Proposal_GRP16: US Exchange Research Database

INFO7290 - Fall2020

Presented by: Dimple Bapna, Shreshtha Jha, Wasinee Opal Sriapha, Yufei Wang

Revision Table

Date	Version	Changes	Author
11/05/2020	1.0	• Initial document • Dataset selection • Designing Dimension Tables • Outlining workflow for the project	Team 16
11/15/2020	1.1	• Updated information about each dataset • Redefined schemas • Updated dimension tables • Added Additional Details for ETL architecture • Added Additional Details for Error handling	Team 16
12/01/2020	1.2	• Edited entire document to reflect the current design (refer to the project's design document for more details) • Elaborated on the objective • Implemented a lookup task on currency	Team 16
12/03/2020	1.3	• AWS lambda function for extracting daily stiock data • SCD on Dest_companyInfo and Dest_ComapnyOverview tables • Erro handling for bad prefix and bad sector on Dest_companyInfo • Visualizations for a quick analysis in Tableau	Team 16
12/15/2020	1.4	• Updated SSIS package with appropriate names • Updated dimension tables section • Updated OLAP/star schema section and included a brief introduction to the project's cube design	Team 16

Table of Contents

Objective

With various sources of daily/quarterly updated exchange information available across market research platforms, the data warehouse in this project serves as a repository for integrated and transformed sets of information as well as providing robust information access that enables prompt and profitable market analyses. The aim of the project is to build a system in AWS that would automate daily API data-extraction for the daily-updated dataset. Through SSIS, this data will then be integrated and transformed along with other financial data (e.g. valuation ratio, earnings and income statement, etc.) The data will pass through a number of lookup tasks for data validating, SCD capturing, and error handling purposes. The clean and transformed data will be loaded in a database with an entity–relationship model. A star schema, which is the simplest data mart schema, will also be constructed for specific technical analyses. Further analyses will also be presented in a visualization format using Tableau dashboard.

Datasets

Please note: Four companies from four different sectors; Health Care, Technology, Financials, and Energy (1 company/sector) will be analyzed using the five datasets below.

Stock Time Series Daily Data	
Source	Alpha Vantage Inc. - the company is in partnership with major exchange institutions and a lead stock APIs provider. https://www.alphavantage.co/documentation/#dailyadj
Description of Data File	This **daily-updated** dataset contains raw stock quote prices. Information for each ticker may be traced back over 20+ years of historical data.
Usage	Daily volumes, prices (open, close, high, and low) as well as dividend occurrences will be captured for each stock ticker starting from January 2, 2018 – Present for time-series plotting, patterns mining, and other analyses.
Extraction Method	API Call - explained in the 'ETL Architecture' section

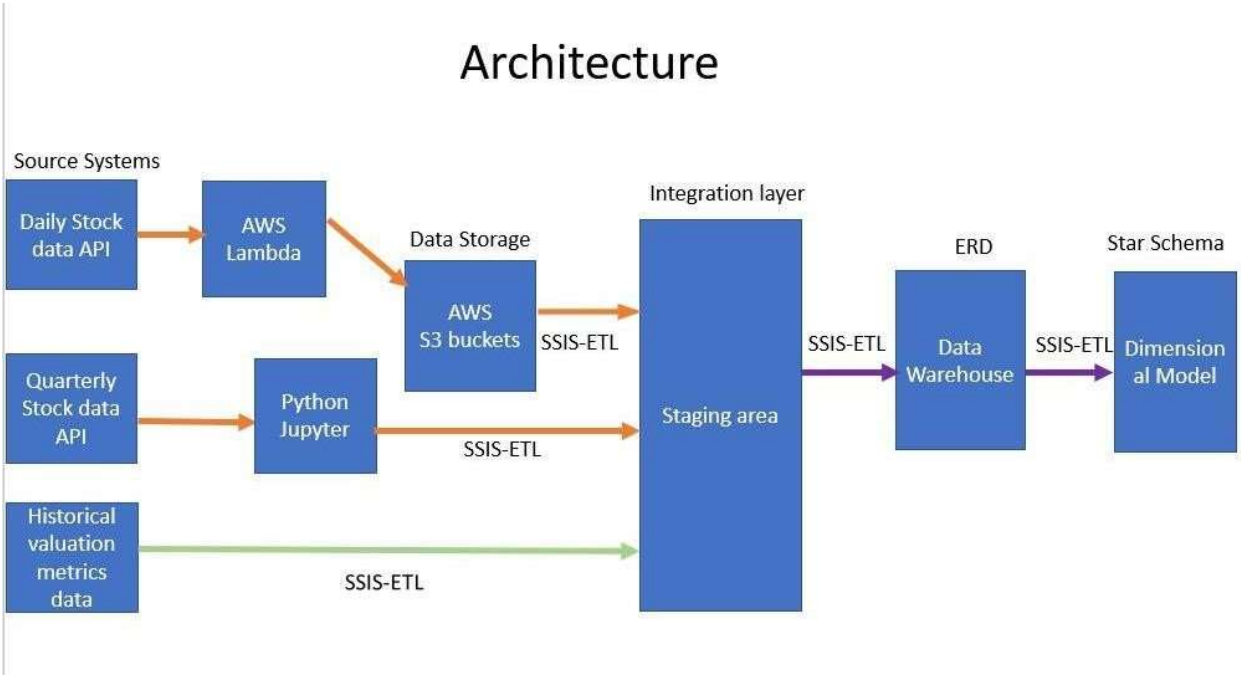
Company Overview	
Source	Alpha Vantage Inc. https://www.alphavantage.co/documentation/#company-overview
Description of Data File	**Quarterly** updated dataset on financial ratios and other key metrics as well as any changes in the general information on the company
Usage	Quarterly data of the following columns of the dataset will be captured for each stock ticker starting from January 2, 2018 – Present. The data will help investors understand the brief company's performance & growth, the stock's current value, and dividend offered. Company's General Info: company_name, sector, address, and FullTimeEmployees Key Metrics: Market Capitalization, EBITDA (Earnings Before Interest Tax Depreciation and Amortization), Dividend Per Share, DividendYield, QuarterlyEarningsGrowth, and QuarterlyRevenueGrowth
Extraction Method	API Call - explained in the 'ETL Architecture' section

Quarterly Earning	
Source	Alpha Vantage Inc. https://www.alphavantage.co/documentation/#earnings
Description of Data File	**Quarterly** updated dataset on company earnings, analyst estimates and surprise metrics
Usage	For every fiscal quarter from January 2, 2018 – Present, data on reported EPS, estimated EPS, EPS surprise (difference between the reported EPS and estimated EPS), and surprise percentage will be captured for each stock ticker to understand how the company has been performing compared to Wall Street estimates. It will also be observed if any positive surprises had led to sharp increases in the company's stock price, or if any negative surprises had led to sudden declines.
Extraction Method	API Call - explained in the 'ETL Architecture' section

Income Statement	
Source	Alpha Vantage Inc. https://www.alphavantage.co/documentation/#income-statement
Description of Data File	**Quarterly** updated dataset that summarizes a company's revenues, profitability and expenses over a quarter period.
Usage	For every fiscal quarter from January 2, 2018 – Present, data on total revenue, gross profit, net income, total operating expense, and research and development cost will be captured for each stock ticker to help analyze the profitability, expenses and future growth of a company.
Extraction Method	API Call - explained in the 'ETL Architecture' section

Valuation Data	
Source	YCharts Inc.- investment research company with a full-service fundamental research platform. https://ycharts.com/stocks
Description of Data	Historical and daily-updated valuation metrics for any searched ticker.
Usage	Financial ratios; Price-to-earnings (P/E) ratio, Price-to-Sales (P/S) Ratio, Price-to-book (P/B) ratio will be captured for each stock ticker starting from January 2, 2018. These ratios, computed using data found in financial statements, are integral indicators of fundamental analysis and serve as a summary of the quarterly financial statements as well as revealing the financial health and prospects of the company.
Extraction Method	Data will be weekly exported from the platform into a csv file.

Data Architecture



Due to the high data-update frequency of the stock price dataset, the process to automate Daily API calls in AWS using a built-in Lambda function was built and the response would be stored in a 'csv' format in the project's Amazon S3 cloud database. A trigger (Cloud Watch event) was added to the lambda function to automatically call the API on a daily basis. For the quarterly updated datasets, due to the lower update frequency, API calling using Python outside of the AWS was opted for. Lastly, the valuation data from Ychart would be manually exported and stored in a local desktop.

Moving into the integration layer in the middle of the diagram, the main actions are now in SQL Server Integration Services and all the data would pass through a number of lookups for data validation as well as the slowly changing dimension and error handling processes. The clean and transformed data would then be loaded in a DW with an entity–relationship model. A star schema to store a certain set of columns to conduct specific technical analyses was also constructed. The analyses would later be presented in a visualization format using Tableau dashboard.

Joining Methods

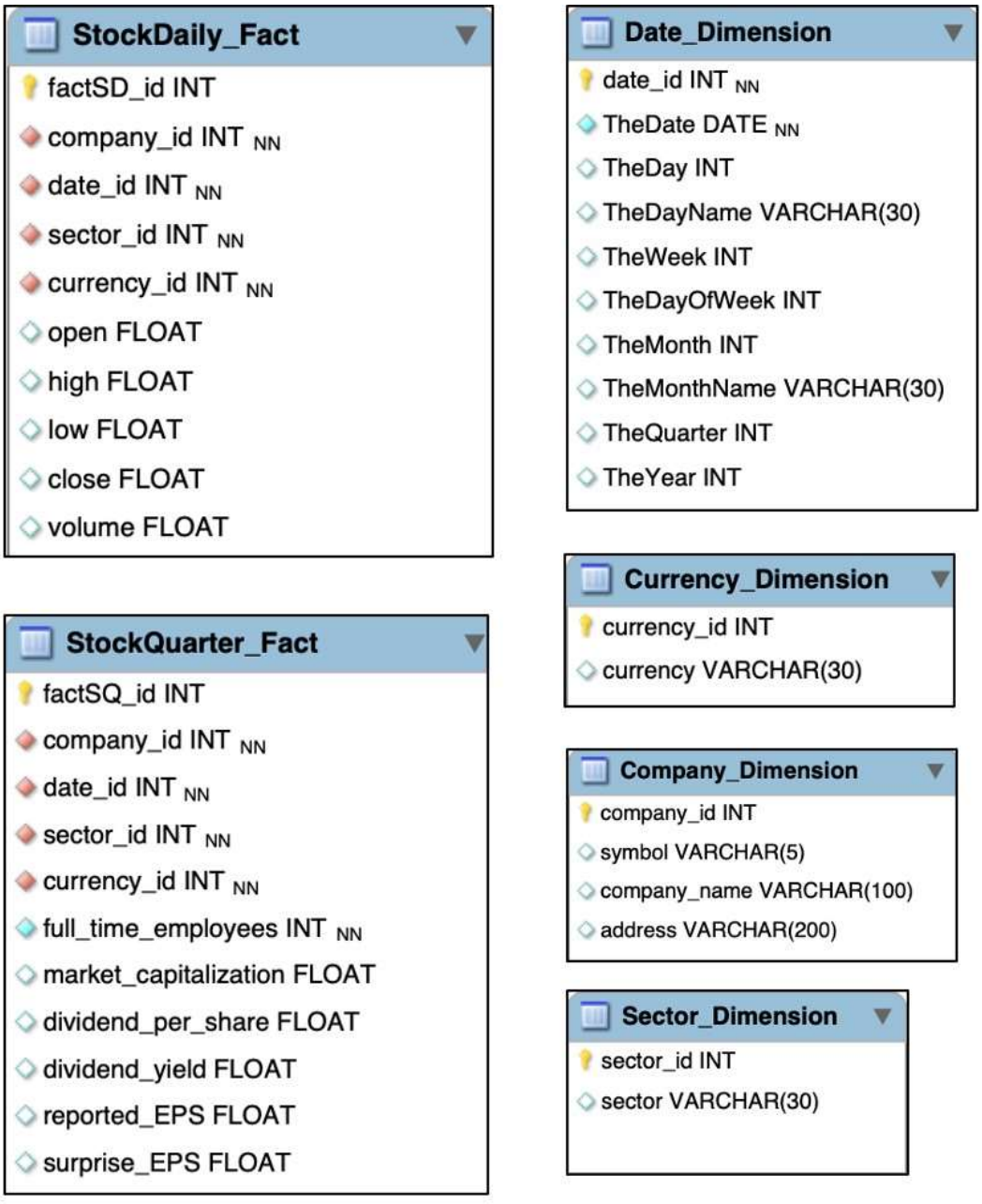
1. Create shells of all tables in SSMS including staging, fact, dimension, and destination tables for both ERD and Star Schema models.
2. For ERD, all tables were joined on symbols and have the common attribute of company_id.
3. For the dimensional model, the fact table has its own PK for each row and contains PKs of all dimension tables (as FKs). After creating PKs for dimension tables, join staging tables with dimension tables on common columns to copy unique keys from dimension tables to staging tables. After the fact table has been loaded with unique keys from the staging table, querying a report for any stock tickers can be easily done via the fact table.

Data Transformation

1. Dates/times formatting
2. Percentage/decimal formatting
3. Change data lengths and data types if necessary
4. Replace space in the column headers with underscore "_"
5. Replace "None" with 0 or blank

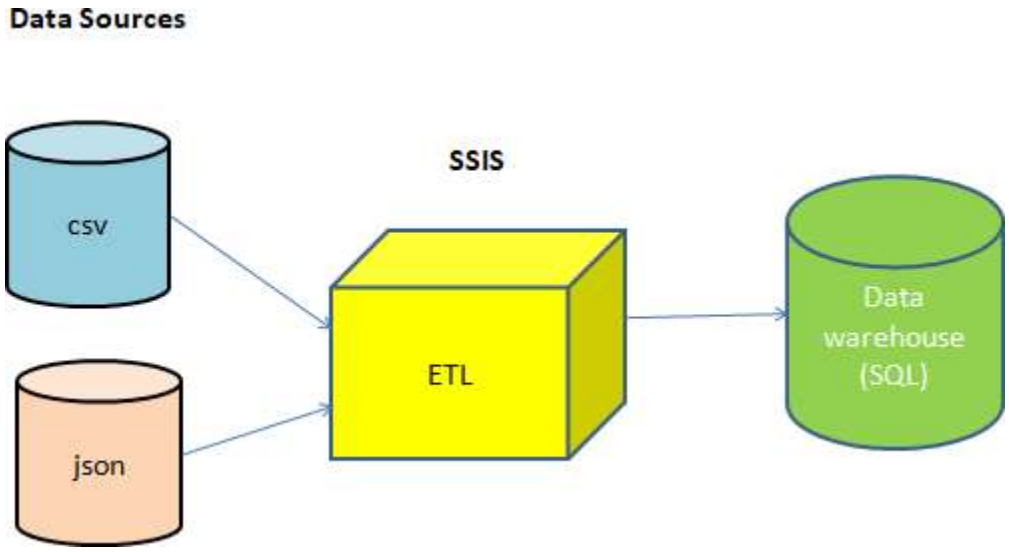
--- Dimension tables next page ---

Dimension Tables



High-level data flow

1. Diagram of the data flow



ETL process in SSIS:

- Data integration in SSIS
- Staging / archive areas in SSIS
- Transformation in SSIS
- Load into Target destination(SSMS) in SSIS

Extraction and loading it into Staging Tables:

Datasets extracted using APIs are Stock Time Series Daily Data, Company Overview, Quarterly Earnings. The data will be in json format. So, 3 staging tables are created to store three different sources named as Stock_Time_Series_Daily_Data_Staging, Company_Overview_Staging and Quarterly_Earning_Staging respectively. Similarly, the valuation data (in csv format) will be staged into Valuation_Data_Staging table.

Transformation:

Once all the data is in the staging tables, data transformation will be performed (data cleaning, data type changes, column name updates, etc.), and then this is loaded into another table. Also, the data will be split here into Facts and Dimension tables.

Load:

Once done, the SSIS package will be executed to load the data directly into the tables in SQL (Data Warehouse).

2. Describe error handling

Introduce errors into source systems to force SSIS to fail:

- Bad string characters in "CHAR" columns
- Bad prefix in "CHAR" columns

Possible Columns in test source files to be introduced errors:

- CompanyOverview: Sector, Name

Error handling in SSIS:

- SSIS (look up task)
- SSIS (data conversion)

Introducing errors into source systems will help test the SSIS packages`elasticity. Good/sample data in the staging area will be used into the look-up task to validate columns that contain errors. The good/matched rows with no errors will be loaded into the final table, and the data conversion task may be added in order to transform the outcome with appropriate data types/lengths. The error rows will be logged into a flat file with clear indications.

3. Logging process

SSIS makes it easy to log everything by its capture everything mode. However, logging everything can consume a lot of storage space, network bandwidth and processing overhead, etc.
The following events and metrics were logged:

- Start and stop events
- Status
- Errors and exceptions
- Audit Information
- Testing and debugging information

4. First time vs Continuous loads

Data from all the five sources will be updated in a periodic fashion as mentioned below.

- Stock Price Data- Daily update.
- Company overview- Quarterly update
- Income Statement - Quarterly update
- Earning - Quarterly update
- Valuation data - Data is daily updated, but the load frequency will based on the manually exported of dataset.

In SSIS, for the first-time loading, the initial loading package was designed to load the data from the source systems (the quarterly company overview data, historical stock price data, the quarterly earnings data, and the historical valuation data) into its corresponding staging table. The data conversion tasks were designed to change datatype and length, data derived tasks to update the current record date, and lookup tasks to verify the foreign key before loading it into its corresponding destination table.

In SSIS, for the continuous loading, the continuous loading package was created. After the initial loading, error handling task was added into the data flow and used the first-load no-error data to validate different error types. In addition, Slowing Changing Dimension task was added to handle company`s address and full-time employee information update.

5. Data warehouse design along with history

This data warehouse is designed through a dimensional model, and the dimensional model is a database structure that consists of fact and dimension tables. A star schema is considered to join the facts with dimensions in order to organize the data. A dimensional technique--slow changing dimension (SCD) is used in the dimensional data warehouse to capture the changing data with the dimension over time.

In the company dimension table, companies' addresses may be updated along with time, therefore, several columns (old_address, start_date, end_date and update_date) are added into the company dimension table to provide historical changes.
Original record example:

Stg_CompanyOverview									
company_id	record_id	company_name	sector	address	full_time_employees	old_address	start_date	end_date	update_date
IBMZY_012	1233	IBM	technology	1200 5th Ave #1100, Seattle, WA, 98101	40000	NULL	2010-08-03	NULL	NULL

Updated record example:

Dest_ComapnyInfo									
company_id	record_id	company_name	sector	address	full_time_employees	old_address	start_date	end_date	update_date
IBMZY_012	1233	IBM	technology	1100 Olive Way, Seattle, WA 98101	40000	1200 5th Ave #1100,Seattle, WA, 98101	2010-08-03	NULL	2010-08-05

Similarly, the number of full-time employees might change over time. The historical data can also be tracked using SCD technique, as follows:

Updated record if the full_time_employees count for a company changed from 40000 to 45000:

Dest_CompanyOverview									
company_id	record_id	company_name	sector	address	full_time_employees	old_address	start_date	end_date	update_date

IBMZY_012	1233	IBM	technology	1200 5th Ave #1100, Seattle, WA, 98101	40000	NULL	2010-08-03	2010-08-06	NULL
IBMZY_012	1233	IBM	technology	1200 5th Ave #1100, Seattle, WA, 98101	45000	NULL	2010-08-06	NULL	NULL

6. Data mart design with aggregates

Data marts are similar to DWs in the way that they use the same tools, data models and design principle. However, data marts have more specific business audiences. Data marts with aggregations will be created in this project to store data that will be used in fast reporting, BI analytics, and data mining for specific analysis tasks.

Data validation

1. How will you validate data has been loaded correctly?

The objective of data validation is to assure that data is accurately loaded from source systems to the destination and transformed as expected. Validation conditions :

- Validate the source and target table structure corresponding mapping structure
- Validate the match of source data type and target data type
- Validate the match of the length of data type of source data and target data
- Validate the match of the name of columns of source data and target data
- Validate the constraints are defined for specific tables
- Validate correctness issues, null, non-unique or out of range data through lookup tasks
- Validate the currency types by using the lookup validation table

2. How will validations be logged / recorded?

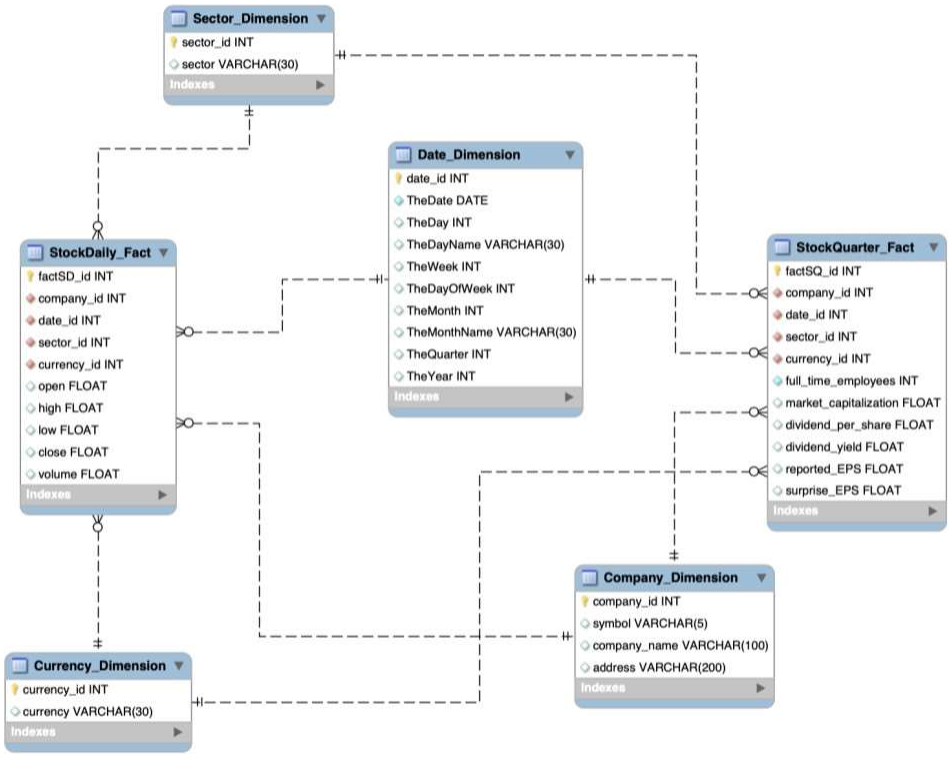
Validation Data using SSIS lookup tasks:

Add validations when loading the data from the Staging tables into the Destination tables (Data Warehouse). The lookup no match output will go into the error table with error types indication. The lookup match output will pass into the final destination table.

Analytics Design

1. OLAP cubes from star schema

OLAP cubes' biggest value lies in its multidimensional approach to organizing and analyzing data. It breaks down data into logical and useful dimensions that aims to focus on easy data management and rapid data analysis. In the study providing users with searchable access to key data points with important metrics like company name, sector, date, currency is the fundamental purpose of cubes so that the data can be sliced, and diced as needed to handle the widest variety of questions that are relevant to a user's area of interest.



2. Reporting tools and visualizations

Create analysis views/reports in SSMS and visualization dashboard in Tableau

Analytics outcomes

1. Why is this data interesting?

- Today's businesses need timely information that helps the organizations to make important decisions in business, meanwhile also keep a check on how their competitors are performing. The datasets that were used in the study will help in predicting the market's next move.

- Stock market time series data and its analysis enables investors to identify the intrinsic worth of a market move even before investing in it. By analyzing this, investors and traders can arrive at buying and selling decisions. Understanding this market acts as is the main source of knowledge for the companies that want to raise funds for their expansion. It can also help a company to launch new products and pay its debt.

- The company overview and historical valuation data will showcase the financial position of the organization and will help in forecasting insightful results to leverage the business.

- Financial analytics helps in shaping up tomorrow's business goals. A new business model can be developed based on the study of historical finance data and evaluating balance sheets, earning metrics, etc. The changing needs of the traditional financial department and the advancement in finance analysis will lead to supportive results.

- The earning metrics are typically interesting as it focuses on the tangible assets of an organization such as cash, company earnings, estimates and on measuring and managing these can demonstrate profitability, cash flow and value of the business.

2. What questions do you hope to answer?

There are various reasons why financial analytics is becoming more important these days. Through the study the aim is to answer the questions and throw light on certain areas that are listed below:

- Measure the profitability and the strength of the organization.

- Predict the stock market value for the future market using historical data by identifying patterns in individual sectors by exploring time series forecasting.

- It will help understand the performance and financial condition of the company. The stock market analysis is extremely intuitive in terms of showcasing the financial position of the organization.

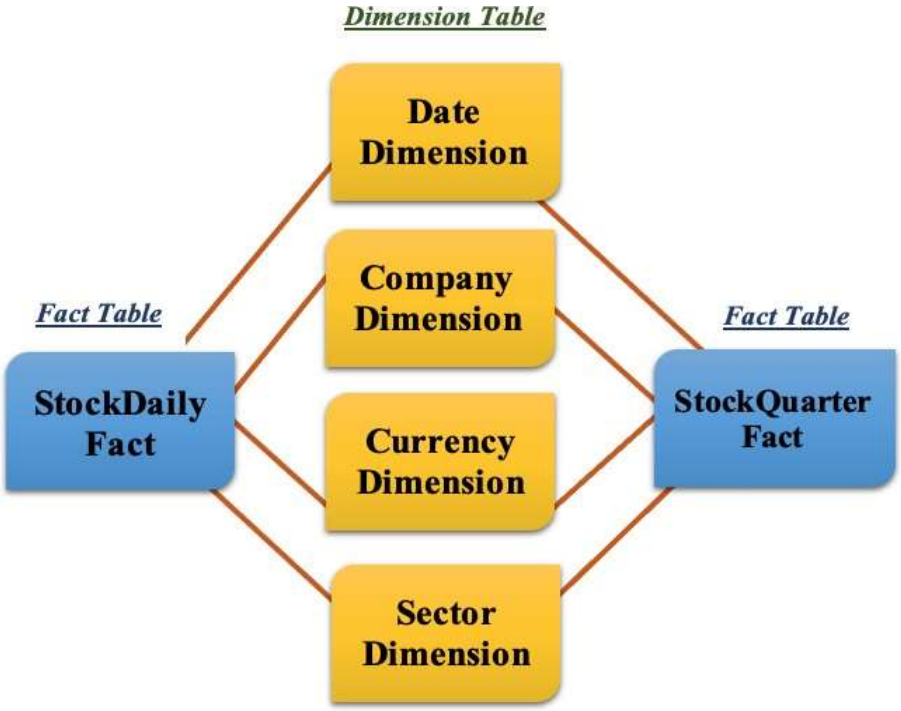
- Both the internal and external users will be able to use the analysis in numerous ways. The external users such as banks who might want to give a loan and/ or investors who look at the organization to see whether or not they want to do business with the organization.

- The management of the organization who are the internal users can use this analysis and help with the decision-making processes to improve the health of the company and optimize its operations.

- Forecasting the variations in the market can give us a deeper understanding of the economic condition in general of each of the sectors in the country.

High-level conceptual model

To create a data warehouse that serves as a repository to provide robust information access and enable prompt exchange/market analyses, a star schema model will be adopted to build a simplified schema with two fact tables indexing all dimensional tables. This model will optimize the querying writing process due to its straightforwardness in joining the tables. Reports with queried information can then be retrieved faster.



Appendix

Professor Comments

Version 1.0

You need to tell me more about the data files I think you might be ok but need some more info
You should include links to the files
It would be good to see if you can pull in some additional data
How are you going to call the APIs
Logging looks good
I like the update process for adding data
You will likely need to introduce your own errors into the json or other files to force SSIS to fail
That really isn't a high-level conceptual model

Version 1.1 & 1.2

Prof: Objective doesn't tell me anything

- Should we elaborate more on the purposes of choosing stock datasets and designing this type of data warehouse?
    - **Prof: Yes tell me what you intend on building**

Prof: You should tell me what will have scd's

- Could you please elaborate this comment? Are you asking which of our table and what column will have the scd? If so, we have already included this on page 10 - 11 (section 5 'Data warehouse design along with history'). Please let us know if this is not what you're expecting.
    - **Prof: No that should be fine**

Prof: You could add a lookup table of States to use in validation steps

- Thank you for suggesting this. However, our data includes companies outside of the US and they may not have States in their addresses. If this type of lookup validation doesn't meet our project's design requirements, would it be alright not to include it?
    - **Prof: You could still do a lookup for us based companies.**

Version 1.3

Watch the naming of ssis objects such as derived column and ole db command
Very good