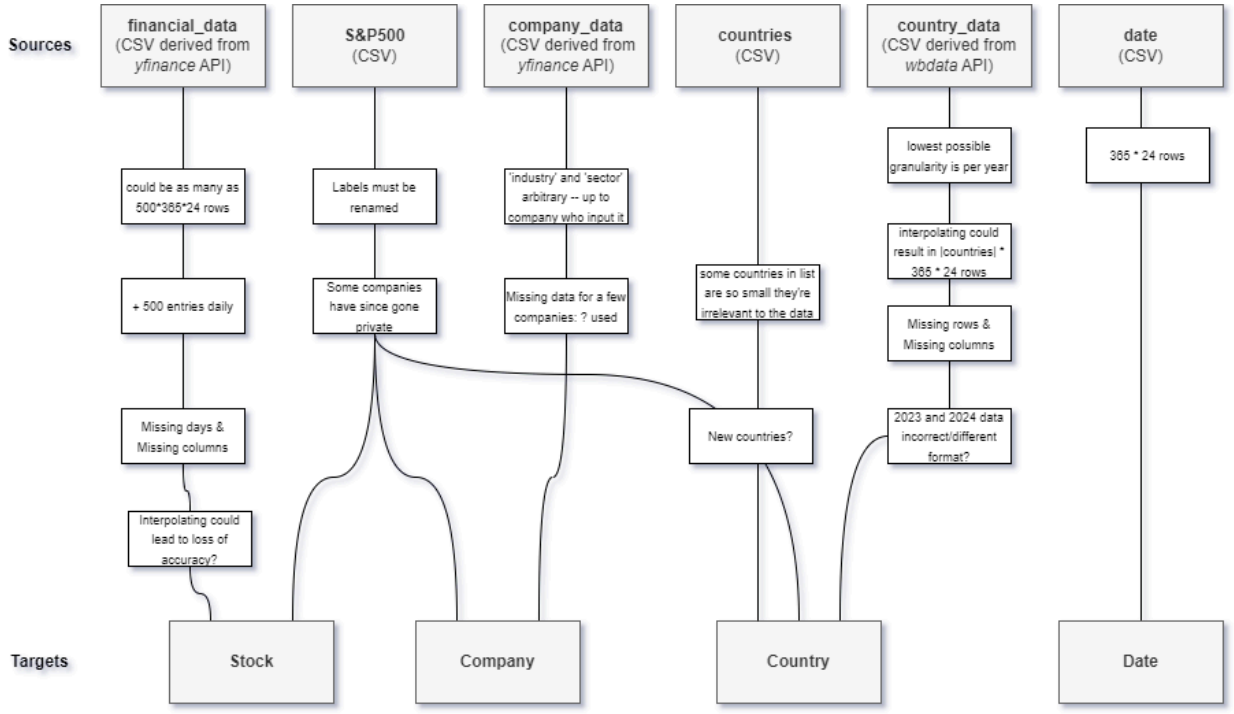
CSI 4142: Phase 2 Deliverable

Physical Design and Data Staging

https://github.com/c-stev/CSI-4142-Project

Cole Stevens (300171413)
William Beaupre (300174392)
Emiliano Bustamante (300229811)

High-Level Data Staging Plan



Additional Project Details

Removed Dimensions

The initially proposed `Exchange Rate` dimension was removed from the project due to the insufficient number of potential analytical queries that could be performed that involved `Exchange Rate` that related to the initial goal of the research project.

Altered Dimensions

- ➔ Attribute `industry` of dimension `Company` was removed because of its arbitrary nature – they will be of no use when organizing data.
- ➔ A large collection of attributes was proposed for the `Financial Data` dimension (e.g., `dividend_rate`, `dividend_yield`, `52_week_low`, `52_week_high`, etc) were removed because either they are aggregate values that could be calculated later, or they are unable to be obtained despite the group's initial assumptions (the current data is accessible,

but not historical). Attributes `ticker`, `date`, `open`, `close`, `high`, `low`, `volume`, `dividends`, and `stock_splits` remain because of their relevance to the research.

➔ Since stock data is collected only on weekdays, additional columns `Day_String` and `Weekday` were added to the `Date` dimension. `Weekday` returns `true/false` depending on if a particular day is when the stock markets are open, which will help with querying.

Added Measures

Alongside the volatility measure, a new measure, *returns* was created to represent the numeric change in a stock's price in a 24 hour period. This measure is calculated by subtracting the current day's closing price by yesterday's closing price.

Removed Measures

Since the `Exchange Rate` dimension is being removed (as mentioned above), the measure that requires data from the Exchange Rate dimension, `inflation_adjusted_price_usd` was removed as well.

Data Quality Issues

Country Data

One of the source datasets, `countries.csv`, contained many countries that were so small that the World Bank repository does not track data on it (e.g., Antarctica, Vatican City, etc.). Since the overall percentage of countries that the World Bank did not have data on was so small, it was decided to remove these country CSV rows manually to prevent any impacts null data would have on the end result data. The country codes that were removed are as follows: `AX`, `AI`, `AQ`, `BQ`, `BV`, `IO`, `CX`, `CC`, `CK`, `FK`, `GF`, `TF`, `GP`, `GG`, `HM`, `VA`, `JE`, `MQ`, `YT`, `MS`, `NU`, `NF`, `PN`, `RE`, `BL`, `SH`, `PM`, `GS`, `SJ`, `TW`, `TK`, `UM`, `WF`, and `EH`.

Another large issue with the World Bank data is that they collect data yearly, not every day – which is the granularity that this project is using. Since it is impractical to assume that another organization exists that collects accurate, daily census data for almost every country in the world, a decision was made to expand the data to 365 entries per year, instead of 1, and then interpolate the missing days' data using built-in functions provided by the python `pandas` library. A column was

added called ‘Interpolated’ that indicates whether the data was sourced directly from the World Bank, or if it was interpolated by the staging python code. Since interpolation results in float numbers with a large number of decimal points, a decision was made to round floats to a maximum of 3 digits.

Financial_Data

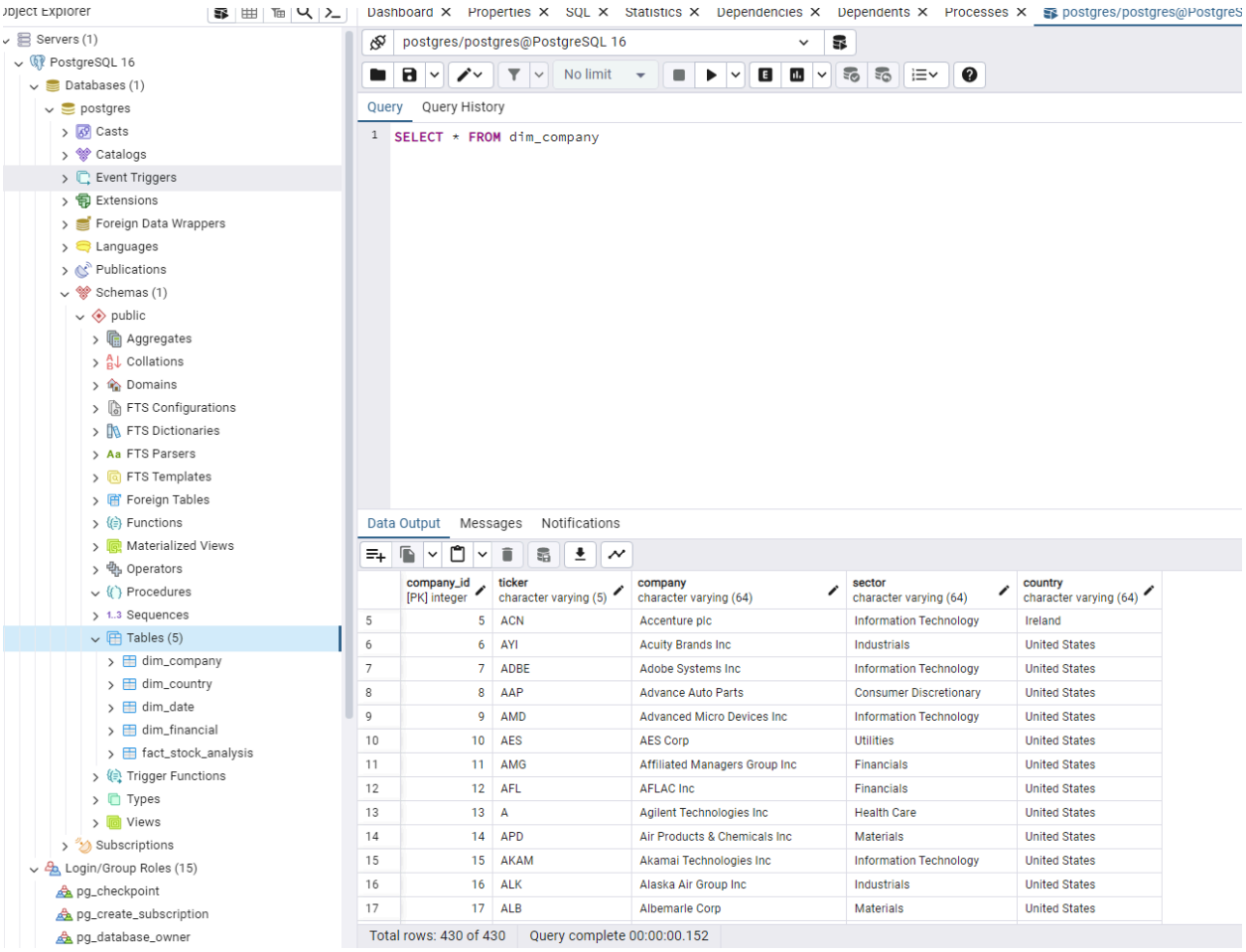
The initial dataset for the financial data contained two extra columns: Dividends, and Stock_Splits. It was noticed that these pieces of data were always marked as 0.0 for every company, on every day. Since 100% of the data was unusable, it was decided to remove these columns as they would serve no purpose in further analysis.

Company Data

The Company dimension was created as a combination of values from the S&P 500 CSV list, which displays the company ticker, company name, and the sector of each company, and company data derived from Yahoo! Finance. When combining data, it was noticed that certain entries from the ‘Country’ column were marked as missing. This indicates that Yahoo! Finance does not contain data pertaining to these companies, presumably because they have become unlisted from the market. Since a small percentage of these companies were missing, a decision was made to just remove these rows from the dataframe.

PostgreSQL Database Screenshots

Company Dimension



Country Dimension

Data Output | Messages | Notifications

	country_id [PK] integer	date date	code character varying (4)	country character varying (64)	population integer	gdp double precision	inflation double precision	employment double precision	unemployment double precision	interpolated boolean
1	1	2000-01-01	US	United States	282162411	10250947997000	3.377	63.297	3.99	false
2	2	2000-01-02	US	United States	282170079	10251852318795.082	3.375	63.295	3.992	true
3	3	2000-01-03	US	United States	282177747	10252756640590.164	3.374	63.293	3.994	true
4	4	2000-01-04	US	United States	282185415	10253660962385.246	3.372	63.291	3.996	true
5	5	2000-01-05	US	United States	282193083	10254565284180.328	3.371	63.288	3.998	true
6	6	2000-01-06	US	United States	282200751	10255469605975.41	3.369	63.286	4	true
7	7	2000-01-07	US	United States	282208419	10256373927770.492	3.368	63.284	4.002	true
8	8	2000-01-08	US	United States	282216088	10257278249565.574	3.366	63.282	4.004	true
9	9	2000-01-09	US	United States	282223756	10258182571360.656	3.365	63.28	4.006	true
10	10	2000-01-10	US	United States	282231424	10259086893155.738	3.363	63.278	4.008	true
11	11	2000-01-11	US	United States	282239092	10259991214950.82	3.362	63.276	4.01	true
12	12	2000-01-12	US	United States	282246760	10260895536745.902	3.36	63.274	4.012	true
13	13	2000-01-13	US	United States	282254428	10261799858540.984	3.359	63.271	4.014	true

Total rows: 1000 of 32148 | Query complete 00:00:00.100

Date Dimension

Data Output | Messages | Notifications

	date_id [PK] integer	date date	year integer	quarter integer	month integer	month_string character varying (32)	day integer	day_string character varying (32)	weekday boolean
1	1	2000-01-01	2000	1	1	January	1	Saturday	false
2	2	2000-01-02	2000	1	1	January	2	Sunday	false
3	3	2000-01-03	2000	1	1	January	3	Monday	true
4	4	2000-01-04	2000	1	1	January	4	Tuesday	true
5	5	2000-01-05	2000	1	1	January	5	Wednesday	true
6	6	2000-01-06	2000	1	1	January	6	Thursday	true
7	7	2000-01-07	2000	1	1	January	7	Friday	true
8	8	2000-01-08	2000	1	1	January	8	Saturday	false
9	9	2000-01-09	2000	1	1	January	9	Sunday	false
10	10	2000-01-10	2000	1	1	January	10	Monday	true
11	11	2000-01-11	2000	1	1	January	11	Tuesday	true
12	12	2000-01-12	2000	1	1	January	12	Wednesday	true
13	13	2000-01-13	2000	1	1	January	13	Thursday	true

Total rows: 1000 of 8037 | Query complete 00:00:00.092

Financial Dimension

Data Output | Messages | Notifications

	financial_data_id [PK] integer	ticker character varying (5)	date date	open double precision	close double precision	high double precision	low double precision	volume bigint
1	1	MMM	2000-01-03	24.419517229092836	23.99054718017578	24.530731686219475	23.91110828222818	2173400
2	2	MMM	2000-01-04	23.609264573722047	23.037303924560547	24.10178624383334	23.037303924560547	2713800
3	3	MMM	2000-01-05	23.164399228850257	23.7045841217041	24.467198088086004	23.164399228850257	3699400
4	4	MMM	2000-01-06	23.97467232223776	25.611114501953125	26.055972570225265	23.97467232223776	5975800
5	5	MMM	2000-01-07	25.70645276879452	26.119535446166992	26.38962796598745	25.404584658406943	4101200
6	6	MMM	2000-01-10	25.531674361753577	25.992420196533203	26.31017594465708	25.42045984991022	3863800
7	7	MMM	2000-01-11	25.611120442252847	25.547569274902344	26.05597861370637	25.547569274902344	2357600
8	8	MMM	2000-01-12	25.91298247685208	25.611114501953125	26.34195275697164	25.61111450195313	2868400
9	9	MMM	2000-01-13	25.754104595326314	25.611114501953125	25.897094688699504	25.51578777303767	2244400
10	10	MMM	2000-01-14	25.611120042746887	25.24570083618164	25.65878341751627	25.15037408664288	2541800
11	11	MMM	2000-01-18	24.72140227505599	24.689626693725582	25.229811576342488	24.689626693725582	2114800
12	12	MMM	2000-01-19	24.848502702610467	25.118595123291016	25.404575333423363	24.848502702610467	2628200
13	13	MMM	2000-01-20	25.182138065784265	24.18120765686035	25.261576987127437	24.02232981417401	3104600

Financial Analysis fact table

Data Output | Messages | Notifications

	date_id [PK] integer	company_id [PK] integer	country_id [PK] integer	financial_data_id [PK] integer	returns double precision	volatility double precision
1	3	1	3	1	0.4289700489170549	0.015364489643610571
2	4	1	4	2	0.5719606491615004	0.027128036604769888
3	5	1	5	3	-0.5401848928538442	0.03286077700235525
4	6	1	6	4	-1.6364421797153668	0.04999615696959172
5	7	1	7	5	-0.41308267737247206	0.022846187066601395
6	10	1	10	6	-0.46074583477962605	0.020660162262206492
7	11	1	11	7	0.06355116735050359	0.011834101220057932
8	12	1	12	8	0.3018679748989541	0.0168976378555768
9	13	1	13	9	0.14299009337318935	0.008908364145310965
10	14	1	14	10	0.36541920656524596	0.012019136947614629
11	18	1	18	11	0.031775581330407476	0.01299801273660161
12	19	1	19	12	-0.2700924206805482	0.013291495495966772
13	20	1	20	13	1.0009304089239137	0.030208672080294003

Group Contributions

Deliverable checklist	Responsible team member(s)	Expected completion date	Actual completion date	Estimated time (hours) to complete	Actual time (hours) to complete	Notes (if any)
Create database instance	William	2024/03/19	2024/03/20	1.5	1.5	
Create Date dimension	William	2024/03/20	2024/03/21	0.5	0.5	
Create Country dimension	William	2024/03/20	2024/03/21	0.5	1	
Create Company dimension	William	2024/03/20	2024/03/21	0.5	0.5	
Create Stock dimension	William	2024/03/20	2024/03/21	0.5	0.5	
Staging of dimension Date	Cole	2024/03/18	2024/03/19	1	1	includes extracting & staging
Staging of dimension Country	Cole	2024/03/18	2024/03/19	1	3	includes extracting & staging
Staging of dimension Company	Cole	2024/03/08	2024/03/19	1	2	includes extracting & staging
Staging of dimension Stock	Cole	2024/03/08	2024/03/19	2	2	includes extracting & staging
Surrogate key pipeline	Cole	2024/03/21	2024/03/21	0.5	1	
Staging of fact table – including FKs and measures	Cole & William	2024/03/22	2024/03/24	1	2	
Data quality handling and reporting	Cole	2024/03/19	2024/03/19	1	1	
Others – if any						

Higher quality version attached alongside submission.