uOttawa
Faculté de génie
Faculty of Engineering

CSI4142

Fundamentals of Data Science

Course Professor**: Yazan Otoum**
Supervisor**: Lansu Dai**

Phase 1: Conceptual Design

Group 15
Team Members:

First Name, Last Name:	Student ID:
Lixiong Wei	300145970
Zhiyuan Lin	300126813

Due Date: Feb 8th

Grain of the data mart:

At weekly level from January 2021 to December 2023, focusing on COVID-19 case number and vaccination situation (No Dosed, Partial Dosed, Fully Dosed and Booster Dosed) across different provinces in Canada.

Assumptions:

1. COVID-19 usually recovers in **one to two weeks**. For severe cases, recovery can take six weeks or more; therefore, we take the data weekly for better analysis.
2. False Positive and True negative will not be considered.
3. To make the data mart more readable, the data will be in unit of thousand (ex. positive 65k)

Dimensions and Dimensional Attributes:

Date Dimension:

- DateID: Integer
- Month: String (January to December)
- Year: Integer (2021-2023)

Vaccination Dimension:

- VacID: Integer
- DateID (Foreign Key)
- Not: Integer
- Partial: Integer
- Full: Integer
- Booster: Integer

COVID19 Metrics Dimension:

- covidID: Integer
- DateID (Foreign Key)
- Positive: Integer
- Active: Integer

Province Dimension:

- proid (province id): Integer
- DateID (Foreign Key)
- name: String (province name ex: ON, BC…)

Fact Table:

- **DateID**
- **VacID**
- **covidID**
- **proid**

Checklist:

1. Place text attributes in the Fact table:
   Fortunately, we don't have text attributes in our fact table. They are mostly numeric. Only province names are string.

2. Limit verbose descriptions to save space:
   not applicable

3. Normalize to have space (leads to slower queries):
   not applicable

4. Ignore the need to track changes:
   The custom ID and other data ID are stored in our project. Both attributes help tracking the historical data and trace back.

5. Add new hardware to solve all query performance issues (Not Applicable)
   The scope of hardware is not covered in our project.

6. Use operational keys as the primary key:
   The primary keys we used are mostly related to date which would not change over time. A custom ID dimension is built to generate a unique primary key to maintain consistency and uniqueness.

7. Neglect to declare the grain:
   Our grain is clearly defined at a weekly level from January 2021 to December 2023, focusing on COVID-19 case numbers and vaccination situations (No Dosed, Partial Dosed, Fully Dosed and Booster Dosed) across different provinces in Canada. Time period, areas, proportion affected by COVID-19 and different vaccination situations are explicitly stated.

8. Neglect a detailed design:
   The design is detailed and reliable

9. Expert users to query normalized data:
   not applicable for now

10. Fail to conform facts and dimensions: Each table has dateID as foreign key to make sure dimensions match.

Team work summary:

Divided work for this design:
Zhiyuan Lin: Search for data set and design mistakes check
Lixiong Wei: Grain declaration, Dimensional and Fact Table design
Meet with the TA: By appointment
Meeting time: Every two weeks

Additional references:

Title: Coronavirus Diagnosis: What Should I Expect?

https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/diagnosed-with-covid-19-what-to-expect