Project Phase 1: Conceptual Design – Dimensional Model

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: Feb 12th, 2023

Table of Contents

Grain

The grain of the conceptual model is: “The correlation between daily COVID-19 cases of a country and greenhouse emissions, supported by relevant economic and demographic characteristics of each country”.

Dimensions and Dimensional Attributes

1. Country

- **Country_key:** Integer, min = 1, max = 227, sample = 1
- **Response_Key:** (Foreign Key to Daily Government Responses Dimension)
- **Country_Name:** String, sample = "Canada"
- **Region:** String, sample = "North America"
- **Area:** Integer, min = 2, max = 17075200 sample = 9147593
- **Population:** Integer, min = 7026, max = 1313973713 sample = 331002651
- **GDP_per_capita:** min = 500, max = 55100 sample = 32700
- **Net_Migration:** min = -20,99, max = 23,06 sample = 7.75
- **Literacy:** min = 17,6, max = 100 sample = 99.9

2. Date

- **Date_Key:** Date, min = 01-01-2019, max = 12-31-2022, sample = 01-01-2022
- **Date:** min = 01-01-2019, max = 12-31-2022, sample = 01-01-2022
- **Year:** Integer, min = 2019, max = 2022, sample = 2022
- **Month:** Integer, min = 1, max = 12, sample = 1
- **Week:** Integer, min = 1, max = 52, sample = 1

3. Government Response (Outrig)

- **Response_Key:** Integer, min = 1, max = unlimited, sample = 1
- **Date:** min = 01-01-2019, max = 12-31-2022, sample = 01-01-2022
- **Num_vaccinated:** Integer, min = 0, max = 8000000000, sample = 4
- **Num_fully_vaccinated:** Integer, min = 0, max = 8000000000, sample = 4
- **Stringency_Index:** Real, min = 0, max = 100, sample = 86.4

4. World Economic Indicator

- **Economics_Key:** Integer, min = 1, max = 3000, sample = 2022
- **World_GDP:** Real, min = 0, max = unlimited, sample =

1.39228E+12

- **Year:** Integer, min = 2019, max = 3000, sample = 2022
- **Unemployment_rate:** Real, min = 0, max = 100, sample = 86.4

Fact Table

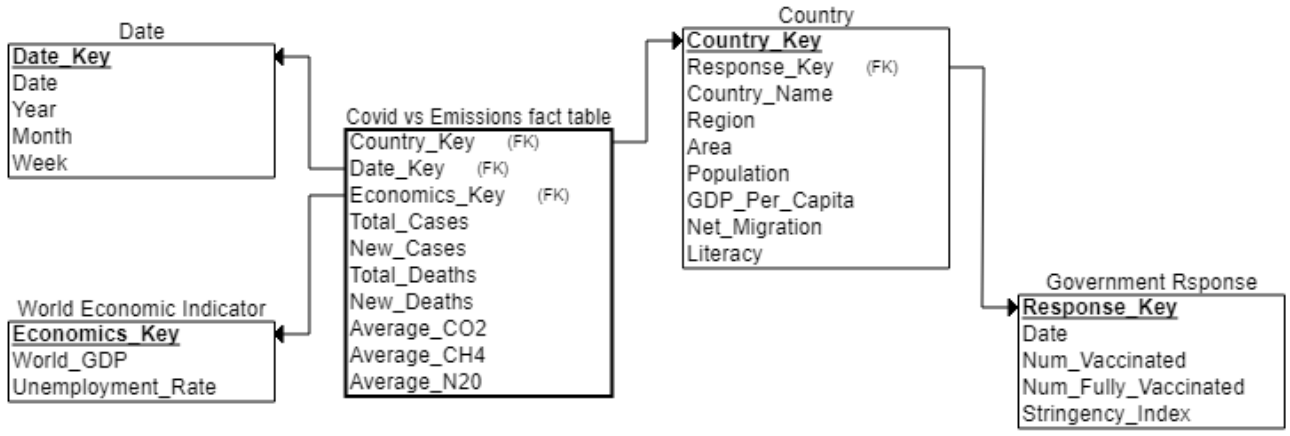
Covid vs Emissions Fact Table

- **Country_ID:** Foreign Key to Country dimension
- **Date_Key:** Foreign Key to Time dimension
- **Economics_Key:** Foreign Key to World Economic Indicators
- **Total_Cases:** Total number of Covid cases for a country (Integer)
    - min = 1, max = 445129499, sample = 1
- **New_Cases:** Number of Covid cases occured on a specific date for a country (Integer)
    - min = 0, max = 4206334, sample = 20
- **Total_Deaths:** Total number of Covid deaths for a country (Integer)
    - min = 1, max = 5995245, sample = 1
- **New_Deaths:** Number of Covid deaths occured on a specific date for a country (Integer)
    - Integer that represents, min = 1, max = 18020, sample = 1
- **Average_CO2:** Weekly average atmospheric CO2 levels in parts per million (Float)
    - min = 0.0, max = unlimited, sample = 331.73
- **Average_CH4:** Monthly average atmospheric CH4 levels in parts per million (Float)
    - min = 0.0, max = unlimited, sample = 10.92
- **Average_N2O:** Monthly average atmospheric N2O levels in parts per million (Float)
    - min = 0.0, max = unlimited, sample = 300.21

Measures/Facts

The facts/measures are Total_cases, New_cases, Total+deaths, New_deaths, Average_CO2, Average_CH4, and Average_N2O. Details for each fact can be found in the “Fact Table” section above.

Diagram



Avoiding Design Mistakes

1. **Placing text attributes in the table:**

    - Most of the numeric measurements from the operations are in teh fact table.
    - Text attributes usually belong to the dimensional table.
    - It can be seen from our diagram.

2. **Limit verbose descriptions to save space:**

    - Our design should be easy to use and consists of detailed readable texts that are accompanied with the dimensional models.
    - Text attributes in the dimensional tables provide all the detailed crucial informations.

3. **Normalise to save space:**

    - According to our diagram, the current design is not recording the same information twice. It also limits the possibility of deleting data which could cause unintended data removals.
    - I believe our design conforms/adheres to the third normal form.

4. **Ignore the need to track changes:**

    - Surrogate keys in our diagram tracks dimension attributes that are changed with a new dimension record for each profile change.

5. **Add new hardware to solve all query performance issues:**

    - Summary table can be a counter example.

6. **Use operational key's as the primary keys:**

    - Operational keys are only used as a tool to be mapped to common profile.

7. **Neglect to declare (and comply with) the grain:**

    - The grain is declared in a greater detail and it fully describes a "single row" in the fact table. We have chosen the lowest meaningful and the most feasible grain of each dimension.
    - It ensures that all facts in a single fact table are at the same grain.

8. **Neglect a detailed design:**

- Our design is detailed as it tries to resolve many to one relationships in dimension tables.
- It also tries to ensure that every fact table has an associated date dimension table.
- Stores filter domain values in dimension tables.

9. **Expect users to query normalised data:**

    - According to our design, the data does not seem to be aggregated in any way.

10. **Fail to conform Facts and Dimensions:**

    - Conformed dimensions are created with a sole purpose of integrating datas.
    - Measured numeric facts are sourced from single system. There is no conflicting definitions.

Team's Work Plan

- Our team consists of three members who live relatively close to each other and have an access to a reliable internet.

- Daily meeting was held on Microsoft teams, Discord and in person since the beginning of the creation of the team to discuss, plan and to refine our project. Furthermore, we have been meeting 18 times starting from the January 19th, 2023 up until the February 11th, 2023.

- All three of us worked together on devising/coming up with the topic in the first deliverable. On the second deliverable, Xia Meng Li worked from questions 1 to 2, Maaz Mazharul worked on the next two questions (questions 3 and 4), and Bill Battushig completed questions 5 and 6. The list of additional references were done by all of the three.

- Eventhough we have met multiple times, the meeting with TA were held only once or twice through email.

References

Dr. Pieter Tans (2023), *Dry Air Mole Fractions from quasi-continuous measurements at Barrow, Alaska (BRW); Mauna Loa, Hawaii (MLO); American Samoa (SMO); and South Pole (SPO), 1973-2023* 5 Feb, 2023

https://gml.noaa.gov/aftp/data/trace_gases/co2/in-situ/surface/README_surface_insitu_co2.html [1]

The World Bank (2021), *GDP (current US$)* 5 Feb, 2023

https://data.worldbank.org/indicator/NY.GDP.MKTP.CD [2]

International Labour Organization (2022), *Unemployment, total (% of total labor force) (modeled ILO estimate)* 6 Feb, 2023

https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS [3]