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 17, 2021

Table of Contents

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.

1.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.

1.1 S&P 500 component stocks

This is the first data source provided in the project statement, it is a Wikipedia page 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.

1.1.1 Company Information

The list of corporations on this page was first **made available on July 21, 2005**[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**. The table consists of 505 rows, each row represents one corporation, and 8 attributes are recorded for each corporation.

[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

Attributes:

- **Symbol**: This is an abbreviation used to uniquely identify publicly traded shares of a particular stock on a particular 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.

Symbol	Security	SEC filings	GICS Sector	GICS Sub-Industry	Headquarters Location	Date first added	CIK	Founded
MMM	3M Company	reports	Industrials	Industrial Conglomerates	St. Paul, Minnesota	1976-08-09	0000066740	1902

Figure 1 A sample of the Company Information table

Note: we found the data in these reports irrelevant to our business case.

- **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.

Classification[1] [edit]

Sector		Industry Group		Industry		Sub-Industry	
10	Energy	1010	Energy	101010	Energy Equipment & Services	10101010	Oil & Gas Drilling
						10101020	Oil & Gas Equipment & Services
				101020	Oil, Gas & Consumable Fuels	10102010	Integrated Oil & Gas
						10102020	Oil & Gas Exploration & Production
						10102030	Oil & Gas Refining & Marketing
						10102040	Oil & Gas Storage & Transportation
						10102050	Coal & Consumable Fuels

Figure 2 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.

1.1.2 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. 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.

Attributes:

- **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

Date	Added		Removed		Reason
	Ticker	Security	Ticker	Security	
January 21, 2021	TRMB	Trimble Inc.	CXO	Concho Resources	S&P 500/100 constituent ConocoPhillips acquired Concho Resources.[6]

Figure 3 A sample of the Market Change table

1.2 Stock Daily Statistics

The second data source we are using in our project is a dataset downloaded from Kaggle, 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. The version we are using in this project was made available on February 10, 2018.

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.

Attributes:

- **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

Note: All prices are in USD.

date	open	high	low	close	volume	Name
08-02-13	15.07	15.12	14.63	14.75	8407500	AAL

Figure 4 A sample of the daily data of S&P500 stocks

1.3 Stock Yearly Statistics

The third data source utilized in this data model is obtained from DataHub. 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 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.

Attributes:

- **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**.

$$\text{EPS} = \frac{\text{Total Earnings}}{\text{Outstanding Shares}}$$

- **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.

$$\text{P/E Ratio} = \frac{\text{Market value per share}}{\text{Earnings per share}}$$

- **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.

$$\textit{Dividend yield} = \textit{annual dividends per share} \div \textit{price per share}$$

- **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.

$$\text{Market cap} = \text{share price} \times \text{no. of shares outstanding}$$

- **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.

$$\text{EBITDA} = \text{Net Income} + \text{Interest} + \text{Taxes} + \text{Depreciation} + \text{Amortization}$$

- **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.

$$\text{P/S Ratio} = \frac{MVS}{SPS}$$

where:

MVS = Market Value per Share

SPS = 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.

$$P/B\ Ratio = \frac{Market\ Price\ per\ Share}{Book\ Value\ per\ Share}$$

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.

$$\text{Statistic in 2015} = random\ value\ between(\text{statistic in 2014}, \frac{statistic\ in\ 2014 + statistic\ in\ 2017}{2})$$

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.

$$\text{Statistic in 2016} = random\ value\ between(\frac{statistic\ in\ 2014 + statistic\ in\ 2017}{2}, \text{statistic in 2017})$$

1.4 Index Daily Statistics

This data source[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.

Attributes:

- **MOM, MOM1, MOM2, MOM3:** The momentum which is also referred to as the **Rate of Return**, is the net gain or loss of an

$$\text{Rate of return} = \left[\frac{(\text{Current value} - \text{Initial value})}{\text{Initial value}}\right]$$

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.

[2] Hoseinzade, E. (2020). CNNpred: CNN-based stock market prediction using a diverse set of variables. Expert Systems with Applications

- **ROC_5, ROC_10, ROC_15, and 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, and 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)

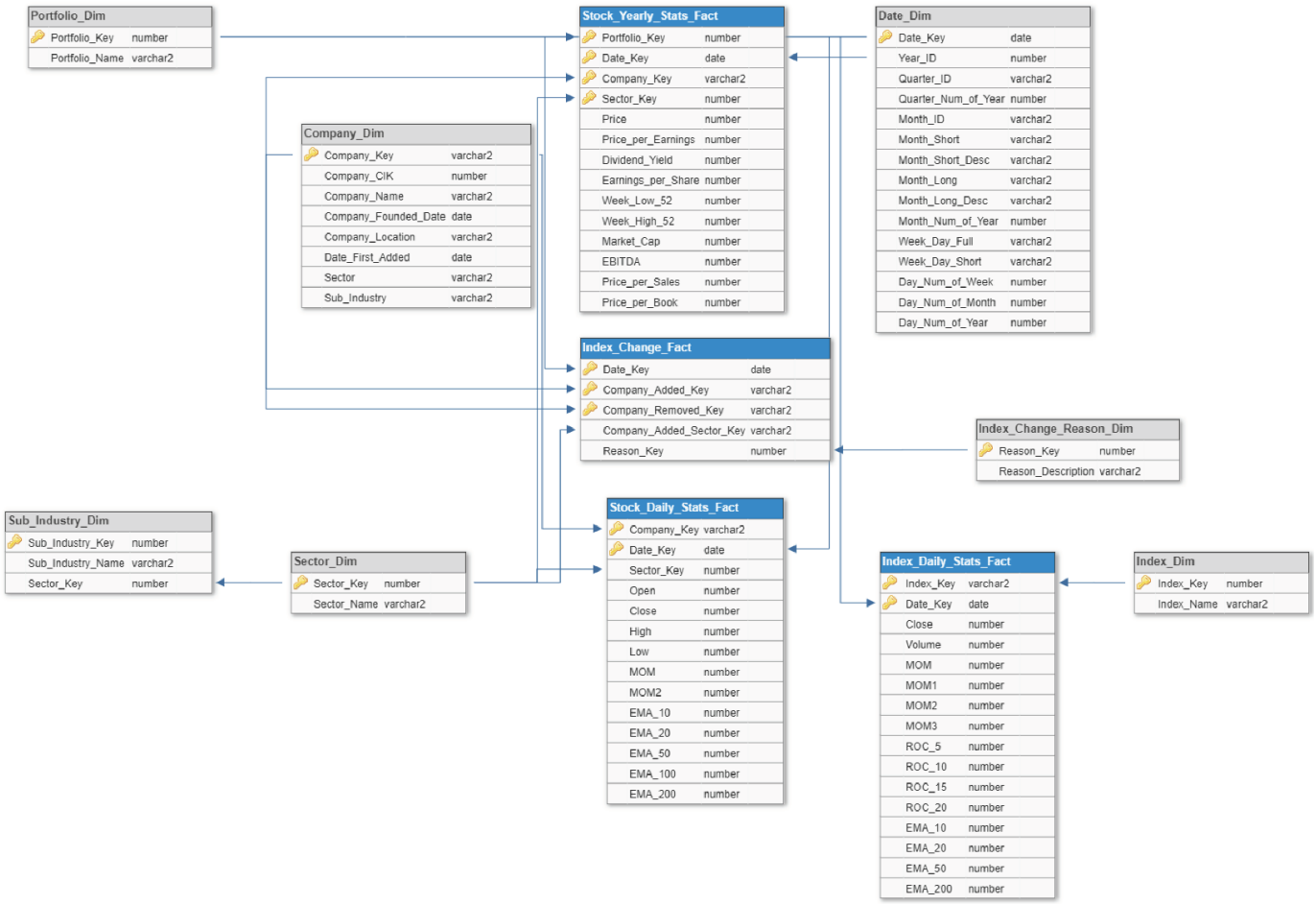
Multiplier = [2 / (Time Periods +1)]

Initial EMA: Simple moving Average

2.Data Warehouse Data Model

2.1 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
COMPANY_DIM	580	8	128
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

Dimension
Fact

2.2 Dimensional Model Structure

- **Business process:** Stock market analysis and portfolio recommendation
- **Granularity:** Granularity level is one day
- **Dimension:** Sector_Dim,Sub_Industry_Dim,Company_Dim, Date_Dim, Index_Change_Reason_Dim, Index_Dim , Portfolio_Dim
- **Fact:** Stock_Yearly_Stats_Fact, Index_Change_Fact, Stock_Daily_Stats_Fact , Index_Daily_Stats_Fact

2.3 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.

2.4 Facts

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.

- **Index Daily Stats Fact**: 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.
- **Index Change Fact:** 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.
- **Stock Yearly Stats Fact**: 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.
- **Stock Daily Stats Fact:** 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]	Index	Index Change Reason	Portfolio
Stock Yearly Stats	x	x	x				x
Stock Daily Stats	x	x	x				
Index Daily Stats	x				x		
Index Change	x	x				x	

Dimension
Fact

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

[3] Although the Sub Industry dimension is not used in any fact table, it is being used in the Sector dimension.

2.5 Data Sources and Data Model relationship matrix

	Company Information	Market Change	Stock Daily Statistics	Stock Yearly Statistics	Index Daily Statistics
Stock Yearly Stats				x	
Stock Daily Stats			x		
Index Daily Stats					x
Index Change		x			
Company	x	x	x	x	
Sector	x				
Sub Industry	x				
Index					x
Index Change Reason		x			
Portfolio[4]					
Date					

Dimension
Fact
Data Source

Table 3 Relationship Matrix showing the relationships between Data sources and Dimensional model

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.

[4] We created this dimension and manually populated it, a detailed explanation about its importance is provided in the conclusion section.

3.Logical Data Map

In order to populate the data correctly on the data warehouse to be suitable for data analysis and queries, a substantial transformations and cleaning have been conducted on different tables as below:

3.1 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 scripts[1, 2] to generate pseudorandom data.

3.2 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 procedure[3] 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, and EMA_200). This procedure takes the measure name as a parameter and then calls the corresponding procedure[3.1, 3.2, 3.3, 3.4, 3.5, 3.6, and 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 script[4].

3.3 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 integrity[5].

Afterwards, the reason is replaced[6] with a numeric reason key from the reason dimension and then set as a foreign key.

3.4 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.

3.5 Date_Dim

The date dimension was created and populated using a readymade script found here. The script was edited to modify the date range, some columns were removed after creating the table due to their irrelevance[7].

3.6 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.

3.7 Index_Dim

The index dimension is created using the following create statement.

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

```
CREATE TABLE Index_dim
( Index_key   VARCHAR2(100) primary key,
  Index_name  VARCHAR2(200));
```

3.8 Sub_Industry_Dim

To populate the sub industry dimension, a PL/SQL script[8] is developed to select unique sub industries related to each sector from the company dimension and add a new row for each sub industry.

3.9 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 scripts[9.1, 9.2].

4. Application of Data Warehouse

After finishing data transformation and loading. The data has been used to extract valuable insights for all business users using SQL and Tableau.

Queries have been divided into sections as below according to business user's different interests:

4.1 Indexes

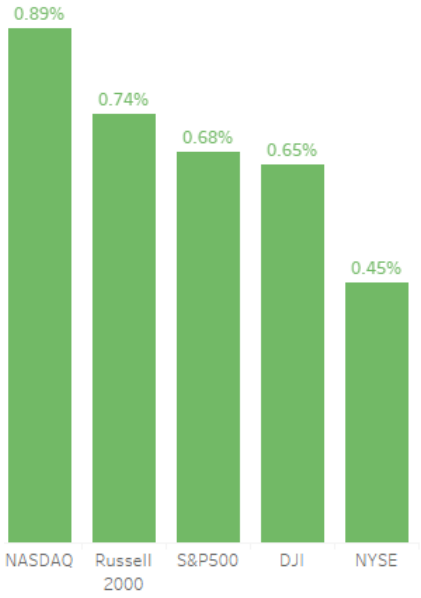
4.1.1 Question 1

Which index has the highest short term growth rate?

Data Source: Index Daily Statistics

Note: Short term Growth Rate measures are (MOM, MOM1, ROC_5, ROC_10, and ROC_15)

INDEX_KEY	AVERAGE_GROWTH_RATE
NASDAQ	0.890409025886745
Russell 2000	0.741983989842052
S&P500	0.67632940898324
DJI	0.65381873679228
NYSE	0.45069724971356

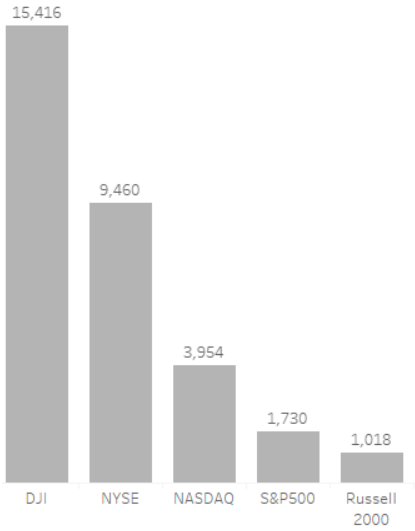


4.1.2 Question 2

Which index has high Long term growth Rate?
Data Source: Index Daily Statistics
Note: Long term Growth Rate measures are (EMA_50, EMA_100)

```
Select index_key, avg (EMA_50) as Average_Growth_LongTerm
From index_daily_stat_fact
Group by index_key
Order by avg (EMA_50) desc;
```

INDEX_KEY	AVERAGE_GROWTH_LONGTERM
DJI	15415.5439821344
NYSE	9459.53218406098
NASDAQ	3953.86126371886
S&P500	1729.5165546522
Russell 2000	1017.73641298832

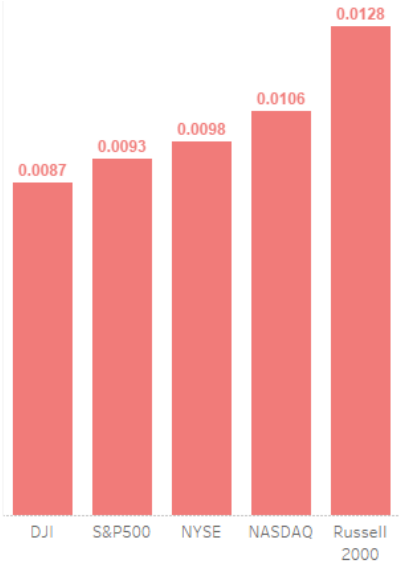


4.1.3 Question 3

Which Index has the lowest risk or is more stable?
Data Source: Index Daily Statistics

```
Select index_key, stddev (MOM) as Risk
From index_daily_stat_fact
Group by index_key
Order by  stddev (MOM);
```

INDEX_KEY	RISK
DJI	0.008680851258453
S&P500	0.00933485875597608
NYSE	0.00977189713019065
NASDAQ	0.0105637827675357
Russell 2000	0.0127665858477924



4.2 Growth Stocks

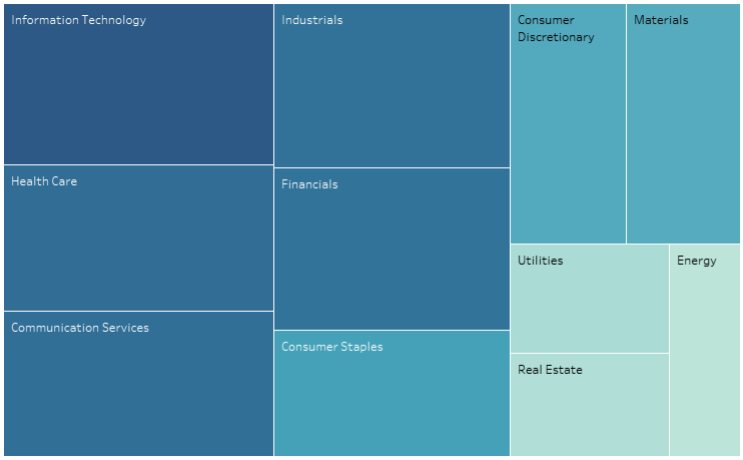
4.2.1 Question 1

Which Sector has highest growth rate?

Data Source: Stock Daily Statistics with our added Calculated Fields and S&P 500 component stocks

```
Select sector_name, trunc (avg (MOM), 6) as Average_Growth_Rate
From stock_daily_stats_fact, sector_dim
Where sector_dim.sector_key = STOCK_DAILY_STATS_FACT.SECTOR_KEY
Group by stock_daily_stats_fact.sector_key, sector_name
Order by avg (MOM) desc;
```

SECTOR_NAME	ROUND(AVG(MOM),5)
Information Technology	0.00077
Health Care	0.00069
Industrials	0.00064
Communication Services	0.00063
Financials	0.00062
Consumer Discretionary	0.0005
Consumer Staples	0.00047
Materials	0.00044
Utilities	0.0003
Real Estate	0.00024
Energy	0.00018



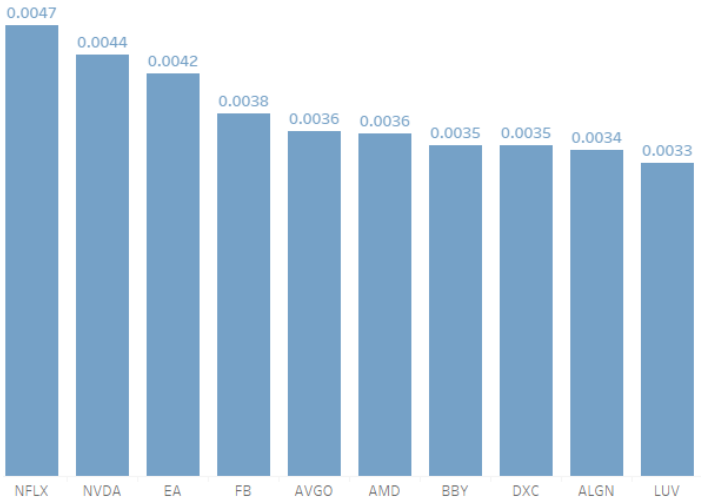
4.2.2 Question 2

What are the top 10 stocks according to Growth Rate performance?

Data Source: Stock Daily Statistics with our added Calculated Fields

```
Select *
From
(Select STOCK_DAILY_STATS_FACT.COMPANY_KEY, round (avg (MOM2), 5) as
Average_Growth_Rate
From stock_daily_stats_fact
Group by stock_daily_stats_fact.COMPANY_KEY
Order by avg (MOM2) desc)
Where rownum <= 10;
```

COMPANY_KEY	AVERAGE_GROWTH_RATE
NVDA	0.0051
NFLX	0.00449
AMD	0.00372
ALGN	0.00357
EA	0.0035
MU	0.00349
AVGO	0.00344
APTV	0.00343



4.3 Dividend Stocks

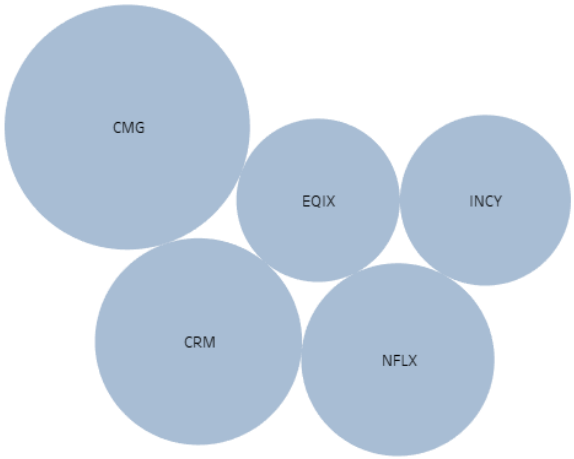
4.3.1 Question 1

Which Company has highest average dividend yield?

Data Source: Stock Yearly Statistics

```
Select  *
From
(Select company_key,  round (avg (STOCK_YEARLY_STATS_FACT.DIVIDEND_YIELD),  5) as
company_avg
From stock_yearly_stats_fact
Group by company_key
Having avg (STOCK_YEARLY_STATS_FACT.DIVIDEND_YIELD)  is not null
Order by company_avg desc)
Where rownum <= 5;
```

COMPANY_KEY	COMPANY_AVG
CMG	239.26679
CRM	170.46692
NFLX	148.31629
INCY	116.10284
EQIX	106.30865



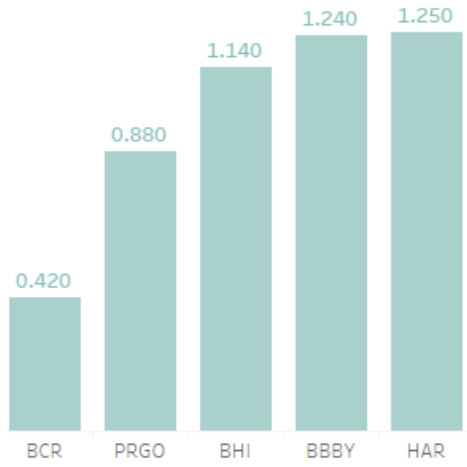
4.3.2 Question 2

Which company has the lowest Price per Earnings?

Data Source: Stock Yearly Statistics

```
Select  *
From
(Select company_key,  avg (STOCK_YEARLY_STATS_FACT.PRICE_PER_EARNINGS)  as
company_avg
 From stock_yearly_stats_fact
 Group by company_key
 Having  avg (STOCK_YEARLY_STATS_FACT.PRICE_PER_EARNINGS)  is not null and avg
(STOCK_YEARLY_STATS_FACT.PRICE_PER_EARNINGS)  > 0
 Order by company_avg  asc)
Where rownum  <= 5;
```

COMPANY_KEY	COMPANY_AVG
BCR	0.42
PRGO	0.88
BHI	1.14
BBBY	1.24
HAR	1.25



4.4 Index Change

4.4.1 Question 1

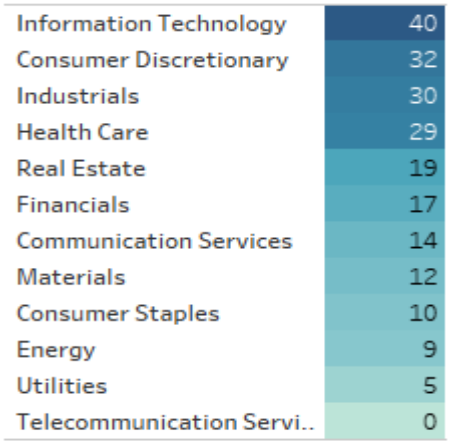
Which industry has the highest change Rate in S&P500 (either adding or removing)?

Data source: S&P 500 component stocks and Market Change

Adding:

```
Select  SECTOR_DIM.SECTOR_NAME,  count (ICF.COMPANY_ADDED_sector_KEY)
From index_change_fact  ICF, sector_dim
Where company_added_sector_key  is not null
And
SECTOR_DIM.SECTOR_KEY  = ICF.company_added_sector_key
Group by SECTOR_DIM.SECTOR_NAME
Order by count (ICF.COMPANY_ADDED_sector_KEY)  desc;
```

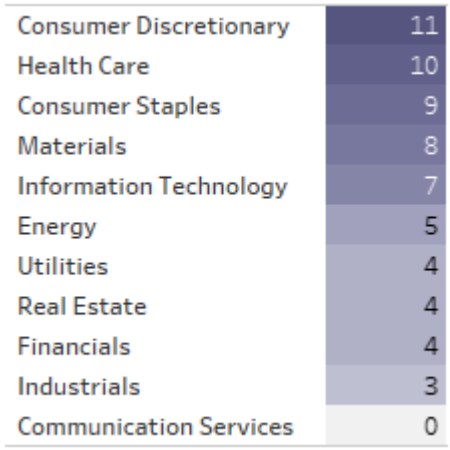
SECTOR_NAME	COMPANIES_COUNT
Information Technology	40
Consumer Discretionary	32
Industrials	30
Health Care	29
Real Estate	19
Financials	17
Communication Services	14
Materials	12
Consumer Staples	10
Energy	9
Utilities	5



Removing:

```
Select CD.SECTOR, count (CD.SECTOR) as Companies_count
From index_change_fact  icf , COMPANY_DIM  CD
Where CD.company_key  = ICF.COMPANY_REMOVED_KEY
And CD.SECTOR is not null
Group by CD.SECTOR
Order by count (CD.SECTOR) desc;
```

SECTOR	COMPANIES_COUNT
Consumer Discretionary	11
Health Care	10
Consumer Staples	9
Materials	8
Information Technology	7
Energy	5
Financials	4
Real Estate	4
Utilities	4
Industrials	3



4.4.2 Question 2

What is the most common reason for index change?
Data Source: Market Change

```
Select REASON_DESCRIPTION,  count (REASON_DESCRIPTION)
From index_change_fact  ICF, reason_dim
Where ICF.REASON_KEY  = REASON_DIM.REASON_KEY
Group by REASON_DESCRIPTION
Order by count (REASON_DESCRIPTION)  desc;
```

REASON_DESCRIPTION	COUNT(REASON_DESCRIPTION)
Acquisition	122
Market Capitalization Change	74
Spin-off	28
Other	13
Merger	6
Stock Taken Private	6
Bankruptcy	2

5. Conclusion

To sum up, stock market consists of several indexes, each one of them includes major companies based on many criteria such as market capitalization. The data warehouse model collects historical information about all the details of the stock market using different data sources which have been transformed in order to be ready for the designed dimensional model.

Then the dimensional model has been populated to answer most of the business questions related to the stock market starting with choosing indexes, comparing them with each other, choosing the best stock based on several measurements, filtering and recommend the best stock portfolio according to the return rate and the request of the investor.

Moreover, the model keeps tracking for all the daily changes that happen to stocks and the rates of increase and decrease in their daily prices.

Finally all analyses have been presented using tableau to visualize the answers of all business questions and to be used as recommendation tool for investors.

Appendix