

DWBI Project: Architect, Populate and Explore a Data Warehouse for Stock Market Analysis

ITI - Data Visualization Track - Graduation Project



Abdelwahab Soliman

Wael Mohammad

Yusuf Madkour

February 12, 2021

Table of Contents

[Introduction 2](#_Toc64418321)

[Data Sources 3](#_Toc64418322)

[S&P 500 component stocks 3](#_Toc64418323)

[Company Information 3](#_Toc64418324)

[Market Change 4](#_Toc64418325)

[Stock Daily Statistics 5](#_Toc64418326)

[Stock Yearly Statistics 5](#_Toc64418327)

[Data Fabrication for the Missing Years 7](#_Toc64418328)

[Index Daily Statistics 7](#_Toc64418329)

[Data Warehouse Data Model 9](#_Toc64418330)

[Description and Schema 9](#_Toc64418331)

[What Dimensions and Why? 10](#_Toc64418332)

[Mapping the Data Sources to the Data Model 12](#_Toc64418333)

[Logical Data Map 13](#_Toc64418334)

[Stock\_yearly\_stats fact table 13](#_Toc64418335)

[Stock \_Daily\_Stats\_Fact 13](#_Toc64418336)

[Application of Data Warehouse 18](#_Toc64418337)

[Conclusion 19](#_Toc64418338)

[Appendix 21](#_Toc64418339)

# Introduction

On daily basis, various companies issue their shares in the stock market in order to collect money from investors. Although many companies use the market to enhance their growth or pay their debts. Everyday Sellers and buyers participate in it exchanging shares in order to increase their income.

There are three main ways to raise money using the stock market as an investor which are capital gain investments also called growth investments, collect dividends from high value stocks or buying derivatives from companies. Even it’s applicable to any investor to take part in the daily market stock exchange. Investors tends to deal with professional brokers or consulting companies to maximize their profit, and to avoid the risk associated with it since the stock market is not always stable and may vary from one day to another.

Consulting companies and stock brokers have a substantial experience to tell investors which stock index to choose from and then help investors either interested in capital gain investments or dividend-based investments to choose the most suitable companies for them and may also encourage them to have a collection of both growth stocks and dividend stocks in the form of a stock portfolio to optimize their profit with respect to investment return and risk.

Therefore, the project simulates a stock broker company which uses multiple data sources and stores them in a data warehouse with a customized dimensional model, in order to use them in its stock analysis and queries and finally support its consulting process.

# Data Sources

In addition to the 3 data sources provided, we obtained an extra data source consisting of historical data of the 5 major stock indexes in the New York Stock Exchange Market. We have also fabricated data to extend the range of one of the datasets provided in the project statement. Each data source is explained in detail in the following sections.

## S&P 500 component stocks

This is the first data source provided in the project statement, it is a [Wikipedia page](https://en.wikipedia.org/wiki/List_of_S%26P_500_companies) listing the details of the 505 corporations that are components of the S&P500 index in the meantime. It also contains changes to the list of S&P500 components. The Wikipedia page reports that there have been **1,186** changes between **January 1, 1963 and December 31, 2014**. However, **only 266 changes** are recorded on the page.

To summarize, we have acquired two datasets from this Wikipedia page. We are calling the first one **Company Information** and the second one **Market Change**. Detailed information about the two datasets is given in the following subsections.­­­­

### Company Information

The list of corporations on this page was first **made available on** [**July 21, 2005**](https://en.wikipedia.org/w/index.php?title=List_of_S%26P_500_companies&oldid=23205669)**[[1]](#footnote-1)**, this old version only listed the names of these corporations without any extra information. However, the current page contains a table listing the stock symbols along with descriptive information about each corporation. The table with its current structure was **first made available on** [**March 1, 2007**](https://en.wikipedia.org/w/index.php?title=List_of_S%26P_500_companies&oldid=112958830). The table consists of 505 rows, each row represents one corporation, 8 attributes are recorded for each corporation.

The 8 attributes are:

* **Symbol**: This is an abbreviation used to uniquely identify publicly traded [shares](https://en.wikipedia.org/wiki/Stock) of a particular [stock](https://en.wikipedia.org/wiki/Stock) on a particular [stock market](https://en.wikipedia.org/wiki/Stock_market). A stock symbol may consist of letters, numbers or a combination of both.
* **Security**: The term "security" refers to a financial instrument that holds some type of monetary value. It represents an ownership position in a publicly-traded corporation via stock. However, in this context, it is sufficient to say it is just the name of the corporation.
* **SEC filings**: These are links to financial statements submitted to the U.S. Securities and Exchange Commission (SEC) by the corporations. Investors and financial professionals rely on these filings for information about companies they are evaluating for investment purposes.   
  Note: we found the data in these reports irrelevant to our business case.

Figure A sample of the Company Information table

* **GICS Sector**: The Global Industry Classification Standard (GICS) is an industry taxonomy developed in 1999 by MSCI and Standard & Poor's (S&P) for use by the global financial community. The GICS structure consists of **11 sectors**, **24 industry groups**, **69 industries and 158 sub-industries** into which S&P has categorized all major public companies.



Figure A sample of the classification from the Wikipedia page of the GICS

* **GICS Sub-Industry:** This is related to the previous attribute; the sub-industry is a higher level of detail attribute for the same classification mentioned earlier.
* **Headquarters Location:** The main location of the corporation.
* **Date First Added:** The date this stock was first added to the public market, in technical terms, it is called the date of the **IPO**.
* **CIK:** A Central Index Key or CIK number is a number given to an individual, company, or foreign government by the United States Securities and Exchange Commission. The number is used to identify its filings in several online databases.
* **Founded:** This contains the year the company was founded.

It is worth noting that some cells contained two years. After some research, we found out that this happens in case the corporation split into two corporations each having its own stock, so two years are included in this case, one for when the mother company was first founded and the other for when the split happened.

### Market Change

The second dataset found on this Wikipedia page is a table recording changes to the components included in the S&P500 index between December 7, 1999 and January 21, 2021. The table was first made available on this Wikipedia page on [March 10, 2011](https://en.wikipedia.org/w/index.php?title=List_of_S%26P_500_companies&oldid=418136367). It consists of 266 rows, each row represents a stock replacement, an addition or a removal from the index. For each row, there are 4 attributes.

The 4 attributes are:

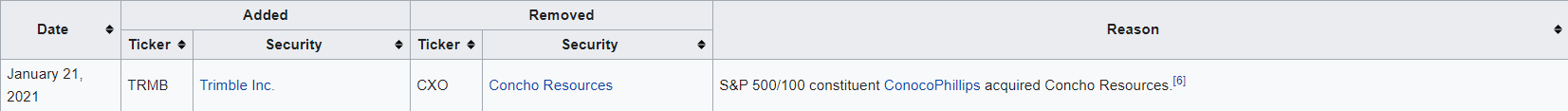
* **Date:** The date on which the change happened
* **Added:** The symbol of the stock and the corporation name of the stock added
* **Removed:** The symbol of the stock and the corporation name of the stock removed
* **Reason:** This is the reason as to why this change happened

Figure A sample of the Market Change table

## Stock Daily Statistics

The second data source we are using in our project is a dataset downloaded from [Kaggle](https://www.kaggle.com/camnugent/sandp500), it is a large and well-structured csv file containing daily stock exchange data pertaining to the aforementioned 505 components of the S&P500 index. The first version of this dataset was made available on [August 11, 2017.](https://www.kaggle.com/camnugent/sandp500/version/1) The version we are using in this project was made available on [February 10, 2018](https://www.kaggle.com/camnugent/sandp500/version/4).

It consists of 619,040 rows; each row represents the daily numbers of one stock. There are 505 companies in this dataset, the data is collected over 1825 days between February 8, 2013 and February 7, 2018. There are 6 attributes recorded in each row.

The 6 attributes are:

* **date**: The date in which the numbers are collected
* **open:** The price at which the stock started this day
* **high:** The highest price this stock reached on this day
* **low:** The lowest price this stock reached on this day
* **close:** The price of the stock at the end of the day
* **volume:** The number of traded shares for this stock on this day
* **Name:** The symbol name of the stock



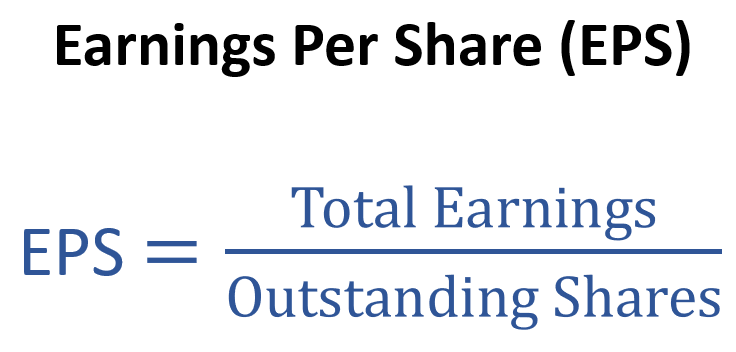
Figure A sample of the daily data of S&P500 stocks

**Note:** All prices are in USD.

## Stock Yearly Statistics

The third data source utilized in this data model is obtained from [DataHub](https://datahub.io/core/s-and-p-500-companies-financials#readme). It was first made available 2 years ago. It contains statistics about each stock at the end of one year. Although it was not clear which year these statistics were reported, the year 2014 was mentioned in the readme, so we assumed that these statistics were collected at the end of 2014. Another data source of the same nature was acquired from [DataHub](https://datahub.io/JohnSnowLabs/list-of-companies-in-the-new-york-stock-exchange#pandas) as well, it contains the same statistics for the year 2017.

The data source is comprised of two csv files; **constituents** and **constituents-financials**. The file constituents.csv contains descriptive information about the corporations, namely, the stock symbol, the corporation name and the GICS sector. The file constituents-financials.csv contains the statistics measured for each stock. Each csv file contains 505 rows, each row represents one stock. The measures in the constituents-financials are described below.

**Earnings/Share:** It is a company's net profit divided by the number of common shares it has outstanding. The resulting number serves as an indicator of a company's **profitability**.

**Price/Earnings:** It is the ratio for valuing a company that measures its current share price relative to its Earnings/Share. A high P/E ratio could mean that a company's stock is over-valued, or else that investors are expecting high growth rates in the future.

**Dividend Yield:** It is the amount of money a company pays shareholders for owning a share of its stock divided by its current stock price.

**52 Week High:** It is the highest price at which a stock, has traded during the year.

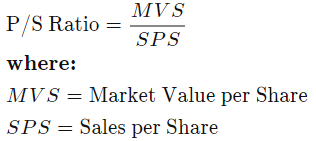
**52 Week Low:** It is the lowest price at which a stock, has traded during the year.

**Market Cap:** it refers to the total dollar market value of a company's outstanding shares of stock. It is calculated by multiplying the total number of a company's outstanding shares by the market price of one share of the stock at the end of the year.

**Market cap = share price x no. of shares outstanding**

**EBITDA:** Earnings before interest, taxes, depreciation, and amortization, is a measure of a company's overall financial performance and is used as an alternative to net income in some circumstances.

**EBITDA = Net Income + Interest + Taxes + Depreciation + Amortization**

**Price per Sales:** It is a key analysis and valuation tool that shows how much investors are willing to pay per dollar of sales for a stock. It is typically calculated by dividing the stock price by the underlying company's sales per share.

**Price per Book:** It measures the market's valuation of a company relative to its book value. P/B ratios under 1 are typically considered solid investments.

In this equation, book value per share is calculated as follows: (total assets - total liabilities) / number of shares outstanding).

### Data Fabrication for the Missing Years

To cover the missing data in the years 2015 and 2016, we developed two PL/SQL scripts to generate pseudorandom data, one for each year. The reason two scripts were used is to ensure the smoothness of the transition between 2014 and 2017, a slightly different calculation is used to generate the data in each year.

For **2015**, a random number bounded by the value of the statistic in **2014** and the average of the statistic in 2014 and 2017 was generated.

For **2016**, a random number bounded by the average of the statistic in 2014 and 2017 and the value of the statistic in **2017** was generated.

## Index Daily Statistics

This data source[[2]](#footnote-2) contains several daily features of S&P 500, NASDAQ Composite, Dow Jones Industrial Average, RUSSELL 2000, and NYSE Composite from 2010 to 2017. It should help our customers compare the performance of different indices and take well-informed investment decisions. It consists of 5 csv files, one file for each index. The dataset along with the cited article were published on **February 3, 2020**.

Each file has 1984 rows, each row represents one day. The authors of the article created this dataset to train a Convolutional Neural Network, so it is a clean well-structured dataset. They have calculated technical indicators that will be utilized in the Data Model. The indicators are explained below.

**mom, mom1, mom2, mom3:** The momentum which is also referred to as the **Rate of Return**, is the net gain or loss of an investment over a specified time period, expressed as a ratio of the investment’s initial cost. mom is the rate of return over one day, mom2 is the rate of return over two days... etc.

**ROC\_5, ROC\_10, ROC\_15, ROC\_20:** It is the **Rate of Change** over a specified time period. It is calculated the same way as the Rate of Return but multiplied by 100 and expressed as a percentage.

**EMA\_10, EMA\_20, EMA\_50, EMA\_200**: An exponential moving average (EMA) is a type of moving average that places a greater weight and significance on the most recent data points. The exponential moving average is also referred to as the exponentially weighted moving average. An exponentially weighted moving average reacts more significantly to recent price changes than a simple moving average (SMA), which applies an equal weight to all observations in the period.

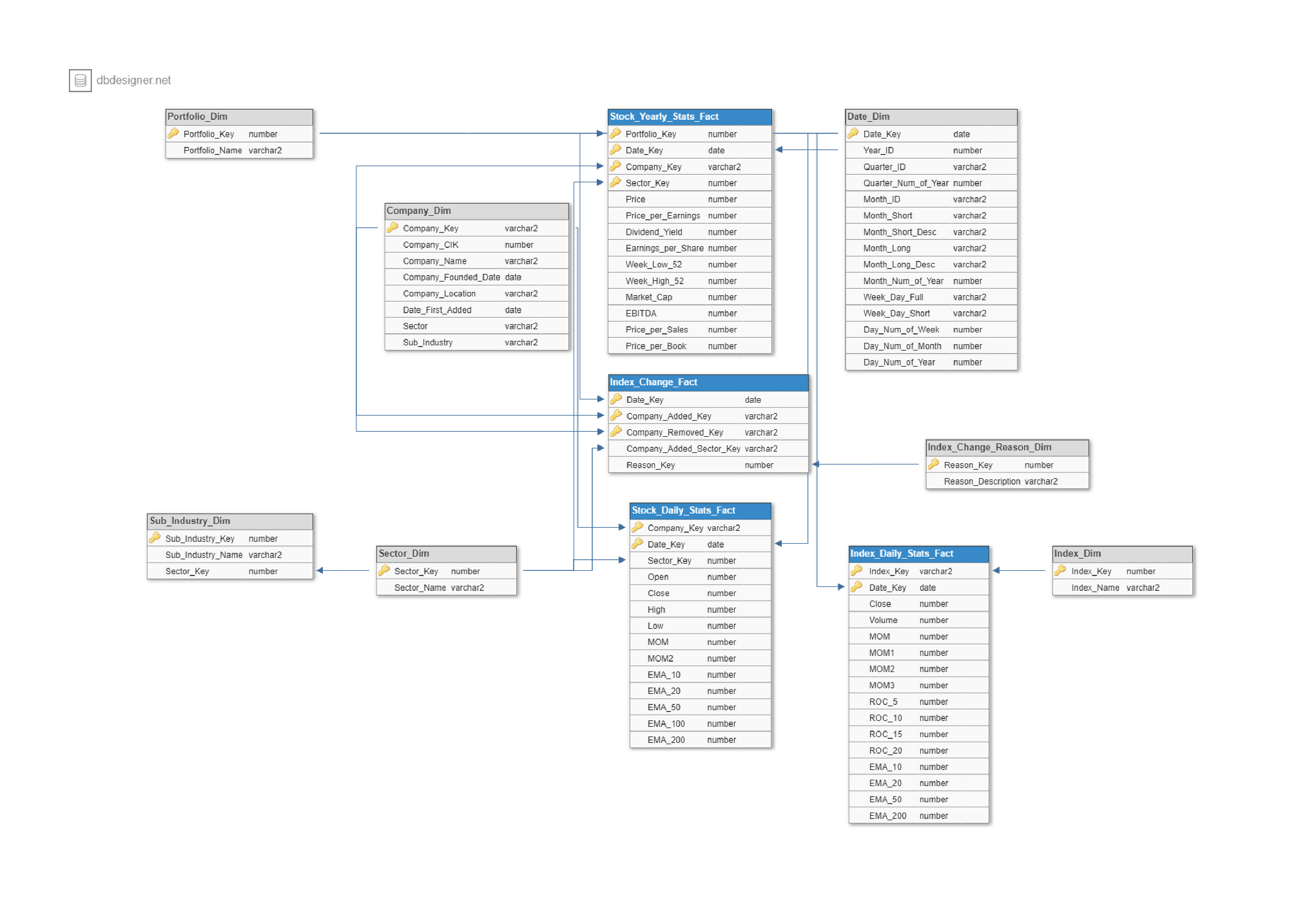
EMA: [Close – EMA(Previous day) ] \* Multiplier + EMA(Previous day)

Multiplier = [2 / (Time Periods +1)]

Initial EMA: Simple moving Average

# Data Warehouse Data Model

## Description and Schema

A **Galaxy Schema** is used to design the Data Model, the model is composed of **four** fact tables and **seven** dimensions. The diagram of the model is shown below, tables with blue headers are for facts and tables with grey headers are for dimensions.

The total size of the schema after populating the data is **78.7MB**, the table below shows the number of rows and the number of attributes in each table along with the size of each table in KB.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table | No. of rows | No. of attributes | Size in KB |  |  |
| STOCK\_DAILY\_STATS\_FACT | 619,040 | 15 | 61,440 |  |  |
| DATE\_DIM | 73,050 | 15 | 16,384 |  |  |
| INDEX\_DAILY\_STAT\_FACT | 9,920 | 16 | 2,048 |  |  |
| STOCK\_YEARLY\_STATS\_FACT | 2,020 | 14 | 192 |  | Dimension |
| COMPANY\_DIM | 580 | 8 | 128 |  | Fact |
| INDEX\_CHANGE\_FACT | 264 | 5 | 64 |  |  |
| SUB\_INDUSTRY\_DIM | 124 | 3 | 64 |  |  |
| SECTOR\_DIM | 12 | 2 | 64 |  |  |
| REASON\_DIM | 7 | 2 | 64 |  |  |
| INDEX\_DIM | 5 | 2 | 64 |  |  |
| PORTFOLIO\_DIM | 3 | 2 | 64 |  |  |

## What Dimensions and Why?

* **Date**

This dimension is essential in almost any Data Model. In this data model, this dimension is used to track the performance of stocks and indices over time. The date is used as the primary key in this table. It contains attributes such as Year\_ID and Month\_ID.

* **Company**

This dimension is used to keep information about corporations that are/were components of the S&P500 index. This helps investors take decisions with the companies’ history in mind. It contains attributes such as the date the company is founded and the date it was first added to the Stock Market (IPO Date).

* **Sector**

This dimension allows investors to look at different stock from the point of view of industries, whether certain industries are on the rise while other industries are falling behind.

* **Sub Industry**

This dimension is complementary to the previous dimension, it offers a higher level of detail allowing investors to specifically invest in sub industries that are generally doing better in the stock market.

* **Index**

The index dimension contains the full names of different indices incorporated in the data model. This dimension is used to help investors compare the performance of different indices.

* **Portfolio**

The portfolio dimension supports the product we provide to our customers, portfolios of curated stocks that ensure we can serve different needs. Some portfolios minimize **risk**, others maximize growth rate (**profitability**) and others are diversified for **balance** **between risk and profitability**.

* **Reason**

The reason dimension contains common reasons companies are added and removed from the S&P500 index. Investors should be able to use this information to keep track of the patterns in stock prices of companies around the time they get added/removed from the index and use their understanding with other stocks.

We found this composition to be optimal for the purposes of the data warehouse for the following reasons:

* Four fact tables are included because four different facts are measured, each is related to different dimensions. Each fact table contributes to fulfilling one of the needs of our customers that will be answered later in the BI queries section.
* The Sector, company, reason, date and portfolio needed to be repeated numerous times in more than one fact table, so they were put in separate dimensions so that they can be referenced by shorter keys. This way the queries will not consume long time to be executed.
* The **Index Daily Stats Fact** table contains information about 5 different indices in the New York Stock Exchange market, this should help investors compare different indices to decide which is better for their investment needs.
* The **Index Change Fact** table contains information pertaining to changes in the S&P500 index, which companies are added to the index, which are removed and the reasons of these changes. This table should help investors specifically interested in the S&P500 index understand the nature of these changes and the most occurring reasons they happen.
* The **Stock Yearly Stats Fact** table contains yearly measurements about S&P500 stocks between 2014 and 2017. This should help investors take a bird eye view on different stocks to understand the stocks’ performance over longer periods.
* The **Stock Daily Stats Fact** table contains daily measurements and statistics calculated based on some of the measurements for individual stocks between 2013 and 2018. This should help investors track the performance of specific stocks they are interested in.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Date | Company | Sector | Sub Industry[[3]](#footnote-3) | Index | Index Change Reason | Portfolio |  |  |
| Stock Yearly Stats | **x** | **x** | **x** |  |  |  | **x** |  | Dimension |
| Stock Daily Stats | **x** | **x** | **x** |  |  |  |  |  | Fact |
| Index Daily Stats | **x** |  |  |  | **x** |  |  |  |  |
| Index Change | **x** | **x** |  |  |  | **x** |  |  |  |

Table 2 Bus Matrix showing which dimensions are used in each fact table

## Mapping the Data Sources to the Data Model

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Company Information | Market Change | Stock Daily Statistics | Stock Yearly Statistics | Index Daily Statistics |  |  |
| Stock Yearly Stats |  |  |  | **x** |  |  | Dimension |
| Stock Daily Stats |  |  | **x** |  |  |  | Fact |
| Index Daily Stats |  |  |  |  | **x** |  | Data Source |
| Index Change |  | **x** |  |  |  |  |  |
| Company | **x** | **x** | **x** | **x** |  |  |  |
| Sector | **x** |  |  |  |  |  |  |
| Sub Industry | **x** |  |  |  |  |  |  |
| Index |  |  |  |  | **x** |  |  |
| Index Change Reason |  | **x** |  |  |  |  |  |
| Portfolio[[4]](#footnote-4) |  |  |  |  |  |  |  |
| Date |  |  |  |  |  |  |  |

The **company dimension** was populated from four different data sources because some of the data sources contained companies that were not existent in the others. For example, the **Market Change** data source contained companies that are not components of the S&P500 index anymore.

The **Stock Daily Stats** table were mainly populated from the Stock Yearly Statistics and Stock Daily Statistics data sources respectively, however, missing information about the company such as the sector was populated from the **Company Information** data source.

# Logical Data Map

## Stock\_Yearly\_Stats\_Fact

First, data from Stock Yearly Statistics data source (years 2014 and 2017) without company name column was extracted to a single excel workbook. Then, it was loaded in Stock yearly stats Fact table using Import data feature from toad tool.

Then some **transformations** were made to standardize the format.

* An update Statement replaces SECTOR\_NAME with SECTOR\_KEY
* Fixing MARKET\_CAP AND EBITDA attributes for 2017 by multiplying by one Billion

Second, to cover the missing data in the years 2015 and 2016, we developed two PL/SQL scripts1, 2 to generate pseudorandom data.

## Stock \_Daily\_Stats\_Fact

First, data from the **stock daily statistics** data source is extracted and loaded in Stock\_Daily\_Stats\_fact table using the Import data feature in Toad.

Second, a PL/SQL procedure3 is used to calculate the derived measures that are going to be used to answer the BI queries (MOM, MOM2, EMA\_10, EMA\_20, EMA\_50, EMA\_100, EMA\_200). This procedure takes the measure name as a parameter and then calls the corresponding procedure3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.7 to calculate the required field and insert it into to the table.

Finally, the sector name was replaced with the numeric sector key from the sector dimension using another PL/SQL script4.

Index\_Change\_Fact

First, data from Market Change data source is extracted and loaded in Index\_Change\_Fact table using the import data feature in Toad. Then, the values in company\_added\_sector\_key column is replaced with numeric sector keys from the company dimension before being set as a foreign key to ensure referential integrity5.

Afterwards, the reason is replaced6 with a numeric reason key from the reason dimension and then set as a foreign key.

Index\_Daily\_Stats\_Fact

This data source was already clean and well-structured, it was extracted and loaded directly into the data model using Toad import data feature.

Date\_Dim

The **date dimension** was created and populated using a readymade script [found here](http://oracleolap.blogspot.com/2011/01/script-for-time-dimension-table.html). The script was edited to modify the date range, some columns were removed after creating the table due to their irrelevance7.

Reason\_Dim

The **reason dimension** is created using the following create statement.

CREATE TABLE Reason\_Dim

(

Reason\_key NUMBER(3) primary key,

Reason\_Description VARCHAR2(30)

);

Reasons were extracted from the Market Change csv file, the reasons were standardized in MS Excel smoother analysis later.

The reason dimension was then populated manually since there are **only 7 reasons** extracted from the Market Change data source.

Index\_Dim

The index dimension is created using the following create statement.

CREATE TABLE Index\_dim

(

Index\_key VARCHAR2(100) primary key,

Index\_name VARCHAR2(200)

);

It was also populated manually since there are **only 5 indices** in the data model.

Sub\_Industry\_Dim

To populate the sub industry dimension, a PL/SQL script8 is developed to select unique sub industries related to each sector from the company dimension and add a new row for each sub industry.

Company\_Dim

The company dimension data is extracted from the company information data source csv file and loaded into the data model using Toad’s import data feature. Some companies from other data sources were missing. They were added using two PL/SQL scripts9.1, 9.2.

# Application of Data Warehouse

# Conclusion

# Appendix

[1 PL/SQL Script to fabricate data for the year 2015](ETL/1-%202015%20fabrication.txt)

[2 PL/SQL Script to fabricate data for the year 2016](ETL/2-%202016%20fabrication.txt)

[3 PL/SQL procedure to calculate the derived measures for the daily data](ETL/3-%20calculate%20field%20procedure.txt)

[3.1 PL/SQL procedure to calculate MOM](ETL/3.1-%20calculate%20mom.txt)

[3.2 PL/SQL procedure to calculate MOM2](ETL/3.2-%20calculate%20mom2.txt)

[3.3 PL/SQL procedure to calculate EMA\_10](ETL/3.3-%20calculate%20ema_10.txt)

[3.4 PL/SQL procedure to calculate EMA\_20](ETL/3.4-%20calculate%20ema_20.txt)

[3.6 PL/SQL procedure to calculate EMA\_100](ETL/3.6-%20calculate%20ema_100.txt)

[3.7 PL/SQL procedure to calculate EMA\_200](ETL/3.7-%20calculate%20ema_200.txt)

[4 PL/SQL script to replace sector names with numeric sector keys in stock daily stats](ETL/4-%20sector%20name%20to%20sector%20key%20in%20stock%20daily%20stats%20table.txt)

[5 PL/SQL script to replace sector names with numeric sector keys in index change fact table](ETL/5-%20change%20sector%20name%20to%20sector%20key%20in%20index%20change%20fact.txt)

[6 PL/SQL script to replace reasons with numeric reason keys](ETL/6-%20change%20reason%20name%20to%20reason%20key.txt)

[7 PL/SQL script to create and populate the date dimension table](ETL/7-%20date%20dimension%20creation.txt)

[8 PL/SQL script to populate the sub industry dimension](ETL/8-%20sub%20industry%20dimension%20population.txt)

[9.1 PL/SQL script to add companies from index change fact table that are not existent in company dimension](ETL/9.1-%20Companies%20from%20index%20change%20fact%20table%20that%20are%20not%20existent%20in%20company%20dimension.txt)

[9.2 PL/SQL script to add companies from stock yearly stats fact table that are not existent in company dimension](ETL/9.2-%20Companies%20from%20stock%20yearly%20stats%20fact%20table%20that%20are%20not%20existent%20in%20company%20dimension.txt)

1. By clicking the view history button at the top right corner of a Wikipedia page, we could view the history of the page going back to its first version [↑](#footnote-ref-1)
2. [Hoseinzade, E. (2020). CNNpred: CNN-based stock market prediction using a diverse set of variables. Expert Systems with Applications](https://zenodo.org/record/3634200#.YCfwBmgzaUl) [↑](#footnote-ref-2)
3. Although the Sub Industry dimension is not used in any fact table, it is being used in the Sector dimension. [↑](#footnote-ref-3)
4. We created this dimension and manually populated it, a detailed explanation about its importance is provided in the conclusion section. [↑](#footnote-ref-4)