University of Ottawa

School of Electrical Engineering and Computer Science

CSI4142 Fundamentals of Data Science

Project Phase 2: Physical Design and Data Staging

Instructions:

A. This is a team assignment.
B. Submit your documentation via BrightSpace using your team locker.
C. For your source code, you may either submit a zipped file or provide a link to a GitHub repository.
D. Demonstrate your work during a Zoom meeting with the TA, in the timeslot allocated to you. Note that all team members are required to attend this demonstration and you will be asked to turn your cameras on.

Project Description - Covid-19 Tracking and Lifestyle Trends Data Mart

Data science and artificial intelligence (AI) have been successfully used to study trends in our behaviours over time. However, our daily routines changed abruptly with the onset of the COVID-19 pandemic. In Canada, and many other countries, lockdown procedures were implemented, thus leaving citizens with little choice to adapt their lifestyles accordingly. For instance, people increasingly turned to online shopping, while participation in outdoor activities increased. Many non-essential businesses, notably in the hospitality sector, also adapted by offering only delivery or curbside pickup, leading to changes in consumer behaviour and traffic patterns.

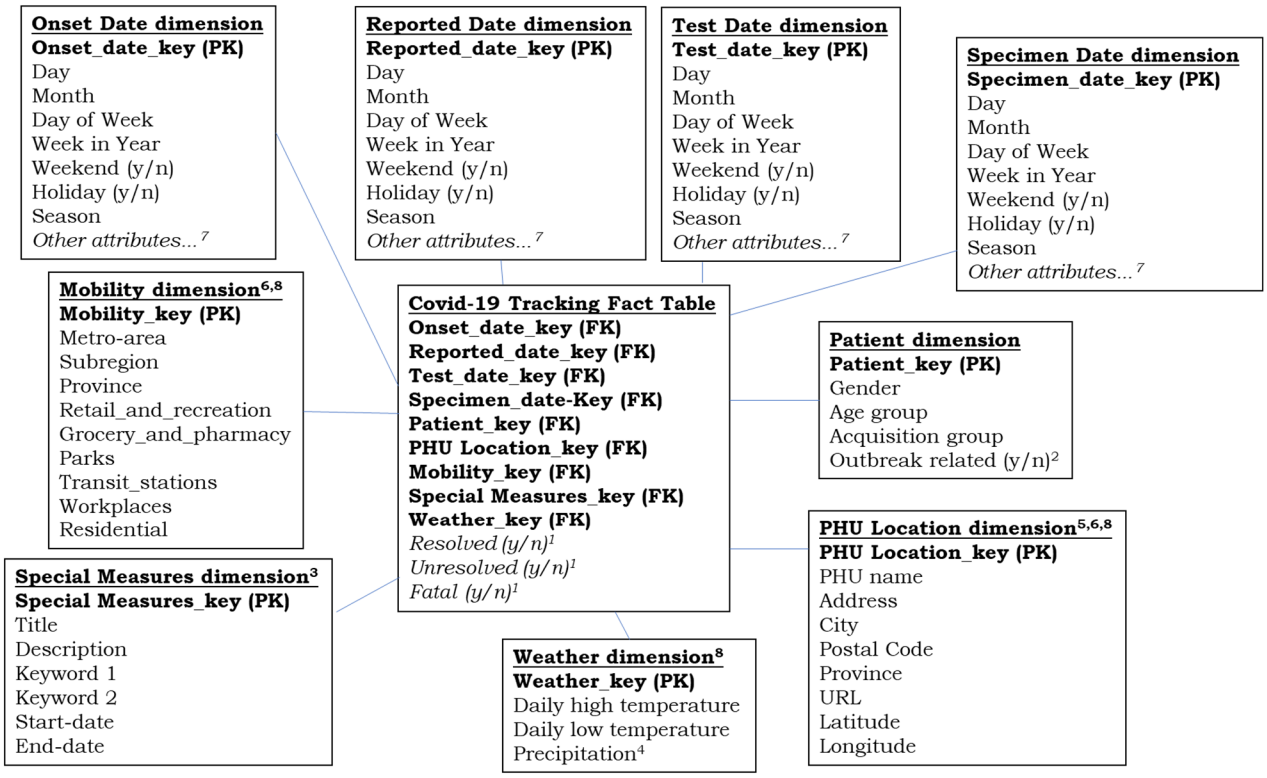
The current shift in Canadians' habits, especially while tracking the ebb and flow in the number of Covid-19 cases, warrants further study. Indeed, it is important to explore change in behaviours, to assess the impact of the pandemic and to plan for our future.

Suppose that your team was hired to design and to implement a data mart to map not only the details of individuals who tested positive for Covid-19, but also to assess lifestyle changes of Canadians during the Covid-19 pandemic.

As a first step, you completed a conceptual design for a Covid-19 case tracking and lifestyle trends data mart. As a proof of concept, you decide to focus on the Greater Toronto Area (GTA) – i.e., the City of Toronto and the regional municipalities of

Durham, Halton, Peel, and York – as well as the City of Ottawa, using historic data from 2020.

Below, find a suggested dimensional model of the Covid-19 case tracking and lifestyle trends data mart. This data mart tracks the outcomes of individual patients. Here, date acts as a role-playing dimension, reflecting the date of first onset (if known), the date a specimen was taken, the date the test result was received and the date it was reported that the patient tested positive for Covid-19. In addition, information about the mobility of the population and the daily weather are incorporated into the model. Finally, the dimensional model includes details about announcements of special measures taken by the provincial or federal governments.



Please note the following (the numbers below correspond to the numbers in the figure above).

1. The three measures – *Resolved, Unresolved* and *Fatal* - are binary attributes that indicate a patient's outcome.
2. The *Outbreak related* attribute refers to individuals who are associated with an outbreak in a facility such as a long-term care home or a food processing plant. This attribute would typically be used as a query constraint, and it is therefore included in the **Patient** dimension.
3. The **Special Measures** dimension refers to announcements of measures taken by the federal or provincial government in their attempts to curb the spread of the pandemic. This dimension could typically be created by analyzing press releases of government agencies and may used to track the impact of special measures (such as a lockdown of all non-essential businesses) on the number of cases and outcomes, and vice versa. (Note that this dimension corresponds to the **Promotion** dimension in the Sales Fact data mart as covered in the lectures.)
4. In the **Weather** dimension, the precipitation may also have been modelled as detailed attributes (e.g., (rain = 10 mm) or (snow = 5 cm)) or alternatively, binary attributes (e.g., rain (y/n) or snow (y/n)).

5. The **PHU Location** dimension contains details about the public health unit (PHU) location where a case was reported. This implies the location of the patients' place of residence is approximated and that we assume that a patient visits the PHU closest to where they live. Similarly, the mobility data refer to general trends in larger regions. However, this data may still be used to reflect on the impact of the number of cases (and outcomes) on mobility and lifestyle choices, and vice versa.
6. We use the **Reported Date** dimension, i.e., the date a patient was referred to the PHU, to connect with the **Mobility** and **PHU Location** dimensions through the fact table. (Another option would have been the **Onset Date** dimension; however, you will notice that this attribute contains missing values.) Note that the **Mobility** dimension use as baseline the median value, for the corresponding day of the week, during the 5-week period from 3 January 2020 to 6 February 2020. More details may be found at: https://www.google.com/covid19/mobility/data_documentation.html?hl=en
7. The **Date** dimensions are often used as query constraints. You are encouraged to use a full-fledged version containing numerous attributes. This is an example of a static dimension that may simply be loaded into the data mart.
8. The intersection between the **Mobility, Weather** and **PHU Location** dimensions may be obtained using postal codes, latitude, and longitude, and/or by consulting regional maps.

Note: Refer to the data dictionary, as obtained from the Ontario Open Data website, for further details of the source attribute domains and values.

Some useful links:
Below find some useful links to Open Data and other resources.
Link to the individual cases in Ontario (includes data dictionary):
https://data.ontario.ca/dataset/confirmed-positive-cases-of-covid-19-in-ontario/resource/455fd63b-603d-4608-8216-7d8647f43350
Google mobility data (download the CSV files for CA:
https://www.google.com/covid19/mobility/ - provides the change, as a percentage, to a baseline.
Weather data: https://climate.weather.gc.ca/historical_data/search_historic_data_e.html
News sources: https://www.covid-19canada.com/ or https://www.cbc.ca/news/covid-19
Government of Canada: https://www.canada.ca/en/public-health/services/diseases/coronavirus-disease-covid-19.html
Ontario Government: https://covid-19.ontario.ca/

Deliverables:

A. Create the **Covid-19 tracking and lifestyle trends data mart** using the PostgreSQL database management system.
B. Follow the data staging steps, as discussed during the lectures, to populate the data mart. You should use the confirmed positive cases data for any four months, for patients located in the GTA and Ottawa regions. Your model should incorporate daily weather and daily mobility data that correspond to this period. In addition, your database should also include the details of at least ten (10) special measures, as announced by the federal or Ontario governments during these four months.
C. Submit your source code, in a zipped file, or submit a link to a GitHub repository.
D. Demonstrate your work during a Zoom meeting with the TA, in the time slot allocated to you. Note that all team members are required to attend this demonstration and that you will be asked to turn your cameras on.
E. Submit a PDF file with the following details.
    1. A one-page schematic with your high-level data staging plan.
    2. A list of data quality issues you encountered and how you handled them. (For instance, how did you detect and handle missing values.)
    3. A table containing the following information (please use the Excel or CSV file attached):

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 and tables						
Create Date dimension						
Create Patient dimension						
Create PHU dimension						
Create Mobility dimension						
Create Weather dimension						
Map PHU, Mobility and Weather dimensions						
Create Special Measures dimension						

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)
Surrogate key pipeline – including role-playing dates						
Staging of dimensions						
Staging of fact table – including FKs and measures						
Data quality handling and reporting						
Others – if any						

An important note about the dimensional model

In the design above, we implicitly assume that there are no two patients with exactly the same profile. That is, there is no-one with the same onset-date, reported-date, test-date, specimen-date at the same PHU unit, that has the same patient profile and outcome.

For instance, we have only one person with a profile where the gender is female, age group is in her 30s, acquisition group is close contacts and outbreak related is equal to no.

In the case where we do have two patients where these are the same, the design as shown below offers alternative measures. In this design, we count the **numbers of resolved, unresolved and fatal cases. These values may range from 0 to n, where n > 1.**