# University of Ottawa

# School of Electrical Engineering and Computer Science

# CSI4142 Fundamentals of Data Science

## Project Phase 2: Physical Design and Data Staging
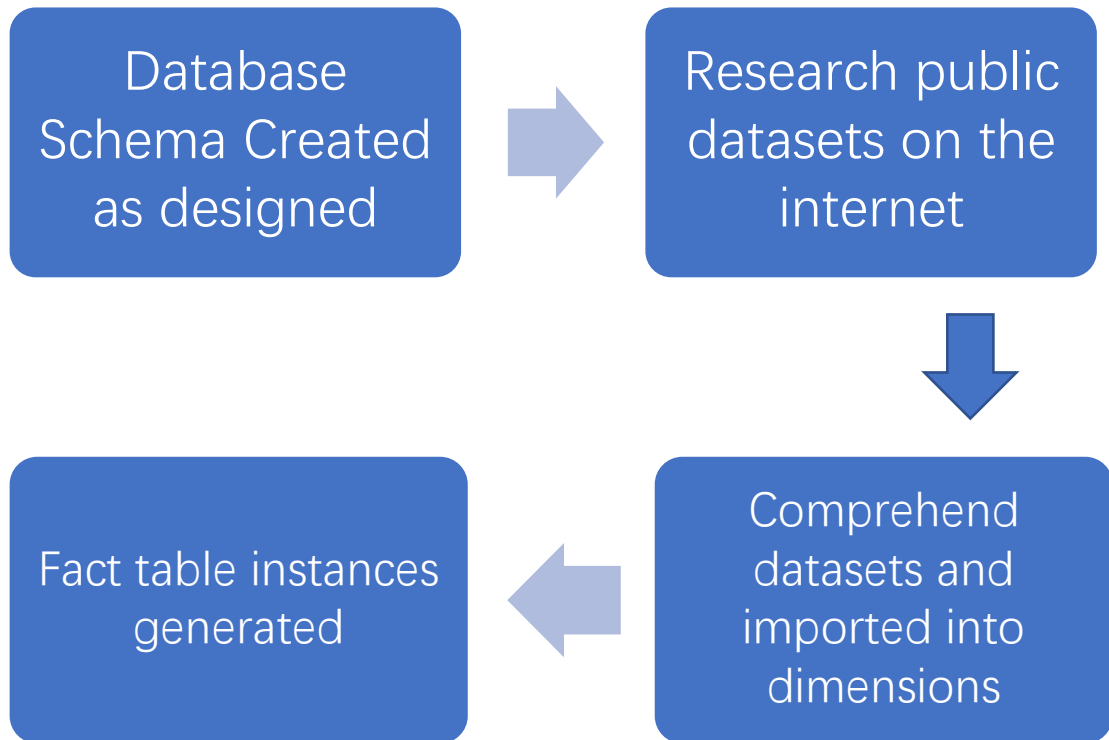
Student Name: Xia (Cynthia) Sheng
Student ID: 300091655
Student Name: Xinyu Ji
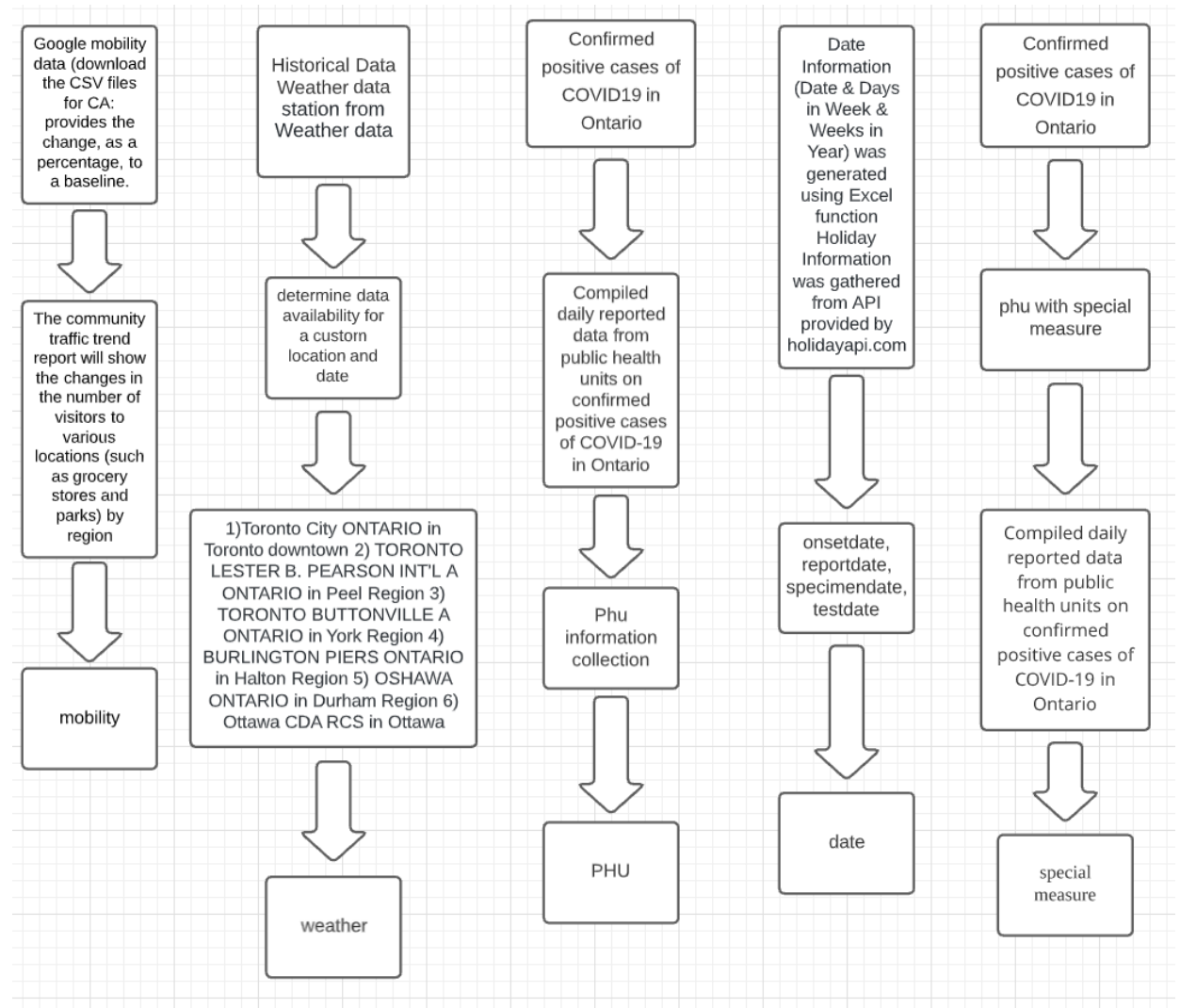Student ID: 300061354
Student name: Heng (Scott) Zhang
Student number: 300067988

Database Schema Created as designed → Research public datasets on the internet → Comprehend datasets and imported into dimensions → Fact table instances generated

# high-level data staging plan

**Column 1:**

Google mobility data (download the CSV files for CA: provides the change, as a percentage, to a baseline.)

↓

The community traffic trend report will show the changes in the number of visitors to various locations (such as grocery stores and parks) by region

↓

mobility

**Column 2:**

Historical Data Weather data station from Weather data

↓

determine data availability for a custorn location and date

↓

1)Toronto City ONTARIO in Toronto downtown 2) TORONTO LESTER B. PEARSON INT'L A ONTARIO in Peel Region 3) TORONTO BUTTONVILLE A ONTARIO in York Region 4) BURLINGTON PIERS ONTARIO in Halton Region 5) OSHAWA ONTARIO in Durham Region 6) Ottawa CDA RCS in Ottawa

↓

weather

**Column 3:**

Confirmed positive cases of COVID19 in Ontario

↓

Compiled daily reported data from public health units on confirmed positive cases of COVID-19 in Ontario

↓

Phu information collection

↓

PHU

**Column 4:**

Date Information (Date & Days in Week & Weeks in Year) was generated using Excel function Holiday Information was gathered from API provided by holidayapi.com

↓

onsetdate, reportdate, specimendate, testdate

↓

date

**Column 5:**

Confirmed positive cases of COVID19 in Ontario

↓

phu with special measure

↓

Compiled daily reported data from public health units on confirmed positive cases of COVID-19 in Ontario

↓

special measure

| Data quality issues | Who solved | Solutions |
|---|---|---|
| Missing data in weather dimension | Xia (Cynthia) Sheng | - All missing 'snow_on_ground' and 'total_recipitation' are filled with 0.<br>- all missing temperature are filled with mean value of the day before and after of the missing date. |
| Missing Data in Fact Table and Positive Case | Heng (Scott) Zhang | 1. Some data may not have specimen date or test date, in case of this circumstance, we use report date as their specimen data/test date<br><br>2. Some cases have no age info. available, we use 0 during data staging |

| | | 3. Gender may have 3 types of value: M for Male, F for Female, U for Unknown |
| --- | --- | --- |

## CSI4142 Physical Design and Data Staging - Project 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 | Heng (Scott) Zhang ,Xia (Cynthia) Sheng, Xinyu Ji | 2021-02-19 | 2021-02-18 | 2 | 1h45min |
| Staging of Date dimension | Heng (Scott) Zhang | 2021-02-20 | 2021-02-19 | 0.5 | 0.5 |
| Staging of Patient dimension | Heng (Scott) Zhang | 2021-02-20 | 2021-02-19 | 1 | 45min |
| Staging of PHU dimension | Xinyu Ji | 2021-02-20 | 2021-02-19 | 0.5 | 0.5 |
| Staging of Mobility dimension | Heng (Scott) Zhang | 2021-02-20 | 2021-02-19 | 1 | 0.5 |
| Staging of Weather dimension | Xia (Cynthia) Sheng | 2021-02-21 | 2021-02-19 | 1 | 45min |
| Map PHU, Mobility and Weather dimensions | Heng (Scott) Zhang | 2021-02-25 | 2021-02-19 | 2 | 1h45min |
| Staging of Special Measures dimension | Xinyu Ji | 2021-02-20 | 2021-02-19 | 0.5 | 0.5 |
| Surrogate key pipeline – including role-playing dates | Xia (Cynthia) Sheng | 2021-02-28 | 2021-02-19 | 2 | 1h45min |
| Staging of fact table – including FKs and measures | Heng (Scott) Zhang ,Xia (Cynthia) Sheng | 2021-02-28 | 2021-02-19 | 2 | 1h45min |
| Data quality handling and reporting | Xinyu Ji | 2021-03-01 | 2021-02-19 | 1 | 1 |