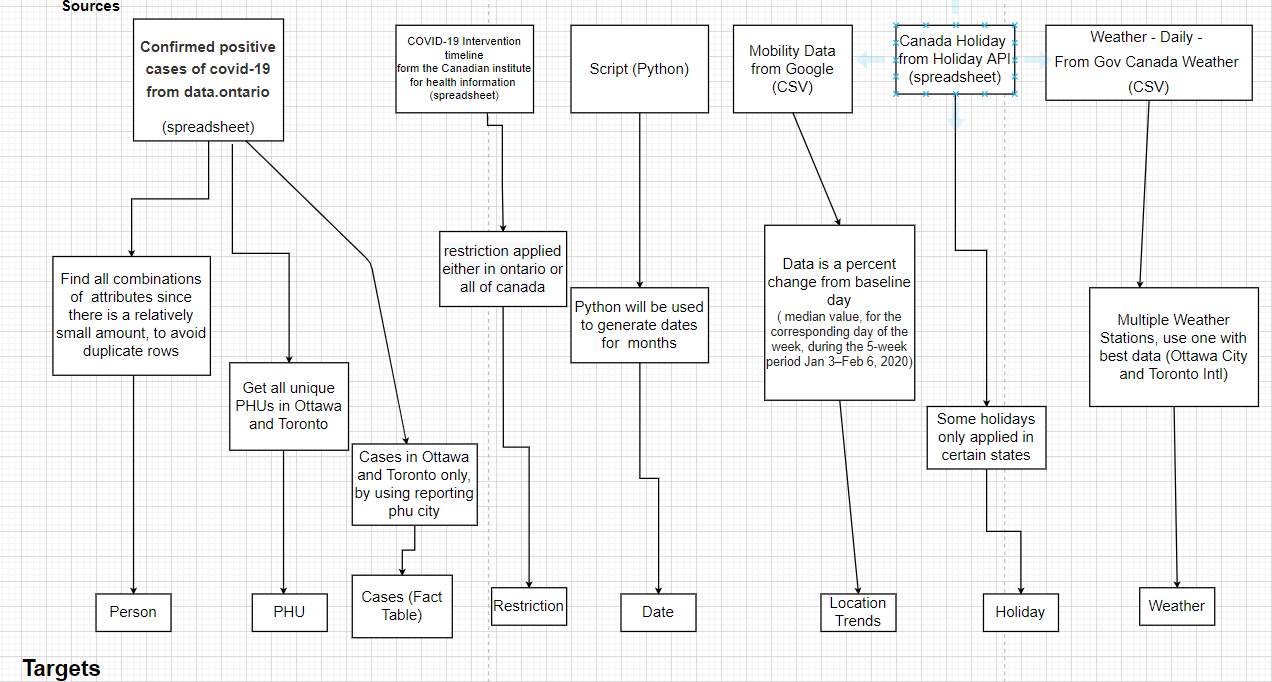
Group 12: Deliverable 2

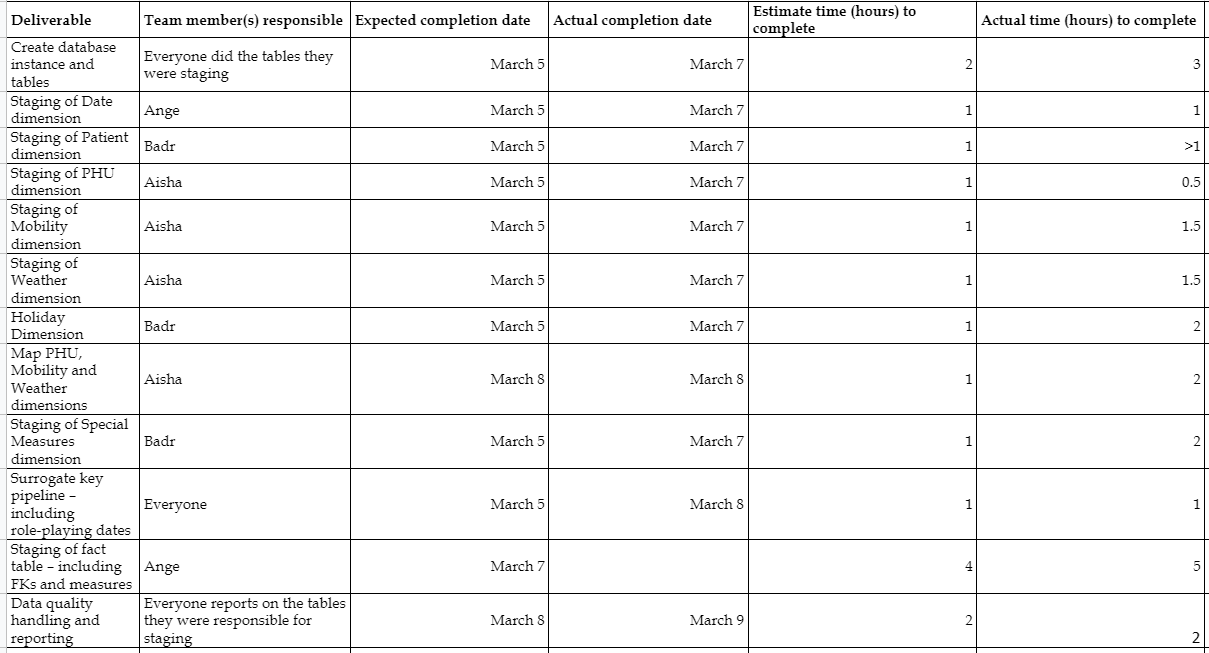
1. A one-page schematic with your high-level data staging plan.



2. A list of data quality issues you encountered and how you handled them. (For instance, how did you detect and handle missing values.)

|  |  |  |
| --- | --- | --- |
| **Data Source** | **Data Target** | **Notes** |
| Positive Cases Data: Accurate\_Episode\_Date | Fact Table: Onset\_date | Blank values may exist, replace with null  Date entered in the iPHIS to provide an estimate of onset date |
| Weather Data | Weather Dimension | ‘T’ used for “Trace”, replace with 0.1  ‘M’ used for “Missing”, replace with NULL |
| COVID-19 testing centre location |  | The postal code and other reporting location data in the positive cases csv is for the PHU location only, not for testing centre locations. So, testing centre data has turned out to be not useable with covid patient data. Will instead map to PHU |
| Mobility Data: sub\_region\_2 | Location\_trends dimension | Can ignore nulls in sub\_region\_2 because we will only be keeping rows with values for ottawa and toronto.  Change “Ottawa Division” and “Toronto Division” values to just “Ottawa” and “Toronto”  Nulls in \*\_baseline\* means data doesn’t meet quality, usually for low traffic areas (unlike ottawa and toronto) |
| Assessment centre and mobility data |  | Columns for province should be consistent. Currently one has full name (“Ontario”) while other has “ON” |
| Canada holidays API | Holiday dimension | Changed the data type of the statutory column to boolean  Many blank values encountered in the statutory column, manually checked (since the amount of rows is small) and assigned the correct boolean value. |
| Government interventions timeline (csv) | Restriction Dimension | Many rows were deleted since they were not considered as ‘restrictions’  The end date of the restrictions was not available in the data source, so we added it manually.  Since some restrictions are still applied till today, they logically can’t have an end date, and for that we assigned to them a ‘default’ end date that is outside of the 4 months we are analysing. |
| Patient data (csv) | Patient Dimension | Due to the large amount of duplicate rows, we decided to only keep the unique combinations between the age column and the gender one. |
| Patient data (csv) | Fact Table | Some of the dates were missing for some patients. The rows with missing values represented roughly 0.82%, so we decided to remove them from the data |

3. A table containing the following information:



4. Meeting Dates and Notes

* Feb 28th, 2021: 12pm - 2pm
  + Edited conceptual model from feedback, divided work using excel sheet
* March 5th, 2021: 6pm
  + Addressed/discussed specific issues we had with implementations to solve them as a group, revision of dimensional model, naming standards
* March 7th, 2021: 3pm
  + Data quality issues
* March 8th, 201: 6pm
  + More discussion/solving of problems with data
* March 9: 8:30
  + Going over everything before submitting

5. Github Link:

<https://github.com/angemichaella/CSI4142_project>

6. Conceptual Model Used

