

ĐẠI HỌC KHOA HỌC TỰ NHIÊN – ĐHQG TPHCM

KHOA CÔNG NGHỆ THÔNG TIN

CHUYÊN NGÀNH HỆ THỐNG THÔNG TIN

-----o0o-----

ĐỒ ÁN THỰC HÀNH

HỆ THỐNG THÔNG TIN PHỤC VỤ TRÍ TUỆ KINH DOANH

XÂY DỰNG VÀ KHAI THÁC KHO DỮ LIỆU



|SINH VIÊN THỰC HIỆN|

Lớp: 19HTTT2 - Nhóm 2

Hoàng Lê Khanh – 19127173

Nguyễn Thị Ngọc Diệu – 19127361

Bùi Đăng Khoa – 19127645

|GIẢNG VIÊN HƯỚNG DẪN|

Giảng viên lý thuyết: Hồ Thị Hoàng Vy

Giảng viên thực hành: Tiết Gia Hồng

Thành phố Hồ Chí Minh, 2022

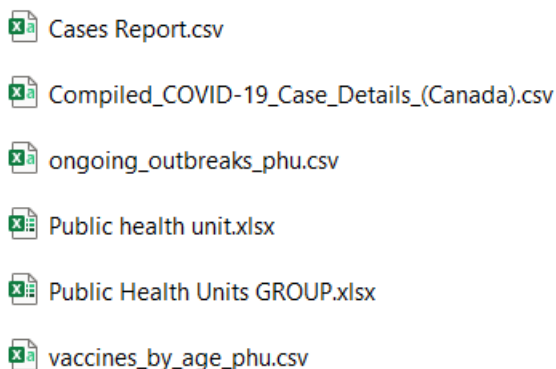
Mục lục

1. Thông tin nhóm	3
2. Phân tích dữ liệu nguồn.....	3
3. Tiền xử lý dữ liệu	13
3.1 Xử lý missing value.....	13
3.2 Xử lý dữ liệu và ý tưởng mining:	15
3.2.1 Xử lý dữ liệu	15
3.2.2 Data mining	16
4 Phân tích NDS	18
5 Phân tích DDS	27
6 System	33
6.1 Thiết kế pipeline	33
6.2 ETL.....	34
6.2.1 Overview Data Flow	34
6.2.2 Source to Stage.....	35
6.2.3 Stage to NDS.....	36
6.2.4 NDS to DDS.....	37
7 Trục quan hóa dữ liệu	39
7.1 Dashboard TotalCase	39
7.2 Dashboard TotalDeath_TotalRecovery_TotalInfection.....	40
7.3 Dashboard Overview	41
7.4 Dashboard Statistics Of Serious Level.....	43
7.5 Data Mining:.....	48
8 Tham khảo và chú thích	49

1. Thông tin nhóm

Mã nhóm	GROUP 2	
Số lượng	3	
MSSV	Họ tên	Email
19127173	Hoàng Lê Khanh	19127173@student.hcmus.edu.vn
19127361	Nguyễn Thị Ngọc Diệu	19127361@student.hcmus.edu.vn
19127645	Bùi Đăng Khoa	19127645@student.hcmus.edu.vn

2. Phân tích dữ liệu nguồn



Hình: Các file dữ liệu nguồn ban đầu

❖ Source Cases Report.csv

Cases report: Dữ liệu ca nhiễm ở tỉnh bang Ontario		
Tên thuộc tính	Mô tả	Data type
Outcome	Kết quả của ca nhiễm: - Resolved: Đã điều trị - Not Resolved: Chưa điều trị - Fatal: Tử vong	Varchar

Age	Độ tuổi của đối tượng, bao gồm: <ul style="list-style-type: none"> - < 20 (Dưới 20 tuổi) - 20s, 30s,... 80s (trong khoảng 20 – 29, 30 – 39,...) - 90+ (Trên 90 tuổi) 	Varchar
Gender	Giới tính đối tượng: <ul style="list-style-type: none"> - FEMALE: nữ - MALE: nam - GENDER DIVERSE: đa giới tính - UNSPECIFIED: chưa xác định 	Varchar
Reporting PHU	Các đơn vị Public Health Unit ghi nhận các báo cáo ca nhiễm	Varchar
SpecimenDate	Ngày lấy mẫu thử	Date
CaseReported Date	Ngày ghi nhận kết quả ca nhiễm	Date
PHUCity	Thành phố của PHU	Varchar
TestReported Date	Ngày trả kết quả test	Date
CaseAcquisition info	Thông tin của ca nhiễm: <ul style="list-style-type: none"> - CC: (Closed contact) Dương tính, xác định được nguồn lây - NO KNOW EPI LINK: Dương tính, không xác định được nguồn lây - OB: (Outbreak) Bùng phát - TRAVEL 	Varchar

	<ul style="list-style-type: none"> - UNSPECIFIED EPI LINK: Dương tính, nguồn lây chưa được xác thực - MISSING INFORMATION: Chưa xác định 	
AccurateEpisode Dt	Ngày khởi phát	Date
PHU Address	Địa chỉ PHU được báo cáo	Varchar
PHU Website	Website PHU được báo cáo	Varchar
OutbreakRelated	Có liên quan đến đợt bùng phát hay không	Varchar
PHU Latitude	Vĩ tuyến PHU	Float
PHU Longitude	Kinh tuyến PHU	Float
PHU Postal Code	Mã bưu điện của PHU	Varchar

⇒ Bảng dữ liệu ghi nhận báo cáo về các ca nhiễm, tuy nhiên không ghi nhận đối tượng cụ thể (cmnd/cccd định danh người) nên dữ liệu hoàn toàn có thể xảy ra trường hợp duplication.

../Datasource/Cases Report.csv	
Outcome	28500
Age	28500
Gender	28500
Reporting PHU	28500
SpecimenDate	28292
CaseReported Date	28500
PHUCity	28500
TestReported Date	28082
CaseAcquisition info	28500
AccurateEpisode Dt	28500
PHU Address	28500
PHU Website	28500
OutbreakRelated	8710
PHU Latitude	28500
PHU Longitude	28500
PHU Postal Code	28500

Duplication
1528

⇒ Có thể thấy được có nhiều dòng dữ liệu có các attributes có giá trị rỗng và xuất hiện các dòng dữ liệu bị duplicate.

❖ Source ongoing_outbreaks_phu.csv

ongoing_outbreaks_phu: Dữ liệu về việc bùng phát dịch tại các đơn vị chăm sóc sức khỏe của Ontario

Tên thuộc tính	Mô tả	Data type
date	Ngày ghi nhận báo cáo	Date

<i>phu_num</i>	Định danh của đơn vị chăm sóc y tế cộng đồng	bigint
outbreak_group	Nhóm khu vực bùng phát dịch: <ul style="list-style-type: none"> - 1 Congregate Care - Chăm sóc cộng đồng - 2 Congregate Living - Lưu trú cộng đồng - 3 Education - Giáo dục - 4 Workplace - Nơi làm việc - 5 Recreational - Cơ sở giải trí - 6 Other/Unknown - Không xác định 	Varchar
number_ongoing_outbreaks	Số đợt bùng phát đang diễn ra	Int

⇒ Bảng dữ liệu ghi nhận báo cáo về số đợt bùng dịch ở các đơn vị chăm sóc sức khỏe. Dữ liệu được lưu trữ theo ngày, PHU và theo loại cơ sở => có thể chọn khóa chính là (**date, phu_num, outbreak_group**)

../Datasource/ongoing_outbreaks_phu.csv	
date	56987
phu_num	56987
outbreak_group	56987
number_ongoing_outbreaks	56987

⇒ Ở Datasource này ta có thể thấy không có dữ liệu bị null hay duplicate.

❖ Source vaccines_by_age_phu.csv

vaccines_by_age_phu: Dữ liệu tiêm vắc-xin tại các đơn vị chăm sóc sức khỏe của Ontario

Tên thuộc tính	Mô tả	Data type
date	Ngày ghi nhận báo cáo	Date
<i>PHU ID</i>	Định danh của đơn vị chăm sóc y tế cộng đồng	bigint
Agegroup	<p>Nhóm tuổi, được phân loại gồm:</p> <ul style="list-style-type: none"> - 05-11yrs: 5-11 tuổi - 12-17yrs - 18-29yrs - 30-39yrs - 40-49yrs - 50-59yrs - 60-69yrs - 70-79yrs - 80+ : Từ 80 tuổi trở lên - Adults_18plus: Người lớn trên 18 tuổi - Ontario_12plus - Ontario_5plus - Undisclosed_or_missing: Chưa xác định 	Varchar
At least one dose_cumulative	Số người tiêm được ít nhất 1 mũi	bigint
Second_dose_cumulative	Số người tiêm được 2 mũi	bigint
fully_vaccinated_cumulative	Số người tiêm đủ vaccin. Tiêm đầy đủ nghĩa là:	bigint

	<ul style="list-style-type: none"> - Tiêm 1 mũi Janssen (Johnson & Johnson) - Tiêm 2 mũi trong danh mục vaccin được Bộ y tế Canada phê duyệt - Tiêm 1 mũi trong danh mục được Bộ ý tế phê duyệt + 1 mũi trong danh mục không được phê duyệt - Tiêm 3 mũi vaccin thuộc loại bất kỳ 	
third_dose_cumulative	Số người tiêm được 3 mũi	bigint

⇒ Dữ liệu tiêm vắc xin được lưu theo ngày, PHU, và tuổi của các nhóm đối tượng. Nhận thấy, với mỗi dòng dữ liệu thì nhóm các giá trị này là duy nhất, do đó có thể chọn khóa chính là **(date, phu_id, agegroup)**.

../Datasource/vaccines_by_age_phu.csv		
Date		171182
PHU ID		171182
Agegroup		171182
At least one dose_cumulative		170971
Second_dose_cumulative		48642
fully_vaccinated_cumulative		121912
third_dose_cumulative		106015

⇒ Ta có thể thấy ở Datasource này không có dữ liệu duplicate và có một số attribute có dữ liệu null.

❖ Source Public health unit.xlsx

Public Health Units: Dữ liệu các đơn vị chăm sóc sức khỏe của Ontario		
Field	Kiểu dữ liệu	Ý nghĩa

PHU_ID	bigint	Mã đơn vị y tế
Reporting_PHU	Varchar	Tên đơn vị y tế
Reporting_PHU_Address	Varchar	Địa chỉ đơn vị y tế
Reporting_PHU_City	Varchar	Đơn vị y tế thuộc thành phố nào
Reporting_PHU_Postal_Code	Varchar	Mã bưu điện của đơn vị y tế (theo vùng/thành phố)
Reporting_PHU_Website	Varchar	Website của đơn vị y tế
Reporting_PHU_Longitude	float	Kinh độ của đơn vị y tế
Reporting_PHU_Latitude	float	Vĩ độ của đơn vị y tế

⇒ Hai attributes Reporting_PHU_Longitude và Reporting_PHU_Latitude giúp ta xác định được vị trí của trung tâm y tế (kinh độ, vĩ độ) để thuận tiện trong việc vận chuyển/ di chuyển.

../Datasource/Public
health unit.xlsx

PHU_ID	35
Reporting_PHU	35
Reporting_PHU_Address	35
Reporting_PHU_City	35
Reporting_PHU_Postal_Code	35
Reporting_PHU_Website	35
Reporting_PHU_Latitude	35
Reporting_PHU_Longitude	35

Duplication
1

⇒ Bảng dữ liệu chứa thông tin chi tiết các đơn vị y tế và có PHU_ID làm khóa chính, nhưng dữ liệu nguồn bị duplicate.

❖ Source Public Health Units GROUP.xlsx

Public Health Units Group: Nhóm các PHU		
Field	Kiểu dữ liệu	Ý nghĩa
PHU_Group	Varchar	Nhóm khu vực của đơn vị y tế
PHU_City	Varchar	Tên thành phố đơn vị y tế trực thuộc
PHU_region	Varchar	Tên đơn vị y tế

⇒ Theo dữ liệu nguồn có thể sử dụng cột PHU_Group + PHU_City để định danh cho nguồn dữ liệu này.

../Datasource/Public
Health Units
GROUP.xlsx

PHU_Group	14
PHU_region	14

⇒ Dữ liệu nguồn không có duplicate và missing value.

❖ Source Compiled_COVID-19_Case_Details_(Canada).csv

Compiled_COVID-19_Case_Details_(Canada): Dữ liệu ca nhiễm của tất cả các tỉnh bang ở canada		
Field	Kiểu dữ liệu	Ý nghĩa
ObjectID	bigint	Mã định danh mỗi hồ sơ
row_ID	bigint	Mã dòng
Date_reported	Datetime	Thời điểm báo cáo (Time create at)
health_region	Varchar	Tên đơn vị y tế
age_group	Varchar	Thuộc nhóm tuổi

gender	Varchar	Giới tính
exposure	Varchar	Thông tin ca nhiễm: - Close Contact - Outbreak - Not Reported - Travel-Related
case_status	Varchar	Tình trạng bệnh nhân: - Active: Đang nhiễm bệnh - Deceased: Đã chết - Recovered: Đã hồi phục - Not Reported: Chưa ghi nhận
province	Varchar	Nơi sinh sống của bệnh nhân (địa bàn/khu vực)

⇒ Chọn ObjectID làm khóa chính.

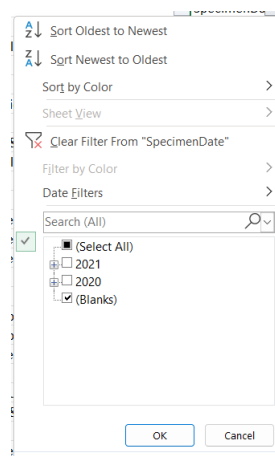
../Datasource/Compiled_COVID-19_Case_Details_(Canada).csv	
Objectid	1048575
row_id	1048575
date_reported	1048575
health_region	1048575
age_group	1048575
gender	1048575
exposure	1048575
case_status	1048575
province	1048575

⇒ Không có missing value và duplicate trong dữ liệu nguồn.

3. Tiền xử lý dữ liệu

3.1 Xử lý missing value

- Sử dụng Filter để tìm những cột có chứa các ô bị thiếu thông tin (Blanks)



❖ Cases Report

- Những cột bị thiếu dữ liệu: SpecimenDate, TestReported Date, OutbreakRelated
- Đề xuất giải pháp:
 - Quan sát bảng dữ liệu, có thể thấy có 3 cột ghi lại dữ liệu Date, trong đó cột **CaseReported Date** không bị miss dữ liệu. Xem xét ý nghĩa các cột, nhận thấy 3 cột có mối liên quan với nhau => có thể fill dữ liệu cho **SpecimenDate, TestReported Date** dựa vào **CaseReported Date**

SpecimenDate	CaseReported Date	TestReported Date
9/30/2020	12/9/2020	12/9/2020
10/17/2020	10/19/2020	10/18/2020
10/5/2020	10/9/2020	10/8/2020
10/7/2020	10/9/2020	10/9/2020
10/6/2020	10/9/2020	10/8/2020
10/6/2020	10/9/2020	10/9/2020
10/5/2020	10/9/2020	10/9/2020
10/1/2020	10/9/2020	10/9/2020
10/5/2020	10/9/2020	10/9/2020
10/6/2020	10/9/2020	10/9/2020
10/5/2020	10/9/2020	10/9/2020
10/7/2020	10/9/2020	10/9/2020
10/7/2020	10/9/2020	10/9/2020
10/2/2020	10/8/2020	10/8/2020
10/6/2020	10/8/2020	10/8/2020
10/7/2020	10/8/2020	10/8/2020
10/6/2020	10/8/2020	10/8/2020
10/7/2020	10/8/2020	10/8/2020
10/7/2020	10/8/2020	10/8/2020
10/4/2020	10/7/2020	10/7/2020
10/4/2020	10/7/2020	10/7/2020
10/6/2020	10/6/2020	10/8/2020
10/6/2020	10/5/2020	10/8/2020
10/6/2020	10/5/2020	10/8/2020
9/28/2020	10/5/2020	10/5/2020
10/1/2020	10/5/2020	10/5/2020
10/1/2020	10/5/2020	10/5/2020

- Sau khi so sánh các cột dữ liệu với nhau, nhận thấy phần lớn các dòng dữ liệu có:

+ **TestReported Date** và **CaseReported Date** giống nhau.

+ **SpecimenDate** nhỏ hơn **CaseReported Date** 2 - 4 ngày.

⇒ Do đó, có thể fill dữ liệu bằng cách:

[TestReported Date] = null? : [TestReported Date] = [CaseReported Date]

[SpecimenDate] = null? : [SpecimenDate] = [CaseReported Date] – 3 ngày

- Quan sát cột **OutbreakRelated** và xem xét ý nghĩa, nhận thấy các ô không có dữ liệu có ý nghĩa là No

OutbreakRelate
Yes
Yes

⇒ Do đó:

[OutbreakRelated] = null? : [OutbreakRelated] = No

❖ Compiled_COVID-19_Case_Details_(Canada)

Khác với các cột mang dữ liệu kiểu Date ở các bảng khác, cột **date_reported** ở bảng này mang dữ liệu datetime. Tuy nhiên time đều mang dữ liệu giống nhau.

date_reported
2020/03/23 12:00:00+00
2020/04/02 12:00:00+00
2020/03/25 12:00:00+00
2020/03/28 12:00:00+00
2020/03/30 12:00:00+00
2020/03/31 12:00:00+00
2020/03/26 12:00:00+00

⇒ Đề xuất **ép kiểu về Date** cho đồng bộ với tập dữ liệu.

3.2 Xử lý dữ liệu và ý tưởng mining:

3.2.1 Xử lý dữ liệu

❖ Biến đổi age:

```
df_case_report = pd.read_csv("../Datasource/Cases Report.csv")
unique_age = df_case_report['Age'].unique()
df_case_report_age = pd.DataFrame(unique_age, columns = ['Age'])

def func_age(x):
    start_age = 0
    end_age = 1000
    if x.Age[0].isnumeric() == True:
        start_age = int(x.Age[0]) * 10
    elif x.Age[0] == '<':
        end_age = 20

    if x.Age[2] == 's':
        end_age = start_age + 9
    x['start_age'] = start_age
    x['end_age'] = end_age
    return x

df_case_report_age = df_case_report_age.apply(func_age,axis=1)

df_case_report_age
```

	Age	start_age	end_age
0	80s	80	89
1	<20	0	20
2	20s	20	29
3	30s	30	39
4	40s	40	49
5	70s	70	79
6	60s	60	69

⇒ Tương tự với source complied_covid19 và **những bảng khác**, sau đó sẽ merge lại để sử dụng dễ dàng.

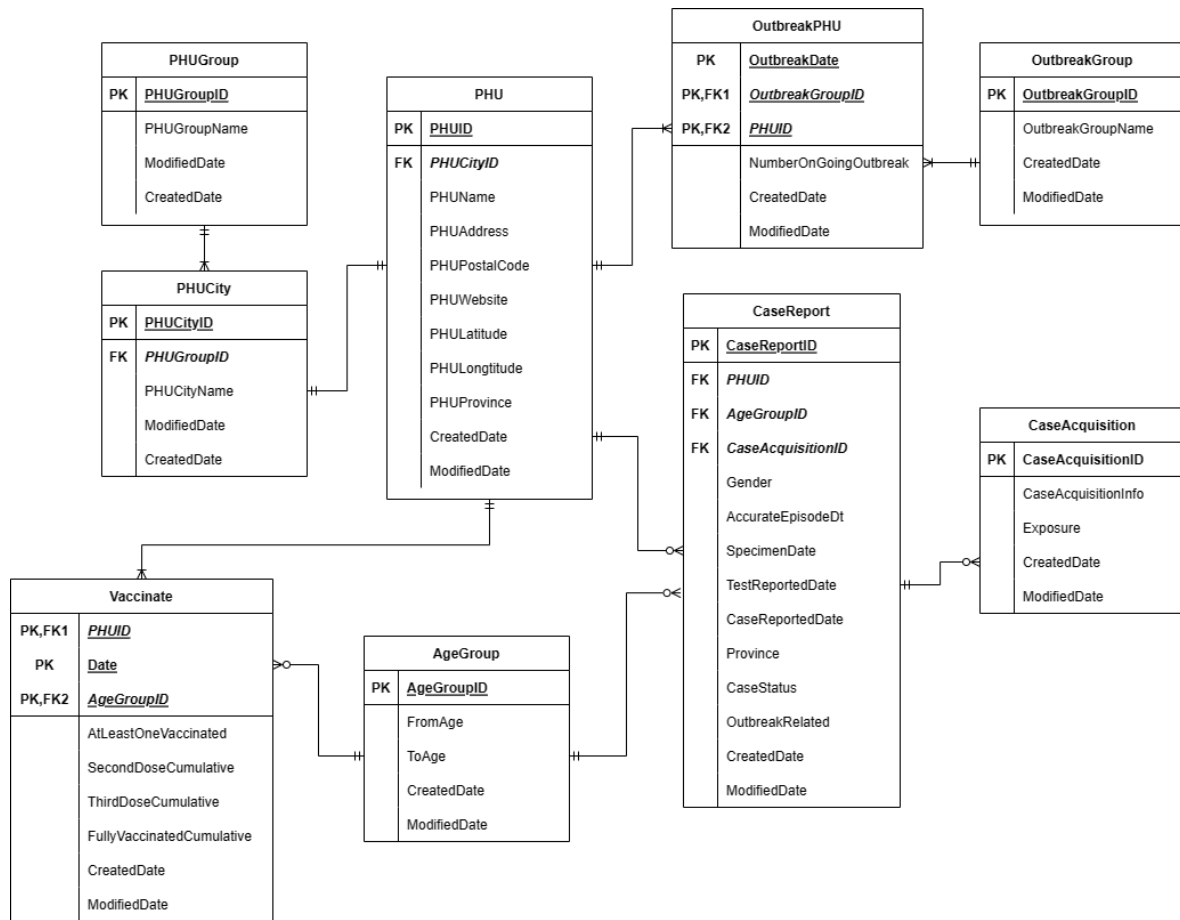
3.2.2 Data mining

- Number report:

- Mức độ nguy hiểm tử vong theo độ tuổi: tổng số người chết ở độ tuổi A / tổng số người nhiễm ở độ tuổi A
- Mức độ nguy hiểm tử vong theo khu vực: tổng số người chết ở khu vực A / tổng số người nhiễm ở khu vực A
- Mức độ nguy hiểm lan truyền bệnh:
 - + Tổng số ca nhiễm / đơn vị diện tích (thu thập thêm data diện tích các khu vực)
 - + Tổng số đợt bùng phát trong 1 khu vực
- Nếu các chỉ số cao hơn median thì được coi là nguy hiểm

- Prediction:
 - Dự đoán outcome (ở bảng case report) - supervisor learning
 - + Feature: age group, city, specimen date
 - + Label: outcome
- Clustering:
 - Gom thành phổ theo 2 feature:
 - + Số lượng người tử vong (trục Y)
 - + Số lượng người tiêm vắc xin đầy đủ ít nhất 2 mũi tính tới thời điểm cuối cùng được report (trục X)
 - Gom PHU theo 2 feature:
 - + Số người tử vong
 - + Số người hồi phục

4 Phân tích NDS



Theo dữ liệu nguồn **Public health unit** và **Public Health Units Group**

- Bảng **PHUGroup**: thông tin của các nhóm đơn vị y tế.
- Bảng **PHU**: thông tin chi tiết của các đơn vị y tế.
- Bảng **PHUCity**: thông tin thành phố - mỗi thành phố có 1 đơn vị y tế
- 1 nhóm đơn vị y tế gồm nhiều thành phố, 1 thành phố có 1 đơn vị y tế

❖ **Bảng PHU**

Field	Kiểu dữ liệu	Ý nghĩa
<u>PHUID</u>	Bigint	Mã đơn vị y tế
<i>PHUCityID</i>	Bigint	Khóa ngoại tham chiếu
PHUName	Varchar	Tên đơn vị y tế
PHUAddress	Varchar	Địa chỉ đơn vị y tế
PHUCity	Varchar	Đơn vị y tế thuộc thành phố nào
PHUPostalCode	Varchar	Mã bưu điện của đơn vị y tế (theo vùng/thành phố)
PHUWebsite	Varchar	Website của đơn vị y tế
PHULatitude	Varchar	Kinh độ của đơn vị y tế
PHULongitude	Varchar	Vĩ độ của đơn vị y tế
CreatedDate	Datetime	Ngày tạo
ModifiedDate	Datetime	Ngày cập nhật
PHUProvince	Varchar	Bang trực thuộc

❖ **Bảng PHUGroup**

Field	Kiểu dữ liệu	Ý nghĩa
<u>PHUGroupID</u>	Bigint	Mã nhóm đơn vị y tế
PHUGroupName	Varchar	Tên nhóm các đơn vị y tế.
CreatedDate	Datetime	Ngày tạo
ModifiedDate	Datetime	Ngày cập nhật

❖ **Bảng PHUCity**

Field	Kiểu dữ liệu	Ý nghĩa
<u>PHUCityID</u>	Bigint	Mã thành phố trực thuộc
<i>PHUGroupID</i>	Bigint	Tham chiếu tới PHUGroupID của bảng PHUGroup
PHUCityName	Varchar	Tên thành phố PHU trực thuộc
CreatedDate	Datetime	Ngày tạo
ModifiedDate	Datetime	Ngày cập nhật

Theo dữ liệu nguồn thì **vaccines_by_age_phu**: Dữ liệu tiêm vắc-xin tại các đơn vị chăm sóc sức khỏe:

- Ta thấy AgeGroup và PHUID có dữ liệu trùng lặp nên ta chuẩn hóa bằng cách tạo khóa ngoại FK PHUID (Tham chiếu tới PHUID của bảng **PHU**).
- Tách thuộc tính AgeGroup ra thành bảng riêng: bảng **AgeGroup** chứa thông tin nhóm tuổi và tạo khóa ngoại AgeGroupID (của bảng Vaccinate) tham chiếu tới AgeGroupID của bảng AgeGroup.

❖ **Bảng Vaccinate**

Field	Kiểu dữ liệu	Ý nghĩa
<u>Date</u>	Date	Ngày ghi nhận báo cáo
<u>PHU ID</u>	Bigint	Tham chiếu tới PHUID của bảng PHU
<u>AgeGroupID</u>	Bigint	Tham chiếu tới AgeGroupID của bảng AgeGroup
AtLeastOneVaccinated	Int	Số người tiêm được ít nhất 1 mũi
SecondDoseCumulative	Int	Số người tiêm được 2 mũi

FullyVaccinatedCumulative	Int	Số người tiêm đủ vaccin. Tiêm đầy đủ nghĩa là: <ul style="list-style-type: none"> - Tiêm 1 mũi Janssen (Johnson & Johnson) - Tiêm 2 mũi trong danh mục vaccin được Bộ y tế Canada phê duyệt - Tiêm 1 mũi trong danh mục được Bộ ý tế phê duyệt + 1 mũi trong danh mục không được phê duyệt - Tiêm 3 mũi vaccin thuộc loại bất kỳ
ThirdDoseCumulative	Int	Số người tiêm được 3 mũi
CreatedDate	Datetime	Ngày tạo
ModifiedDate	Datetime	Ngày cập nhật

❖ **Bảng AgeGroup**

Field	Kiểu dữ liệu	Ý nghĩa
<u>AgeGroupID</u>	bigint	Mã định danh của bảng
FromAge	Varchar	Nhóm tuổi, được phân loại gồm: <ul style="list-style-type: none"> - 05-11 yrs: 5 - 11 tuổi - 12-17 yrs: 12 – 17 tuổi - 18-29yrs: 18 – 29 tuổi - 30-39yrs: 30 – 39 tuổi - 40-49yrs: 40 – 49 tuổi
ToAge	Varchar	

		<ul style="list-style-type: none"> - 50-59yrs: 50 – 59 tuổi - 60-69yrs: 60 – 69 tuổi - 70-79yrs: 70 – 79 tuổi - 80+ : Từ 80 tuổi trở lên - Adults_18plus: Người lớn trên 18 tuổi - Ontario_12plus - Ontario_5plus <p>Undisclosed_or_missing: Chưa xác định</p>
CreatedDate	Datetime	Ngày tạo
ModifiedDate	Datetime	Ngày cập nhật

Theo dữ liệu nguồn **ongoing_outbreaks_phu**: Dữ liệu về việc bùng phát dịch tại các đơn vị chăm sóc sức khỏe của Ontario:

- Các thuộc tính PHU_Num và outbreak_group có giá trị lặp nên ta đặt khóa ngoại OutbreakPHUID (bảng OutbreakPHU) tham chiếu tới PHUID (bảng PHU).
- Tách outbreak_group ra thành bảng riêng **OutbreakGroup** và có khóa ngoại OutbreakgroupID (bảng OutbreakPHU) tham chiếu tới OutbreakgroupID (bảng Outbreakgroup).

❖ **Bảng OutbreakPHU**

Field	Kiểu dữ liệu	Ý nghĩa
<u>OutbreakDate</u>	bigint	Ngày ghi nhận báo cáo
<u>OutbreakGroupID</u>	bigint	Tham chiếu tới OutbreakgroupID (bảng Outbreakgroup)
<u>OutbreakPHUID</u>	bigint	Tham chiếu tới PHUID của bảng PHU
NumberOnGoingOutbreak	bigint	Số đợt bùng phát đang diễn ra
CreatedDate	Datetime	Ngày tạo
ModifiedDate	Datetime	Ngày cập nhật

❖ **Bảng OutbreakGroup**

Field	Kiểu dữ liệu	Ý nghĩa
<u>OutbreakGroupID</u>	bigint	Mã định danh cơ sở bùng phát
OutbreakGroupName	Varchar	Tên cơ sở bùng phát
CreatedDate	Datetime	Ngày tạo
ModifiedDate	Datetime	Ngày cập nhật

Theo dữ liệu nguồn **Cases report**: Dữ liệu ca nhiễm ở tỉnh bang Ontario và dữ liệu nguồn **Compiled_COVID-19_Case_Details_(Canada)**:

- Dữ liệu của hai nguồn này khá tương đồng và có cùng ý nghĩa nên ta gộp thành bảng **CaseReport**. Đồng thời thêm các thuộc tính riêng của từng nguồn vào bảng.
- Thuộc tính **CaseAcquisitionInfo** và **Exposure** mang cùng ý nghĩa, *thông tin ca nhiễm*, chỉ khác Exposure là *ghi rõ thông tin ca nhiễm*, còn Case AcquisitionInfo là *viết tắt của thông tin ca nhiễm* => Tạo thêm bảng **CaseAcquisition** có khóa chính là CaseAcquisitionID và bảng CaseReport sẽ có khóa ngoại là CaseAcquisitionID tham chiếu tới bảng CaseAcquisition.
- Tạo bảng **AgeGroup** để quản lý các nhóm tuổi => Thay thế bằng thuộc tính AgeGroupID (FK) tham chiếu tới bảng AgeGroup.
- Các thuộc tính như PHU Website, PHU Latitude, PHU PostCode, PHU Longitude, PHUCity đã được quản lý ở bảng **PHU** => Thay thế các thuộc tính đó bằng khóa ngoại PHUID tham chiếu tới bảng PHU.

❖ **Bảng CaseReport**

Field	Kiểu dữ liệu	Ý nghĩa
<u>CaseReportID</u>	bigint	Mã định danh cho mỗi hồ sơ được ghi nhận
<i>AgeGroupID</i>	bigint	Tham chiếu tới bảng AgeGroup
<i>PHUID</i>	bigint	Tham chiếu tới PHUID của bảng PHU
<i>CaseAcquisitionID</i>	bigint	Tham chiếu tới bảng CaseAcquisition
Gender	Varchar	Giới tính đối tượng: <ul style="list-style-type: none"> - FEMALE: nữ - MALE: nam - GENDER DIVERSE: đa giới tính - UNSPECIFIED: chưa xác định
CaseStatus	Varchar	Kết quả của ca nhiễm: <ul style="list-style-type: none"> - Resolved, Recovered: Đã điều trị - Not Resolved: Chưa điều trị - Fatal, Deceased: Tử vong - Active: Còn nhiễm bệnh - Not Reported: Chưa ghi nhận báo cáo
Province	Varchar	Nơi sinh sống của bệnh nhân (địa bàn/khu vực)

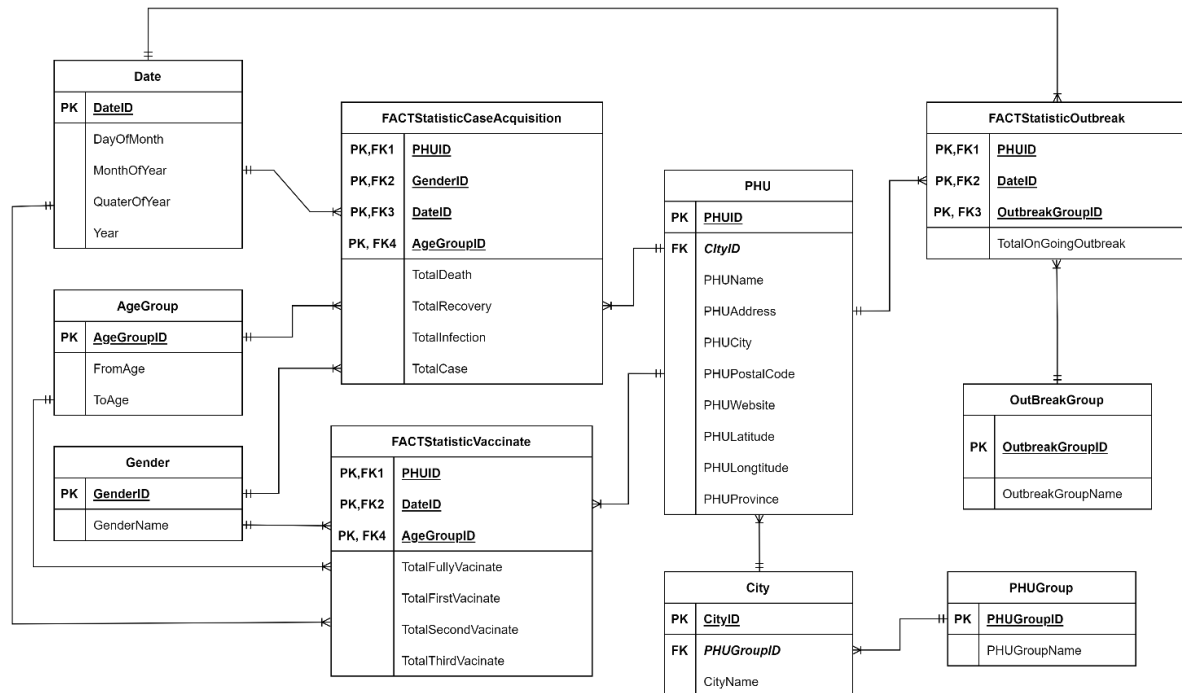
SpecimenDate	Date	Ngày lấy mẫu thử
CaseReported Date	Date	Ngày ghi nhận kết quả ca nhiễm
TestReported Date	Date	Ngày trả kết quả test
AccurateEpisode Dt	Date	Ngày khởi phát
OutbreakRelated	smallint	Có liên quan đến đợt bùng phát hay không?
CreatedDate	datetime	Ngày tạo
ModifiedDate	datetime	Ngày cập nhật

❖ **Bảng CaseAcquisition**

Field	Kiểu dữ liệu	Ý nghĩa
<u>CaseAcquisitionID</u>	bigint	Mã định danh dữ liệu trong bảng
CaseAcquisitionInfo	Varchar	<p>Thông tin của ca nhiễm:</p> <ul style="list-style-type: none"> - CC: (Closed contact) Dương tính, xác định được nguồn lây - NO KNOW EPI LINK: Dương tính, không xác định được nguồn lây - OB: (Outbreak) Bùng phát - TRAVEL - UNSPECIFIED EPI LINK: Dương tính, nguồn lây chưa được xác thực - MISSING INFORMATION: Chưa xác định

CreatedDate	Datetime	Ngày tạo
ModifiedDate	Datetime	Ngày cập nhật

5 Phân tích DDS



❖ Event

- Thống kê số ca nhiễm, số ca tử vong, số ca hồi phục theo từng PHU, độ tuổi và giới tính trong một mốc thời gian.
- Thống kê số lượt tiêm phòng theo từng PHU, độ tuổi trong một mốc thời gian.
- Thống kê tổng số lượt bùng dịch theo nhóm bùng dịch theo từng PHU trong một mốc thời gian.

❖ Grain

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

- Thống kê **Mức Độ Nghiêm Trọng** (theo tỉ lệ tử vong, hồi phục, tiêm đầy đủ vaccine, tiêm ít nhất 1 mũi, tiêm 2 mũi, tiêm 3 mũi) của dịch Covid-19 **theo PHU** và **theo các Quý trong từng năm**.
- Thống kê **tổng số người tử vong** theo **Giới Tính** và **Nhóm Tuổi** theo **các năm**.
- Thống kê **số ca nhiễm, tử vong** theo **Mức Độ Nghiêm Trọng** theo **Ngày Trong Tháng của các năm**.
- Thống kê **số ca nhiễm, tử vong** theo **Mức Độ Nghiêm Trọng, khu vực (PHU_Group, City)**, và **số người đã được tiêm vaccin trong các năm**.
- Thống kê **số ca nhiễm** theo **Mức Độ Nghiêm Trọng, nhóm bùng phát của từng khu vực trong các năm**.

* **Mức độ nghiêm trọng** được đánh giá dựa trên tổng số ca nhiễm và mức độ tử vong. Tổng số ca nhiễm và mức độ tử vong càng cao => mức độ nghiêm trọng càng lớn

❖ **Measure (không được cộng, được phép cộng, phải cộng theo công thức định nghĩa)**

- **Additive (được phép cộng)**

- TotalDeath: SUM
- TotalRecovery: SUM
- TotalInfection: SUM
- TotalOnGoingOutbreak: SUM
- TotalCase: SUM
- TotalFullyVaccinate: SUM
- TotalFirstVaccinate: SUM
- TotalSecondVaccinate: SUM
- TotalThirdVaccinate: SUM

❖ **Phân cấp**

- Phân cấp chiều **Geography**: PHUGroup => City => PHU
- Phân cấp chiều **DATE**: Year => Quarter => Month => Day

❖ **Bảng Date**

Field	Ý nghĩa	Kiểu dữ liệu
<u>DateID</u>	Khóa chính, sinh tự động	bigint
DayOfMonth	Ngày trong tháng	int
MonthOfYear	Tháng trong Năm	int
QuarterOfYear	Quý trong năm	int
Year	Năm	int

⇒ Bảng lưu thông tin về ngày có sự phân cấp giữa Ngày < Tháng < Quý < Năm.

❖ **Bảng AgeGroup**

Field	Ý nghĩa	Kiểu dữ liệu
<u>AgeGroupID</u>	Khóa chính, sinh tự động	bigint
FromAge	Độ tuổi từ	Varchar(12)
ToAge	Độ tuổi đến	Varchar(12)

⇒ Bảng lưu thông tin các nhóm tuổi khác nhau.

❖ **Bảng Gender**

Field	Ý nghĩa	Kiểu dữ liệu
<u>GenderID</u>	Khóa chính, sinh tự động	bigint
GenderName	Tên giới tính	Varchar(12)

⇒ Bảng lưu thông tin giới tính.

❖ **Bảng OutBreakGroup**

Field	Ý nghĩa	Kiểu dữ liệu
<u>OutbreakGroupID</u>	Khóa chính, sinh tự động	bigint
OutbreakGroupName	Nhóm địa điểm xảy ra bùng nổ	Varchar(128)

⇒ Bảng lưu thông tin nhóm địa điểm xảy ra các vụ bùng nổ dịch.

❖ **Bảng PHUGroup**

Field	Ý nghĩa	Kiểu dữ liệu
<u>PHUGroupID</u>	Khóa chính, sinh tự động	bigint
PHUGroupName	Nhóm y tế cộng đồng	Varchar(128)

⇒ Bảng lưu thông tin nhóm y tế cộng đồng.

❖ **Bảng City**

Field	Ý nghĩa	Kiểu dữ liệu
<u>CityID</u>	Khóa chính, sinh tự động	bigint
<i>PHUGroupID</i>	Khóa phụ tham chiếu	bigint
CityName	Tên thành phố	Varchar(128)

⇒ Bảng lưu thông tin thành phố trực thuộc của trung tâm y tế cộng đồng.

❖ **Bảng PHU**

Field	Ý nghĩa	Kiểu dữ liệu
<u>PHUID</u>	Khóa chính, sinh tự động	bigint
<i>CityID</i>	Khóa phụ tham chiếu	bigint

PHUName	Tên trung tâm y tế	Varchar(128)
PHUAddress	Địa chỉ trung tâm y tế	Varchar(128)
PHUPostalCode	Mã vùng	Varchar(128)
PHUWebsite	Website trung tâm y tế	Varchar(128)
PHULatitude	Vĩ độ trung tâm y tế	Varchar(128)
PHULongtitude	Kinh độ trung tâm y tế	Varchar(128)
PHUProvince	Bang trực thuộc	Varchar(128)

⇒ Bảng lưu thông tin trung tâm y tế cộng đồng.

❖ **Bảng *FACTStatisticCaseAcquisition***

Field	Ý nghĩa	Kiểu dữ liệu
<i>PHUID</i>	Khóa phụ tham chiếu	bigint
<i>GenderID</i>	Khóa phụ tham chiếu	bigint
<i>AgeGroupID</i>	Khóa phụ tham chiếu	bigint
<i>DateID</i>	Khóa phụ tham chiếu	bigint
TotalDeath	Tổng số ca chết	bigint
TotalRecovery	Tổng số ca hồi phục	bigint
TotalInfection	Tổng số ca nhiễm	bigint
TotalCase	Tổng số ca	bigint

⇒ Fact sử dụng để lưu thông tin thống kê số ca nhiễm, số ca chết, số ca hồi phục, mức độ nguy hiểm của một PHU theo ngày, giới tính và nhóm tuổi.

❖ **Bảng *FACTStatisticVaccinate***

Field	Ý nghĩa	Kiểu dữ liệu
<i>PHUID</i>	Khóa phụ tham chiếu	bigint
<i>AgeGroupID</i>	Khóa phụ tham chiếu	bigint
<i>DateID</i>	Khóa phụ tham chiếu	bigint
TotalFullyVaccinate	Số lượng tiêm đủ mũi vaccine theo tiêu chí	Decimal(34,2)
TotalFirstVaccinate	Số lượng tiêm đủ 1 mũi vaccine	Decimal(34,2)
TotalSecondVaccinate	Số lượng tiêm đủ 2 mũi vaccine	Decimal(34,2)
TotalThirdVaccinate	Số lượng tiêm đủ 3 mũi vaccine	Decimal(34,2)

⇒ Fact sử dụng để lưu thông tin thống kê tỷ lệ tiêm vaccine của một PHU theo ngày và nhóm tuổi.

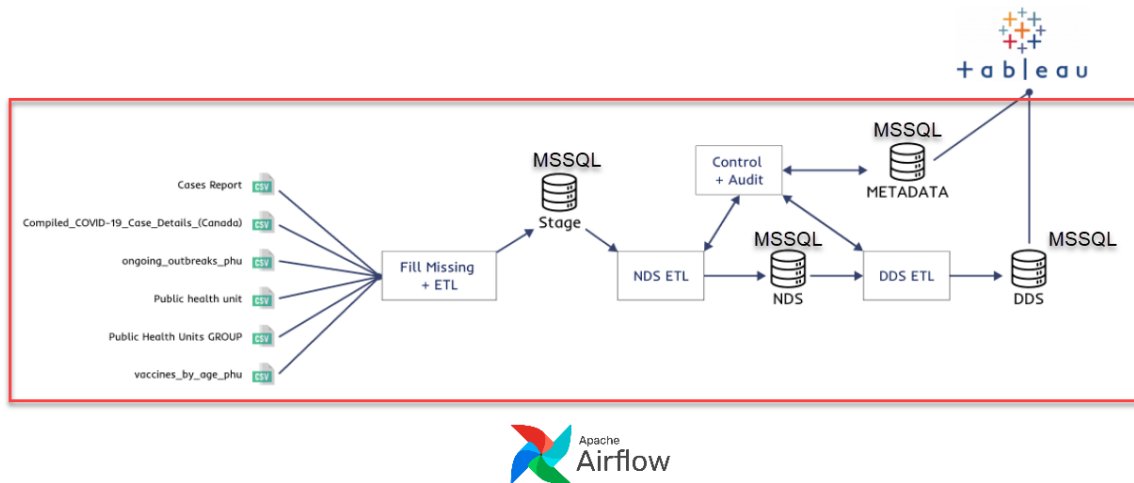
❖ **Bảng *FACTStatisticOutbreak***

Field	Ý nghĩa	Kiểu dữ liệu
<i>PHUID</i>	Khóa phụ tham chiếu	bigint
<i>DateID</i>	Khóa phụ tham chiếu	bigint
<i>OutbreakGroupID</i>	Khóa phụ tham chiếu	bigint
TotalOnGoingOutBreak	Tổng số đợt bùng nổ	bigint

⇒ Fact sử dụng để lưu thông tin thống kê số lượng bùng nổ của một PHU theo ngày.

6 System

6.1 Thiết kế pipeline



- Luồng ETL sử dụng Airflow.
- Database sử dụng MS SQL Server.
- Source các file .csv, .xlsx
- DAG trong airflow

6.2 Lập lịch

- Mỗi ngày, code sẽ chạy ETL vào lúc 0h
- Thời gian ETL dự tính: 15 phút (45 phút nếu data tầm 1 triệu dòng)
- Với mỗi task bị chạy lỗi sẽ tự động chạy lại sau 15 phút.

```
default_args = {
    'owner': 'khoabui',
    'retries': 3,
    'retry_delay': timedelta(minutes=15),
    'start_date': days_ago(0),
}

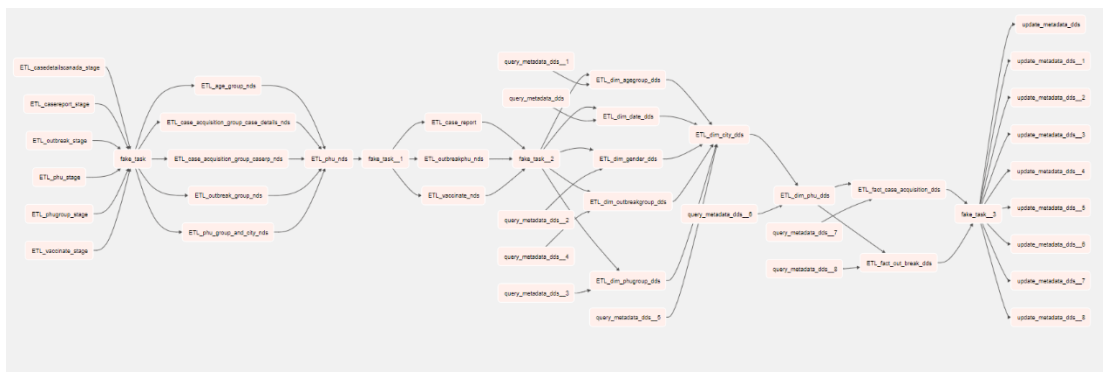
@dag(
    dag_id='covid_19_system',
    default_args=default_args,
    schedule="@daily",
    dagrun_timeout=timedelta(minutes=60),
)
```

Setup schedule trên airflow

6.3 ETL

6.3.1 Overview Data Flow

❖ Overview Data Flow



❖ Các dữ liệu được ETL lần 2

PHU_ID	Reporting_PHU	Reporting_PHU_Address	Reporting_PHU_City	Reporting_PHU_Postal_Code	Reporting_PHU_Website	Reporting_PHU_Latitude	Reporting_PHU_Longitude
2240	Chatham-Kent Health Unit	435 Grand Avenue West,	Chatham	N7M 5L8	www.ckphu.com	4,241,925,325	-8,212,808,769

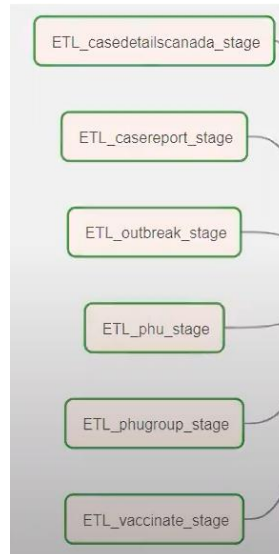
Source Public Health Unit

A	B	C	D	E	F	G	H	I
Objectid	row_id	date_repo	health_reg	age_group	gender	exposure	case_status	province
1048576	1047571	2022/12/0	Peel Public	20-29	FEMALE	Close Cont	Recovered	Ontario
1048577	1047572	2022/12/0	Peel Public	40-49	FEMALE	Close Cont	Recovered	Ontario
1048578	1047573	2022/12/0	Kingston, F	30-39	MALE	Close Cont	Recovered	Ontario

Source Compiled_COVID-19_Case_Details_(Canada)

6.3.2 Source to Stage

❖ Data Flow Source to Stage



- Extract all source table và load all table to stage
- Truncate tất cả table trước khi thực hiện etl

❖ ETL lần 1

PHUStage x

Properties Data ER Diagram master Databases Stage_Covid Sc

PHUStage Enter a SQL expression to filter results (use Ctrl+Space)

	PHUIDKey	PHUID	ReportingPHU	ReportingPHUAddress
1	1	2,244	Middlesex-London Health Unit	50 King Street
2	2	2,236	Halton Region Health Department	1151 Bronte Road
3	3	3,895	Toronto Public Health	277 Victoria Street, 5th Floor
4	4	5,183	Huron Perth District Health Unit	653 West Gore Street
5	5	2,266	Wellington-Dufferin-Guelph Public Health	160 Chancellors Way
6	6	2,263	Timiskaming Health Unit	247 Whitewood Avenue, Unit 43
7	7	2,262	Thunder Bay District Health Unit	999 Balmoral Street
8	8	2,253	Peel Public Health	7120 Hurontario Street
9	9	2,261	Sudbury & District Health Unit	1300 Paris Street
10	10	2,230	Durham Region Health Department	605 Rossland Road East
11	11	2,237	Hamilton Public Health Services	110 King St. West, 2nd Floor
12	12	2,268	Windsor-Essex County Health Unit	1005 Ouellette Avenue
13	13	2,265	Region of Waterloo, Public Health	99 Regina Street South
14	14	2,260	Simcoe Muskoka District Health Unit	15 Sperling Drive
15	15	4,913	Southwestern Public Health	1230 Talbot Street
16	16	2,226	Algoma Public Health Unit	294 Willow Avenue
17	17	2,243	Leeds, Grenville and Lanark District Health U	458 Laurier Boulevard
18	18	2,238	Hastings and Prince Edward Counties Health	179 North Park Street
19	19	2,247	North Bay Parry Sound District Health Unit	345 Oak Street West
20	20	2,270	York Region Public Health Services	17250 Yonge Street
21	21	2,249	Northwestern Health Unit	210 First Street North
22	22	2,255	Peterborough Public Health	185 King Street
23	23	2,233	Grey Bruce Health Unit	101 17th Street East
24	24	2,241	Kingston, Frontenac and Lennox & Addington	221 Portsmouth Avenue
25	25	2,256	Porcupine Health Unit	169 Pine Street South
26	26	2,257	Renfrew County and District Health Unit	7 International Drive
27	27	2,251	Ottawa Public Health	100 Constellation Drive
28	28	2,227	Brant County Health Unit	194 Terrace Hill Street

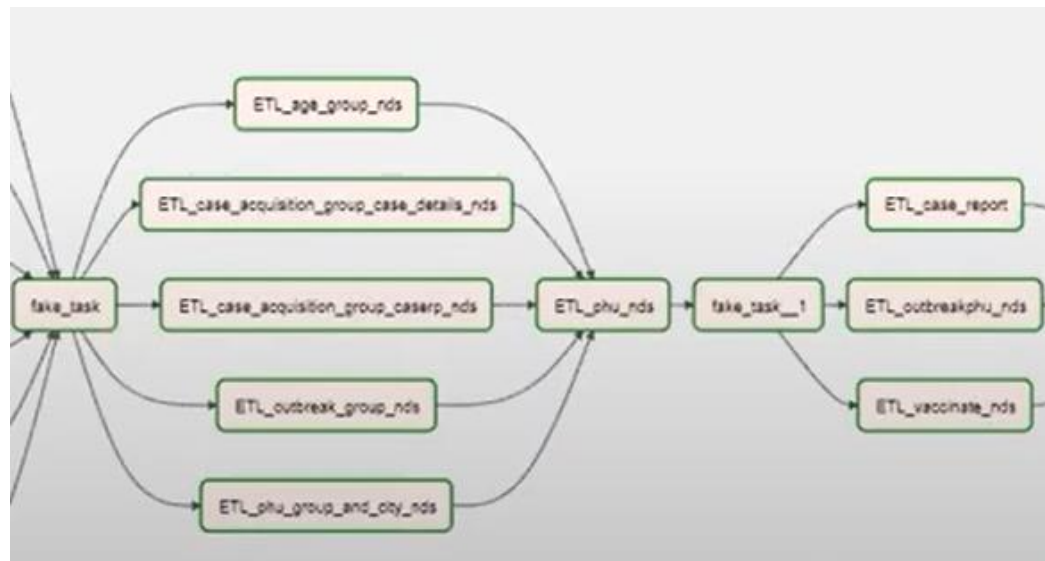
❖ ETL lần 2

- Dữ liệu mới ở bảng CaseDetailsCanadaStage

Grid	DateReported	HealthRegion	AgeGroup	Gender	Exposure	CaseStatus
1	2022-12-05 12:00:00.000	Peel Public Health	20-29	MALE	Close Contact	Recovered
2	2022-12-05 12:00:00.000	Peel Public Health	40-49	FEMALE	Close Contact	Recovered
3	2022-12-05 12:00:00.000	Kingston, Frontenac and Lennox & Addington	30-39	MALE	Close Contact	Recovered

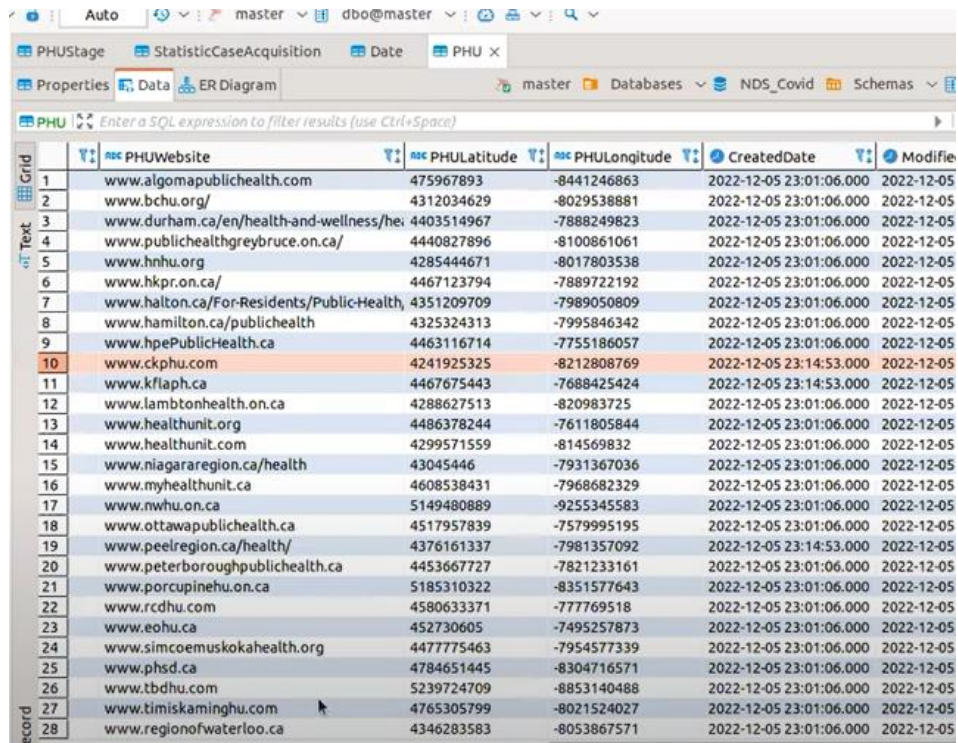
6.3.3 Stage to NDS

❖ Data Flow Stage to NDS



- Extract all stage table to NDS
- Thực hiện phép upsert ở một số và một số bảng như case report được insert thẳng.

❖ ETL lần 1

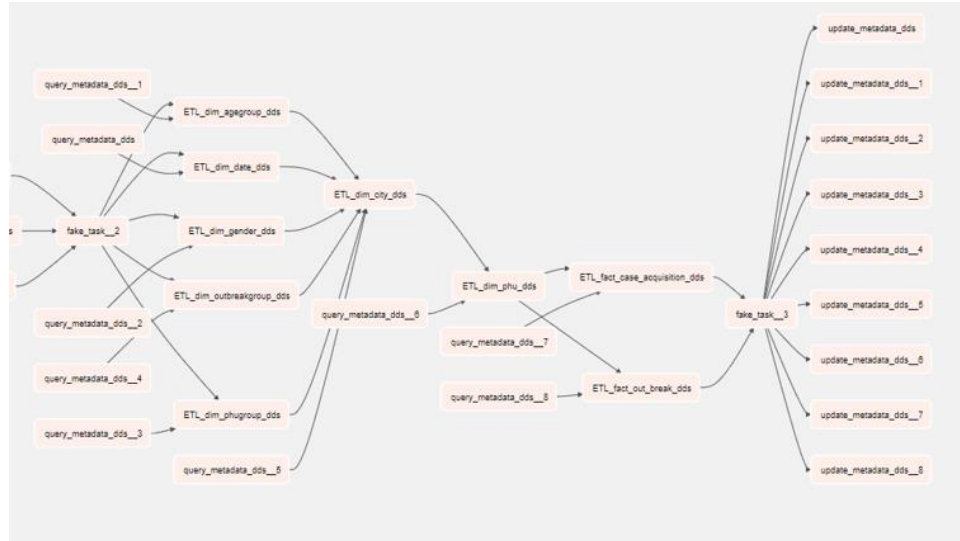


Grid	PHUWebsite	PHULatitude	PHULongitude	CreatedDate	Modified
1	www.algomapublichealth.com	475967893	-8441246863	2022-12-05 23:01:06.000	2022-12-05
2	www.bchu.org/	4312034629	-8029538881	2022-12-05 23:01:06.000	2022-12-05
3	www.durham.ca/en/health-and-wellness/he	4403514967	-7888249823	2022-12-05 23:01:06.000	2022-12-05
4	www.publichealthgreybruce.on.ca/	4440827896	-8100861061	2022-12-05 23:01:06.000	2022-12-05
5	www.hnhu.org	4285444671	-8017803538	2022-12-05 23:01:06.000	2022-12-05
6	www.hkpr.on.ca/	4467123794	-7889722192	2022-12-05 23:01:06.000	2022-12-05
7	www.halton.ca/For-Residents/PublicHealth	4351209709	-7989050809	2022-12-05 23:01:06.000	2022-12-05
8	www.hamilton.ca/publichealth	4325324313	-7995846342	2022-12-05 23:01:06.000	2022-12-05
9	www.hpePublicHealth.ca	4463116714	-7755186057	2022-12-05 23:01:06.000	2022-12-05
10	www.ckphu.com	4241925325	-8212808769	2022-12-05 23:14:53.000	2022-12-05
11	www.kflaph.ca	4467675443	-7688425424	2022-12-05 23:14:53.000	2022-12-05
12	www.lambtonhealth.on.ca	4288627513	-820983725	2022-12-05 23:01:06.000	2022-12-05
13	www.healthunit.org	4486378244	-7611805844	2022-12-05 23:01:06.000	2022-12-05
14	www.healthunit.com	4299571559	-814569832	2022-12-05 23:01:06.000	2022-12-05
15	www.niagararegion.ca/health	43045446	-7931367036	2022-12-05 23:01:06.000	2022-12-05
16	www.myhealthunit.ca	4608538431	-7968682329	2022-12-05 23:01:06.000	2022-12-05
17	www.nwhu.on.ca	5149480889	-9255345583	2022-12-05 23:01:06.000	2022-12-05
18	www.ottawapublichealth.ca	4517957839	-7579995195	2022-12-05 23:01:06.000	2022-12-05
19	www.peelregion.ca/health/	4376161337	-7981357092	2022-12-05 23:14:53.000	2022-12-05
20	www.peterboroughpublichealth.ca	4453667727	-7821233161	2022-12-05 23:01:06.000	2022-12-05
21	www.porcupinehu.on.ca	5185310322	-8351577643	2022-12-05 23:01:06.000	2022-12-05
22	www.rcdhu.com	4580633371	-777769518	2022-12-05 23:01:06.000	2022-12-05
23	www.eohu.ca	452730605	-7495257873	2022-12-05 23:01:06.000	2022-12-05
24	www.simcoemuskokahhealth.org	4477775463	-7954577339	2022-12-05 23:01:06.000	2022-12-05
25	www.phsd.ca	4784651445	-8304716571	2022-12-05 23:01:06.000	2022-12-05
26	www.tbduh.com	5239724709	-8853140488	2022-12-05 23:01:06.000	2022-12-05
27	www.timiskaminghu.com	4765305799	-8021524027	2022-12-05 23:01:06.000	2022-12-05
28	www.regionofwaterloo.ca	4346283583	-8053867571	2022-12-05 23:01:06.000	2022-12-05

6.3.4NDS to DDS

❖ Data Flow NDS to DDS

- Thực hiện extract incremental theo 2 cột createdDate và modifiedDate.
- Sau đó, thực hiện kéo các data liên quan có sẵn ở trong dds để tiến hành cập nhật



❖ ETL lần 1

PHUStage | StatisticCaseAcquisition X

Properties | Data | ER Diagram | master | Databases | DDS_Covid | Schemas | dbo | Tables

StatisticCaseAcquisition | Enter a SQL expression to filter results (use Ctrl+Space)

Grid	123 PHUID	123 DateID	123 AgeGroupID	123 GenderID	123 TotalDeath	123 TotalRecovery	123 TotalInfection
1	2,237	1	8	4	0	1	0
2	2,238	1	8	4	0	1	0
3	2,251	1	6	4	0	1	0
4	2,253	1	8	4	0	1	0
5	2,256	1	6	4	0	1	0
6	2,256	1	8	4	0	1	0
7	2,265	1	8	4	0	1	0
8	2,266	1	9	4	0	1	0
9	2,268	1	1	4	0	1	0
10	2,268	1	8	4	0	2	0
11	2,270	1	5	4	0	1	0
12	2,270	1	7	4	0	1	0
13	2,270	1	8	4	0	1	0
14	3,895	1	3	4	0	1	0
15	3,895	1	5	4	0	1	0
16	3,895	1	6	4	0	1	0
17	3,895	1	8	4	0	8	0
18	3,895	1	9	4	0	4	0
19	5,206	1	9	4	0	0	0
20	2,227	2	8	4	0	1	0
21	2,230	2	2	4	0	1	0
22	2,242	2	7	4	0	1	0
23	2,251	2	6	4	0	0	1
24	2,253	2	6	4	0	0	1
25	2,253	2	8	4	0	0	1
26	2,253	2	9	4	0	1	0
27	2,256	2	2	4	0	0	1
28	2,256	2	7	4	0	1	0

❖ ETL lần 2

- Dữ liệu mới ở bảng Dim Date

941	930	26	12	4	2,021
942	942	5	12	4	2,022

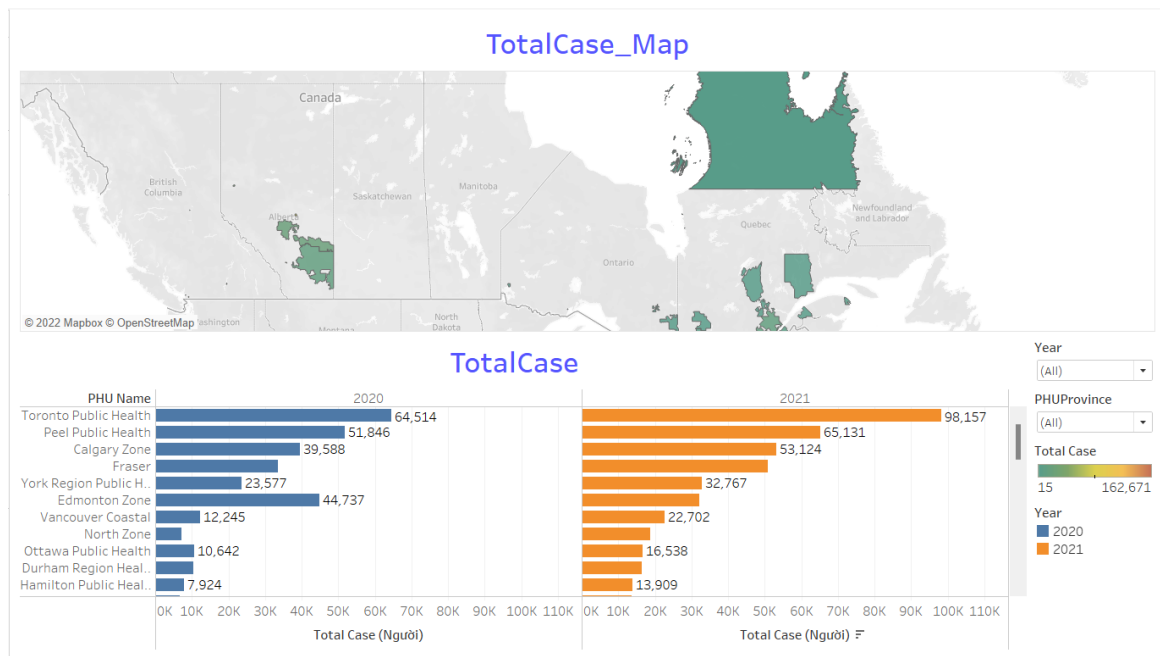
- Dữ liệu mới ở bảng Fact Statistic Case Acquisition

Grid	PHUID	DateID	AgeGroupID	GenderID	TotalDeath	TotalRecovery	TotalInfection	CriteriaD
1	2,241	942	9	2	0	2	0	
2	2,253	942	8	1	0	2	0	
3	2,253	942	3		0	2	0	

7 Trục quan hóa dữ liệu

- Sử dụng dữ liệu ở DDS
- Công cụ trực quan: Tableau

7.1 Dashboard TotalCase



Hình: Dashboard thể hiện tổng số ca nhiễm Covid theo khu vực và năm

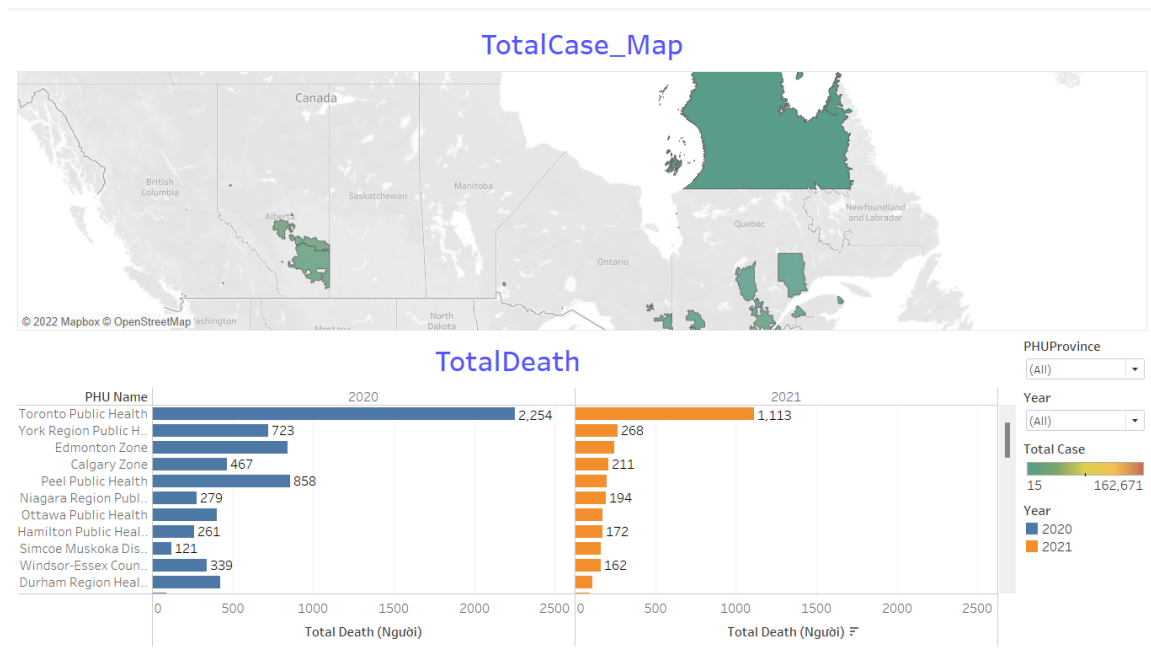
- Ta có thể chọn khu vực (PHUProvince) và Year để hiển thị thông tin theo tùy chọn của filters.
- Đặc biệt ở phần map ta có thể thấy các khu vực/ PHU sẽ được tô màu (theo dải màu từ xanh lá đến đỏ) theo tổng số ca mắc Covid

❖ **TotalCase:** Tổng số ca nhiễm ở từng PHU hoặc khu vực

* **Nhận xét**

- Từ biểu đồ trên ta có thể thấy được Toronto Public Health tiếp nhận số ca nhiễm lớn nhất trong cả hai năm gần đây (năm 2020, 2021). Năm 2020 tiếp nhận 64,514 ca nhiễm và con số này tăng lên vào năm 2021 với 98,157 ca nhiễm.
- Đa số các PHU/ khu vực đều có xu hướng tăng từ năm 2020 đến 2021. Riêng khu vực Edmonton Zone lại giảm từ 44,737 ca nhiễm năm 2020 xuống 32,187 năm 2021.

7.2 Dashboard TotalDeath_TotalRecovery_TotalInfection



Hình: Dashboard thể hiện tổng số ca tử vong do Covid theo khu vực và năm

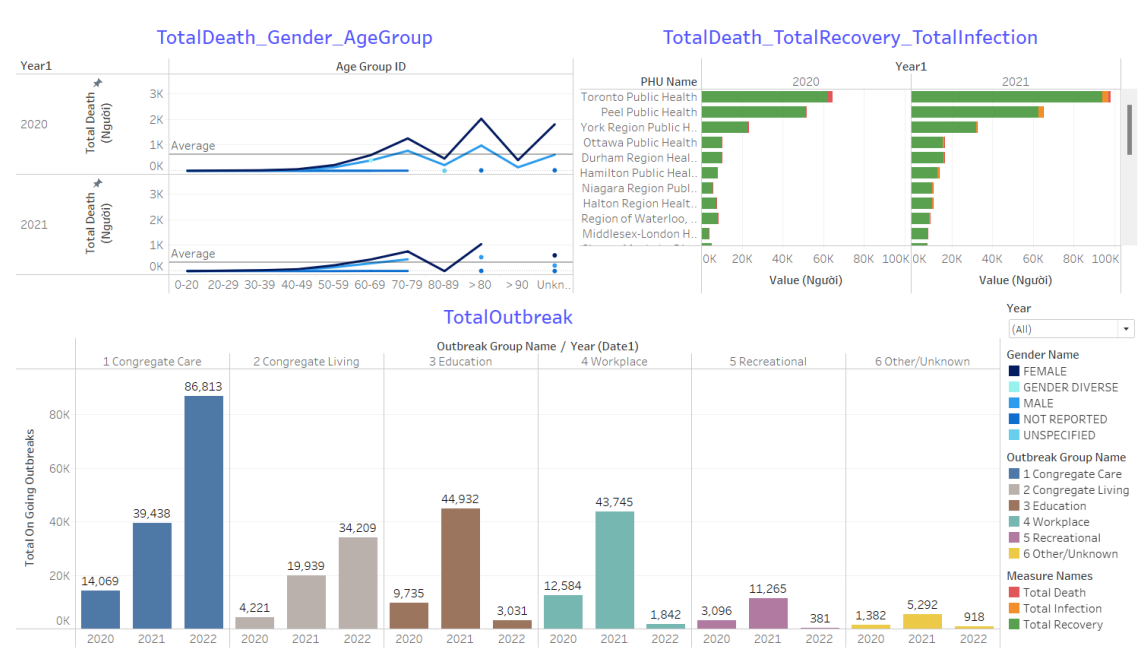
- Ta có thể chọn khu vực (PHUProvince) và Year để hiển thị thông tin theo tùy chọn của filters.
- Đặc biệt ở phần map ta có thể thấy các khu vực/ PHU sẽ được tô màu (theo dải màu từ xanh lá đến đỏ) theo tổng số ca tử vong do Covid.

❖ **TotalDeath:** Tổng số ca tử vong do Covid ở từng PHU hoặc khu vực

* Nhận xét

- Từ biểu đồ trên ta có thể thấy được Toronto Public Health ghi nhận số ca tử vong do Covid lớn nhất trong cả hai năm gần đây (năm 2020, 2021). Năm 2020 ghi nhận 2,254 ca và con số này giảm xuống vào năm 2021 với 1,113 ca.
- Hầu hết các PHU/ khu vực đều có xu hướng giảm từ năm 2020 đến 2021. Có thể đoán được con số thống kê số ca tử vong do Covid giảm vào năm 2021 do lúc này đã xuất hiện các vắc xin phòng chống Covid nên số ca trở nặng/ chết cũng giảm.

7.3 Dashboard Overview



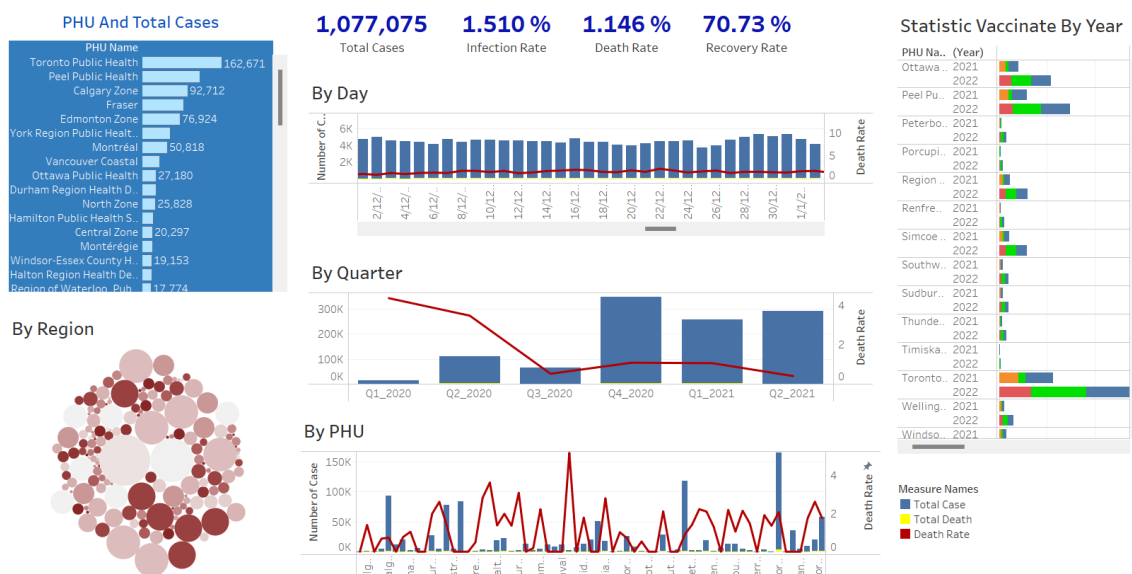
Hình: Dashboard overview

- ❖ **TotalDeath_Gender_AgeGroup:** Thể hiện xu hướng số ca tử vong do Covid theo nhóm tuổi và giới tính (màu sắc).
- ❖ **TotalDeath_TotalRecovery_TotalInfection:** Thống kê số ca tử vong, hồi phục, số ca còn nhiễm Covid theo từng PHU hoặc khu vực theo năm.
- ❖ **TotalOutbreak:** Thống kê tổng số lượt bùng nổ theo nhóm bùng dịch và theo năm.

*** Nhận xét**

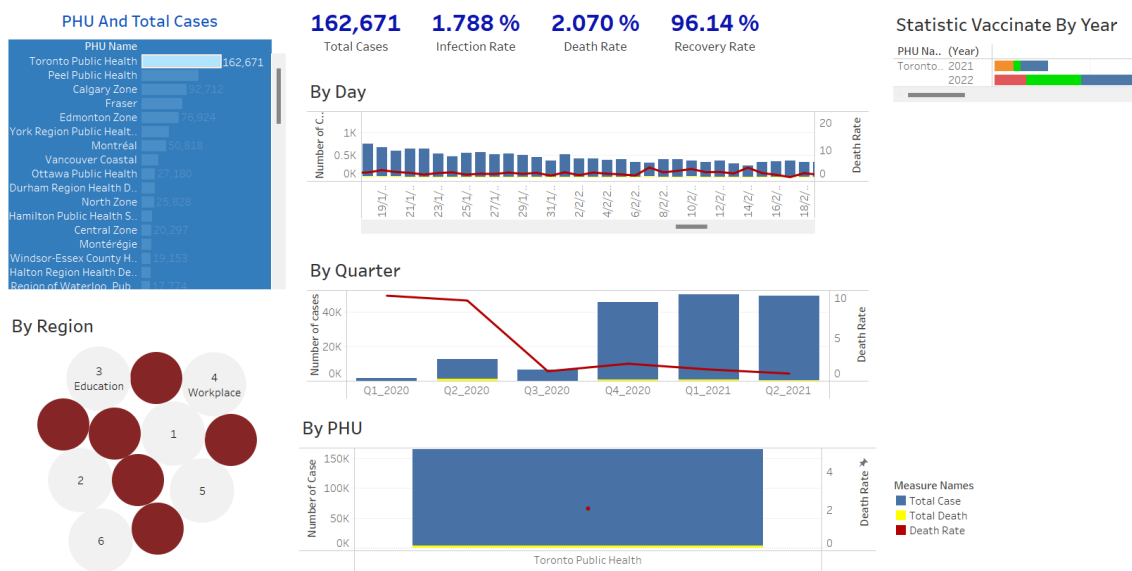
- Năm 2020, số ca tử vong cao ở các nhóm tuổi từ 70 – 79 tuổi và lớn hơn 80 do ở độ tuổi này đa số sẽ mắc bệnh nền và vào thời điểm này vắc xin phòng chống Covid chưa được phổ biến rộng rãi nên khi mắc Covid khả năng tử vong sẽ cao hơn. Vào năm 2021, tuy đã giảm về số ca tử vong nhưng các nhóm tuổi này vẫn cần phải đặc biệt lưu ý cũng như chú ý ở phái nữ.
- Có thể thấy được số ca nhiễm của Toronto Public Heath dẫn đầu ở cả 2 năm 2020, 2021. Song đó có hướng tích cực là số ca hồi phục ở cả 2 năm chiếm thị phần cao không chỉ riêng mỗi Toronto mà các PHU hoặc khu vực khác cũng thế.
- Số lượng lượt bùng nổ ở nhóm chăm sóc tập thể (Congregate Care) đạt cao nhất ở cả 3 năm. Do các ca nhiễm Covid và nghi nhiễm sẽ được cách ly ở cùng địa điểm nhưng khác khu vực nên việc lây nhiễm chéo sẽ xảy ra rất cao. Tuy nhiên đáng chú ý ở năm 2022 số lượng này lại vượt lên xa so với hai năm trước do lúc này việc giãn cách xã hội đã được nới lỏng và mọi người đã tiêm ngừa Covid nên sẽ hoạt động cuộc sống bình thường dẫn đến việc lây nhiễm chéo rất cao và khi có dấu hiệu mọi người sẽ test cá nhân hoặc đến trung tâm chăm sóc sức khỏe để tiến hành test Covid nên số liệu này có thể dự đoán trước.

7.4 Dashboard Statistics Of Serious Level



Hình: Dashboard liên quan đến các tỉ lệ nhiễm, tỉ lệ hồi phục, tỉ lệ tử vong, tình hình tiêm vaccine ở từng PHU

Có thể chọn vào PHU cụ thể ở **PHU And Total Cases** để filter các biểu đồ còn lại theo PHU đã chọn



Hình: Dashboard được filter theo PHU Toronto Public Health

❖ Key Metrics:

- Total Cases: Tổng số ca nhiễm ở tất cả các PHU = Total Case
- Infection Rate: Tỷ lệ nhiễm = $\frac{\text{sum}([\text{Total Infection}])}{\text{sum}([\text{Total Case}])} \times 100$
- Death Rate: Tỷ lệ tử vong = $\frac{\text{sum}([\text{Total Death}])}{\text{sum}([\text{Total Case}])} \times 100$
- Recovery Rate: Tỷ lệ hồi phục = $\frac{\text{sum}([\text{Total Recovery}])}{\text{sum}([\text{Total Case}])} \times 100$

❖ PHU And Total Cases: Thống kê tổng số ca nhiễm ở từng PHU

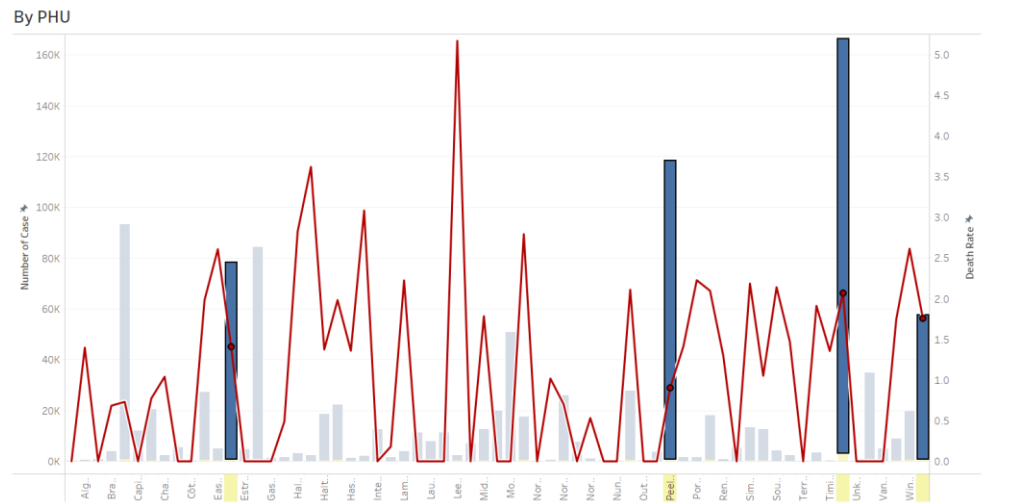
❖ By Day, By Quarter, By PHU: Thống kê tỷ lệ tử vong so với tổng số ca nhiễm theo từng PHU, từng ngày, từng quý để xem xét *mức độ nghiêm trọng*

❖ By Region: Thống kê tỷ lệ tử vong so với tổng số ca nhiễm theo từng cơ sở bùng phát dịch (Outbreak Group) ở từng PHU Group

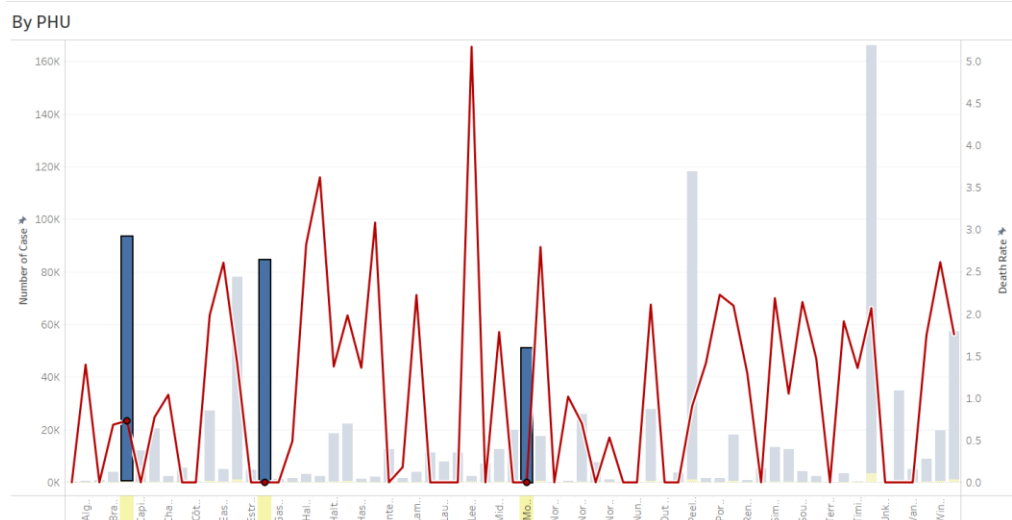
❖ Statistic Vaccinate By Year: Thống kê số người tiêm Vaccine ở từng PHU theo các năm

* Nhận xét:

- Dữ liệu về ca nhiễm được ghi nhận từ ngày 1/1/2020 đến ngày 2/6/2021
- Trong khoảng thời gian này, ghi nhận được 1077075 ca nhiễm, nổi bật là Toronto Public Health với 162671 ca. Các PHU kể đó như: Peel Public Health, Calgary Zone, Fraser, Edmonton Zone cũng chiếm phần lớn số ca nhiễm khảo sát
- Xem xét mức độ nghiêm trọng: Một PHU được xem là chịu ảnh hưởng nghiêm trọng từ Covid khi: Tổng số ca nhiễm cao và tỷ lệ tử vong cao (~1% trở lên)
 - Quan sát By PHU



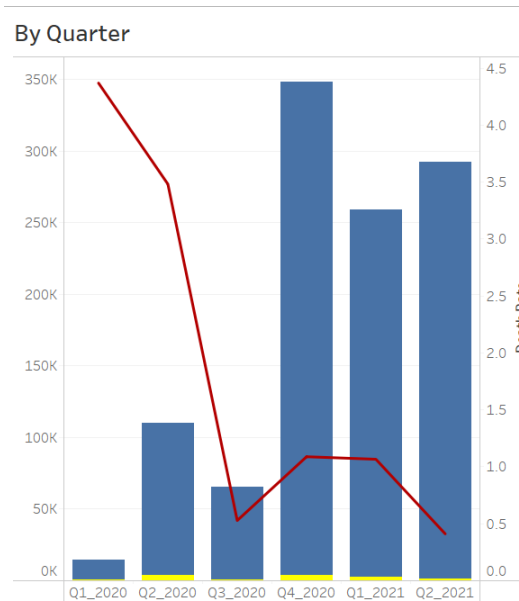
Có thể thấy Edmonton Zone, Toronto Public Health, Peel Public Health, York Region Public Health Services là các PHU có đồng thời tổng số ca nhiễm và tỷ lệ tử vong cao/ tương đối cao => các đơn vị này chịu **ảnh hưởng nghiêm trọng**



Bên cạnh đó, các PHU: Calgary Zone, Fraser, Montréal có tổng số ca nhiễm cao nhưng tỉ lệ tử vong thấp => các đơn vị này chịu **ảnh hưởng ít nghiêm trọng** hơn các PHU trên

Các PHU còn lại có thể xem là **ít chịu ảnh hưởng**

- Quan sát By Quarter



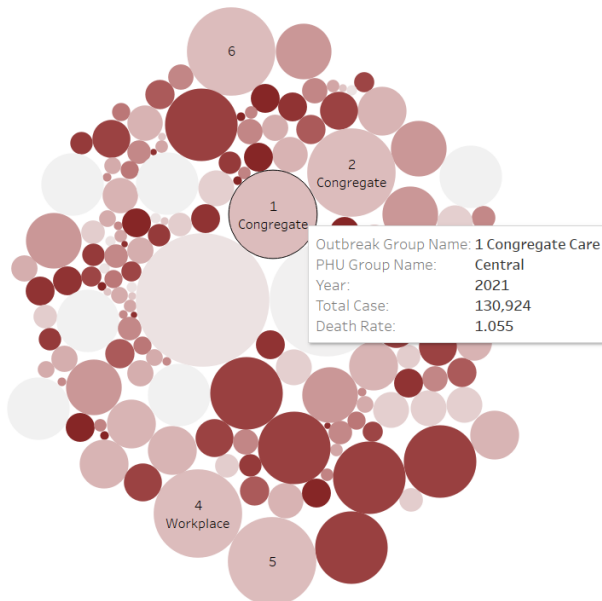
Vào quý 2 năm 2020, tổng số ca nhiễm tương đối cao và tỷ lệ tử vong rất cao (~3.483%) => Đáng báo động

Vào quý 4 năm 2020, tổng số ca nhiễm tăng đột biến (344565 ca) và tỉ lệ tử vong ở mức cao (~1.089%) => Lây nhiễm cao, **ảnh hưởng nghiêm trọng**

Đến Quý 2 năm 2021, mặc dù tổng số ca nhiễm vẫn còn cao nhưng tỉ lệ tử vong có dấu hiệu giảm dần (~0.412%) => **Ảnh hưởng ít nghiêm trọng**

- Quan sát By Region

By Region



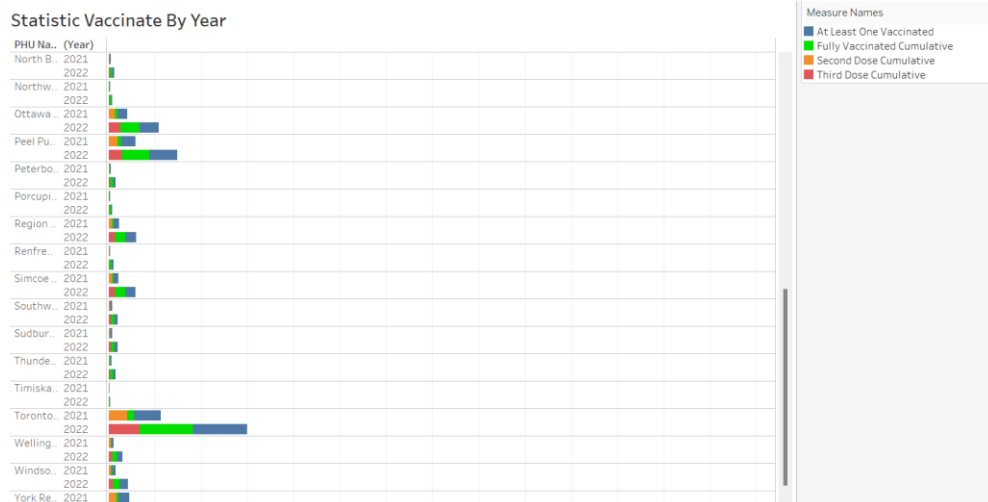
Kích thước vòng tròn càng lớn -> tổng số ca nhiễm càng cao

Màu càng đậm -> Tỷ lệ tử vong càng cao

Quan sát biểu đồ có thể dễ dàng nhận thấy những vòng tròn có kích thước lớn và màu đậm => những nơi bị ảnh hưởng nghiêm trọng

Hầu hết các nơi chịu ảnh hưởng nghiêm trọng đều ở vùng **Central**

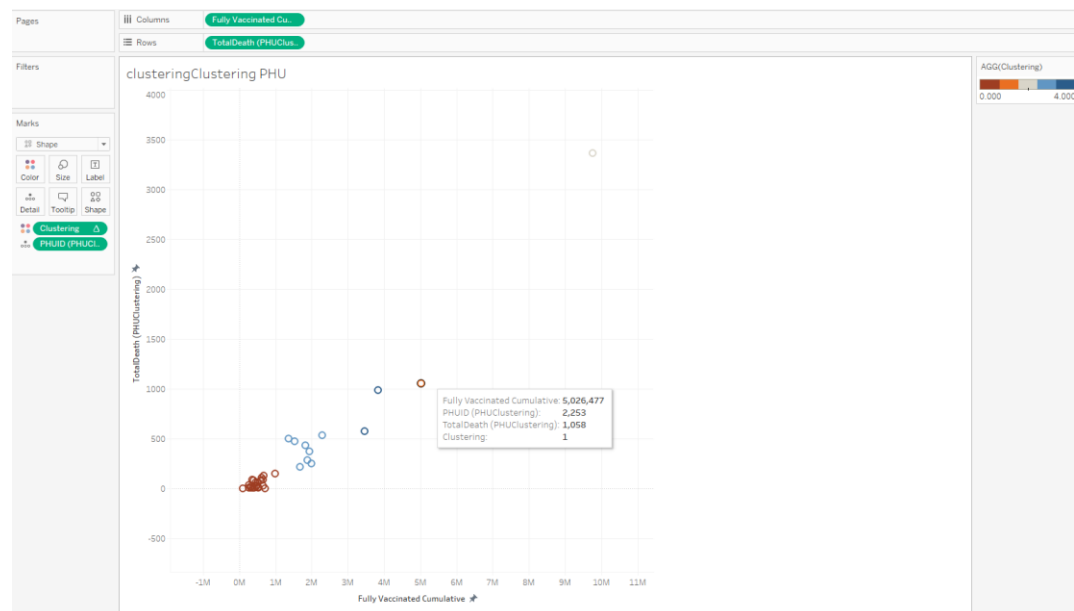
- Quan sát Statistic Vaccinate By Year



Số người tiêm vaccine ngày càng tăng qua các năm -> dấu hiệu tốt đối với việc chống dịch. Đây cũng là một trong những nguyên nhân ảnh hưởng đến việc tỉ lệ tử vong giảm xuống trong thời điểm này

7.5 Data Mining:

- Sử dụng thuật toán clustering để gom cụm các PHU có số lượng người tử vong và số lượng người tiêm đủ mũi vắc xin.



- Trục X: tổng số người được tiêm đầy đủ vắc xin của một PHU
- Trục Y: tổng số người chết
- Nhận xét:
 - Ở Canada tập trung tiêm vaccine cho các PHU có số lượng người chết cao. Hay có thể nói các vùng có người chết cao là những vùng đông dân cư -> cần nhiều vaccine
 - Các vùng có số người chết thấp -> ít tập trung hơn, do số lượng dân cư thấp, hoặc mức độ nghiêm trọng không cao nên ưu tiên những vùng khác trước

- Để thực hiện trực quan hóa một cách dễ dàng, nhóm tạo ra một view trong DDS để thể hiện tổng số lượng chết tại một PHU và tổng số lượng người tiêm vắc xin đầy đủ.

	123 PHUID	123 TotalDeath	123 FullyVaccinatedCumulative
1	2,244	220	1,677,315
2	2,226	6	374,888
3	2,227	27	473,904
4	2,261	32	653,692
5	2,246	473	1,529,861
6	2,251	574	3,453,271
7	3,895	3,367	9,770,564
8	2,240	23	325,730
9	2,241	3	701,030
10	2,255	23	470,444
11	5,183	61	451,519
12	2,253	1,058	5,026,477
13	2,270	991	3,840,541
14	2,234	87	353,726
15	2,266	151	986,728
16	2,249	6	268,748
17	2,242	85	397,288
18	2,247	5	403,874
19	2,238	16	531,309
20	2,237	433	1,829,991
21	2,262	64	502,690
22	2,257	10	323,493
23	2,230	532	2,292,305
24	2,260	285	1,884,150
25	2,236	251	1,986,789
26	2,243	109	614,123
27	2,268	501	1,357,285
28	4,913	87	656,265

* Code:

```
def clustering_phu(TotalDeath, FullyVaccinatedCumulative):
    dict = {'TotalDeath': TotalDeath, 'FullyVaccinatedCumulative': FullyVaccinatedCumulative}
    df = pd.DataFrame(dict)
    kmeans = KMeans(n_clusters=5)
    print(df)
    res = kmeans.fit_predict(df[['TotalDeath', 'FullyVaccinatedCumulative']])
    res = res.tolist()
    return res

client.deploy('clustering_phu', clustering_phu, 'clustering_phu', override=True)
```

8 Tham khảo và chú thích

- Sử dụng thư viện Pandas của Python để kiểm tra missing value và duplicate.
- Link google drive:
<https://drive.google.com/drive/folders/1VoBzcTw5DW29NtcujTyT56fiSVyFgceb?usp=sharing>