Project Phase 2: Physical Design and Data Staging

CSI4142 - Fundamentals of Data Science

Winter 2023



Faculté de génie Faculty of Engineering

School of Electrical Engineering and Computer Science University of Ottawa

Professor Yazan Otoum

Group 10:

Mazharul Maaz - 300128179 Bill Battushig - 300109257 Xiao Meng Li - 300109886

Due date: Mar 24th, 2023

Table of Contents

Table of Contents	1
A. Schematic of High-Level Data Staging Plan	2
B. Additional Details	3
C. Data Quality Issues	3
DBMS Screenshots	4

A. Schematic of High-Level Data Staging Plan

- 1. Create and preprocess dimensions
 - a. Date Dimension
 - 1. Generate date range
 - 2. Create a DataFrame with dates
 - 3. Convert dates to string format
 - b. Country Dimension
 - 1. Load data, remove duplicate rows, check and convert data types and rename columns for consistency
 - 2. Remove non-alphanumeric values from the Country column for merging
 - c. World Economic Indicator Dimension
 - 1. Load data, remove duplicate and Null rows, check and convert data types and rename columns for consistency
 - 2. Convert columns to appropriate data types
 - d. Government Response Dimension
 - 1. Load data, remove duplicate rows, check and convert data types and rename columns for consistency
 - 2. Drop rows without valid vaccination info
 - 3. Keep only the latest vaccination info for each country
 - 4. Remove non-alphanumeric values from the Country column for merging
- 3. Create and preprocess fact tables
 - a. Covid-19 Fact
 - Load data, remove duplicate rows, check and convert data types and rename columns for consistency
 - b. Emissions Fact
 - 1. Load data, remove duplicate rows, check and convert data types and rename columns for consistency
- 4. Merge dimensions and fact tables to create a final fact table
 - a. Merge Dimensions and Facts into one table
- 5. Data normalization/scaling
 - a. Scale World GDP to millions
 - b. Scale Area from square miles to square kilometers
 - c. Calculate emission changes
 - d. Changing emissions so that changes are highlighted
- 6. Feature engineering
 - a. Calculate the percentage of the population affected by Covid-19
 - b. Ensuring the percentage caps at 100
- 7. Data transformation
 - a. Normalize Emission Changes
 - b. Normalize Covid-19 data
- 8. Final steps
 - a. Make columns snake case
 - b. Replace NaN values with "N/A"
 - c. Generate a surrogate key

- d. Reorder columns
- e. Export the final dataset as a CSV file

B. Additional Details

- 1. We used Github to version control the source data sets.
- 2. All the columns that end with n are normalized data columns.
- 3. We added texts in the Jupyter Notebook explaining the purpose of code blocks.

C. Data Quality Issues

- 1. Handling missing or noisy data
 - The fillna("N/A") function was used to replace NaN values in the final dataset with "N/A".
 - The data preprocessing includes dropping rows with missing values in columns such as total_deaths, new_deaths, new_cases, total_cases, average_CO2, average_CO4, and average_N2O.
- 2. Integrating data from different sources
 - The data for the project was gathered from multiple sources, such as CSV files containing country data, COVID-19 data, world economic indicator data, and environment data.
 - The data from these different sources were merged using the pd.merge() function to combine the relevant columns based on common keys like 'Country' and 'Year'.
- 3. Checking for duplicates and handling them
 - The duplicated().sum function was used to check for duplicate rows in the dataset, and the drop_duplicates() function was used to remove any duplicates found.
 - For the government response data, only the latest vaccination information for each country was retained using the drop_duplicates(subset="Country", keep="last", inplace=True) function.
- 4. Data cleaning and transformation
 - Data cleaning included removing non-alphanumeric characters from the 'Country' column using the str.replace() function in order to merge the columns later.
 - The 'Net migration' and 'Literacy (%)' columns were converted to float values by replacing commas with dots and casting the data type accordingly.
 - Data normalization and scaling were performed on several columns, such as emissions changes, COVID-19 cases, and vaccination information, using min-max normalization.
- 5. Feature engineering
 - New features were created, such as the percentage of the population affected by COVID-19, using the existing columns in the dataset.
- 6. Checking data types and converting them
 - The dtypes function was used to check the data types of each column, and the appropriate data types were assigned using the astype() function.

- 7. Renaming columns and reordering the dataset
 - Columns were renamed for clarity and to follow the snake_case naming convention.
 - The dataset columns were reordered, and a surrogate key column 'id' was added.

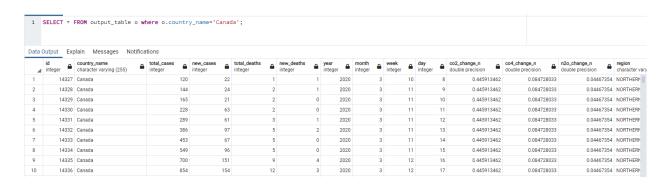
By implementing these steps, the data quality issues were effectively addressed, and the data from different sources were integrated into a single, clean dataset.

DBMS Screenshots

Uploading csv file to the dbms:

```
1 CREATE TABLE output_table (
2
        iд
                                      int,
3
       Country_name
                                     varchar(255),
       Total_cases
5
       New_cases
                                      int,
       Total_deaths
                                      int,
6
7
       New_deaths
                                      int.
8
       Year
                                      int,
9
       Month
                                      int,
10
       Week
                                      int,
11
       Day
                                      int,
12
       Co2_change_n
                                    float,
13
       Co4 change n
                                    float.
       N2o_change_n
14
                                    float.
                                     varchar(255).
15
       Region
16
       Population
                                     int,
17
                                    float,
18
       Net_migration_rate
19
       Gdp_per_capita
                                    float,
       Literacy rate
                                    float,
20
       Num_vaccinated_final
21
                                    float.
       Num_fully_vaccinated_final float,
22
23
       Stringency_index
24
       Unemployment_rate
25
       World_gdp_millions
26
       Covid_case_percent_n
                                    float,
27
       Covid death percent n
                                    float.
       Vaccinated_final_percent_n
                                   float.
28
29
       Full_vaccinated_final_percent_n float
30 );
32 COPY output_table FROM 'C:\Users\Public\Output.csv' csv header;
```

Select all rows where country is Canada:



Select the row where the highest percentage of the population contracted Covid-19:

