CSI4142

Fundamentals of Data Science

Course Professor**: Yazan Otoum**
Supervisor**: Lansu Dai**

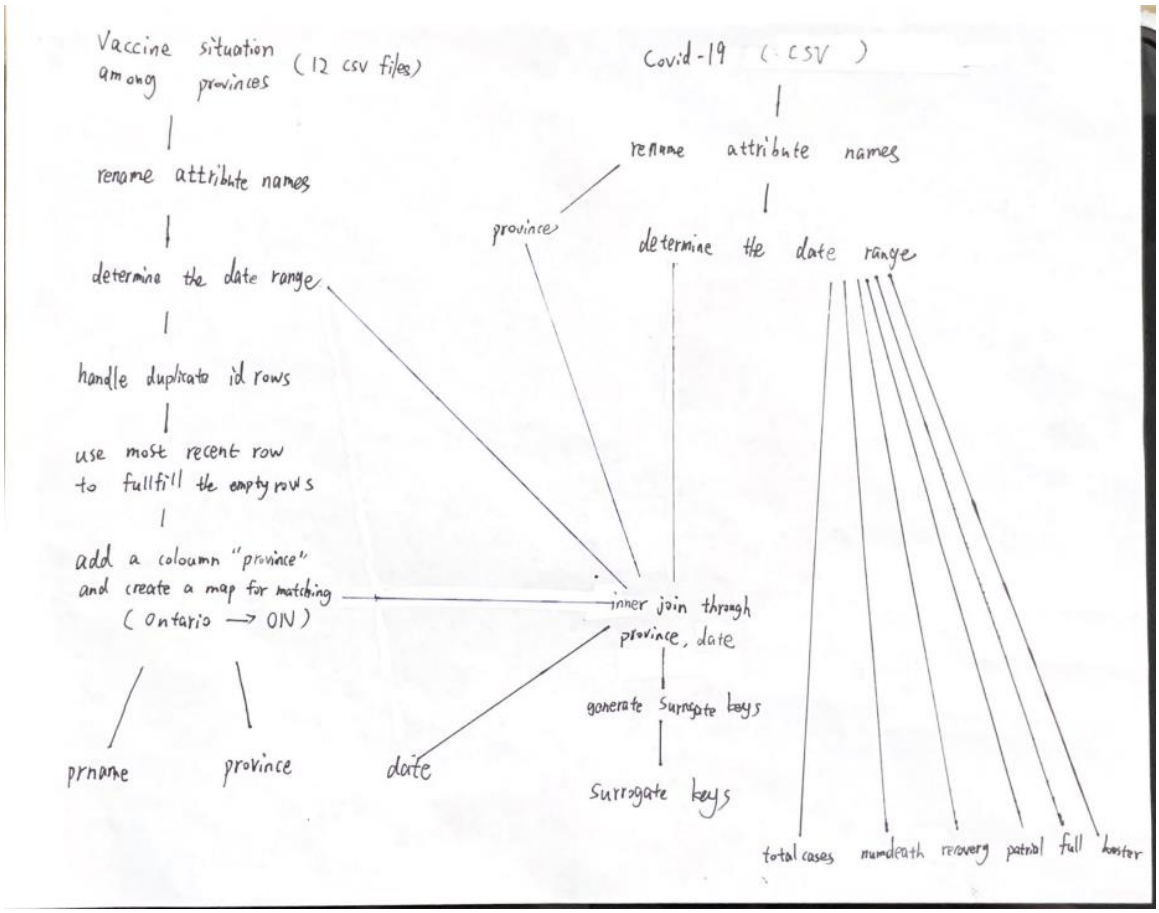
Phase 2: Physical Design and Data Staging

Group 15
Team Members:

First Name, Last Name:	Student ID:
Lixiong Wei	300145970
Zhiyuan Lin	300126813

Due Date: March 22nd

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



Modification for original dimensional table design:

During the data staging process, we notice that some of our original design are not needed; therefore, we remove some features:

- In the Vaccination dimension, there is a column called not, which represents the number of people that are not vaccinated. When doing OLAP, we will try to discover the relationship between vaccination and covid-19 cases, not vaccinated is not needed.
- Active from covidMetric dimension is removed (duplicated with positive)

Issue and Solution

Issue 1:

Data from different sources does not match, in our case, provinces in one table are shown as full names like “Ontario”, the other only with abbreviations like “ON”.

Solution 1: Matching

Create a dictionary for matching, add a new column in one of the table to match with the other

Code example:

```
province_mapping = {
    'Ontario': 'ON',
    'Quebec': 'QC',
    'Nova Scotia': 'NS',
    'New Brunswick': 'NB',
    'Manitoba': 'MB',
    'British Columbia': 'BC',
    'Prince Edward Island': 'PE',
    'Saskatchewan': 'SK',
    'Alberta': 'AB',
    'Newfoundland and Labrador': 'NL',
    'Northwest Territories': 'NT',
    'Yukon': 'YT',
    'Nunavut': 'NU'
}
cases['province'] = cases['prname'].map(province_mapping)
```

Issue 2:

Date is not match. In our case, we will analyze the data weekly, which means a specific date in a week will represent the change in the whole week (Sunday). However, each province updates the data irregularly (some provinces update on Wednesday, some on Friday).

Solution 2: Binning Methods (Smoothing by bin boundaries)

Fill in the table with the closest one. We use the covid-download.csv as standard because the table updates regularly, then for each date, if it can not be found on the other table, add the date with data from the closest date.

For example, if 2022-03-06 is needed but not occurred in table A, in the meantime, 2022-03-08 and 2022-03-01 are stored, use the data in 2022-03-08 to represent the data in 2022-03-06 and add it into table A.

Issue 3:

Data missing: Nunavut's data has not been updated frequently since November 2022. There exists empty data "0" which harms the quality of data analysis.

Solution 3:

Fulfill the empty data by using the most recent row data to maintain the data validity.

Team Planning and Work Distribution

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	Zhiyuan Lin	15-Mar	17-Mar	3 hours	6 hours	
Create Date dimension	Zhiyuan Lin	16-Mar	19-Mar	1 hour	1 hour	
Create Province dimension	Zhiyuan Lin	16-Mar	19-Mar	1 hour	1 hour	
Create vaccination dimension	Zhiyuan Lin	16-Mar	19-Mar	1 hour	1 hour	
create a connection between python and postgre	Zhiyuan Lin	16-Mar	19-Mar	3 hours	3 hour	
Create COVID19Metric dimension	Zhiyuan Lin	15-Mar	17-Mar	3 hours	3 hours	
Staging of dimension Date	Lixiong Wei	17-Mar	18-Mar	0.5 day	1 day	
Staging of dimension Province	Lixiong Wei	17-Mar	18-Mar	0.5 day	1 day	
Staging of dimension Vaccination	Lixiong Wei	17-Mar	18-Mar	0.5 day	1 day	

Staging of dimension COVID19 Metric	Lixiong Wei	17-Mar	18-Mar	0.5 day	1 day	
Surrogate key pipeline	Lixiong Wei	19-Mar	19-Mar	1 hour	1 hour	
Staging of fact table – including FKs and measures	Lixiong Wei	20-Mar	20-Mar	1 day	1 day	
Data quality handling and reporting	Lixiong Wei	20-Mar	19-Mar	3 hours	2 hour	
SQL statement design	Zhiyuan Lin	18-Mar	19-Mar	3 hour	8 hour	