Extraction, Transformation, and Load Technical Report

Nasdaq, News, and Knowledge

Stories, Stocks, and Statistics

TABLE OF CONTENTS

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

This section summarized the final objective of the project, the business problem definition (problem statement) and the expected outcome of ETL.

The project goal is to capture a list of 100 stock ticker symbols from the Nasdaq stock market index. Those symbols will be used to generate the top news stories for each company.

1.2 Scope

This section explicitly outlines the disparate data sources that are to be integrated, which components of the overall data science project is in the scope for this initiative and also lists out the components of the data science project that are not in scope here.

The project utilizes the Nasdaq website for the list of stock ticker symbols. The project also uses a news API and a separate API to capture stock prices.

1.3 Technologies and resource contributions

This section lists out the team members and their contributions towards the ETL initiative. Use this section to also outline (or list) the tech stack used to obtain the final outcome.

Tinu Ngo performed the web scraping to gather stock ticker symbols and company summary data as well as staged the final data in Postgres. David Rojo looped through the stock ticker symbols and passed them as search criteria to the news API. Will Betterton consolidated the final code and wrote the technical report.

1.4 Definitions, Acronyms and Abbreviations

List acronyms and terms that need to be defined in this section, such as ETL: Extract, Transform and Load

pd – pandas

bs4 – beautiful soup

2. ETL DETAILS

This section outlines a more detailed description of the processes utilized/proposed to achieve the objectives of this initiative.

2.1 Data Import/Extract Sources and Method

This section provides information about the data and its source. For example, API names and URLs, key parameters available and its subset which will be preserved (loaded). Data extraction protocols (API, FTP, Web scraping etc.), any permissions required to access the said extraction dataset and any restriction placed on the usage and distribution of the acquired dataset.

The project utilized the following website to web scrape the stock ticker symbols
https://www.nasdaq.com/market-activity/quotes/Nasdaq-100-Index-Components

The project utilized the following Alpha Vantage API to gather the stock prices
https://www.alphavantage.co/

The project utilized the following API to gather top new stories
https://newsapi.org/v2/everything?

The project needed to request an API key in order to use the news API and Alpha Vantage API

2.2 Data Acquisition

This section outlines the data needed, such as range and if the data is static or dynamic and needs continuous update. Outline the process to obtain again or update the dataset. The formatting and any special attributes about the data the one should be mindful of while obtaining and processing the raw dataset. How to decide on the selection of data while re-obtaining or updating. Discuss, here the dimension of the obtained dataset and if updated what is the project growth rate of the data. Lastly, address any issues or pre-requisites that needs to be cleared prior to getting the data?

The stock ticker symbols are dynamic, but they do not change often. The project limited the stock ticker symbols to five due to the API key limit of five calls per minute.

The news API is dynamic and is updated daily with today's date.

The stock price API is dynamic and changes daily.

2.3 Data Transform

In this section address any data transformation that needs to be performed to modify, clean, filter or create existing and new parameters. Address any technical analysis performed, include design

specification or data models used (example linear interpolation etc.), and any calculations performed for any newly derived fields.

The project combines the ticker symbol with their corresponding price into a single dataframe. It creates a dataframe for each stock ticker symbol with closing price and then merges all the individual ticker dataframes into a single dataframe. Then it merges the single dataframe with the news API information. The project outputs a CSV file.

2.4 Data Integrity

In this section discuss the reliability of the extraction source data (e.g., missing data, dates stored as text, invalid code values, text fields with odd characters, etc.). Address the frequency with which the data sources are updated and if it is necessary to update the local data at the same frequency. Lastly, how if any notification can be received when the source data is updated; and what if any notification will be sent to the internal team when the local dataset is updated.

Due to the fact that the code uses stock ticker symbols to web-scrape for news articles, the news articles are not always about the specific company and may just mention the ticker symbol. There are no assurances that the primary topic of the article is about the specific company.

The data sources are all updated daily and it would be necessary to update the local data at the same frequency to get an up-to-date CSV file.

2.5 Data Refresh Frequency

This section explicitly lists the frequency with which this ETL process will refresh the local dataset (Daily, Weekly, Monthly, Quarterly, Semi-Annually, etc.).

The data should refresh on a daily basis so the local dataset can get up-to-date news stories.

2.6 Data Security

This section discusses any data anonymity and security requirements need to be satisfied. Address any federally mandated HIPAA considerations, any need to build in additional privacy, Encryption, Data masking, Auditing, Backups etc.

There is no need for additional data security.

2.7 Data Loading and Availability

This section addresses the data schema and during of data retention. Discuss the interface that will allow your Client/Users to access the data.

The code outputs a CSV file for the client/user to access the data.

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?

The code outputs a CSV file for the client/user to access the data. CSV files can be uploaded into any database tool for their use.