**CSI 4142 – Group 25 - Deliverable 2   
Physical design and data staging**

**High level data staging plan**

Special Measures

Weather

Date

Fact Table

Patient

PHU\_Location

Mobility

Positive Cases (CSV) from Google

Announcements from *news.ontario.ca/mohltc/en*

Weather data from *climate.weather.gc.ca*

Mobility (CSV) from Google

**Team planning**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Deliverable** | **Team member(s) responsible** | **Expected completion date** | **Actual completion date** | **Estimate time (hours) to complete** | **Actual time (hours) to complete** |
| Create database instance and tables | Jonas, Jugraj, Jiaxin | 24/02/2020 | 24/02/2020 | 2 | 1 |
| Staging of Date dimension | Jonas | 1/3/2020 | 1/3/2020 | 1 | 1 |
| Staging of Patient dimension | Jonas | 1/3/2020 | 1/3/2020 | 1 | 1 |
| Staging of PHU dimension | Jiaxin | 3/3/2020 | 3/3/2020 | 1 | 1 |
| Staging of Mobility dimension | Jugraj | 4/3/2020 | 3/3/2020 | 1 | 1 |
| Staging of Weather dimension | Jiaxin | 4/3/2020 | 4/3/2020 | 1 | 1 |
| Map PHU, Mobility and Weather dimensions | Jonas | 5/3/2020 | 5/3/2020 | 2 | 1 |
| Staging of Special Measures dimension | Jugraj | 5/3/2020 | 5/3/2020 | 1 | 1 |
| Surrogate key pipeline – including role-playing dates | Jonas, Jugraj, Jiaxin | 6/3/2020 | 6/3/2020 | 2 | 2 |
| Staging of fact table – including FKs and measures | Jonas | 7/3/2020 | 9/3/2020 | 5 | 3 |
| Data quality handling and reporting | Jonas, Jugraj, Jiaxin | 9/3/2020 | 10/3/2020 | 2 | 2 |

**Issues encountered**

Throughout the data staging process, there was some missing data in the date fields. The gender field also had some instances of “unknown”. This was handled by alternating these cases between male and female. The most common was test date and in some instances specimen dates were missing. This was handled by using the case reported date when the test date was missing (as these are always the same date) and omitting the ones with specimen date missing as there was so few. The data for the dates required cleaning as some dates were stored in a different format, and some converted to float type instead of string. The team ran into one issue where we did not fully understand the surrogate key pipeline and didn’t included columns for both the surrogate key and dimension ID, once this issue was realized we all worked together to add the required columns so that we could properly fill in our fact table.

Also, there are some missing values in the weather daily data. We left the missing value as null values because the null value can still match to the specific date and keep the data frame same.