12	2019-10-15	2	CSCO	46.25	46.36	Cisco Closes CloudCherry Buyout, Boosts Customer Experience
13	2019-10-15	2	CSCO	46.25	46.36	AppDynamics Delivers Latest App Attention Index Report, Revealing Emergence of Th...
14	2019-10-15	2	CSCO	46.25	46.36	Merger Arbitrage Mondays - 7 Deals Close, With Few New Ones In Sight
15	2019-10-15	3	MSFT	140.06	141.57	Momentum investors are now buying shares of Apple, Amazon and Netflix
16	2019-10-15	3	MSFT	140.06	141.57	Is Amazon Stock Ready to Break Out Again?
17	2019-10-15	4	NFLX	283.82	284.25	Netflix Set To Report Earnings As Streaming Wars Intensify
18	2019-10-15	4	NFLX	283.82	284.25	Momentum investors are now buying shares of Apple, Amazon and Netflix
19	2019-10-15	4	NFLX	283.82	284.25	美股盤後－美股迎來開門紅聯合健康、小摩、嬌生財報告捷－鉅亨網財經新聞
28	2019-10-15	5	GOOGL	1221.5	1242.24	This Tesla Rival Is Destroying Elon Musk's Lofty Self-Driving Car Ambitions
29	2019-10-15	5	GOOGL	1221.5	1242.24	Google unveils Pixel 4 and Pixel 4 XL smartphones with enhanced cameras
30	2019-10-15	5	GOOGL	1221.5	1242.24	Warren and Biden Attack Facebook's Most Successful Business - Market Realist
31	2019-10-15	5	GOOGL	1221.5	1242.24	Here's everything Google just announced at its big Pixel event (GOOG, GOOGL)
32	2019-10-15	5	GOOGL	1221.5	1242.24	Momentum investors are now buying shares of Apple, Amazon and Netflix