

BỘ MÔN HỆ THỐNG THÔNG TIN – KHOA CÔNG NGHỆ THÔNG TIN

ĐẠI HỌC KHOA HỌC TỰ NHIÊN THÀNH PHỐ HỒ CHÍ MINH, ĐẠI HỌC QUỐC GIA TP HCM

HỆ THỐNG THÔNG TIN DOANH NGHIỆP



HỌC KỲ I – NĂM HỌC 2022-2023

BẢNG THÔNG TIN CHI TIẾT NHÓM

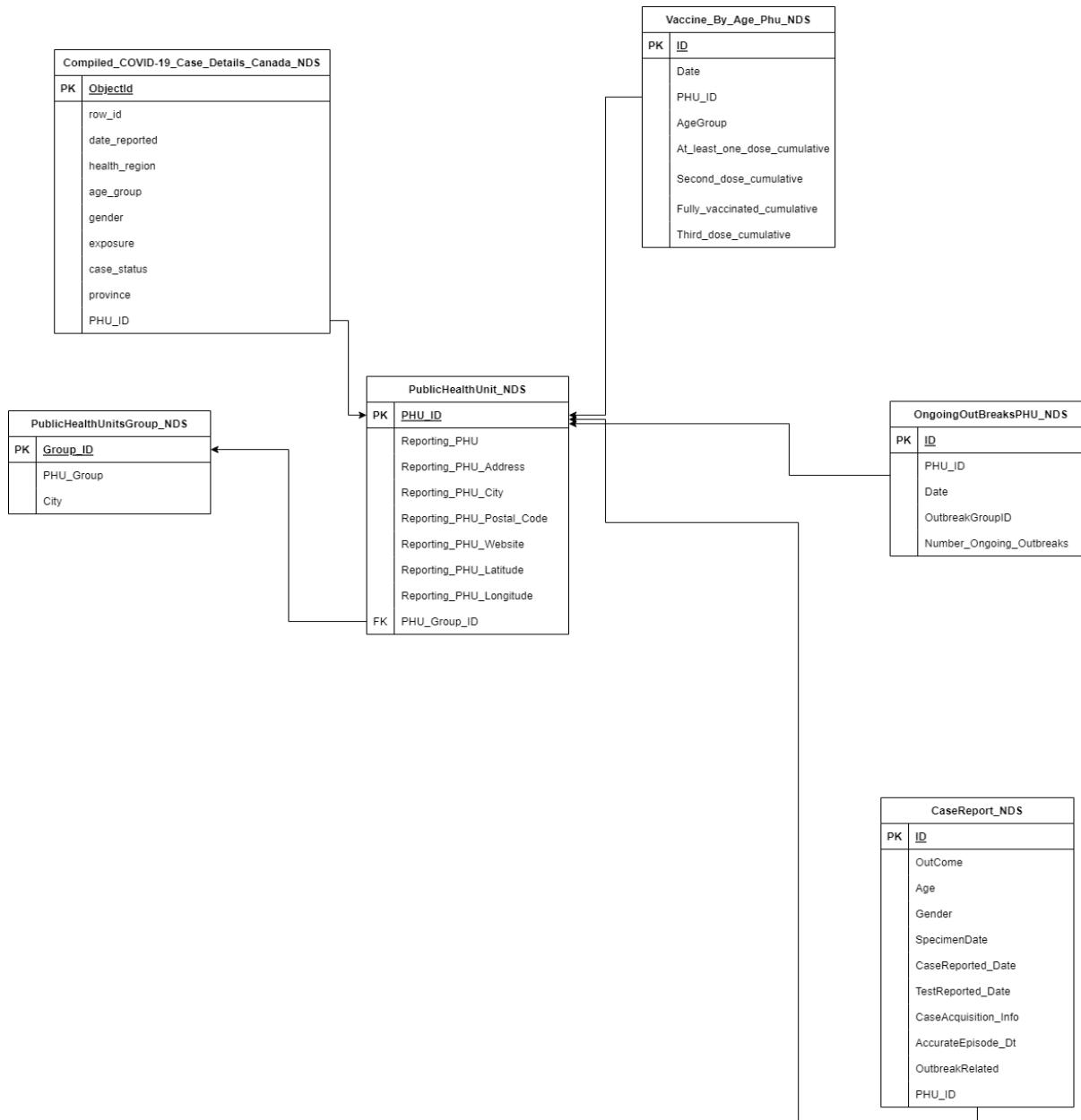
Mã nhóm:	BI#13	
Số lượng:	4	
MSSV	Họ tên	Email
19120555	Nguyễn Chánh Kiệt	19120555@student.hcmus.edu.vn
19120652	Nguyễn Trọng Thái	19120652@student.hcmus.edu.vn
19120699	Ngô Mậu Trường	19120699@student.hcmus.edu.vn
19120733	Lê Hoàng Thịnh Như Ý	19120733@student.hcmus.edu.vn

A. Phân tích và thiết kế kho dữ liệu.

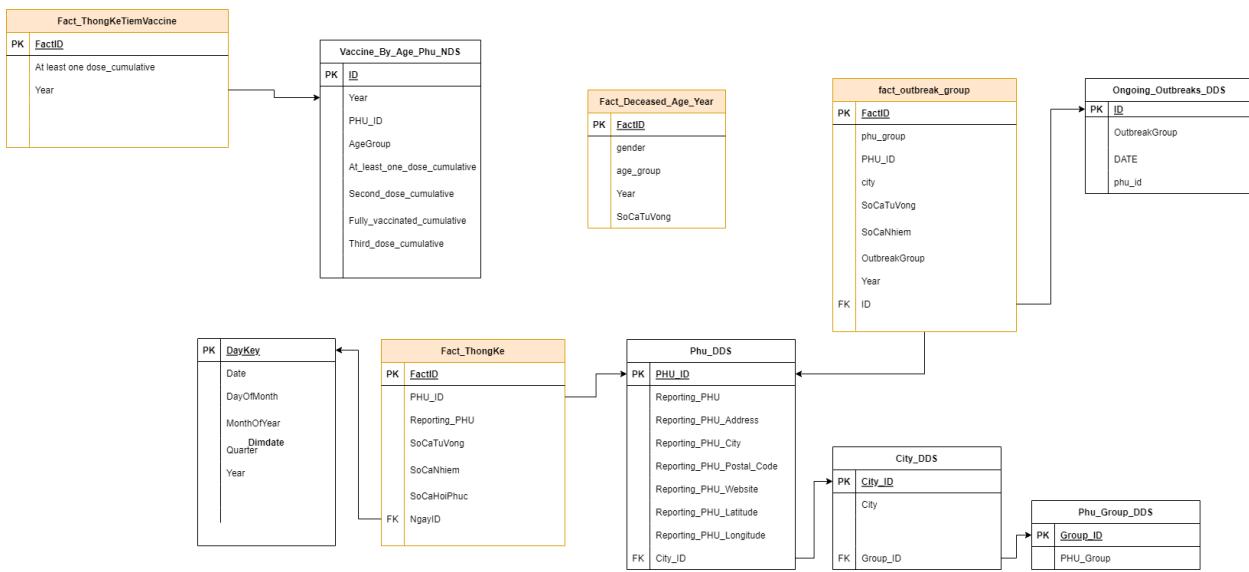
I. Thiết kế kho dữ liệu.

1. Mô hình NDS:

- Ta giữ nguyên dữ liệu không chuẩn hóa hoặc thay đổi thuộc tính dữ liệu.



2. Mô hình DDS:



Phân loại mức độ nghiêm trọng: tính theo tỉ lệ CFR = tử vong / tổng số ca bệnh

- + Cấp độ 1: <0.025
- + Cấp độ 2: <0.05
- + Cấp độ 3: <0.2
- + Cấp độ 4: >=0.2

a. Thống kê **Số ca nhiễm, số ca tử vong, số ca phục hồi** của dịch Covid-19 theo từng **PHU** trong từng năm.

- Sự kiện: Khi một người mắc Covid-19

- Bối cảnh:

- + Ai: người mắc Covid-19
- + Cái gì: Tình trạng ca bệnh
- + Ở đâu: Các đơn vị chăm sóc sức khỏe
- + Khi nào: Ngày báo cáo

- Đo lường: Số lượng, tình trạng ca nhiễm

- Các giá trị có sẵn từ nguồn: Reporting_PHU, PHU_ID,

- Các giá trị phải tính toán:

+ Deceased(SoCaTuVong)

+ Infected(SoCaNhiem) = Deceased + Active + Recovered

+ Recovered(SoCaHoiPhuc)

- Cấp chi tiết dữ liệu (độ mịn): Số ca nhiễm, số ca tử vong, số ca phục hồi của dịch Covid-19 theo từng PHU trong từng năm.

b. Thống kê **Mức Độ Nghiêm Trọng** của dịch Covid-19 theo **PHU** và theo các **Quý** trong từng năm.

- Sự kiện: Khi một người mắc Covid-19 hoặc mất vì Covid-19

- Bối cảnh:

+ Ai: Người mắc Covid-19

+ Ở đâu: Các đơn vị chăm sóc sức khỏe

+ Cái gì: Mức độ nghiêm trọng

+ Khi nào: theo từng quý

- Đo lường: Theo cấp độ từ 1 đến 4 mức độ nghiêm trọng tăng dần

- Các giá trị có sẵn từ nguồn: người bệnh, người tử vong

- Các giá trị phải tính toán: mức độ nghiêm trọng, tổng số tử vong, tổng số ca bệnh.

- Cấp chi tiết dữ liệu (độ mịn): số ca nhiễm theo từng PHU của từng năm

c. Thống kê tổng số người tử vong theo **Giới Tính** và **Nhóm Tuổi** theo các năm.

- Sự kiện: Khi một người tử vong vì Covid 19

- Bối cảnh:

+ Ai: Người bệnh Covid-19

+ Ở đâu: Lãnh thổ Canada

+ Cái gì: Tổng số người tử vong

+ Khi nào: Hằng năm

- Đo lường: số lượng, đơn vị người

- Các giá trị có sẵn từ nguồn: gender, age_group

- Các giá trị phải tính toán: total_death

- Cấp chi tiết dữ liệu (độ mịn): một dòng trong fact tương ứng với tổng số người tử vong theo giới tính và nhóm tuổi theo các năm.

d. Thống kê số ca nhiễm, tử vong theo **Mức Độ Nghiêm Trọng** theo **Ngày Trong Tháng** của các năm.

- Sự kiện: Khi một người nhiễm Covid-19

- Bối cảnh:

+ Ai: Người bệnh Covid-19

+ Ở đâu: Lãnh thổ Canada

+ Cái gì: Số ca nhiễm, số ca tử vong, mức độ nghiêm trọng

+ Khi nào: Mỗi ngày trong tháng của các năm

- Đo lường:

- Các giá trị có sẵn từ nguồn: số ca nhiễm, số ca tử vong

- Các giá trị phải tính toán: Mức độ nghiêm trọng

- Cấp chi tiết dữ liệu (độ mịn): số ca nhiễm, số ca tử vong của từng mức độ nghiêm trọng của từng ngày trong tháng của các năm

e. Thống kê số ca nhiễm, tử vong theo **Mức Độ Nghiêm Trọng, khu vực** (PHU_Group, City), và **số người đã được tiêm vacxin** trong các năm.

- Sự kiện: Khi một người nhiễm Covid-19, khi 1 người tiêm vacxin mũi 1

- Bối cảnh:

+ Ai: Người bệnh Covid-19, người tiêm vacxin

+ Ở đâu: Từng khu vực ở Canada

+ Cái gì: Số ca nhiễm , số ca tử vong, mức độ nghiêm trọng

+ Khi nào: trong các năm

- Đo lường:

- Các giá trị có sẵn từ nguồn: ca nhiễm, tử vong, khu vực, số người được tiêm vaccine mũi 1

- Các giá trị phải tính toán:

- Cấp chi tiết dữ liệu (độ mịn): số ca nhiễm, tử vong, mức độ nghiêm trọng của từng khu vực của từng năm, số người đã vacxin mũi 1 của từng năm

f. Thống kê số ca nhiễm theo **Mức Độ Nghiêm Trọng, nhóm bùng phát** của từng khu vực trong các năm

- Sự kiện: khi có thêm 1 ca nhiễm

- Bối cảnh:

+ Ai: Người bệnh Covid-19

+ Ở đâu: từng khu vực của Canada

+ Cái gì: ca nhiễm, mức độ nghiêm trọng, nhóm bùng phát

+ Khi nào: mỗi năm

- Đo lường:

- Các giá trị có sẵn từ nguồn: số ca nhiễm, khu vực, nhóm bùng phát

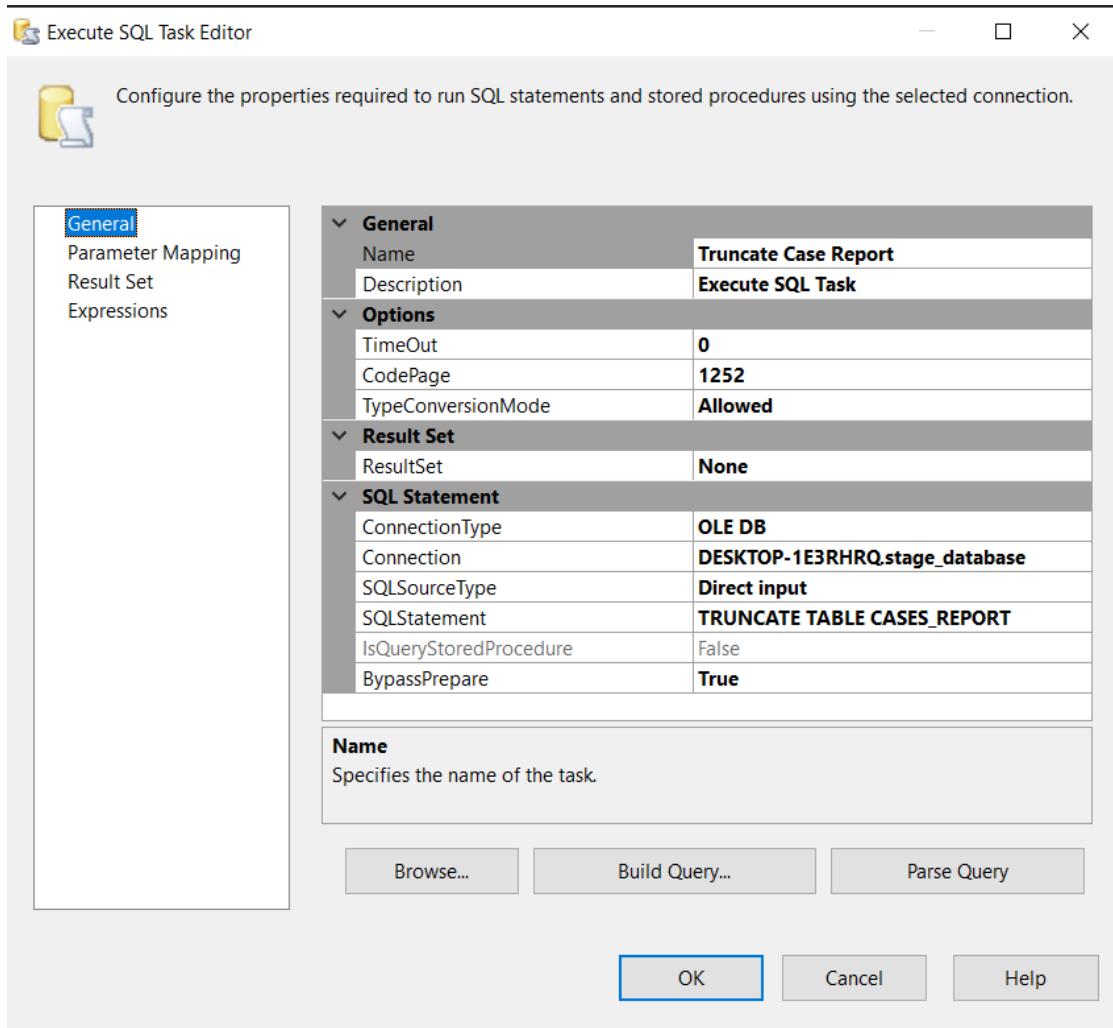
- Các giá trị phải tính toán: mức độ nghiêm trọng

- Cấp chi tiết dữ liệu (độ mịn): số ca nhiễm theo từng mức độ nghiêm trọng theo từng nhóm bùng phát của từng khu vực mỗi năm

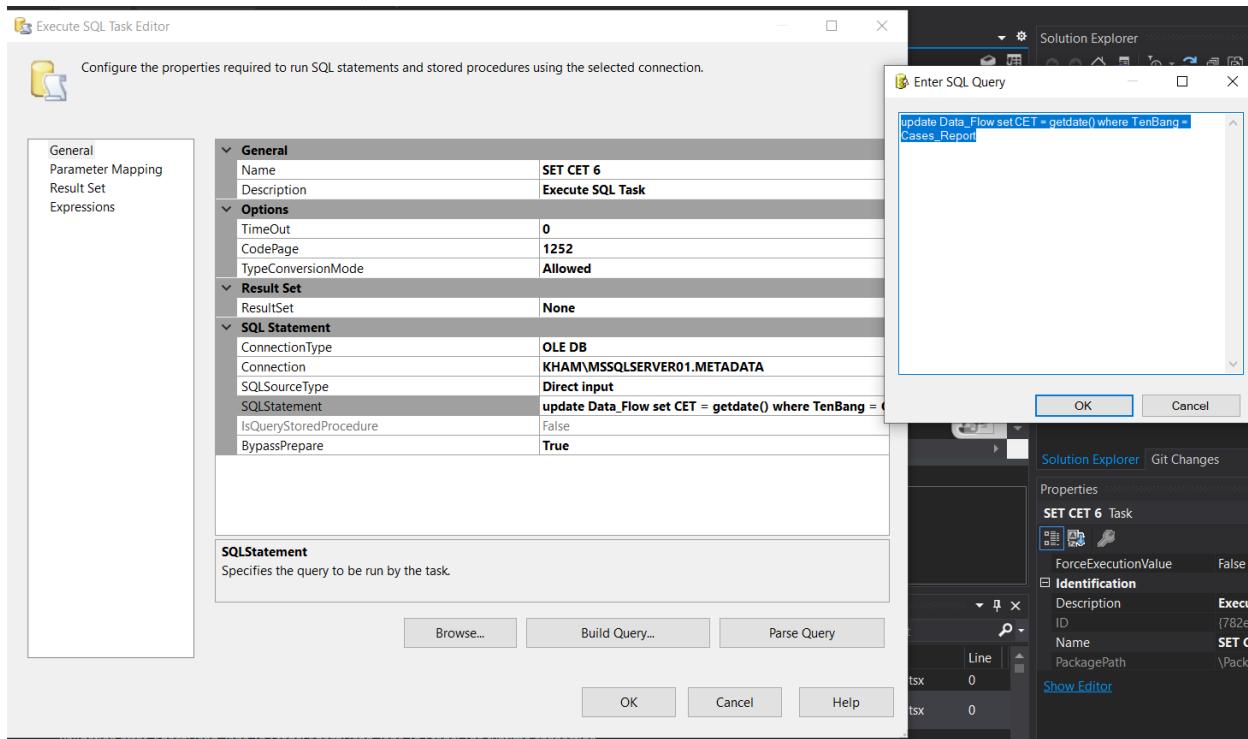
II. Quá trình ETL

- Trong file phu_group có 35 dòng cần xóa bớt dòng cuối do dữ liệu trùng
 - a. Source to Stage

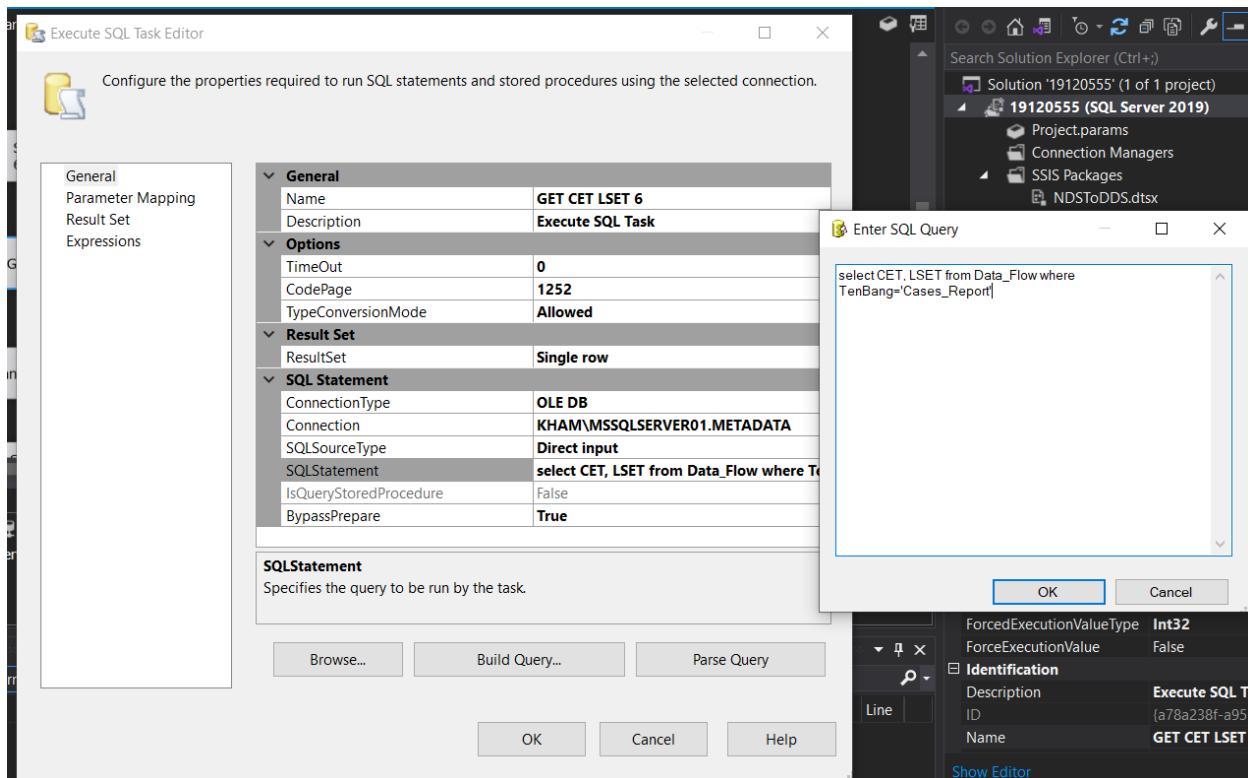
B1: Truncate table trong Stage



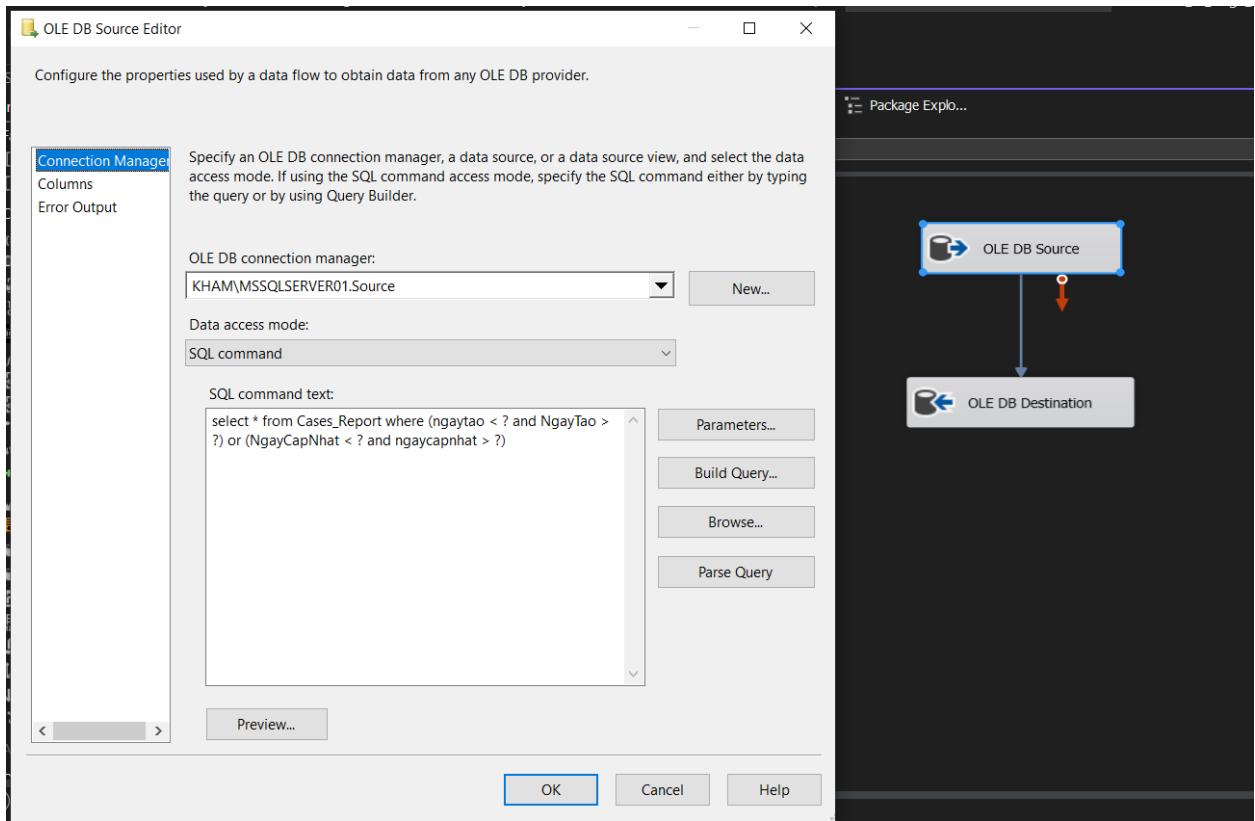
B2: Cập nhật CET = GETDATE()



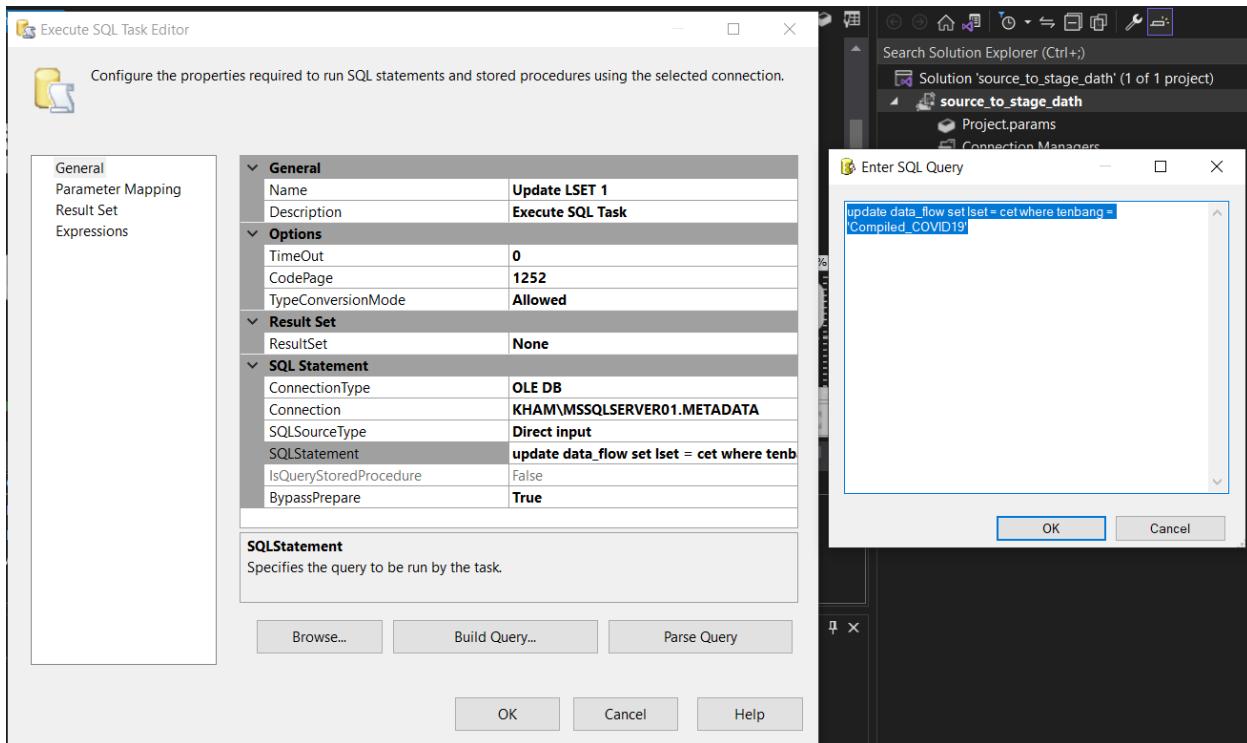
B3: Lấy CET, LSET: thời gian khởi động ETL package



B4: Rút trích dữ liệu: đồ dữ liệu từ Source sang Stage



B5: Cập nhật LSET = CET



b. Stage to NDS

B1: Lấy dữ liệu từ Stage

B2: Biến đổi dữ liệu sao cho phù hợp (nếu có)

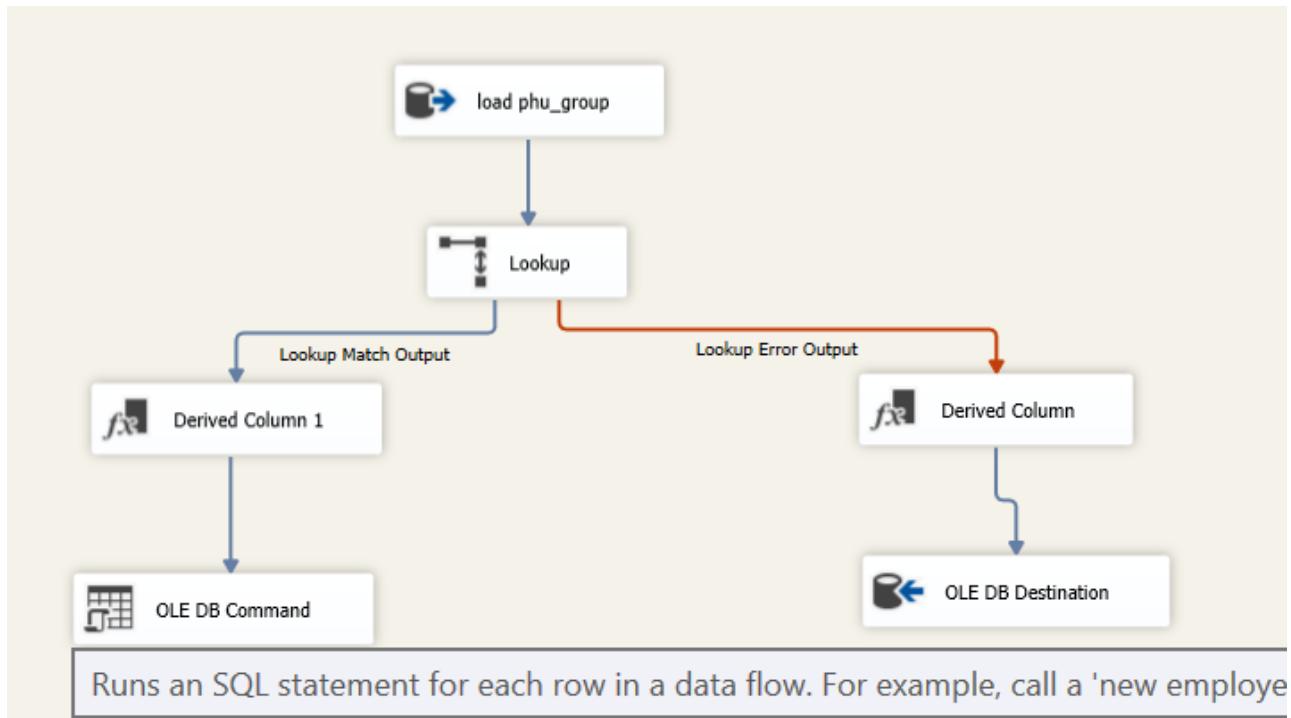
B3: Thêm nguồn dữ liệu

B4: Kiểm tra tồn tại trong NDS:

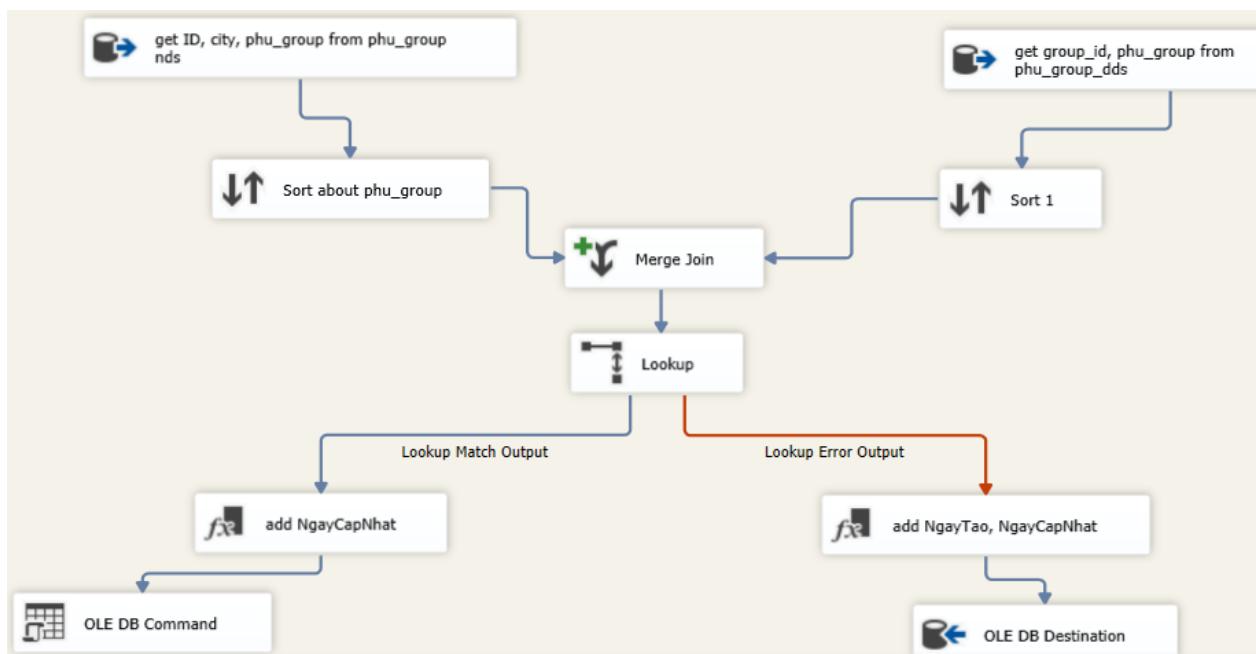
- Nếu có, cập nhật lại dữ liệu và ngày cập nhật
- Nếu không có, thêm thuộc tính ngày tạo và ngày cập nhật và tiến hành thêm mới dữ liệu

c. NDS to DDS

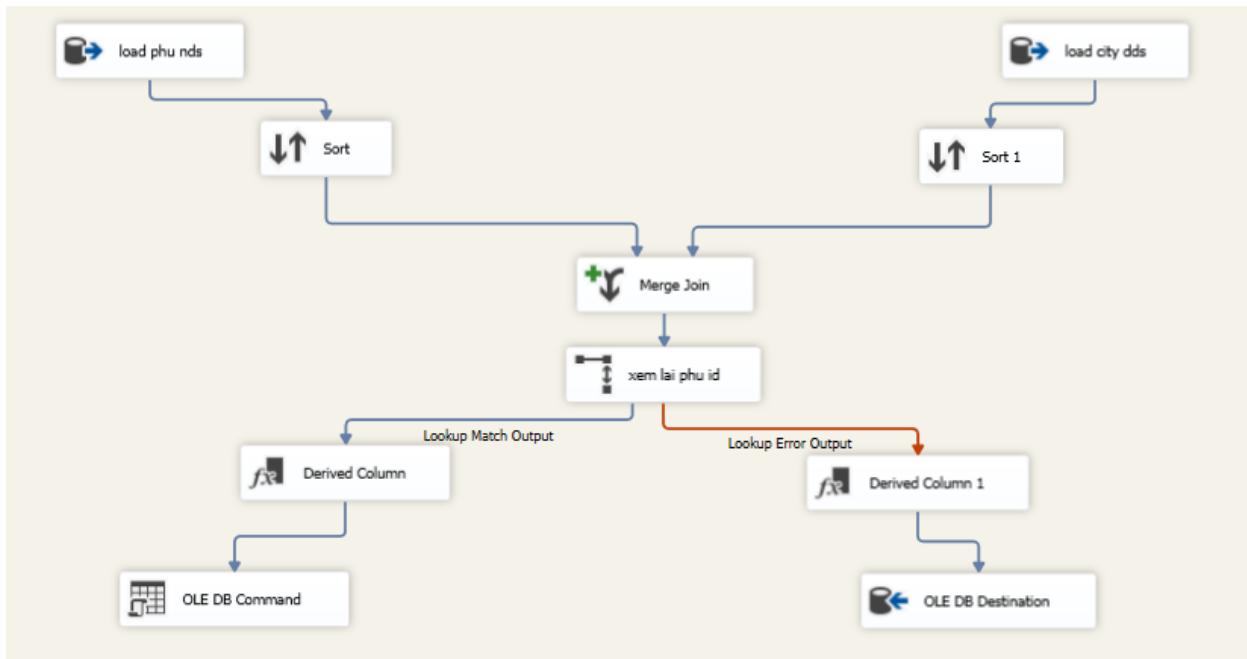
- Phân loại mức độ nghiêm trọng: tính theo tỉ lệ CFR = tử vong(Deceased) / tổng số ca bệnh(Infected)
 - + Cấp độ 1: <0.025
 - + Cấp độ 2: <0.05
 - + Cấp độ 3: <0.2
 - + Cấp độ 4: >=0.2
- Phân cấp Phu-> Phu city -> Phu group
- Phân cấp dimdate sẽ làm trong ssas
- Tạo bảng phu_group_dds theo bảng phu_group trong nds



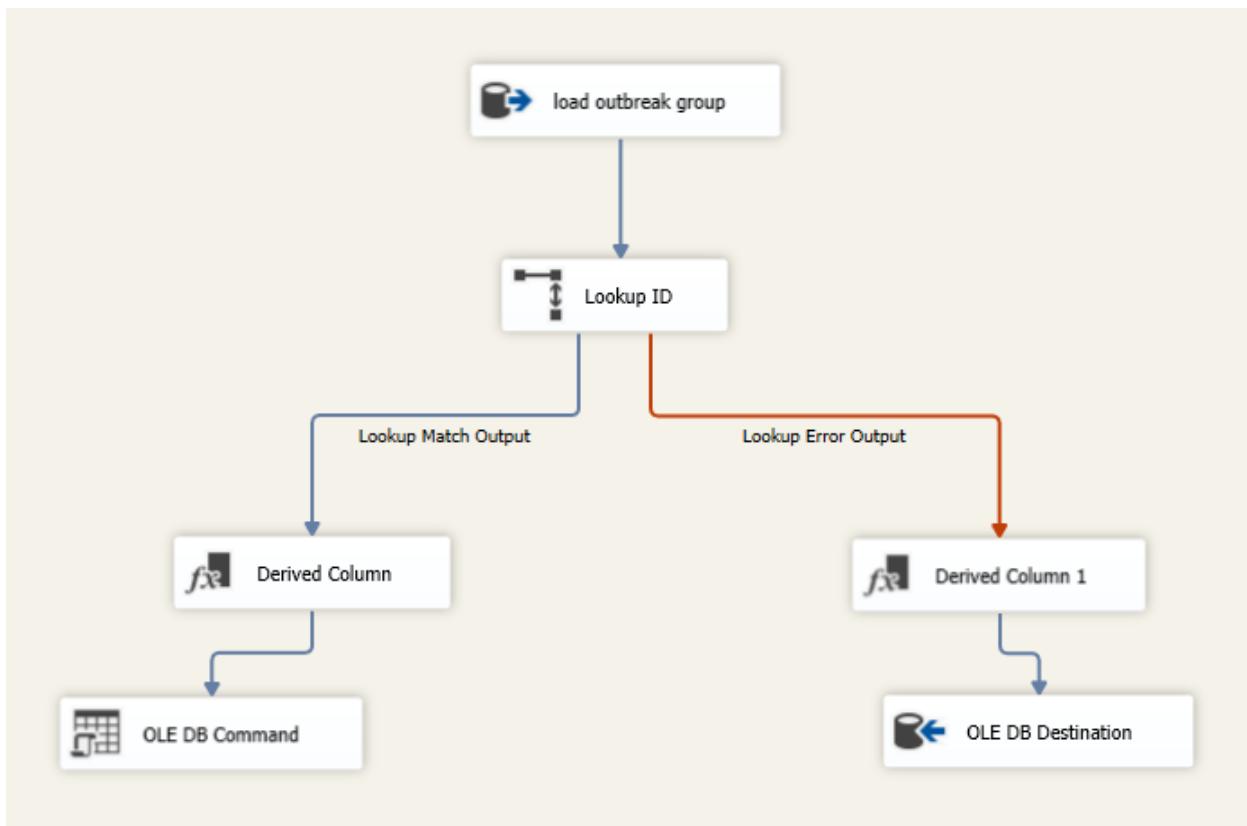
- Tạo bảng city từ bảng phu_group_dds trong nds và bảng phu trong dds



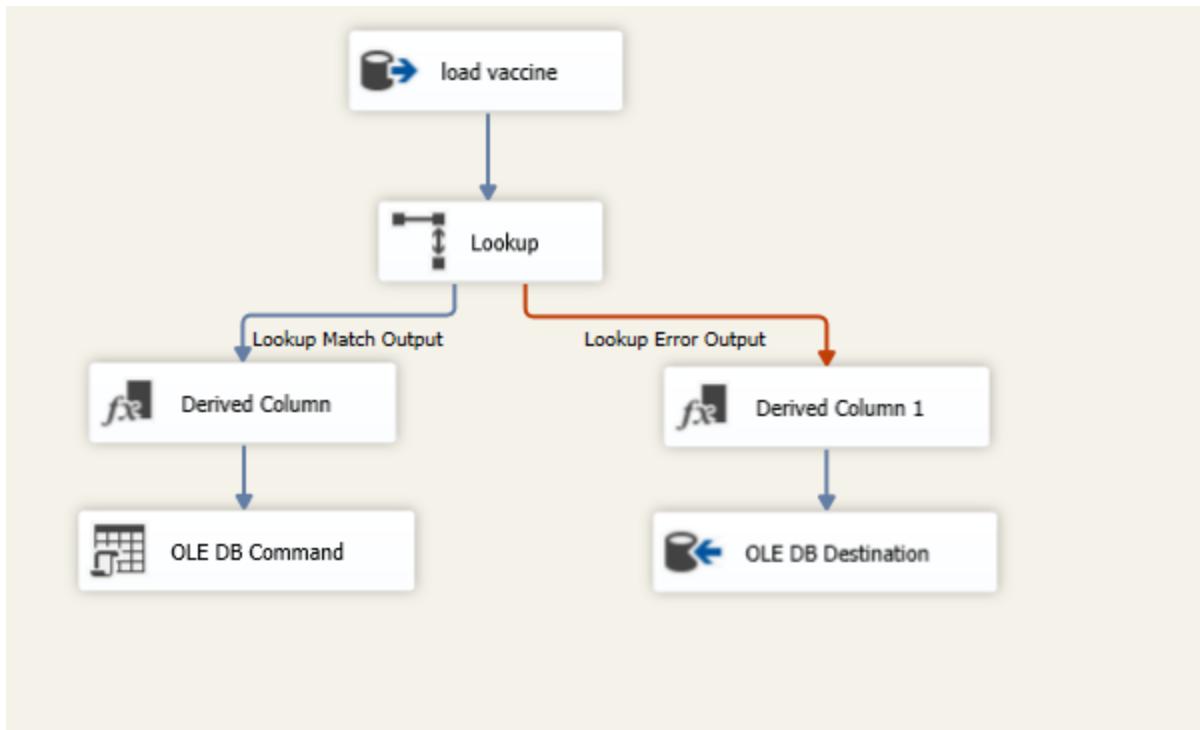
- Tạo bảng phu_dds dựa vào bảng phu của dds và join với bảng city có khóa ngoại đến bảng city để lấy city_ID



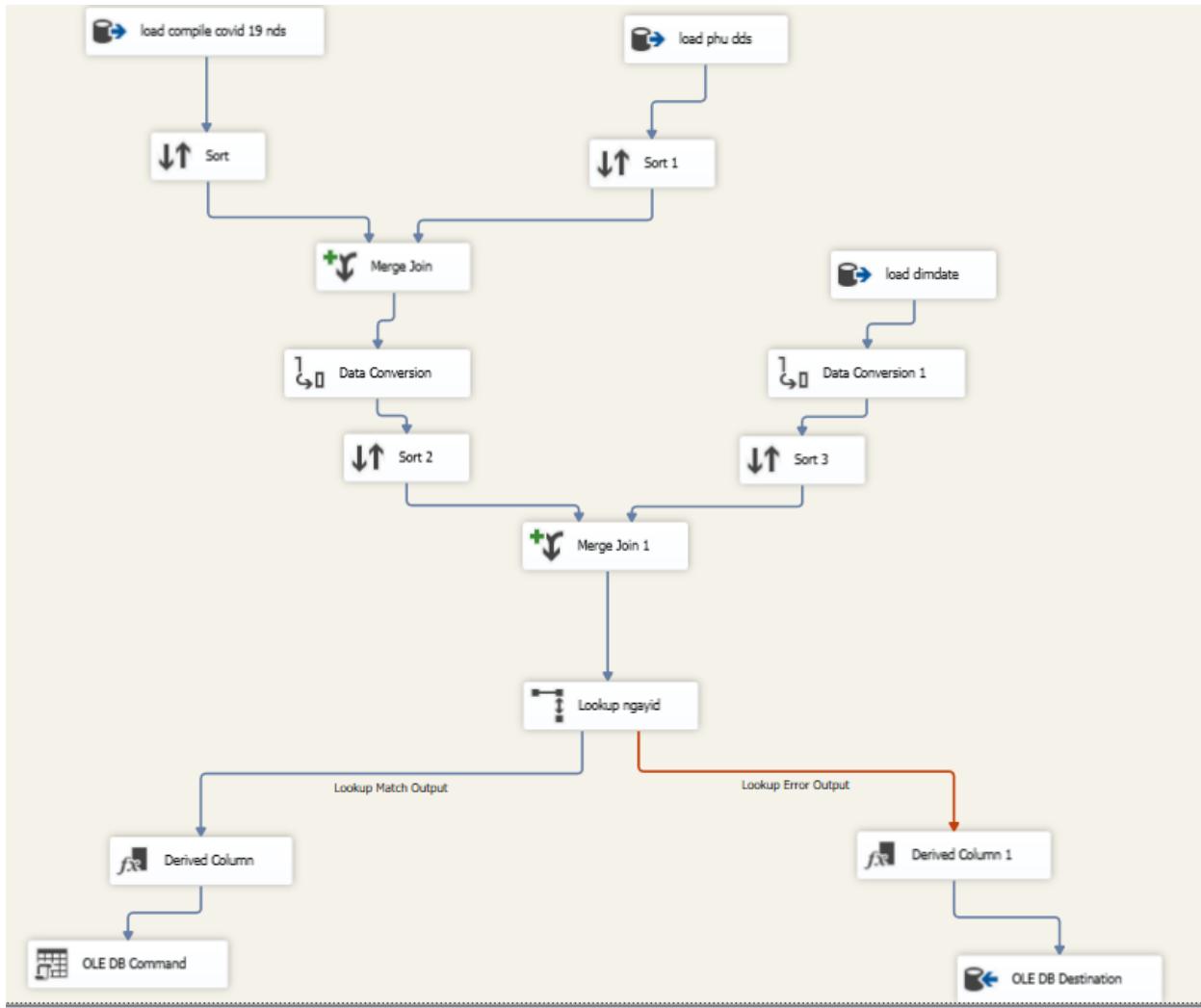
- Load bảng Ongoing_Outbreaks_DDS từ bảng Ongoing_Outbreaks trong NDS



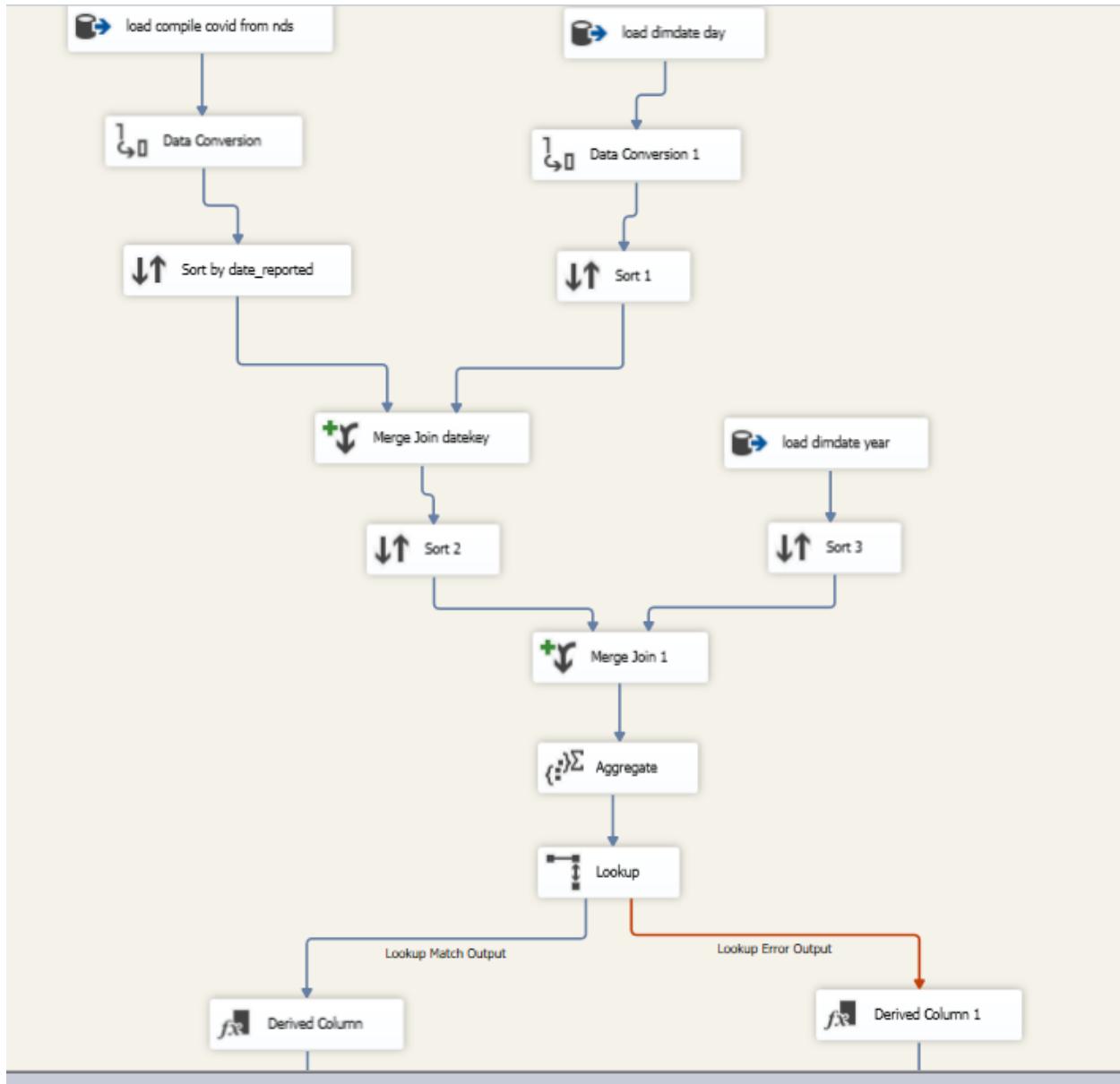
- Load bảng Vaccines_By_Age_PHU từ NDS sang DDS



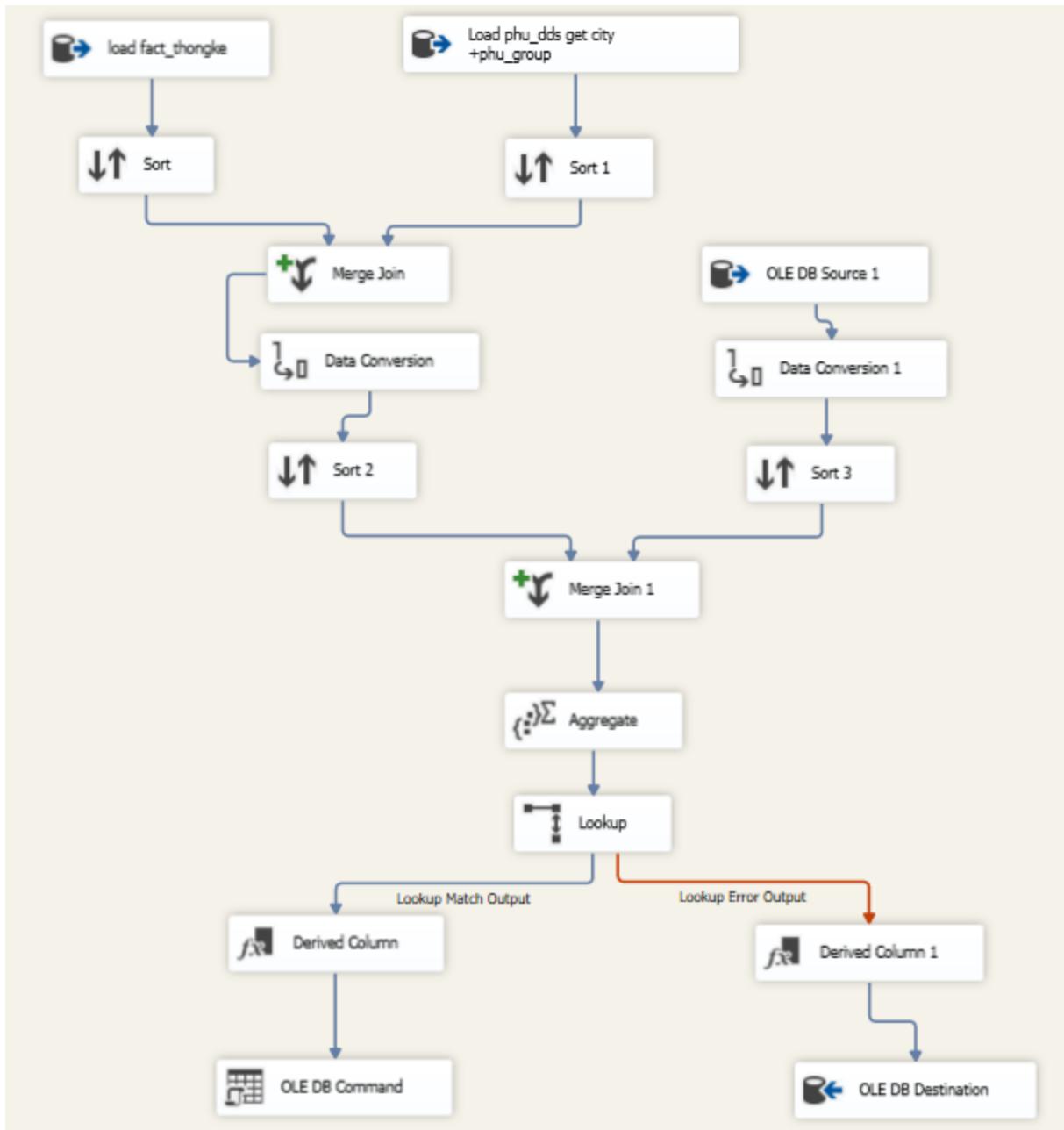
- Tạo bảng fact_thong_ke từ bảng compile covid 19 từ NDS và phu_dds, dimdate từ DDS
 - + Từ bảng compile ta đếm số ca tử vong, hồi phục, số ca nhiễm theo ngày và phu_id
 - + Sau đó lấy reporting_phu từ phu_dds theo phu_id
 - + Chuyển ngày theo dimdate
 - + Kiểm tra tồn tại để biết update hay insert



- Tạo bảng fact_gender_age từ bảng compile covid 19 của NDS và bảng dimdate của DDS
 - + Load bảng compile covid 19: lấy các thông số cần thiết với điều kiện các số đó phải có dữ liệu
 - + Inner join với dimdate
 - + Kiểm tra tồn tại để biết update hay insert



- Tạo bảng fact_outbreak_group từ bảng fact_thong_ke, phu_dds, dimdate của DDS
 - + Từ bảng fact_thong_ke ta có dữ liệu theo ngày, phu_id
 - + Sau đó ta inner join với bảng city để có (phu_group, city) là khu vực
 - + inner join với dimdate để lấy year từ day
 - + Sau đó group by để đạt được yêu cầu



B. OLAP và Report

- Thống kê Số ca nhiễm, số ca tử vong, số ca phục hồi của dịch Covid-19 theo từng PHU trong từng năm.

Row Labels	Infected - Fact Thong Ke	Deceased - Fact Thong Ke	Recovered
2226			
2020	394	6	382
2021	394	6	382
2227			
2020	3737	23	3603
2021	3737	23	3603
2233			
2020	1302	7	1284
2021	1302	7	1284
2234			
2020	2621	53	2522
2021	2621	53	2522
2235			
2020	2067	62	1922
2021	2067	62	1922
2236			
2020	17017	223	16415
2021	17017	223	16415
2237			
2020	20671	390	19797

2. Thống kê Mức Độ Nghiêm Trọng (tiêu chí nghiêm trọng sinh viên tự định nghĩa) của dịch Covid-19 theo PHU và theo các Quý trong từng năm.
- Xét mức độ nghiêm trọng theo: Deceased/ Infected

Row Labels	Deceased - Fact Thong Ke	Infected - Fact Thong Ke
2226		
2020		
1	0	9
2	0	16
3	0	11
4	1	52
2021		
1	3	150
2	2	156
2227		
2020		
1	2	56
2	3	77
3	1	83
4	3	914
2021		
1	7	1186
2	7	1421
2233		

3. Thông kê tổng số người tử vong theo Giới Tính và Nhóm Tuổi theo các năm.

Row Labels	So Nguoi
Female	
\square <20	
2020	11596
2021	24485
\square 20-29	
2020	18477
2021	29341
\square 30-39	
2020	14583
2021	23528
\square 40-49	
2020	14188
2021	21552
\square 50-59	
2020	14394
2021	20358

4. Thông kê số ca nhiễm, tử vong theo Mức Độ Nghiêm Trọng theo Ngày Trong Tháng của các năm.

Row Labels	Infected - Fact Thong Ke	Deceased - Fact Thong Ke
2020		
\square 1		
1	3	0
10	1	0
13	1	0
16	1	0
21	1	0
22	2	0
24	1	0
25	1	0
27	1	0
31	1	0
\square 2		
1	1	0
10	2	0
14	2	0
15	1	0
16	1	0
17	2	0
19	1	0
2	1	0
20	4	0
21	1	0
22	5	0

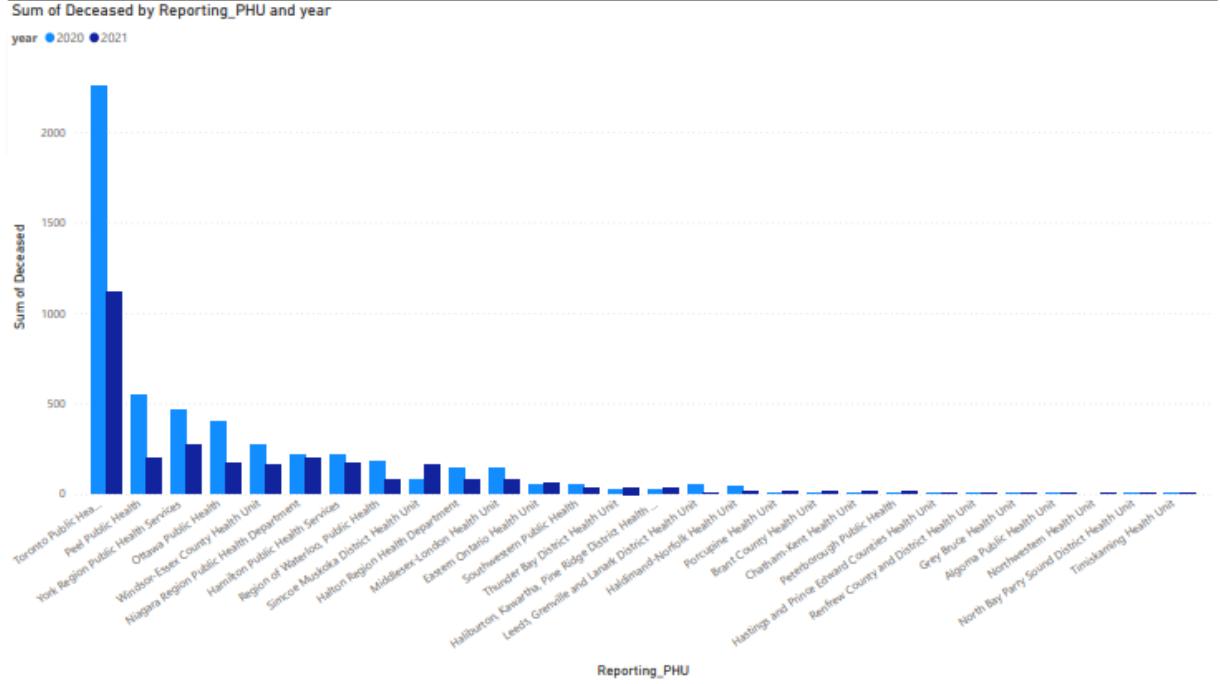
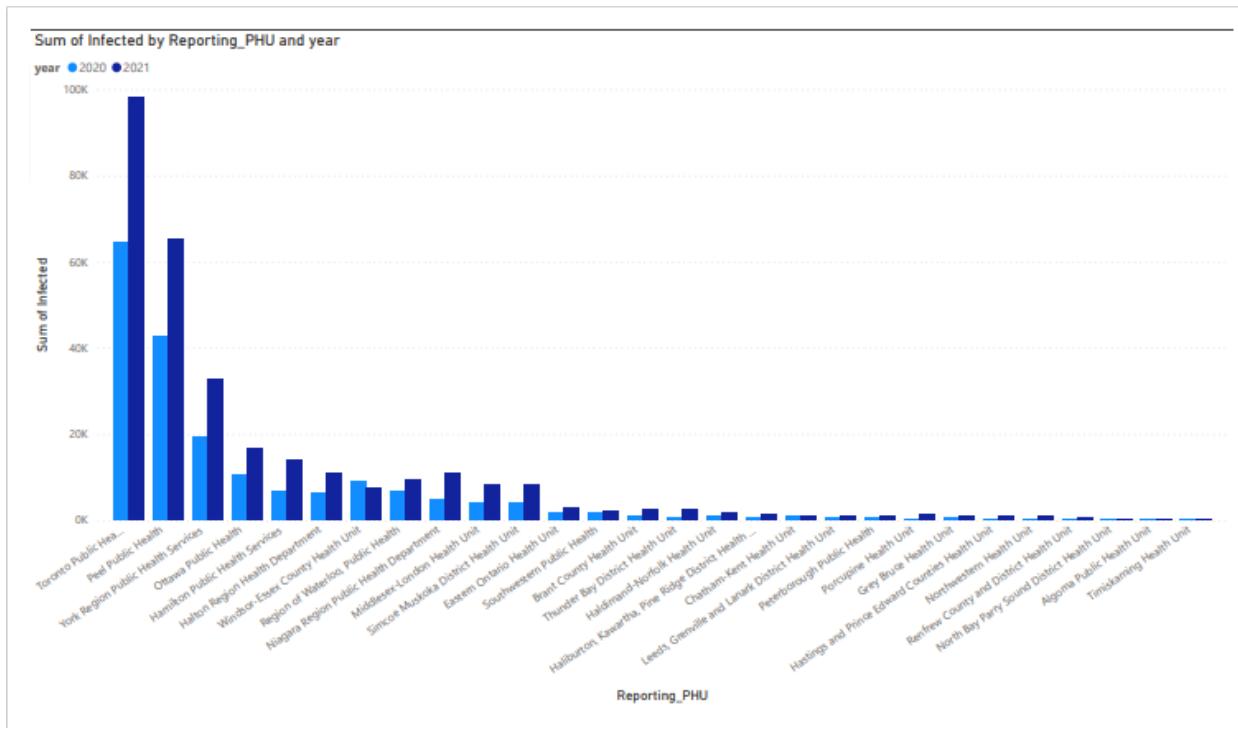
5. Thông kê số ca nhiễm, tử vong theo Mức Độ Nghiêm Trọng, khu vực (PHU_Group, City), và số người đã được tiêm vaccine trong các năm.

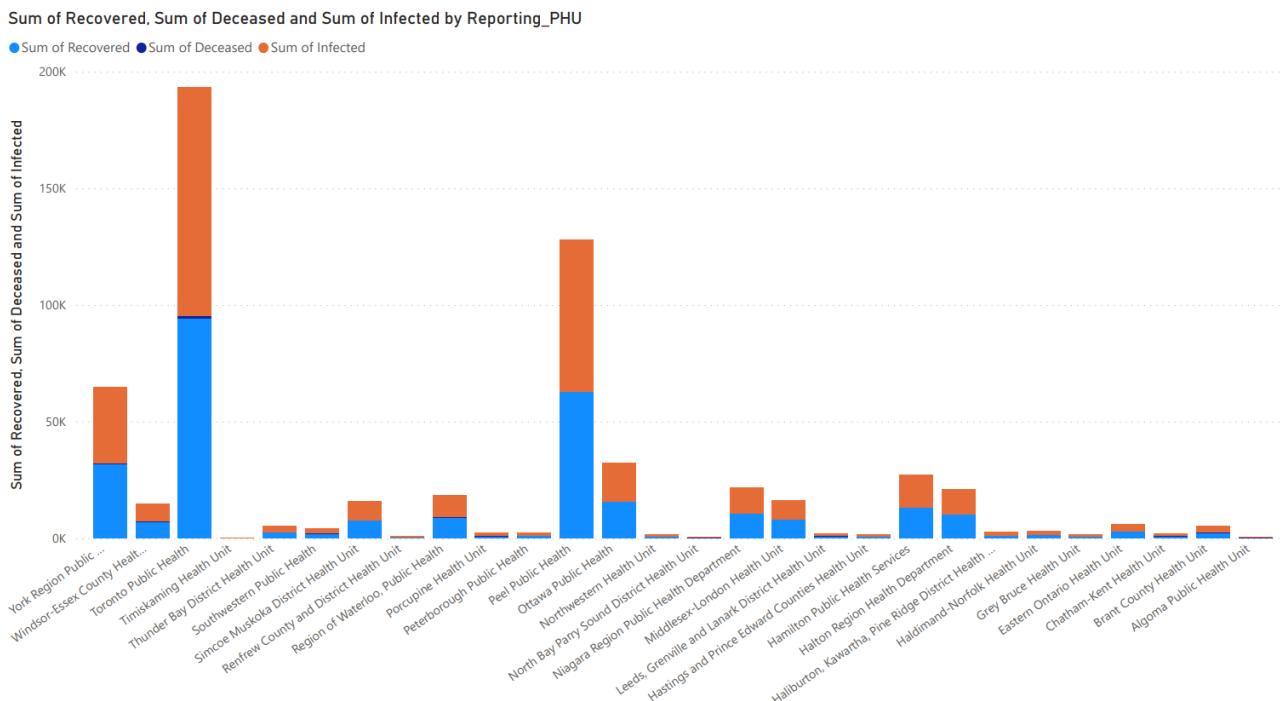
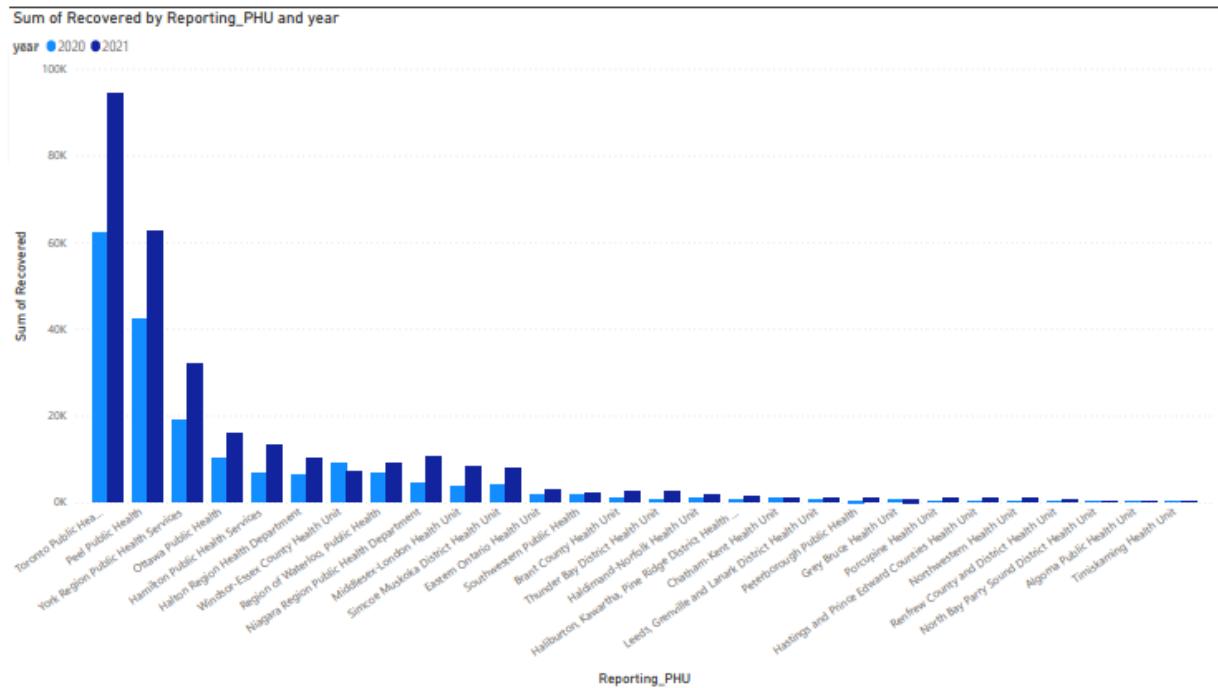
Row Labels	Infected - Fact Thong Ke	Deceased - Fact Thong Ke
⊖ 1		
⊖ 1		
3895	162625	3365
⊕ 2	52057	733
⊖ 2		
⊖ 4		
2235	2067	62
⊕ 5	1508	21
⊖ 4		
⊖ 7		
2258	4613	116
⊖ 8		
2243	1734	60
⊖ 9		
2251	27163	573
⊕ 10	719	9

B. Report:

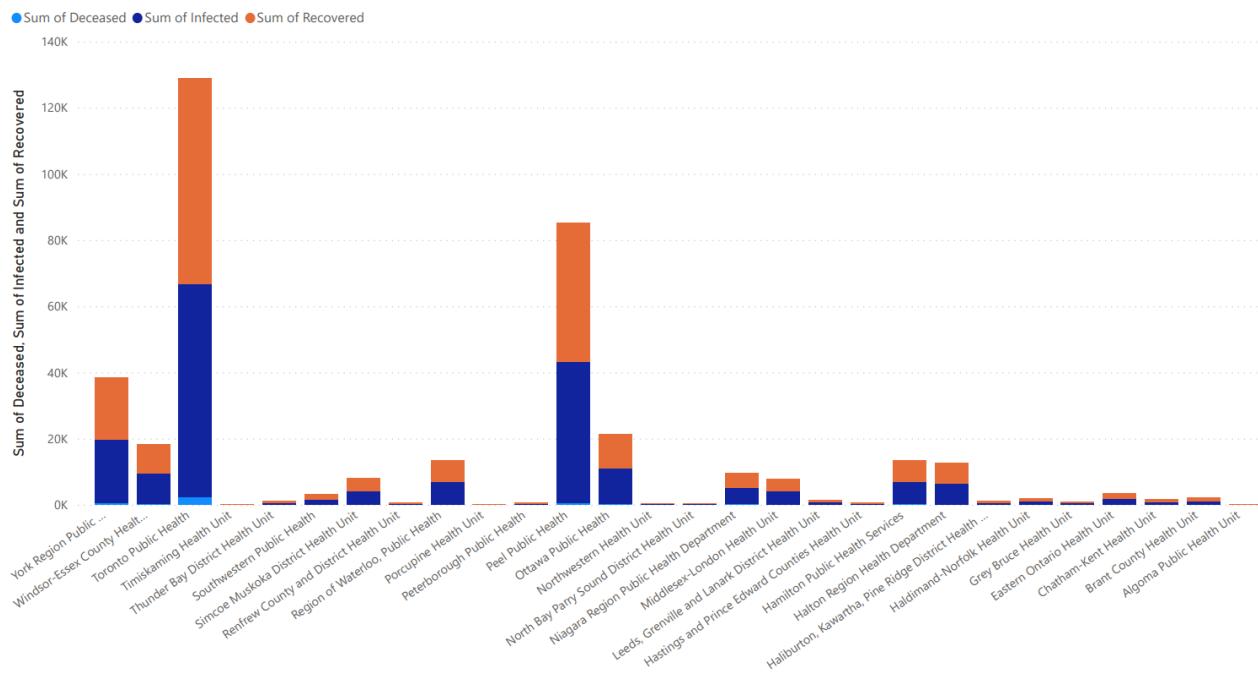
1. Thông kê Số ca nhiễm, số ca tử vong, số ca phục hồi của dịch Covid-19 theo từng PHU trong từng năm.

year Reporting_PHU	2020			2021			Total		
	Sum of Recovered	Sum of Infected	Sum of Deceased	Sum of Recovered	Sum of Infected	Sum of Deceased	Sum of Recovered	Sum of Infected	Sum
Timiskaming Health Unit	78	79	1	124	126	2	202	205	
Algoma Public Health Unit	87	88	1	295	306	5	382	394	
North Bay Parry Sound District Health Unit	152	153	1	281	301	3	433	454	
Renfrew County and District Health Unit	265	267	2	424	452	7	689	719	
Northwestern Health Unit	199	199		848	865	6	1047	1064	
Hastings and Prince Edward Counties Health Unit	301	306	5	794	816	6	1095	1122	
Porcupine Health Unit	128	137	9	976	1310	18	1104	1447	
Grey Bruce Health Unit	507	508	1	777	794	6	1284	1302	
Peterborough Public Health	395	400	5	1046	1108	16	1441	1508	
Leeds, Grenville and Lanark District Health Unit	655	709	54	1012	1025	6	1667	1734	
Chatham-Kent Health Unit	822	826	4	1002	1035	17	1824	1861	
Haliburton, Kawartha, Pine Ridge District Health Unit	578	605	27	1344	1462	35	1922	2067	
Haldimand-Norfolk Health Unit	969	1012	43	1553	1609	10	2522	2621	
Thunder Bay District Health Unit	587	613	26	2544	2637	37	3131	3250	
Brant County Health Unit	1121	1130	9	2482	2607	14	3603	3737	
Southwestern Public Health	1595	1642	47	2076	2157	35	3671	3799	
Eastern Ontario Health Unit	1611	1663	52	2841	2950	64	4452	4613	
Simcoe Muskoka District Health Unit	3890	3971	81	7697	8129	164	11587	12100	
Middlesex-London Health Unit	3804	3945	141	8071	8359	79	11875	12304	
Niagara Region Public Health Department	4558	4773	215	10544	11056	194	15102	15829	
Region of Waterloo, Public Health	6534	6714	180	8961	9323	76	15495	16037	
Windsor-Essex County Health Unit	8890	9157	267	7117	7445	162	16007	16602	
Halton Region Health Department	6136	6279	143	10279	10738	80	16415	17017	
Hamilton Public Health Services	6544	6762	218	13253	13909	172	19797	20671	
Ottawa Public Health	10225	10625	400	15749	16538	173	25974	27163	
Total	183796	189002	5199	290861	303110	2968	474657	492112	



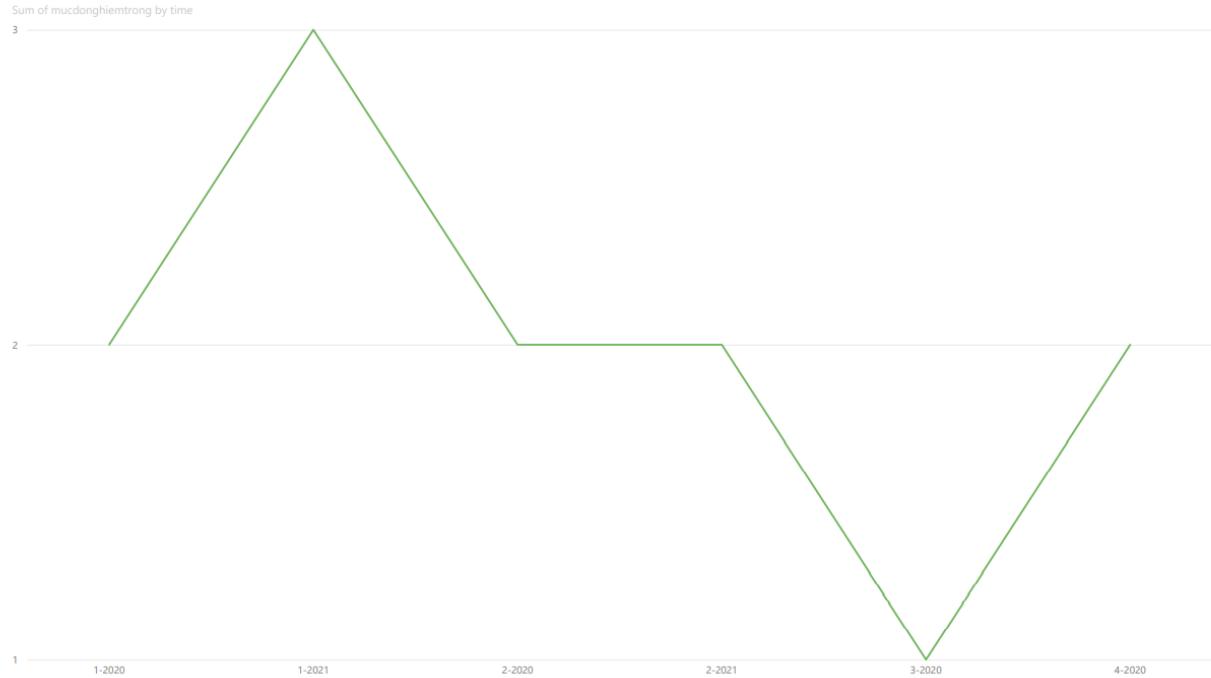


Sum of Deceased, Sum of Infected and Sum of Recovered by Reporting_PHU

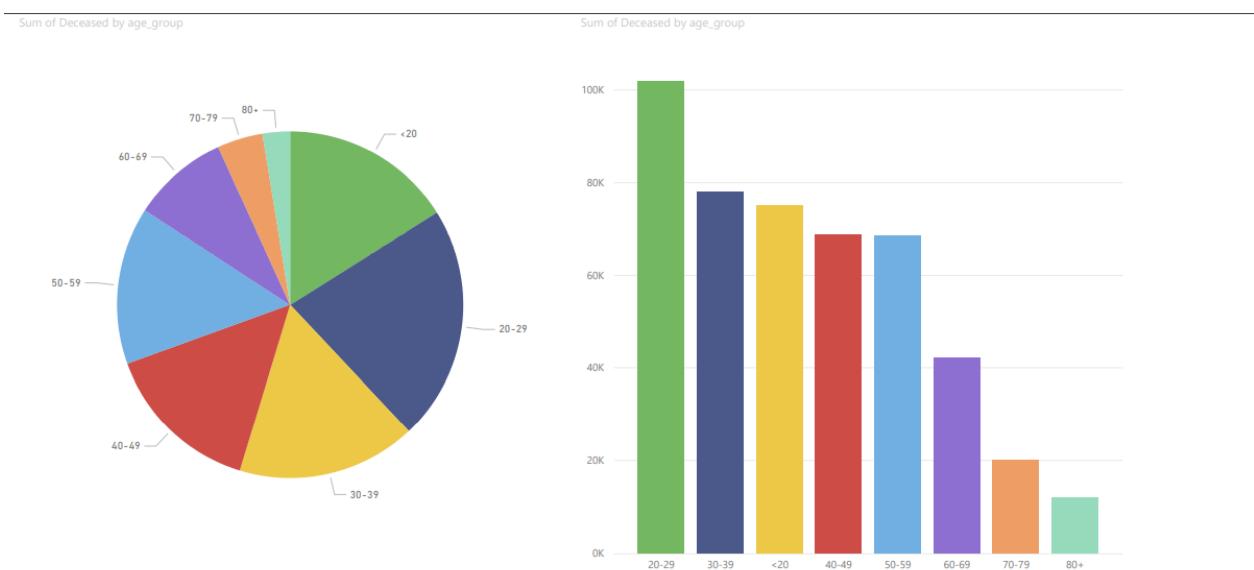


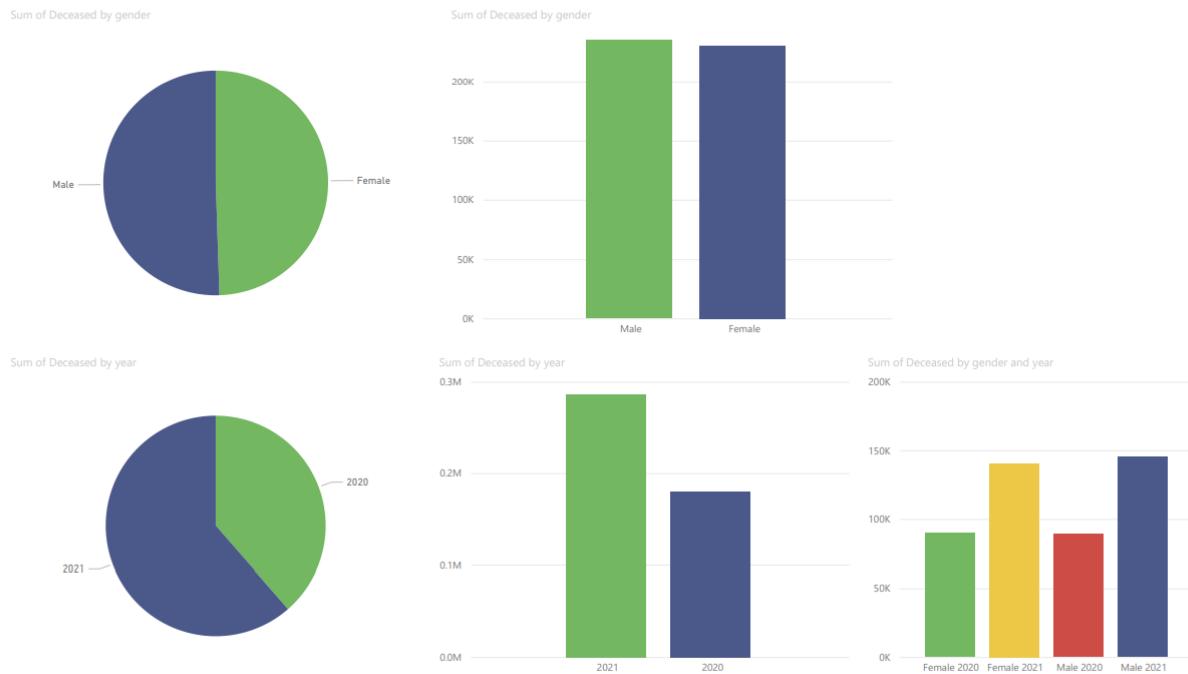
2. Thống kê Mức Độ Nghiêm Trọng (tiêu chí nghiêm trọng sinh viên tự định nghĩa) của dịch Covid-19 theo PHU và theo các Quý trong từng năm.



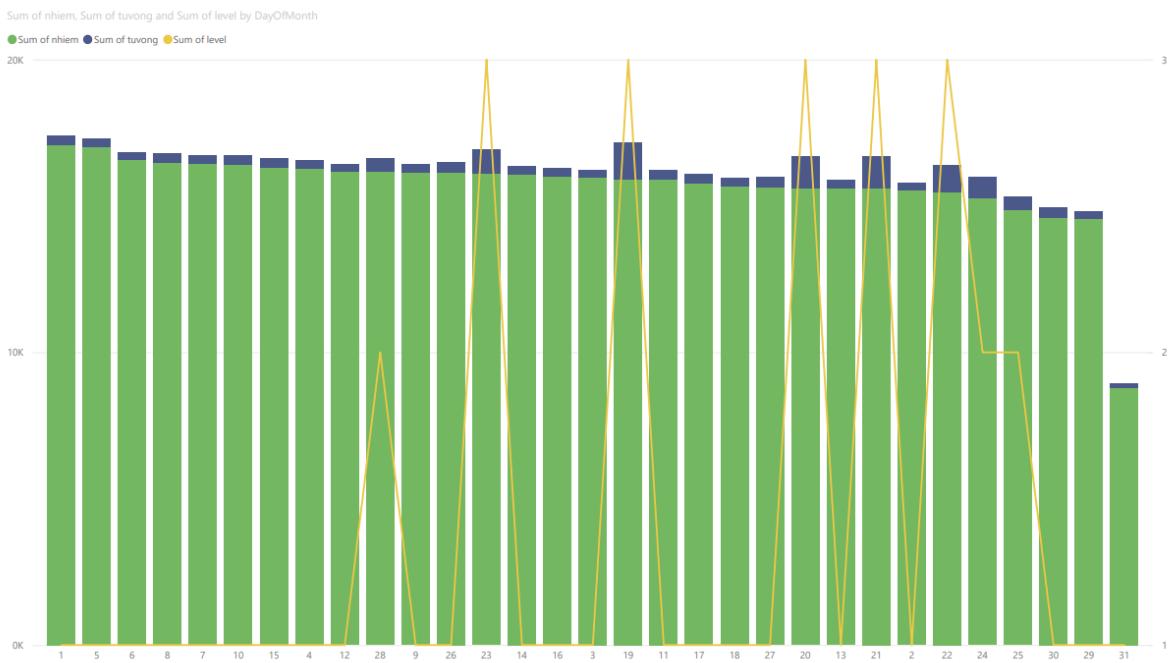


3. Thống kê tổng số người tử vong theo Giới Tính và Nhóm Tuổi theo các năm.

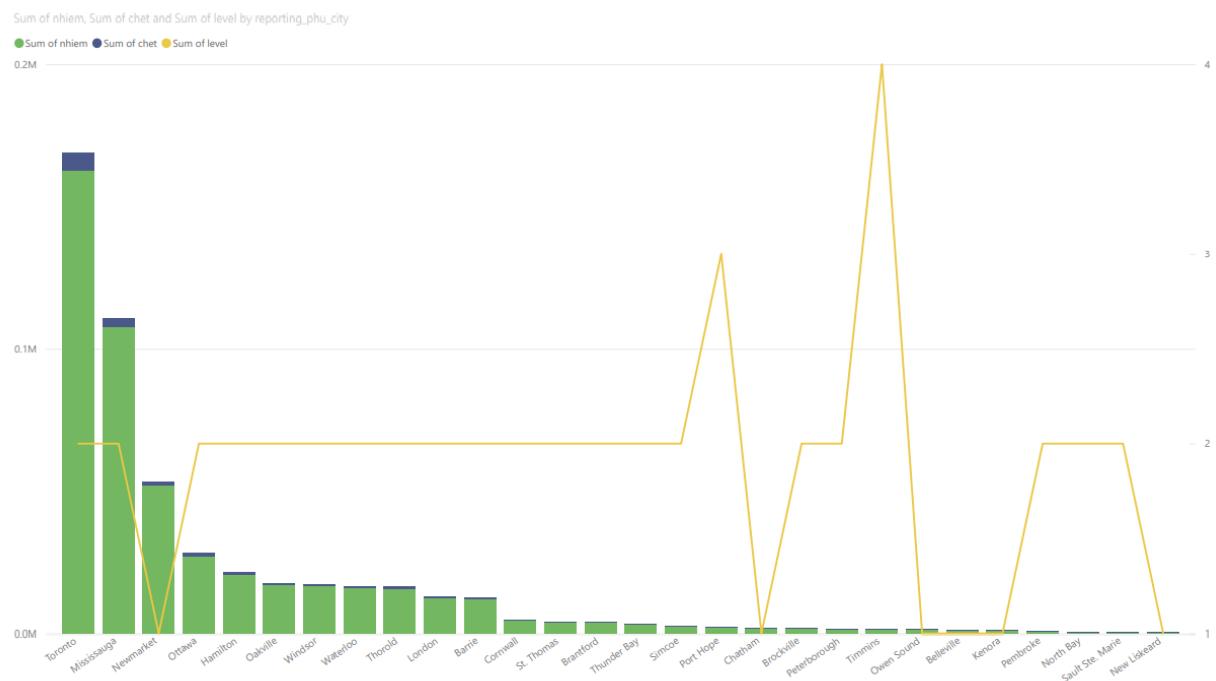




4. Thống kê số ca nhiễm, tử vong theo Mức Độ Nghiêm Trọng theo Ngày Trong Tháng của các năm.



5. Thống kê số ca nhiễm, tử vong theo Mức Độ Nghiêm Trọng, khu vực (PHU_Group, City), và số người đã được tiêm vaccine trong các năm.



9.

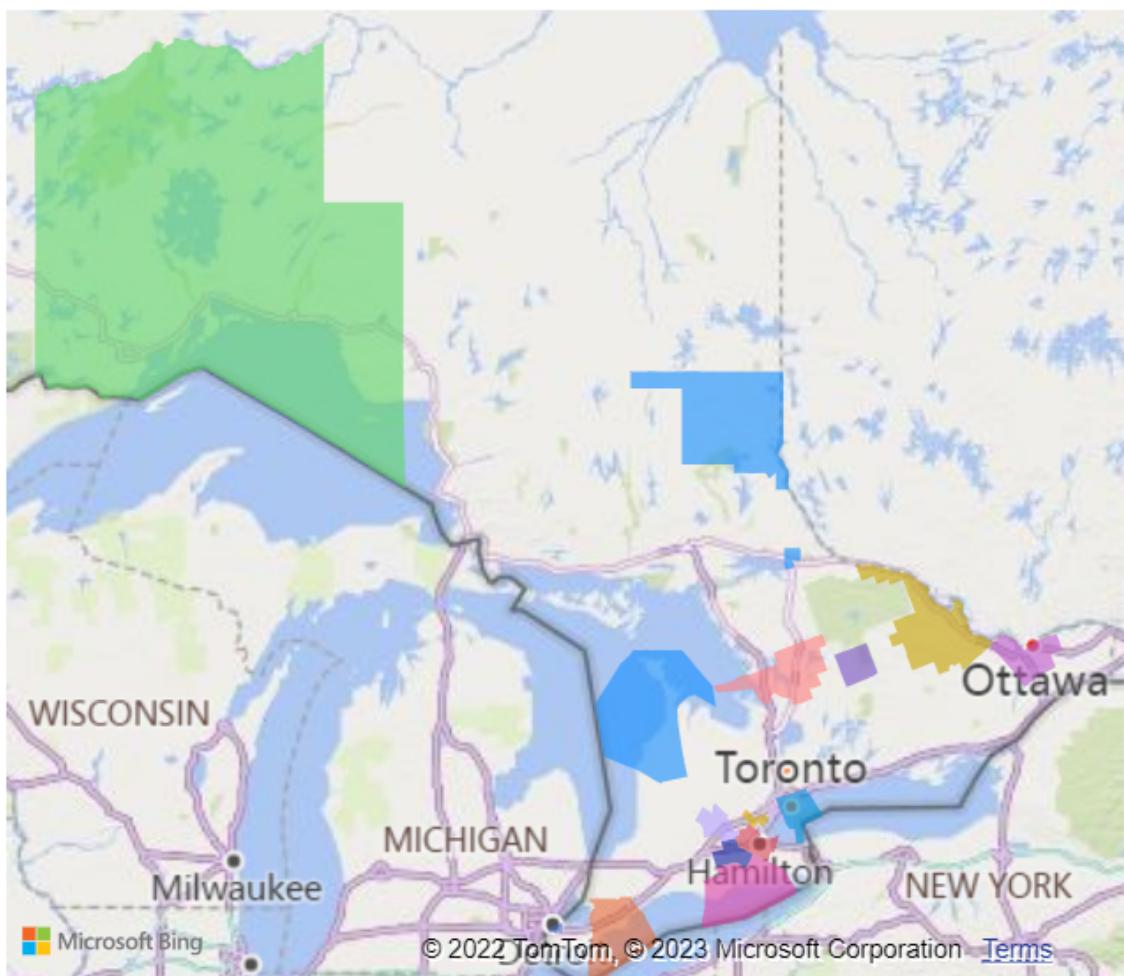
Số người nhiễm năm 2020

Infected ● 79 ● 88 ● 137 ● 153 ● 199 ● 267 ● 306 ● 400

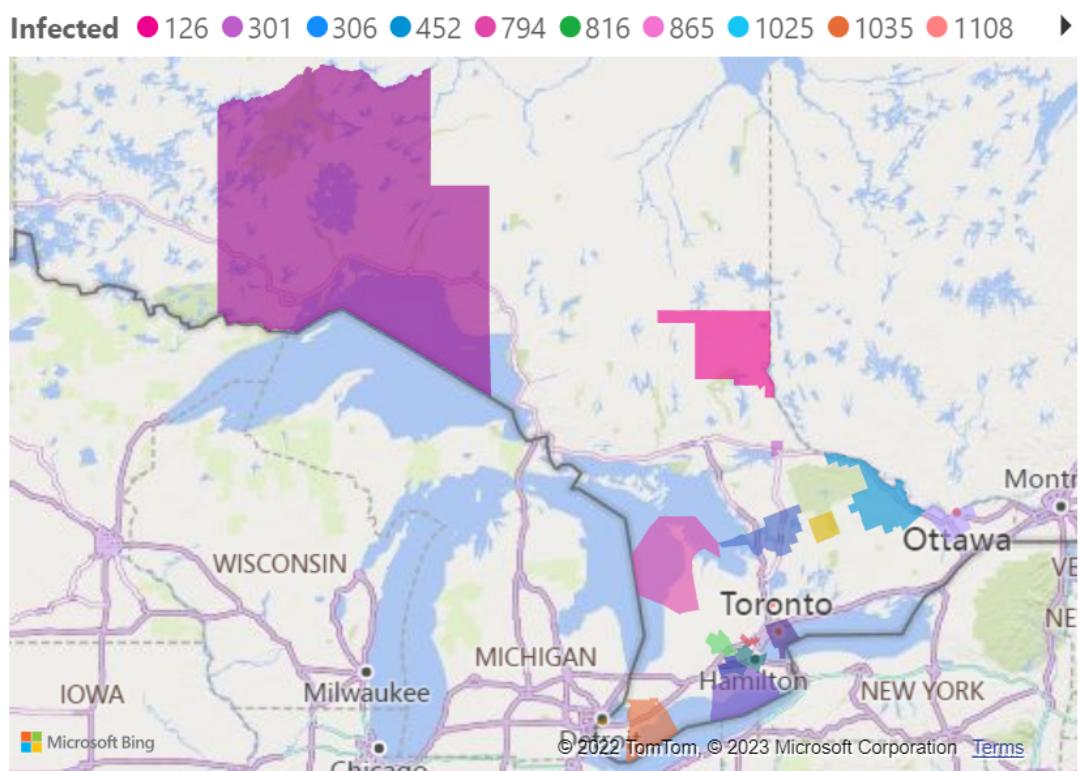


Số người tử vong năm 2020

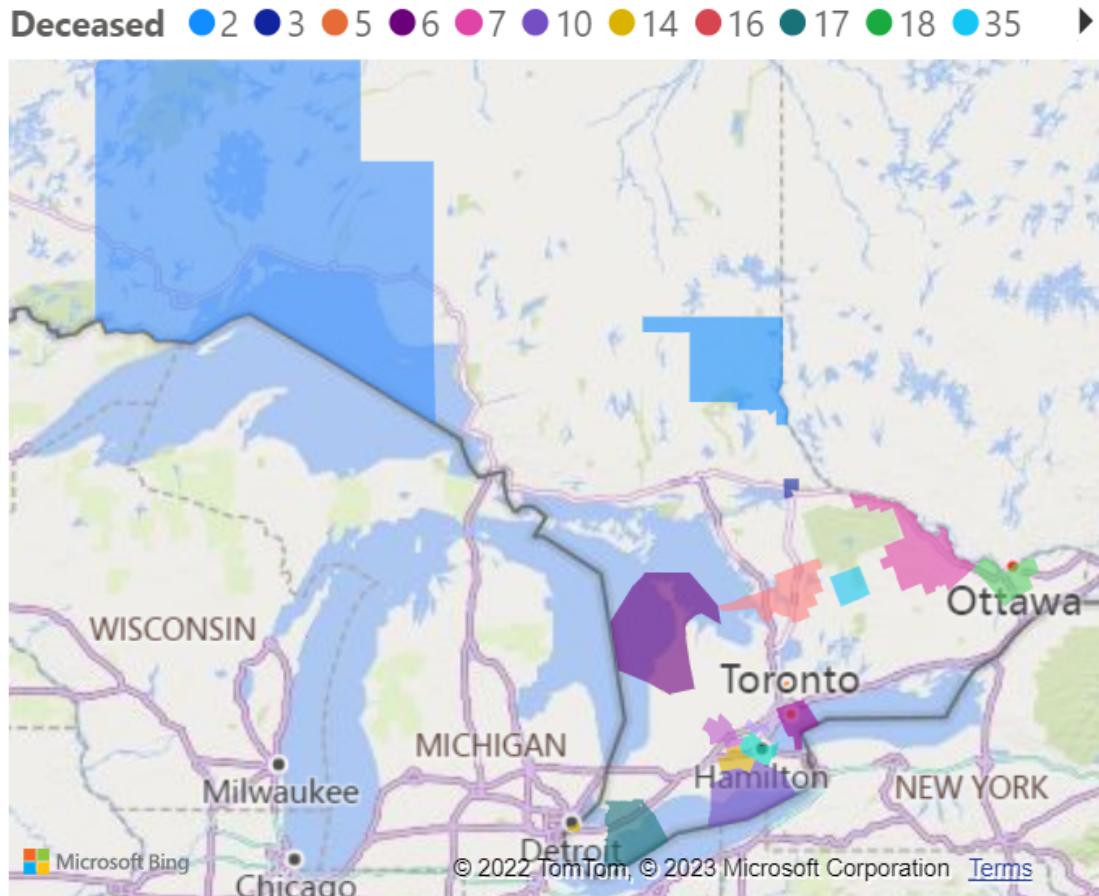
Deceased ● (Blank) ● 1 ● 2 ● 4 ● 5 ● 9 ● 26 ● 27 ● 43



Số người nhiễm năm 2021



Số người tử vong năm 2021



C. Data Mining

- Nguồn dữ liệu:
 - + Database: NDS
 - + Table: Compiled_COVID19_Case_Details_Canada
- Đặt vấn đề: dự đoán tỷ lệ tử vong của người mắc bệnh Covid-19
- Các trường cần lấy: **age_group, gender, exposure, case_status**
- 1. Sử dụng mô hình decision tree
 - Lấy các cột cần thiết: **age_group, gender, exposure, case_status**

```
new_df =
df[['age_group', 'gender', 'exposure', 'case_status']]
```

 - Ta chuyển dữ liệu cột **case_status** về giá trị **Alive** và **Death**

```

case_status_encoder = { 'Recovered': 'Alive',
                       'Active': 'Alive',
                       'Deceased': 'Death',
                     }

new_df['case_status'] =
new_df['case_status'].replace(case_status_encoder)

- Chuyển các giá trị ở các cột còn lại về dạng số( dùng preprocessing của
sklearn)

for column in new_df:

    if column == 'case_status':
        continue

    le = preprocessing.LabelEncoder()

    le.fit(new_df[column].unique())

    new_df[column] = le.transform(new_df[column])

- Chia tập train và tập test (70% train và 30% test)

X = new_df[['age_group', 'gender', 'exposure']]

y = new_df['case_status']

X_train, X_test, y_train, y_test = train_test_split(X,
                                                    y, test_size = 0.7, random_state = 1)

- Xây dựng cây quyết định

# Create Decision Tree classifier object

clf = DecisionTreeClassifier()

# Train Decision Tree Classifier

clf = clf.fit(X_train, y_train)

# Predict the response for test dataset

y_pred = clf.predict(X_test)

```

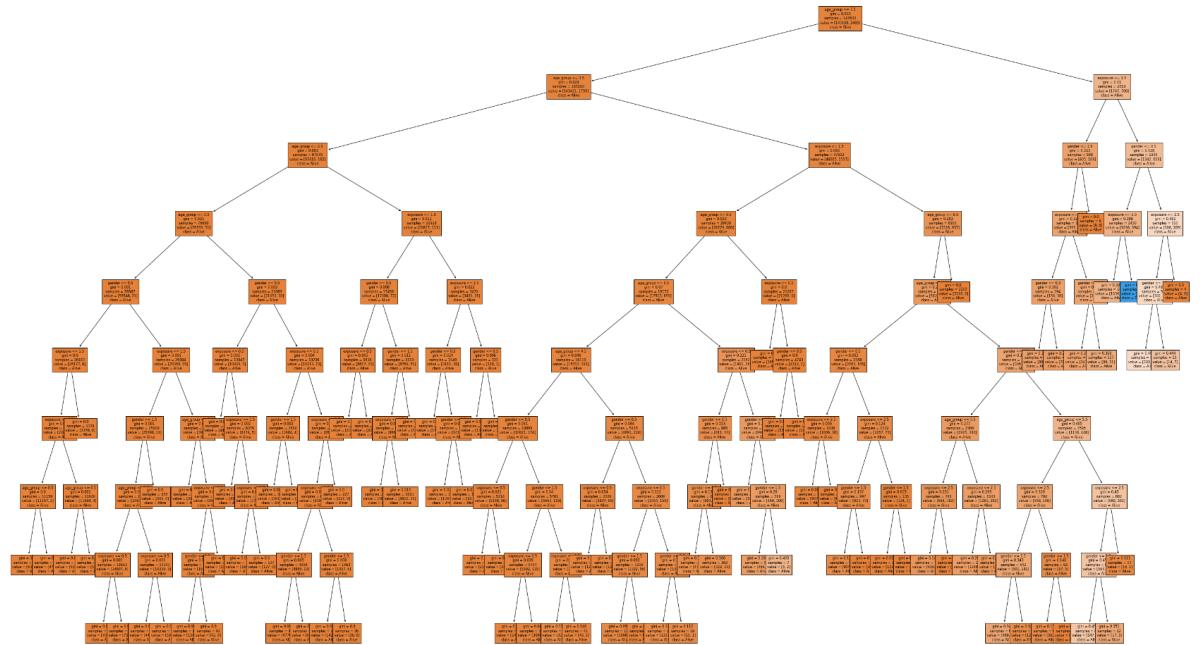
```
score = accuracy_score(y_test, y_pred)
```

2. Visualize

```
plt.figure(figsize=(50, 30))

plot_tree(clf, filled=True,
feature_names=['age_group','gender','exposure'],class_names
=new_df['case_status'].unique() ,fontsize=8)

plt.show()
```



3. Kết quả

```
score = accuracy_score(y_test, y_pred)
```

- Độ chính xác của mô hình ~98,3%. Độ chính xác cao như vậy vì độ phân bố giữa thuộc tính Alive và Death có sự chênh lệch cao.