Table of Contents

1. Introduction

1.1. Introduction about the Project

The Standard and Poor's 500 or S&P 500 is the most famous financial benchmark in the world.

This stock market index tracks the performance of 500 large companies listed on stock exchanges in the United States. As of December 31, 2020, more than $5.4 trillion was invested in assets tied to the performance of this index.

1.2. Business Requirements

This project seeks to analyse the data of the S&P 500 stock market and its firms using a variety of KPIs in order to derive insights that could aid investors in better understanding the market and identifying profitable investment opportunities.

Analysis of stocks will be useful for new investors to invest in stock market based on the various KPIs like Market capitalization, earning per share, and stock price etc. considered by dashboards.

2. Data Sources

2.1. Wikipedia List of S&P 500 companies

Link

Date of creation: Unknown.

Descriptions:

The data source comprises 503 common stocks which are issued by 500 large-cap companies traded on American stock exchanges (including the 30 companies that compose the Dow Jones Industrial Average).

The data shows the symbol, sector, sub-industry, Date first added

ETL:

The data was extracted using https://wikitable2csv.ggor.de/ website and saved as a CSV file.

We dropped the unnecessary columns like (SEC fillings).

We had 2 columns having multi-values rows like (Date first added, Founded) we separated each of them to two columns based on the space delimiter and then we dropped the two second columns.

2.2. Datahub's S&P 500 Companies with Financial Information

Link

Date of creation: 4 years ago.

Description:

List of companies in the S&P 500 (Standard and Poor's 500). The S&P 500 is a free-float, capitalization-weighted index of the top 500 publicly listed stocks in the US (top 500 by market cap). The dataset includes a list of all the stocks contained therein and associated key financials such as price, market capitalization, earnings, price/earnings ratio, price to book etc.

Notes: Market Capitalization and EBIDTA are in Billions.

ETL:

The data was downloaded as a CSV file. No transformation was required.

2.3. Kaggle S&P 500 Stocks

Link

Date of creation: Unknown.

Description:

The data consists of 3 tables (**sp500_companies.csv, sp500_index.csv, sp500_stocks.csv).**
The data is daily updated from 2009 till now.

ETL:

The data was downloaded as a CSV file. The data from 2009 till 2017 was dropped.

3. Data Warehouse Data Model

3.1. Why?

We chose "Galaxy schema" because we have two fact tables linked to four dimensions tables with different granularity levels.

Advantages:

1. Its multidimensional nature helps in structuring complex Database systems efficiently.
2. Minimum or no redundancy, because of Normalization.
3. This is a flexible Schema, considering the complexity of the system.
4. Data Quality will be fine, as Normalization provides the advantage for well-defined tables/ data formats.
5. When queried with Joins, clear & accurate data can be extracted.
6. High Data quality & accuracy helps in creating exceptional Reporting & Analytical results.

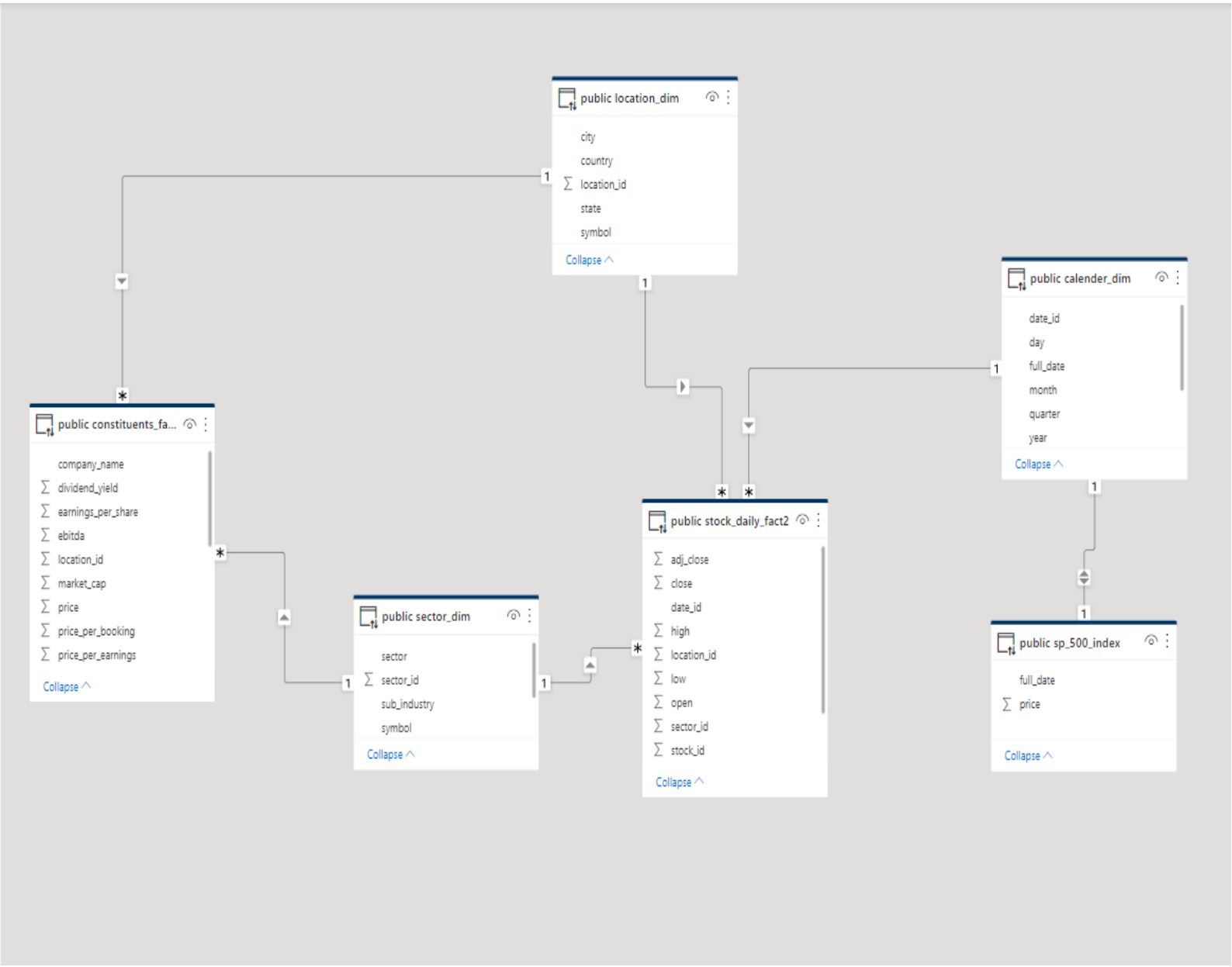
Disadvantages:

1. Galaxy schema can be Complex in structure.
2. Working on this schema is tedious, as the complexity in both Schema and database system makes it more intricate all together.
3. Data retrieval is done with multi-level joins combined with conditional expressions.
4. The number of levels of normalization is expected, depending on the depth of the given database.
5. Maintenance and support tasks get difficult as Galaxy schema is applied for larger database systems with complex structures.

6. Large storage space is required for its larger design arrangement and detailed querying process.
7. The analysis gets difficult, as it has no limitation on how many fact and dimension tables it can have.

3.2. Dimensional Model

1. Calendar dimension table: It contains: (Date_ID, Full date, Month, Quarter, Year) columns.
2. Sector dimension table: It contains: (Symbol, Sector_ID, Sector, Sub-sector) columns.
3. Location dimension table: It contains: (Symbol, Location_ID, Country, State, City) columns.
4. Company dimension table: It contains: (Stock_ID, Symbol, Name, Founded, Date_First_Added, Cik).
5. Stock_Daily_Fact fact table: It contains: (Open, Low, High, Close, Adj_Close, Stock_ID, Sector_ID, Location_ID, Date_ID, Symbol_Location, Symbol_sector, Volume) columns.
6. constituents_Fact fact table: It contains: (Symbol_Location, Company_Name, Sector, Price, Price_Per_Earnings, Dividend_Yield, Earning_Per_Share, Week_high_52, Week_low_52, Market_cap, Ebitda, Price_to_sales, Price_per_booking, Year, Stock_ID, Sector_ID, Location_ID, Symbol_Sector) columns.
7. Sp_500_index_dimension table: It contains: (Full_Date, Price) columns.


public location_dim
city
country
location_id
state
symbol
Collapse
public calender_dim
date_id
day
full_date
month
quarter
year
Collapse
public constituents_fa...
company_name
dividend_yield
earnings_per_share
ebitda
location_id
market_cap
price
price_per_booking
price_per_earnings
Collapse
public stock_daily_fact2
adj_close
close
date_id
high
location_id
low
open
sector_id
stock_id
Collapse
public sector_dim
sector
sector_id
sub_industry
symbol
Collapse
public sp_500_index
full_date
price
Collapse

4. Logical Data Mapping

Source Table Name	Column	Data Type	PK	Table Type	Data Source	Transformation	Target Table Name
Generated	Date_ID	int	(Y)	Dimension	Kaggle	None	Calendar dimension
Sp_500_index	Full date	Date	(N)	Dimension	Kaggle	None	Calendar dimension
Sp_500_index	Month	int	(N)	Dimension	Kaggle	Extracted From Full Date	Calendar dimension
Sp_500_index	Quarter	int	(N)	Dimension	Kaggle	Extracted From Full Date	Calendar dimension
Sp_500_index	Year	int	(N)	Dimension	Kaggle	Extracted From Full Date	Calendar dimension
Generated	Sector_ID	int	(Y)	Dimension	Wiki	None	Sector dimension
Sp_500_wiki	Symbol	string	(Y)	Dimension	Wiki	None	Sector dimension
Sp_500_wiki	Sector	string	(N)	Dimension	Wiki	None	Sector dimension
Sp_500_wiki	Sub-sector	string	(N)	Dimension	Wiki	None	Sector dimension
Sp_500_wiki & Sp_500_Comp	Symbol	string	(Y)	Dimension	Wiki & Kaggle	None	Location dimension
Sp_500_wiki & Sp_500_Comp	Location_ID	int	(Y)	Dimension	Wiki & Kaggle	None	Location dimension
Sp_500_wiki & Sp_500_Comp	Country	string	(N)	Dimension	Wiki & Kaggle	None	Location dimension
Sp_500_wiki & Sp_500_Comp	State	string	(N)	Dimension	Wiki & Kaggle	Extracted from head quarter	Location dimension

Sp_500_wiki & Sp_500_Comp	City	String	(N)	Dimension	Wiki & Kaggle	Extracted from head quarter	Location dimension
Sp_500_index	Full_Date	Date	(Y)	Dimension	Kaggle	None	Sp_500_index
Sp_500_index	Price	float	(N)	Dimension	Kaggle	None	Sp_500_index
Sp_500_wiki	Stock_ID	int	(Y)	Dimension	Wiki	None	Company dimension
Sp_500_wiki	Symbol	string	(Y)	Dimension	Wiki	None	Company dimension
Sp_500_wiki	Name	string	(N)	Dimension	Wiki	None	Company dimension
Sp_500_wiki	Founded	int	(N)	Dimension	Wiki	None	Company dimension
Sp_500_wiki	Date_First_Added	Date	(N)	Dimension	Wiki	None	Company dimension
Sp_500_wiki	Cik	int	(N)	Dimension	Wiki	None	Company dimension
Sp_500_stocks	Open	float	(N)	Fact	Kaggle	None	Stock_Daily_Fact
Sp_500_stocks	Low	float	(N)	Fact	Kaggle	None	Stock_Daily_Fact
Sp_500_stocks	High	float	(N)	Fact	Kaggle	None	Stock_Daily_Fact
Sp_500_stocks	Close	float	(N)	Fact	Kaggle	None	Stock_Daily_Fact
Sp_500_stocks	Adj_Close	float	(N)	Fact	Kaggle	None	Stock_Daily_Fact
Sp_500_stocks	Stock_ID	int	(N)	Fact	Kaggle	None	Stock_Daily_Fact
Sp_500_stocks	Sector_ID	int	(N)	Fact	Kaggle	None	Stock_Daily_Fact
Sp_500_stocks	Location_ID	int	(N)	Fact	Kaggle	None	Stock_Daily_Fact
Sp_500_stocks	Date_ID	Date	(N)	Fact	Kaggle	None	Stock_Daily_Fact
Sp_500_stocks	Symbol_Location	string	(N)	Fact	Kaggle	None	Stock_Daily_Fact
Sp_500_stocks	Symbol_sector	string	(N)	Fact	Kaggle	None	Stock_Daily_Fact
Sp_500_stocks	Volume	int	(N)	Fact	Kaggle	None	Stock_Daily_Fact

constituents	Symbol_Location	string	(N)	Fact	Datahub	None	constituents_Fact
constituents	Company_Name	string	(N)	Fact	Datahub	None	constituents_Fact
constituents	Sector	string	(N)	Fact	Datahub	None	constituents_Fact
constituents	Price	float	(N)	Fact	Datahub	None	constituents_Fact
constituents	Price_Per_Earnings	float	(N)	Fact	Datahub	None	constituents_Fact
constituents	Dividend_Yield	float	(N)	Fact	Datahub	None	constituents_Fact
constituents	Earning_Per_Share	float	(N)		Datahub	None	
constituents	Week_high_52	float	(N)	Fact	Datahub	None	constituents_Fact
constituents	Week_low_52	float	(N)	Fact	Datahub	None	constituents_Fact
constituents	Market_cap	float	(N)	Fact	Datahub	None	constituents_Fact
constituents	Ebitda	float	(N)	Fact	Datahub	None	constituents_Fact
constituents	Price_to_sales	float	(N)	Fact	Datahub	None	constituents_Fact
constituents	Price_per_booking	float	(N)	Fact	Datahub	None	constituents_Fact
constituents	Year	int	(N)	Fact	Datahub	None	constituents_Fact
constituents	Stock_ID	int	(N)	Fact	Datahub	None	constituents_Fact
constituents	Sector_ID	int	(N)	Fact	Datahub	None	constituents_Fact
constituents	Location_ID	int	(N)	Fact	Datahub	None	constituents_Fact
constituents	Symbol_Sector	string	(N)	Fact	Datahub	None	constituents_Fact

5. Queries

```
    SELECT symbol, company_name, start_of_year_price, percent_change
    FROM ytd_stock_change, constituents_fact
    WHERE symbol = symbol_location
    ORDER BY percent_change ASC
    LIMIT 10
)
```

Output | Explain | Messages | Notifications

symbol character varying (10)	company_name character varying (100)	start_of_year_price double precision	percent_change numeric
NFLX	Netflix Inc.	597.3699951171875	-70.73
PYPL	PayPal	194.94000244140625	-64.17
ALGN	Align Technology	648.0499877929688	-63.48
CCL	Carnival Corp.	21.40999984741211	-59.60
RCL	Royal Caribbean Cruises Ltd	80.83000183105469	-56.81
ILMN	Illumina Inc	380.8699951171875	-51.60
NCLH	Norwegian Cruise Line	22.18000030517578	-49.86
NVDA	Nvidia Corporation	301.2099914550781	-49.67
AMD	Advanced Micro Devices Inc	150.24000549316406	-49.10
F	Ford Motor	21.770000457763672	-48.87

```
    SELECT *,
       ROUND(CAST(((current_price - start_of_year_price) / start_of_year_price) *100 As numeric), 2) as percent_change
       FROM (
       SELECT distinct symbol,
       FIRST_VALUE(close) OVER(PARTITION BY symbol
                               ORDER BY (stock_date)
                                ROWS BETWEEN UNBOUNDED PRECEDING AND UNBOUNDED FOLLOWING) As start_of_year_price,
       LAST_Value(close) OVER(PARTITION BY symbol
                               ORDER BY (stock_date)
                                ROWS BETWEEN UNBOUNDED PRECEDING AND UNBOUNDED FOLLOWING) As current_price
        FROM sp_500_stocks
        where extract(year from stock_date) = '2022'
    ) as test
ORDER BY symbol
)
```

Output | Explain | Messages | Notifications

symbol character varying (10)	start_of_year_price double precision	current_price double precision	percent_change numeric
A	156.47999572753906	118.7699966430664	-24.10
AAL	18.75	12.680000305175781	-32.37
AAP	236.77999877929688	173.08999633789062	-26.90
AAPL	182.00999450683594	136.72000122070312	-24.88
ABBV	135.4199981689453	153.16000366210938	13.10
ABC	132.6199951171875	141.47999572753906	6.68
ABMD	366.2900085449219	247.50999450683594	-32.43

```
    select *,
           extract (year from stock_date) as year,
           extract (quarter from stock_date) as quarter,
           TO_CHAR(TO_DATE(extract (month from stock_date)::text, 'MM'), 'Month') as month,
           extract (day from stock_date) as day
    from sp_500_stocks
    WHERE SYMBOL = 'AAPL'
    and extract ( year from stock_date) >='2019'
)
```

Output | Explain | Messages | Notifications

stock_date [PK] date	symbol [PK] character varying (10)	adj_close double precision	close double precision	high double precision	low double precision	open double precision	volume double precision
2019-01-02	AAPL	38.439735412597656	39.47999954223633	39.712501525878906	38.557498931884766	38.72249984741211	148158800
2019-01-03	AAPL	34.6108512878418	35.54750061035156	36.43000030517578	35.5	35.994998931884766	365248800
2019-01-04	AAPL	36.08836364746094	37.064998626708984	37.13750076293945	35.95000076293945	36.13249969482422	234428400
2019-01-07	AAPL	36.00804138183594	36.98249816894531	37.20750045776367	36.474998474121094	37.17499923706055	219111200
2019-01-08	AAPL	36.69446563720703	37.6875	37.95500183105469	37.130001068115234	37.38999938964844	164101200
2019-01-09	AAPL	37.31760025024414	38.32749938964844	38.63249969482422	37.407501220703125	37.8224983215332	180396400
2019-01-10	AAPL	37.43687438964844	38.45000076293945	38.49250030517578	37.71500015258789	38.125	143122800
2019-01-11	AAPL	37.06931686401367	38.0724983215332	38.42499923706055	37.877498626708984	38.220001220703125	108092800
2019-01-14	AAPL	36.511905670166016	37.5	37.817501068115234	37.30500030517578	37.712501525878906	129756800
2019-01-15	AAPL	37.259185791015625	38.26750183105469	38.34749984741211	37.51250076293945	37.567501068115234	114843600
2019-01-16	AAPL	37.71426200914453	38.725006103515625	38.970001220703125	38.25	38.27000045776367	122278800

```
    SELECT sector_dim.sector, sub_industry, company_name,
       SUM(market_cap) as Market_Cap
    FROM constituents_fact, sector_dim
    WHERE symbol_sector = symbol
    And symbol_location = symbol
    GROUP BY sector_dim.sector, sub_industry, company_name
    ORDER BY Market_Cap DESC
    LIMIT 10
)

-- 4. Lowest Market Capitalization in S&P 500 index
```

Output | Explain | Messages | Notifications

sector character varying (100)	sub_industry character varying (100)	company_name character varying (100)	market_cap double precision
Information Technology	Technology Hardware, Storage & Peripherals	Apple Inc.	809508034020
Communication Services	Interactive Media & Services	Alphabet Inc Class A	733823966137
Communication Services	Interactive Media & Services	Alphabet Inc Class C	728535558140
Information Technology	Systems Software	Microsoft Corp.	689978437468
Consumer Discretionary	Internet & Direct Marketing Retail	Amazon.com Inc	685873374731
Financials	Diversified Banks	JPMorgan Chase & Co.	386613611000
Health Care	Pharmaceuticals	Johnson & Johnson	353062464971
Energy	Integrated Oil & Gas	Exxon Mobil Corp.	326148660000
Financials	Diversified Banks	Bank of America Corp	321478200969
Consumer Staples	Hypermarkets & Super Centers	Wal-Mart Stores	304680931618

```
    SELECT sector_dim.sector, sub_industry, company_name,
       SUM(market_cap) as Market_Cap
    FROM constituents_fact, sector_dim
    WHERE symbol_sector = symbol
    And symbol_location = symbol
    GROUP BY sector_dim.sector, sub_industry, company_name
    ORDER BY Market_Cap ASC
    LIMIT 10
)
```

Output | Explain | Messages | Notifications

sector character varying (100)	sub_industry character varying (100)	company_name character varying (100)	market_cap double precision
Financials	Multi-line Insurance	Assurant Inc	4653993594
Industrials	Construction & Engineering	Quanta Services Inc.	5330131216
Consumer Discretionary	Apparel, Accessories & Luxury Goods	Under Armour Class C	5366628950
Consumer Staples	Distillers & Vintners	Brown-Forman Corp.	5498033502
Consumer Discretionary	Apparel, Accessories & Luxury Goods	Under Armour Class A	5856913571
Real Estate	Retail REITs	Kimco Realty	6180487499
Utilities	Independent Power Producers & Energy Traders	AES Corp	6920851212
Industrials	Human Resource & Employment Services	Robert Half International	7047165475
Industrials	Building Products	Allegion	7599609494
Consumer Discretionary	Restaurants	Chipotle Mexican Grill	76852

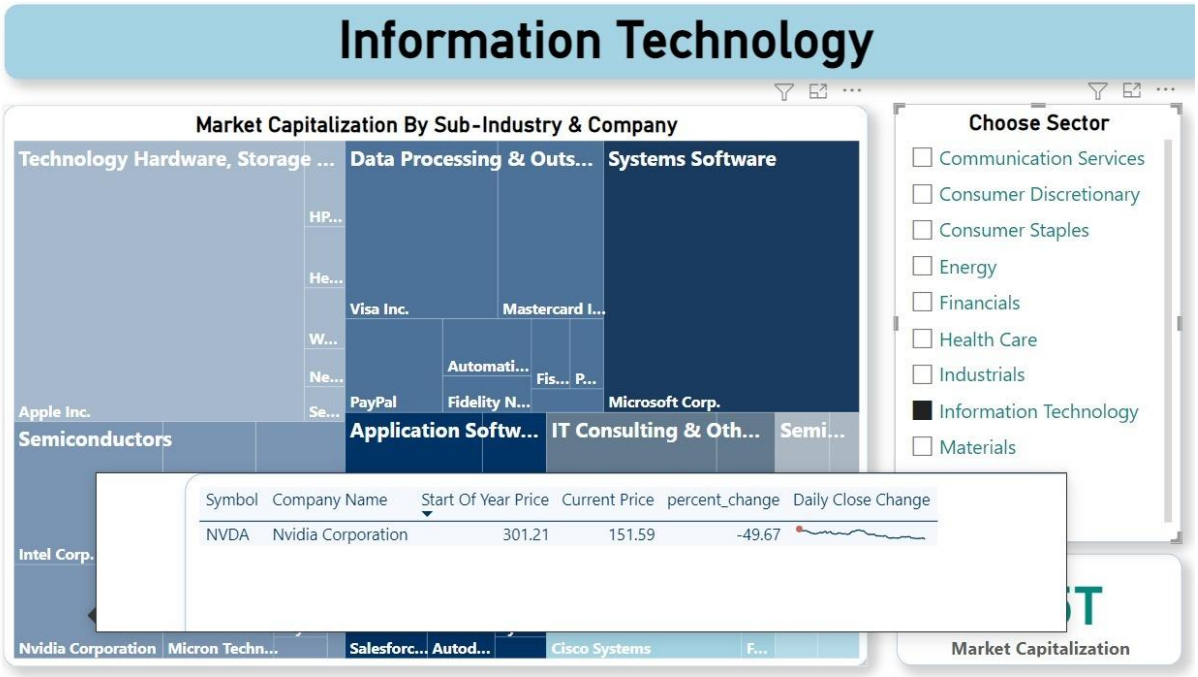
```
    SELECT symbol, company_name, start_of_year_price, percent_change
    FROM ytd_stock_change, constituents_fact
    WHERE symbol = symbol_location
    ORDER BY percent_change DESC
    LIMIT 10
)
```

Output | Explain | Messages | Notifications

symbol character varying (10)	company_name character varying (100)	start_of_year_price double precision	percent_change numeric
OXY	Occidental Petroleum	31.059999465942383	89.57
HES	Hess Corporation	76.79000091552734	37.96
VLO	Valero Energy	77.13999938964844	37.78
XOM	Exxon Mobil Corp.	63.540000915527344	34.78
MRO	Marathon Oil Corp.	16.8700008392334	33.25
MCK	McKesson Corp.	248.10000610351562	31.48
HAL	Halliburton Co.	23.989999771118164	30.72
VRTX	Vertex Pharmaceuticals Inc	222.52999877929688	26.63
MPC	Marathon Petroleum	65.66000366210938	25.21
BMY	Bristol-Myers Squibb	61.880001068115234	24.43

6. Visualization

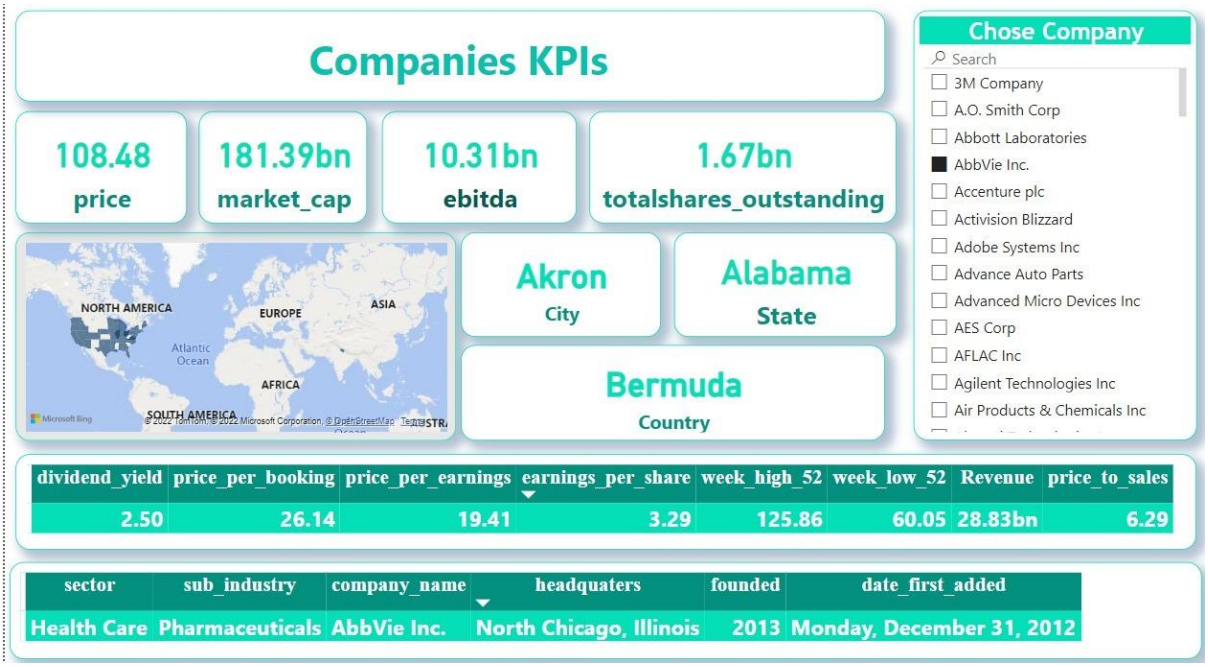
This report page shows the S&P 500 Index's different stocks' daily behaviour in the past 5 years with options to choose the date or the desired companies (stocks). It also shows the most recent price of each stock and the year-to-date change percentage per stock.



This report page shows the different sectors/ sub-industries/ companies of

the stocks and their market capitalization which is visualized through a tree map that highlights the top performing stocks in the S&P 500 index per Sub-Industry and per companies. The investor can select the sector he is interested investing in it and the tree map is changed interactively with his choice showing the top performing stocks. The tree map also contains a very informative tooltip that shows the stock price's behaviour in 2022 in terms of change percent from the beginning of the year till now, and a sparkline showing the trend of the stock price during the year which gives the investor insights

about whether it is worth investing in this stock or not.



This report page has more advanced KPI's targeting more expert investors who understands the stock market terms (Ex. Dividend Yield, Book Per Price,) helping them making the right choice of investment. The dashboard allows the investor to choose or search for a specific company to see

it's KPI's and helping him make the right decision. It also shows information about the company (Location - Headquarters - Date Founded...).

Top And Worst Performers In 2022

Top 10 Performers

Symbol	Company Name	Percent Change
OXY	Occidental Petroleum	89.57
HES	Hess Corporation	37.96
VLO	Valero Energy	37.78
XOM	Exxon Mobil Corp.	34.78
MRO	Marathon Oil Corp.	33.25
MCK	McKesson Corp.	31.48
HAL	Halliburton Co.	30.72
VRTX	Vertex Pharmaceuticals Inc	26.63
MPC	Marathon Petroleum	25.21
BMY	Bristol-Myers Squibb	24.43

Worst 10 Performers

Symbol	Company Name	Percent Change
NFLX	Netflix Inc.	-70.73
PYPL	PayPal	-64.17
ALGN	Align Technology	-63.48
CCL	Carnival Corp.	-59.60
RCL	Royal Caribbean Cruises Ltd	-56.81
ILMN	Illumina Inc	-51.60
NCLH	Norwegian Cruise Line	-49.86
NVDA	Nvidia Corporation	-49.67
AMD	Advanced Micro Devices Inc	-49.10
F	Ford Motor	-48.87

7. Conclusion

As a summary for our work, we defined the business needs, then we collected the datasets from different sources to help us find meaningful solutions to these needs.

We loaded the data into Postgresql RDMS, cleaned it and integrated it to build our model.

We executed non-trivial SQL queries to get the data of interest.

We connected Microsoft power Bi to our database to visualize our data through representative dashboards.

Future work: Index price prediction using RNN in machine learning.