

# **Project Phase 2: Physical Design and Data Staging**

CSI4142 - Fundamentals of Data Science

Winter 2023



uOttawa

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

<b>Table of Contents</b>	<b>1</b>
<b>A. Schematic of High-Level Data Staging Plan</b>	<b>2</b>
<b>B. Additional Details</b>	<b>3</b>
<b>C. Data Quality Issues</b>	<b>3</b>
<b>D. DBMS Screenshots</b>	<b>4</b>
<b>E. Team Contributions</b>	<b>5</b>

## 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
    1. 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.

## D. DBMS Screenshots

Uploading csv file to the dbms:

```
1 CREATE TABLE output_table (  
2     id int,  
3     Country_name varchar(255),  
4     Total_cases int,  
5     New_cases int,  
6     Total_deaths int,  
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,  
14    N2o_change_n float,  
15    Region varchar(255),  
16    Population int,  
17    Area_sq_km float,  
18    Net_migration_rate float,  
19    Gdp_per_capita float,  
20    Literacy_rate float,  
21    Num_vaccinated_final float,  
22    Num_fully_vaccinated_final float,  
23    Stringency_index float,  
24    Unemployment_rate float,  
25    World_gdp_millions float,  
26    Covid_case_percent_n float,  
27    Covid_death_percent_n float,  
28    Vaccinated_final_percent_n float,  
29    Full_vaccinated_final_percent_n float  
30 );  
31  
32 COPY output_table FROM 'C:\Users\Public\Output.csv' csv header;
```

Select all rows where country is Canada:

```
1 SELECT * FROM output_table o where o.country_name='Canada';
```

	id	country_name	total_cases	new_cases	total_deaths	new_deaths	year	month	week	day	co2_change_n	co4_change_n	n2o_change_n	region
	integer	character varying (255)	integer	integer	integer	integer	integer	integer	integer	integer	double precision	double precision	double precision	character vary
1	14327	Canada	120	22	1	1	2020	3	10	8	0.445913462	0.084728033	0.04467354	NORTHERN
2	14328	Canada	144	24	2	1	2020	3	11	9	0.445913462	0.084728033	0.04467354	NORTHERN
3	14329	Canada	165	21	2	0	2020	3	11	10	0.445913462	0.084728033	0.04467354	NORTHERN
4	14330	Canada	228	63	2	0	2020	3	11	11	0.445913462	0.084728033	0.04467354	NORTHERN
5	14331	Canada	289	61	3	1	2020	3	11	12	0.445913462	0.084728033	0.04467354	NORTHERN
6	14332	Canada	386	97	5	2	2020	3	11	13	0.445913462	0.084728033	0.04467354	NORTHERN
7	14333	Canada	453	67	5	0	2020	3	11	14	0.445913462	0.084728033	0.04467354	NORTHERN
8	14334	Canada	549	96	5	0	2020	3	11	15	0.445913462	0.084728033	0.04467354	NORTHERN
9	14335	Canada	700	151	9	4	2020	3	12	16	0.445913462	0.084728033	0.04467354	NORTHERN
10	14336	Canada	854	154	12	3	2020	3	12	17	0.445913462	0.084728033	0.04467354	NORTHERN

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

1
SELECT \* FROM output\_table o where o.covid\_case\_percent\_n=1;

Data Output Explain Messages Notifications															
	id	country_name	total_cases	new_cases	total_deaths	new_deaths	year	month	week	day	co2_change_n	co4_change_n	n2o_change_n	region	
	integer	character varying (255)	integer	integer	integer	integer	integer	integer	integer	integer	double precision	double precision	double precision	character varyin	
1	7908	Bahrain	525171	1517	1458	0	2022	3	9	5	1	0.982217573		1	NEAR EAST

## E. Team Contributions

CSIA142 - Project W23						
Phase 2- Physical design and data staging						
Teamwork - breakdown of duties						
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	Xiao Meng Li	22-Mar	22-Mar	0.5	0.5	
Create Date dimension	Xiao Meng Li	22-Mar	22-Mar	0.5	0.5	
Create Country dimension	Xiao Meng Li	22-Mar	22-Mar	0.5	0.5	
Create World Economic Indicator dimension	Xiao Meng Li	22-Mar	22-Mar	0.5	0.5	
Create Government Response dimension	Xiao Meng Li	22-Mar	22-Mar	0.5	0.5	
Create Covid 19 Fact	Xiao Meng Li	22-Mar	22-Mar	2		
Staging of Date dimension	Bill Battushig	23-Mar	23-Mar	0.5	0.5	
Staging of Country dimension	Bill Battushig	23-Mar	23-Mar	0.5	0.5	
Staging of World Economic Indicator dimension	Bill Battushig	23-Mar	23-Mar	0.5	0.5	
Staging of Government Response dimension	Bill Battushig	23-Mar	23-Mar	0.5	0.5	
Staging of Covid 19 Fact	Bill Battushig	23-Mar	23-Mar	0.5	0.5	
Surrogate key pipeline	Mazharul Maaz	24-Mar	24-Mar	2	2	
Feature Engineering	Mazharul Maaz	24-Mar	24-Mar	1	1	
Data Transformation	Mazharul Maaz	24-Mar	24-Mar	1	1	
Data Normalization	Mazharul Maaz	24-Mar	24-Mar	1	1	
Staging of fact table – including FKs and measures	Mazharul Maaz	24-Mar	24-Mar	2	2	
Data quality handling and reporting	Xiao Meng Li	24-Mar	24-Mar	1	1	
Others – if any						