CSI-4142
Fundamentals Of Data Science

Prof. Herna L. Victor

SUBMISSION DATE
March 10th, 2021

BILLAL ZAZAI, 8572975
YOUSSEF MAKBOUL, 8609614
GRAYDON HOPE, 300045044

Schema Design

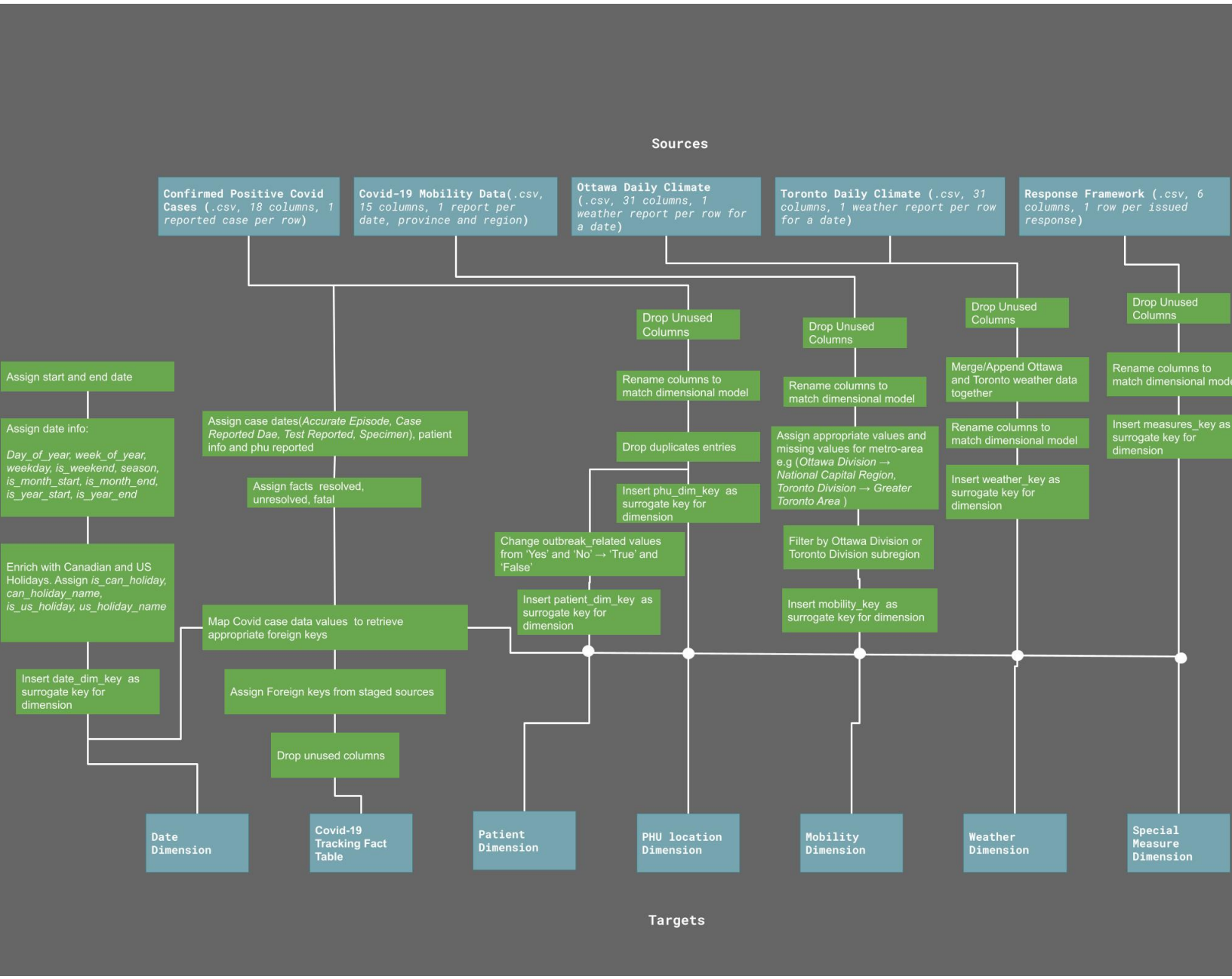

Figure 1. Covid-19 Data Mart Schema- Illustrates a high level overview of the data staging level of Data Mart

Data Quality

Fortunately, the data sets we were provided and found had very little missing data and data quality issues. However, we still encountered poor data quality such as …

- To minimize type conversion we changed outbreak_related values from ‘Yes’ and ‘No’ → ‘True’ and ‘False’ in the Patient Dimension
- The Covid-19 Mobility data set had missing values for the metro_area column. Therefore, utilizing the sub-region2 column we were able to assign the appropriate values and missing values for metro-area e.g (Ottawa Division → National Capital Region, Toronto Division → Greater Toronto Area)
- When merging Ottawa Daily Climate and Toronto Daily climate checking Referential Integrity(RI) was necessary to ensure the correct data is being merged
- To eliminate repetitive data, removal of duplicate data was required for certain dimensions such as Patient Dimension, PHU Location Dimension etc
- Column renaming was necessary to maintain consistency in the naming standards throughout all the dimensions

Tasks were evenly distributed among team members

Sources

Confirmed Positive Covid Cases (*.csv, 18 columns, 1 reported case per row*)

Covid-19 Mobility Data(*.csv, 15 columns, 1 report per date, province and region*)

Ottawa Daily Climate (*.csv, 31 columns, 1 weather report per row for a date*)

Toronto Daily Climate (*.csv, 31 columns, 1 weather report per row for a date*)

Response Framework (*.csv, 6 columns, 1 row per issued response*)

Assign start and end date

Assign date info:

Day_of_year, week_of_year, weekday, is_weekend, season, is_month_start, is_month_end, is_year_start, is_year_end

Enrich with Canadian and US Holidays. Assign *is_can_holiday, can_holiday_name, is_us_holiday, us_holiday_name*

Insert date_dim_key as surrogate key for dimension

Assign case dates(*Accurate Episode, Case Reported Dae, Test Reported, Specimen*), patient info and phu reported

Assign facts resolved, unresolved, fatal

Map Covid case data values to retrieve appropriate foreign keys

Assign Foreign keys from staged sources

Drop unused columns

Drop Unused Columns

Rename columns to match dimensional model

Drop duplicates entries

Insert phu_dim_key as surrogate key for dimension

Change outbreak_related values from ‘Yes’ and ‘No’ → ‘True’ and ‘False’

Insert patient_dim_key as surrogate key for dimension

Drop Unused Columns

Rename columns to match dimensional model

Assign appropriate values and missing values for metro-area e.g (*Ottawa Division → National Capital Region, Toronto Division → Greater Toronto Area*)

Filter by Ottawa Division or Toronto Division subregion

Insert mobility_key as surrogate key for dimension

Drop Unused Columns

Merge/Append Ottawa and Toronto weather data together

Rename columns to match dimensional model

Insert weather_key as surrogate key for dimension

Drop Unused Columns

Rename columns to match dimensional model

Insert measures_key as surrogate key for dimension

Targets

Date Dimension

Covid-19 Tracking Fact Table

Patient Dimension

PHU location Dimension

Mobility Dimension

Weather Dimension

Special Measure Dimension