

**Thông tin thành viên và Phân Công**

| **MSSV** | **Họ và Tên** | **Phân công** | **Đóng góp** |
| --- | --- | --- | --- |
| 19127040 | Trần Ngọc Lam | * Thiết kế quy trình ETL * Cleaning data: file PublicHealthUnit và file PublicHealthUnitGroup * Viết script Stage * Viết script DDS * Cài đặt DDS * Mining * Làm báo cáo | 33.33% |
| 19127088 | Nguyễn Phương Vy | * Thiết kế quy trình ETL * Cleaning data: file Case Report và Compile * Viết script NDS * Cài đặt Stage * Cài đặt NDS * Cube * Làm báo cáo | 33.33% |
| 19127219 | Nguyễn Kim Thị Tố Nga | * Thiết kế quy trình ETL * Cleaning data: file OngoingOutbreak và Vaccine * Viết Script DDS * Cài đặt NDS * Quay clip demo * Dashboard * Làm báo cáo | 33.33% |

**Mục lục**

[**Phân tích dữ liệu nguồn và Làm sạch dữ liệu**](#_heading=h.gjdgxs) **3**

[**Phân tích các giai đoạn ETL**](#_heading=h.30j0zll) **14**

[Liệt kê các yêu cầu nghiệp vụ](#_heading=h.1fob9te) 14

[METADATA](#_heading=h.3znysh7) 14

[Stage Loading](#_heading=h.2et92p0) 14

[Populating NDS](#_heading=h.tyjcwt) 16

[Populating DDS](#_heading=h.3dy6vkm) 23

[**OLAP và Report:**](#_heading=h.aay91ssa6whh) **31**

[OLAP Cube](#_heading=h.mtyctnenpqsv) 31

[Report](#_heading=h.b54sgwfm19eq) 31

[**Data mining:**](#_heading=h.qhkfl2un9lm3) **34**

[Chuẩn bị dữ liệu](#_heading=h.vo0j463wj9bl) 34

[Thực hiện phân cụm với mô hình K-means bằng Python](#_heading=h.s48v4eggabks) 35

[Nhận xét](#_heading=h.kv42gizvql) 36

[Kết luận](#_heading=h.yj5rdy6awqne) 39

[**Link Youtube:**](#_heading=h.n8ubt4m72xe5) **40**

[Tham khảo:](#_heading=h.4d34og8) **40**

# **Phân tích dữ liệu nguồn và Làm sạch dữ liệu**

* **Public health unit:** Dữ liệu các đơn vị y tế cộng đồng của Ontario, Canada

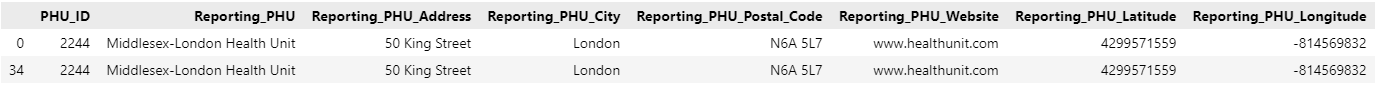
| **Tên thuộc tính** | **Mô tả** | **Số giá trị null** |
| --- | --- | --- |
| PHU\_ID | Mã định danh của đơn vị y tế cộng đồng | 0 |
| Reporting\_PHU | Tên đơn vị y tế cộng đồng được báo cáo | 0 |
| Reporting\_PHU\_Address | Địa chỉ đơn vị y tế cộng đồng được báo cáo | 0 |
| Reporting\_PHU\_City | Thành phố của đơn vị y tế cộng đồng được báo cáo | 0 |
| Reporting\_PHU\_Website | Trang web của đơn vị y tế cộng đồng được báo cáo | 0 |
| Reporting\_PHU\_Latitude | Vĩ tuyến của đơn vị y tế cộng đồng được báo cáo | 0 |
| Reporting\_PHU\_Longitude | Kinh tuyến của đơn vị y tế cộng đồng được báo cáo | 0 |

**Dữ liệu gốc:**

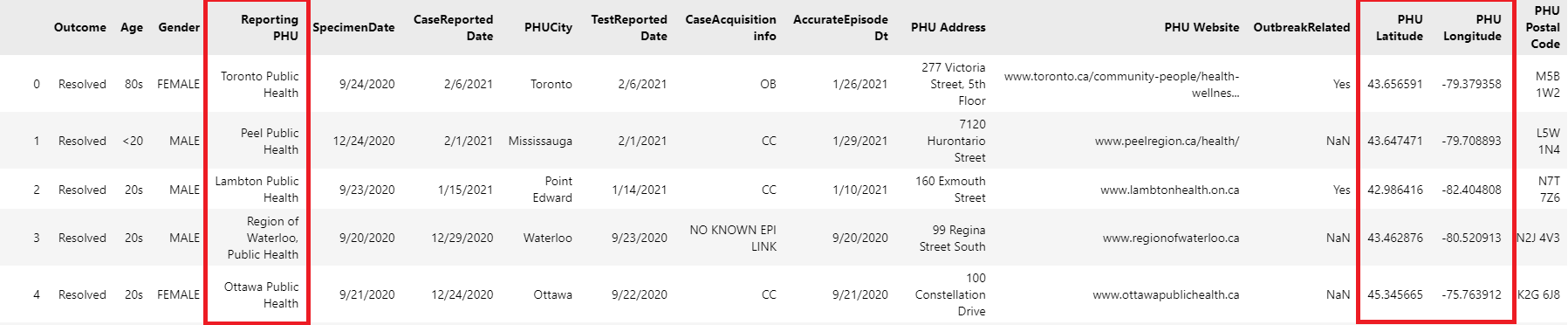


**Làm sạch dữ liệu**

* Dữ liệu có bị lặp, giữ lại dòng đầu tiên và xóa các dòng lặp dữ liệu.



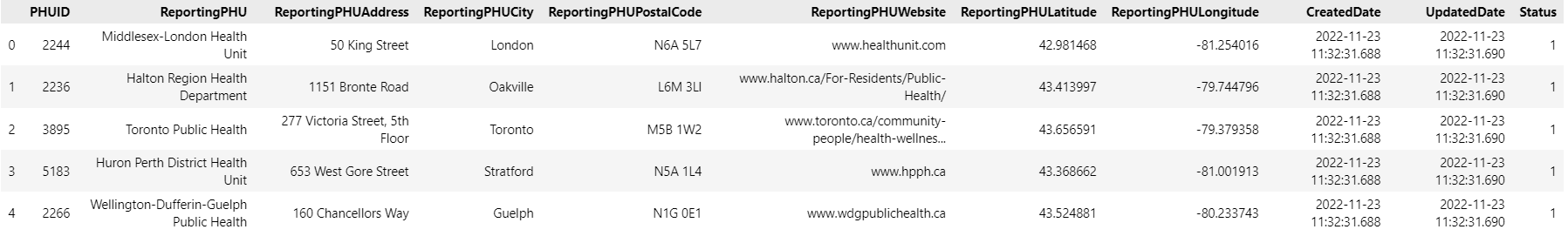
* Dữ liệu về kinh độ, vĩ độ của đơn vị y tế cộng đồng quá lớn, không hợp lý. Thực hiện gán lại giá trị kinh độ, vĩ độ bằng giá trị kinh độ, vĩ độ trong tập tin Case Report so khớp theo tên đơn vị y tế cộng đồng được báo cáo.



*(Dữ liệu của tập tin Case Report)*

* Thêm hai cột UpdatedDate và CreatedDate, tương ứng là ngày cập nhật và ngày tạo của dòng dữ liệu.
* Thêm cột Status là trạng thái của dữ liệu để phục vụ cho delete trong ETL.

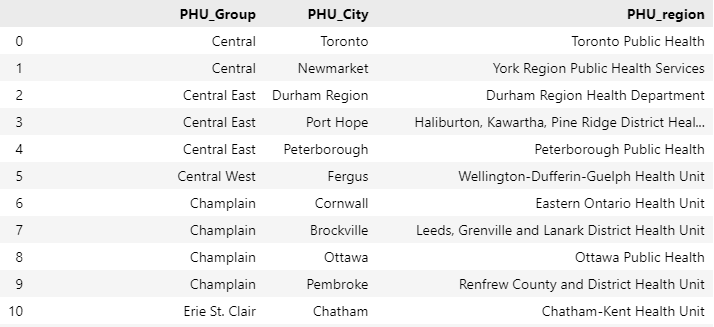
**Dữ liệu sau làm sạch:**



* **Public Health Units Group:** Các đơn vị chăm sóc sức khỏe cộng đồng phân bố theo vùng miền và thành phố tại Ontario, Canada

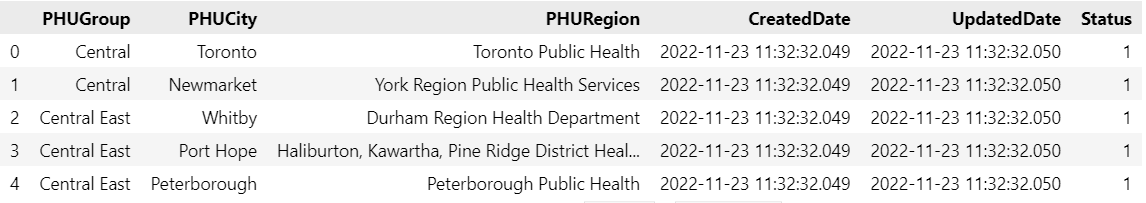
| **Tên thuộc tính** | **Mô tả** | **Số giá trị null** |
| --- | --- | --- |
| PHU\_Group | Tên vùng miền mà đơn vị y tế cộng đồng thuộc về | 0 |
| PHU\_City | Thành phố của đơn vị y tế cộng đồng | 0 |
| PHU\_region | Tên đơn vị y tế cộng đồng thuộc vùng và thành phố tương ứng | 0 |

**Dữ liệu gốc:**



* Tên thành phố và tên PHU không khớp với tập tin *Public health units*, đồng nhất tên thành phố và tên PHU theo tập tin *Public health units* (đã kiểm tra với thông tin thực).
* Thêm hai cột UpdatedDate và CreatedDate, tương ứng là ngày cập nhật và ngày tạo của dòng dữ liệu.
* Thêm cột Status là trạng thái của dữ liệu để phục vụ cho delete trong ETL.

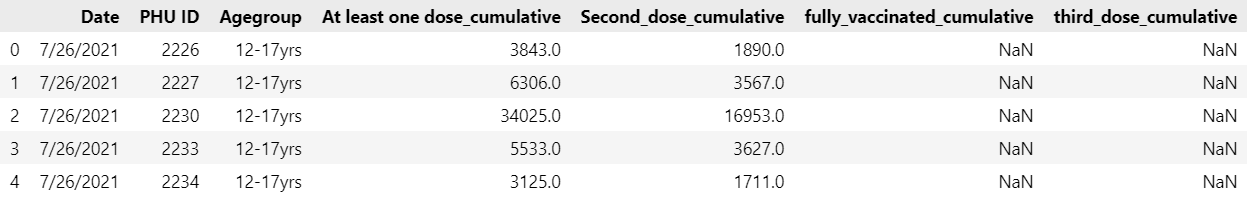
**Dữ liệu sau làm sạch:**

**

* **Vaccine\_by\_age\_phu**: Dữ liệu về vaccine Covid - 19 theo đơn vị chăm sóc sức khỏe cộng đồng và nhóm tuổi

| **Tên thuộc tính** | **Mô tả** | **Số giá trị null** |
| --- | --- | --- |
| Date | Ngày báo cáo về các liều vắc-xin COVID-19 của Ontario được tiêm theo độ tuổi. | 0 |
| PHU\_ID | Số định danh của đơn vị y tế cộng đồng | 0 |
| Agegroup | Phân loại tuổi theo các nhóm sau:   * 5 to 11 years old * 12 to 17 year olds * 18 to 29 years old * 30 to 39 years old * 40 to 49 years old * 50 to 59 years old * 60 to 69 years old * 70 to 79 years old * 80 years and older * Adults\_18plus: tổng số tất cả các cá nhân từ 18 tuổi trở lên. * Ontario\_12plus: tổng số tất cả các cá nhân từ 12 tuổi trở lên. * Ontario\_5plus: tổng số tất cả các cá nhân từ 5 tuổi trở lên. * Undisclosed\_or\_missing: số lượng cá nhân không cung cấp ngày sinh hoặc gửi ngày sinh không hợp lệ. | 0 |
| At least one dose\_cumulative | Số lượng cá nhân đã tiêm được ít nhất một liều. | 211 |
| Second\_dose\_cumulative | Số lượng cá nhân đã tiêm được hai liều. | 122540 |
| fully\_vaccinated\_cumulative | Tổng số người được tiêm phòng đầy đủ. Tiêm phòng đầy đủ được định nghĩa là được tiêm:   * một liều Janssen (Johnson & Johnson), hoặc * hai liều bất kỳ loại vắc xin nào được Bộ Y tế Canada phê chuẩn, hoặc * một liều vắc xin không được Bộ Y tế Canada phê chuẩn, sau đó là một liều vắc xin được Bộ Y tế Canada phê duyệt, hoặc * ba liều vắc xin bất kỳ được Bộ Y tế Canada phê duyệt, hoặc * ba liều vắc xin, cho dù vắc xin đó có được Bộ Y tế Canada phê duyệt hay không. | 49270 |
| third\_dose\_cumulative | Tổng số người được tiêm phòng ba liều. Được định nghĩa là nhận ba liều của bất kỳ loại vắc xin nào được Bộ Y tế Canada phê chuẩn. Những cá nhân này là một tập hợp con của những cá nhân được chọn trong full\_vaccinated\_cumulative. | 65167 |

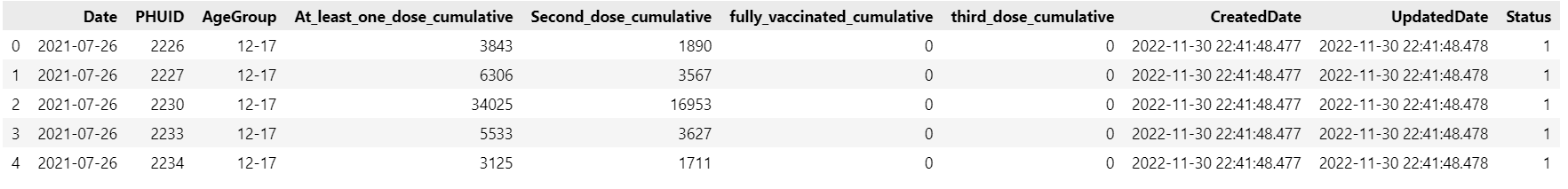
**Dữ liệu gốc:**

****

**Làm sạch dữ liệu:**

* Đồng nhất dữ liệu cột Agegroup
* Điền khuyết các cột có dữ liệu Nan bằng 0 (Không có dữ liệu coi như số mũi tiêm bằng 0).
* Bỏ dữ liệu cả 4 cột các cột At\_least\_one\_dose\_cumulative, Second\_dose\_cumulative, fully\_vacinated\_cumulative, third\_dose\_cumulative đều bằng 0.
* Bỏ dữ liệu có cột PHU ID bằng 9999 do không có dữ liệu về PHU này.
* Thêm cột CreatedDate, UpdatedDate
* Thêm cột Status là trạng thái của dữ liệu để phục vụ cho delete trong ETL.

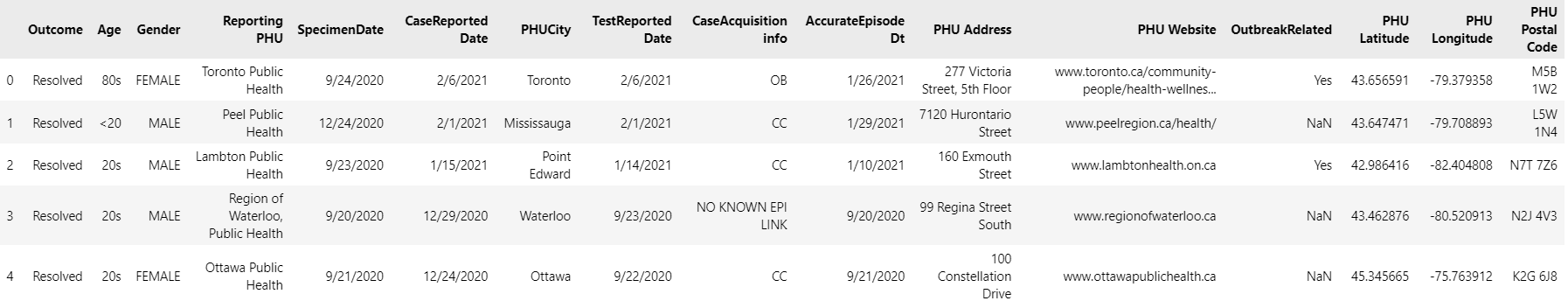
**Dữ liệu sau khi làm sạch:**

****

* **Case Report**: Dữ liệu về các ca nhiễm của tỉnh bang Ontario

| **Tên thuộc tính** | **Mô tả** | **Số giá trị null** |
| --- | --- | --- |
| Outcome | Trạng thái ca nhiễm:   * Resolved: Phục hồi * Fatal: Tử vong * Not Resolved: Chưa phục hồi | 0 |
| Age | Tuổi gồm các nhóm sau: <20, 20s, 30s, 40s, 50s, 60s, 70s, 80s, 90+ | 0 |
| Gender | Giới tính:   * FEMALE: nữ * MALE: nam * UNSPECIFIED: không xác định * GENDER DIVERSE: đa giới tính | 0 |
| Reporting PHU | Tên đơn vị y tế cộng đồng được báo cáo | 0 |
| SpecimenDate | Ngày lấy mẫu | 208 |
| CaseReported Date | Ngày báo cáo đến đơn vị y tế địa phương | 0 |
| PHUCity | Thành phố của đơn vị y tế cộng đồng được báo cáo | 0 |
| TestReported Date | Ngày trả kết quả | 418 |
| CaseAcquisition info | Thông tin về ca nhiễm:   * CC (Closed contact): dương tính xác định được nguồn lây * OB (Outbreak): dương tính trong đợt bùng phát * NO KNOWN EPI LINK/ UNSPECIFIED EPI LINK: Không xác định được nguồn lây * TRAVEL: dương tính do du lịch * MISSING INFORMATION: không có thông tin | 0 |
| AccurateEpisode Dt | Ngày khởi phát bệnh | 0 |
| PHU Address | Địa chỉ đơn vị y tế cộng đồng được báo cáo | 0 |
| PHU Website | Trang web của đơn vị y tế cộng đồng được báo cáo | 0 |
| OutbreakRelated | Ca nhiễm có liên quan đến đợt bùng phát hay không:   * Yes: có liên quan * Nan: không có thông tin | 19790 |
| PHU Latitude | Vĩ tuyến của đơn vị y tế cộng đồng được báo cáo | 0 |
| PHU Longitude | Kinh tuyến của đơn vị y tế cộng đồng được báo cáo | 0 |
| PHU Postal Code | Mã bưu điện của đơn vị y tế cộng đồng được báo cáo | 0 |

**Dữ liệu gốc:**



**Làm sạch dữ liệu:**

* Đồng nhất dữ liệu cột ***CaseAcquisition info***:

| **File: Case Report**  **Column: CaseAcquisition info** | **File: Compiled\_COVID-19\_Case\_Details\_(Canada)**  **Column: exposure** | **Dữ liệu**  **thống nhất** |
| --- | --- | --- |
| OB | Outbreak | OB |
| CC | Close Contact | CC |
| NO KNOWN EPI LINK |  | NO KNOWN EPI LINK |
| TRAVEL | Travel - Related | TRAVEL |
| MISSING INFORMATION | Not Reported | NOT REPORTED |
| UNSPECIFIED EPI LINK |  | NO KNOWN EPI LINK |

* Đồng nhất dữ liệu cột ***Outcome***:

| **File: Case Report**  **Column: Outcome** | **File: Compiled\_COVID-19\_Case\_Details\_(Canada)**  **Column: case\_status** | **Dữ liệu**  **thống nhất** |
| --- | --- | --- |
| Resolved | Recovered | Recovered |
| Fatal | Deceased | Deceased |
| Not Resolved | Active | Active |
|  | Not Reported | Not Reported |

* Đồng nhất dữ liệu cột ***Gender***:

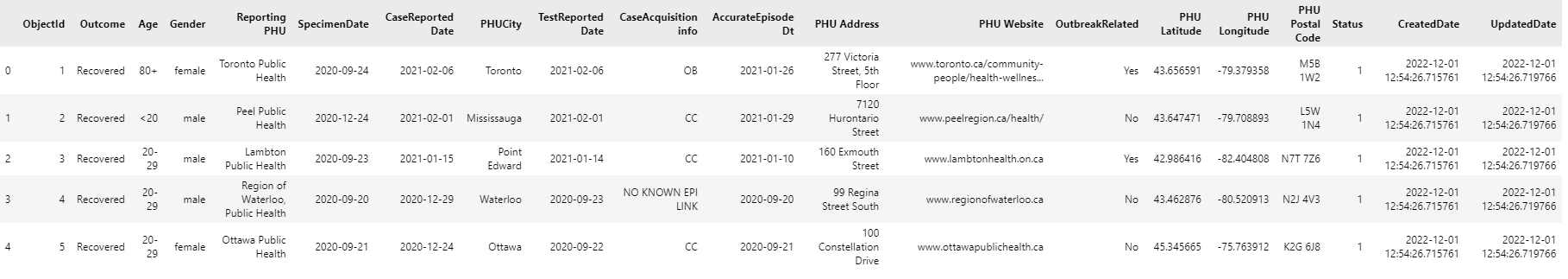
| **File: Case Report**  **Column: Gender** | **File: Compiled\_COVID-19\_Case\_Details\_(Canada)**  **Column: gender** | **Dữ liệu thống nhất** |
| --- | --- | --- |
| FEMALE | Female | female |
| MALE | Male | male |
| UNSPECIFIED |  | unspecified |
| GENDER DIVERSE |  | gender diverse |
|  | Not Reported | not reported |

* Đồng nhất dữ liệu cột ***Age***:

| **File: Case Report**  **Column: Age** | **File: Compiled\_COVID-19\_Case\_Details\_(Canada)**  **Column: age\_group** | **File: vaccines\_by\_age\_phu**  **Column: Agegroup** | **Dữ liệu thống nhất** |
| --- | --- | --- | --- |
|  |  | 05-11yrs | 05-11 |
|  |  | 12-17yrs | 12-17 |
|  |  | 18-29yrs | 18-29 |
| <20 | <20 |  | <20 |
| 20s | 20-29 |  | 20-29 |
| 30s | 30-39 | 30-39yrs | 30-39 |
| 40s | 40-49 | 40-49yrs | 40-49 |
| 50s | 50-59 | 50-59yrs | 50-59 |
| 60s | 60-69 | 60-69yrs | 60-69 |
| 70s | 70-79 | 70-79yrs | 70-79 |
| 80s | 80+ | 80+ | 80+ |
| 90+ |  |  | 80+ |
|  |  | Adults\_18plus | Adults\_18plus |
|  |  | Ontario\_12plus | Ontario\_12plus |
|  |  | Ontario\_5plus | Ontario\_5plus |
|  |  | Undisclosed\_or\_missing | Not Reported |

* Trường Agegroup của file vaccines\_by\_age\_phu có một số giá trị khác biệt là: 05-11yrs, 12-17yrs, 18-29yrs, Adults\_18plus, Ontario\_12plus, Ontario\_5plus. Các giá trị này không thể gộp chung với các giá trị của file khác như <20 do đặc trưng của vaccine có thể được tiêm phụ thuộc vào độ tuổi.
* Trường Age của file Case Report: gộp giá trị 80s và 90+ thành 80+ dựa vì cả hai độ tuổi này đều chịu sự ảnh hưởng nghiêm trọng của covid như nhau.
* Cột SpecimenDate, TestReported Date: điền khuyết bằng giá trị median của cột SpecimenDate (giá trị median không bị ảnh hưởng bởi các giá trị outliers).
* Cột OutbreakRelated: điền ‘Yes’ cho các trường dữ liệu có CaseAcquisition info là OB; điền ‘No’ cho các trường dữ liệu còn lại.
* Thêm cột CreatedDate và UpdatedDate cho dữ liệu.
* Thêm cột Status là trạng thái của dữ liệu để phục vụ cho delete trong ETL.
* Thêm cột ObjectId (có giá trị từ 1) để định danh từng dòng / từng ca nhiễm của dữ liệu.

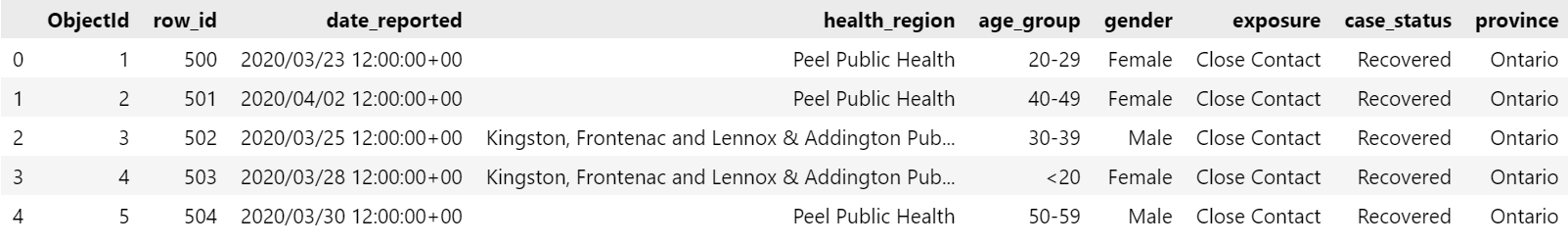
**Dữ liệu sau khi làm sạch**



* **Compiled\_COVID-19\_Case\_Details\_(Canada):** Dữ liệu ca nhiễm của tất cả các tỉnh bang của Canada

| **Tên thuộc tính** | **Mô tả** | **Số giá trị null** |
| --- | --- | --- |
| ObjectId | Mã định danh của dữ liệu trong tập dữ liệu hiện tại | 0 |
| row\_id | Mã định danh của dữ liệu trong tập dữ liệu nguồn (có nhiều nguồn dữ liệu được tổng hợp) | 0 |
| date\_reported | Ngày ca nhiễm được báo cáo lên đơn vị y tế địa phương | 0 |
| health\_region | Tên đơn vị y tế cộng đồng được báo cáo | 0 |
| age\_group | Phân loại tuổi theo các nhóm sau:   * <20 * 20-29 * 30-39 * 40-49 * 50-59 * 60-69 * 70-79 * 80+ * Not Reported | 0 |
| gender | Giới tính bệnh nhân   * Female: nữ * Male: nam * Not Reported: không có dữ liệu báo cáo | 0 |
| exposure | Thông tin về ca nhiễm:   * Closed contact: dương tính xác định được nguồn lây * Outbreak: dương tính trong đợt bùng phát * Travel-Related: dương tính do du lịch * Not Reported: không có thông tin | 0 |
| case\_status | Trạng thái ca nhiễm:   * Resolved: Phục hồi * Deceased: Tử vong * Active: Chưa phục hồi * Not Reported: Không có dữ liệu báo cáo | 0 |
| province | Tỉnh bang của Canada | 0 |

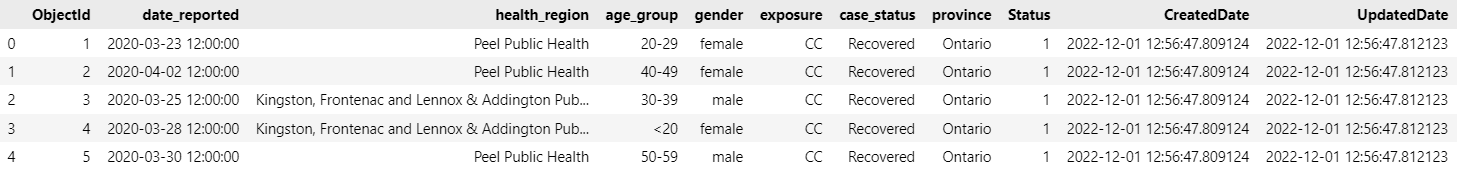
**Dữ liệu gốc:**



**Làm sạch dữ liệu:**

* Đồng nhất dữ liệu age\_group, exposure, gender, case\_status
* Thêm cột CreatedDate và UpdatedDate
* Thêm cột Status là trạng thái của dữ liệu để phục vụ cho delete trong ETL.
* Bỏ cột row\_id do đây là thuộc tính định danh ca nhiễm được thu thập từ nhiều nguồn khác nhau trên toàn Canada, chỉ cần cột ObjectId là định danh ca nhiễm của dữ liệu hiện tại.
* Bỏ các dữ liệu có cột province khác ‘Ontario’ do không có dữ liệu về các health\_region thuộc các tỉnh bang này và dữ liệu không được sử dụng để phục vụ cho việc thống kê số ca nhiễm của Ontario.

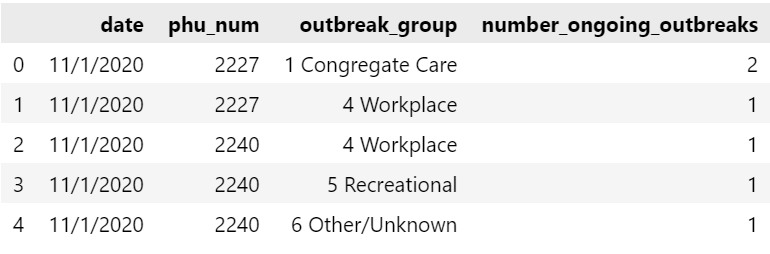
**Dữ liệu sau khi làm sạch:**



* **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** | **Số giá trị null** |
| --- | --- | --- |
| date | Ngày báo cáo | 0 |
| phu\_num | Mã định danh của đơn vị y tế cộng đồng | 0 |
| outbreak\_group | Cơ sở bùng phát dịch:   * 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 | 0 |
| number\_ongoin\_outbreaks | Số đợt bùng phát đang diễn ra | 0 |

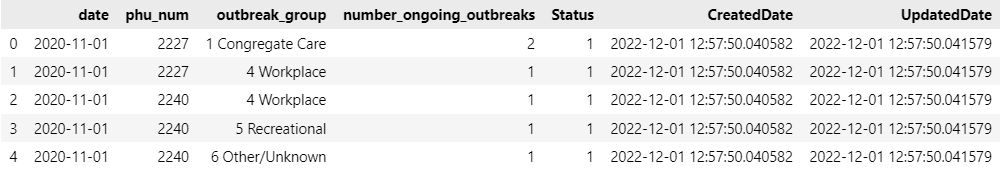
**Dữ liệu gốc:**



**Làm sạch dữ liệu:**

* Thêm cột CreatedDate và UpdatedDate
* Thêm cột Status là trạng thái của dữ liệu để phục vụ cho delete trong ETL.

**Dữ liệu sau khi làm sạch:**



# Phân tích các giai đoạn ETL

## Liệt kê các yêu cầu nghiệp vụ

Lưu trữ dữ liệu thông tin các ca nhiễm vào datawarehouse để phục vụ các thống kê sau:

* **Số ca nhiễm, số ca tử vong, số ca phục hồi** của dịch Covid-19 theo từng **PHU** trong từng **năm.**
* **Mức Độ Nghiêm Trọng** của dịch Covid-19 theo **PHU** và theo các **Quý** trong từng **năm**.
* **Tổng số người tử vong** theo **Giới Tính** và **Nhóm Tuổi** theo các năm.
* **Số ca nhiễm, tử vong** theo **Mức Độ Nghiêm Trọng** theo **Ngày** của các **năm**.
* **Số ca nhiễm, tử vong** theo **Mức Độ Nghiêm Trọng**, **nhóm bùng phát,**  **khu vực (PHU\_Group, City)** trong các **năm.**
* **Số người đã được tiêm vacxin** trong các **năm**.

## **METADATA**

Cơ sở dữ liệu METADATA sẽ có 2 bảng:

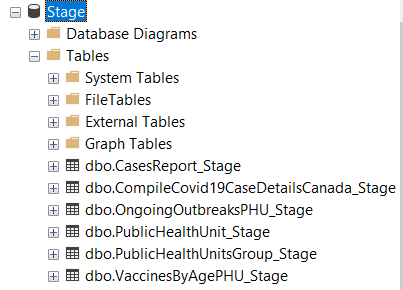
* **Data\_Flow** dùng để chứa thông tin LSET khi đổ dữ liệu từ nguồn vào Stage.
* **DDS** dùng để chứa thông tin LSET khi đổ dữ liệu từ NDS vào DDS.

## **Stage Loading**

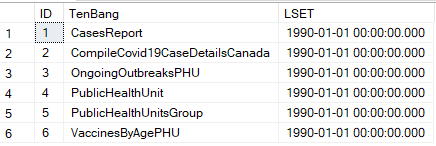
***Tạo cơ sở dữ liệu Stage*** để đổ dữ liệu từ nguồn vào Stage. Số bảng trong Stage và các thuộc tính trong bảng hoàn toàn giống với dữ liệu nguồn. Dữ liệu đổ vào Stage sẽ là các dòng dữ liệu thỏa điều kiện:

(CreatedDate < CET AND CreatedDate >= LSET)

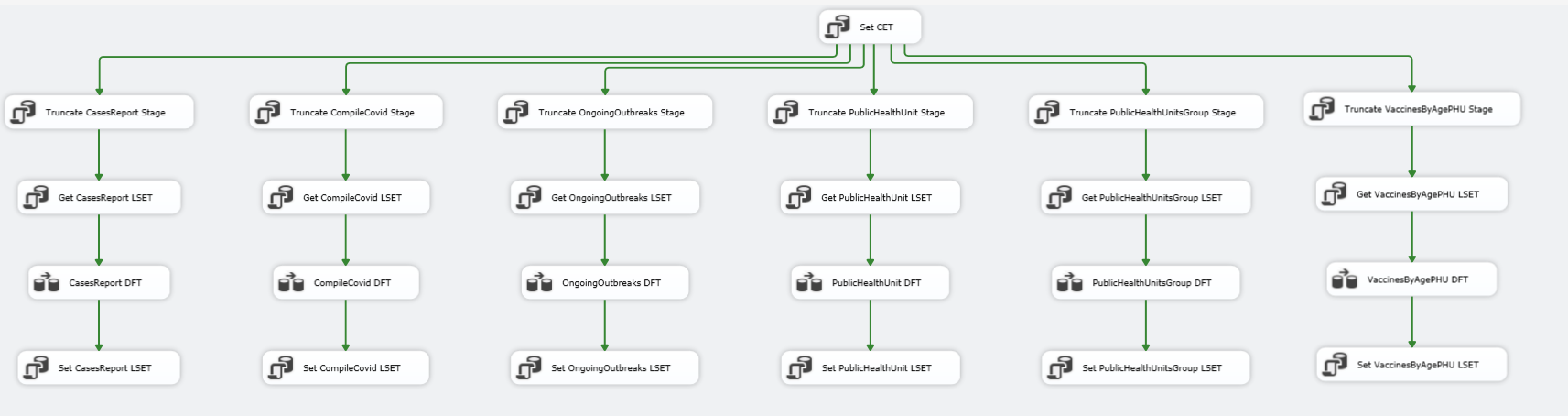
OR ( UpdatedDate < CET AND UpdatedDate >= LSET)



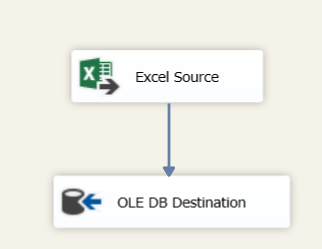
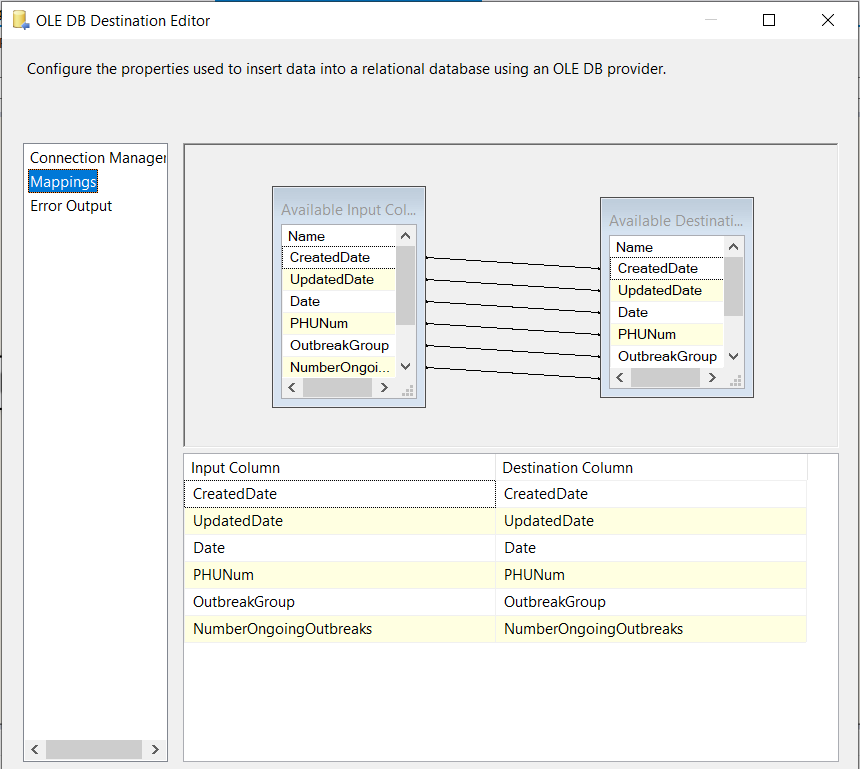
Trong đó, LSET được lưu trữ trong bảng ***Data\_Flow*** trong database ***METADATA***:



* ***Quy trình đổ dữ liệu từ Source vào Stage:***



1. Cập nhật biến ***CET*** với thời gian hiện tại bắt đầu đổ dữ liệu CET = getdate()
2. Truncate các bảng trong Stage
3. Lấy ra ***LSET*** (thời gian cập nhật cuối cùng) từ METADATA
4. Tiến hành đổ dữ liệu từ Source (file excel) vào Stage



*(Đổ dữ liệu file OngoingOutbreaks vào Stage)*

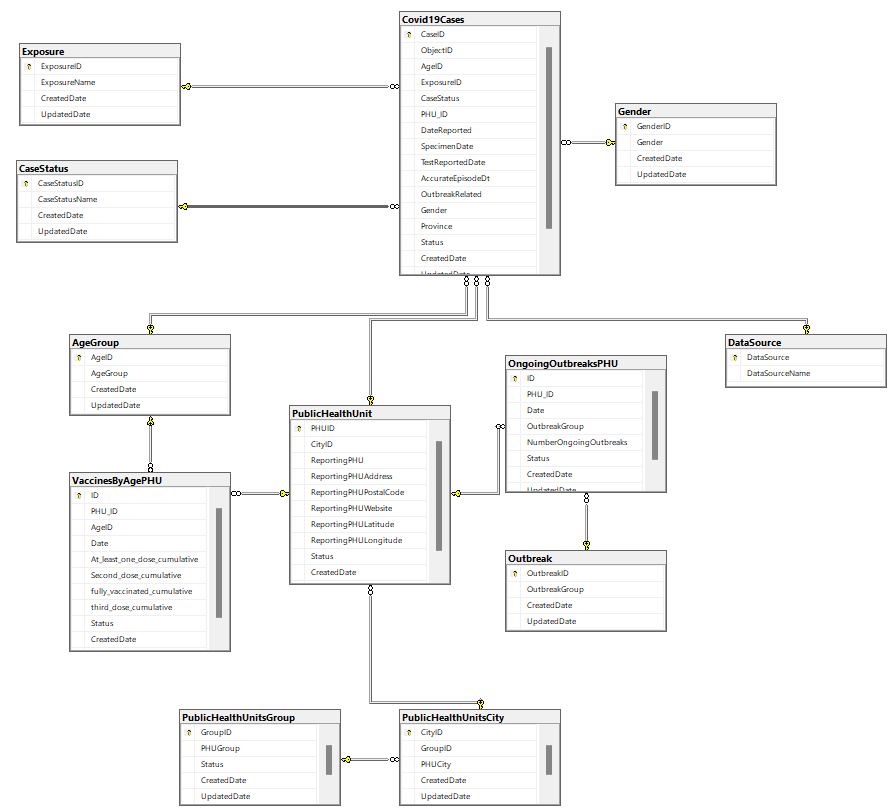
1. Cập nhất lại giá trị LSET = CET trong METADATA

## **Populating NDS**

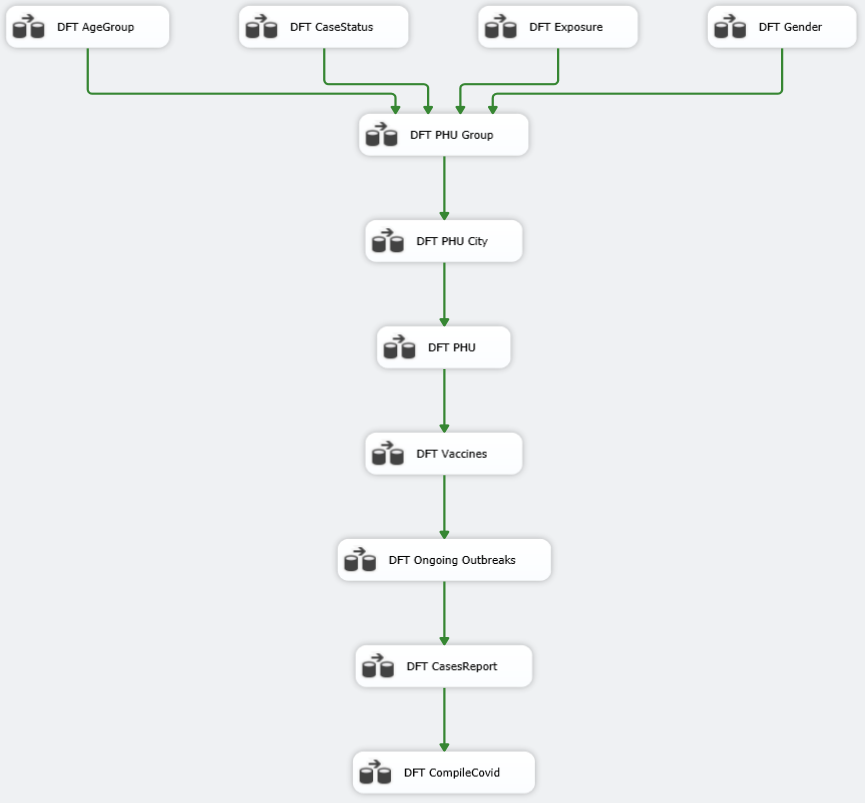
* ***Chuẩn hóa dữ liệu:***
* Gộp dữ liệu ca nhiễm của hai bảng CaseReport và CompileCovid19CaseDetailsCanada thành 1 bảng Covid19Cases.
* Tách các dữ liệu như Age, Gender, Exposure, CaseStatus, Outbreak ra các bảng riêng để phục vụ cho việc chuẩn hóa dữ liệu và quá trình DDS.
* Tạo một bảng DataSource để phân biệt dữ liệu ở file CasesReport.xlsx (1) và Compiled\_COVID-19\_Case\_Details\_(Canada).xlsx (2)
* ***Đặc tả thiết kế NDS:***

| **Tên thuộc tính** | **Kiểu dữ liệu** | **Mô tả** |
| --- | --- | --- |
| **Covid19Cases** | | |
| CaseID | int | Mã định danh của ca nhiễm trong NDS |
| ObjectID | int | Mã định danh của ca nhiễm trong dữ liệu gốc |
| AgeID | int | Mã định danh nhóm tuổi |
| ExposureID | int | Mã định danh thông tin về ca nhiễm |
| CaseStatus | int | Mã định danh trạng thái ca nhiễm |
| PHU\_ID | int | Mã định danh đơn vị y tế công cộng |
| DateReported | datetime | Ngày báo cáo |
| SpecimenDate | datetime | Ngày lấy mẫu |
| TestReportedDate | datetime | Ngày trả kết quả |
| AccurateEpisodeDt | datetime | Ngày khỏi phát bệnh |
| OutbreakRelated | [nvarchar](255) | Liên quan đến đợt bùng phát |
| Gender | [nvarchar](255) | Mã định danh giới tính |
| Province | [nvarchar](255) | Tỉnh bang |
| Status | int | Trạng thái |
| CreatedDate | datetime | Ngày tạo dòng dữ liệu |
| UpdatedDate | datetime | Ngày cập nhật dòng dữ liệu |
| DataSource | int | Bảng nguồn của dữ liệu |
| **OngoingOutbreaksPHU** | | |
| ID | int | Mã định danh của đợt bùng phát |
| PHU\_ID | int | Mã định danh của đơn vị y tế cộng đồng |
| Date | datetime | Ngày báo cáo |
| OutbreakGroup | int | ID Cơ sở bùng phát dịch |
| NumberOngoingOutbreaks | int | Số đợt bùng phát đang diễn ra |
| Status | int | Trạng thái |
| CreatedDate | datetime | Ngày tạo dòng dữ liệu |
| UpdatedDate | datetime | Ngày cập nhật dòng dữ liệu |
| **PublicHealthUnit** | | |
| PHUID | int | Mã định danh của đơn vị y tế cộng đồng |
| CityID | int | Mã định danh thành phố mà đơn vị y tế thuộc về |
| ReportingPHU | [nvarchar](255) | Tên đơn vị y tế cộng đồng |
| ReportingPHUAddress | [nvarchar](255) | Địa chỉ đơn vị y tế cộng đồng |
| ReportingPHUPostalCode | [nvarchar](255) | Mã bưu điện đơn vị y tế cộng đồng |
| ReportingPHUWebsite | [nvarchar](255) | Trang web đơn vị y tế cộng đồng |
| ReportingPHULatitude | [nvarchar](50) | Vĩ độ đơn vị y tế cộng đồng |
| ReportingPHULongitude | [nvarchar](50) | Kinh độ đơn vị y tế cộng đồng |
| Status | int | Trạng thái |
| CreatedDate | datetime | Ngày tạo dòng dữ liệu |
| UpdatedDate | datetime | Ngày cập nhật dòng dữ liệu |
| **VaccinesByAgePHU** | | |
| ID | int | Mã định danh của báo cáo về các liều vắc-xin COVID-19 của Ontario được tiêm theo độ tuổi. |
| PHU\_ID | int | Mã định danh của đơn vị y tế cộng đồng |
| AgeID | int | Mã định danh của nhóm tuổi |
| Date | datetime | Ngày báo cáo |
| At\_least\_one\_dose\_cumulative | int | Số lượng cá nhân đã tiêm được ít nhất một liều. |
| Second\_dose\_cumulative | int | Số lượng cá nhân đã tiêm được hai liều. |
| fully\_vaccinated\_cumulative | int | Tổng số người được tiêm phòng đầy đủ. |
| third\_dose\_cumulative | int | Tổng số người được tiêm phòng ba liều. |
| Status | int | Trạng thái |
| CreatedDate | datetime | Ngày tạo dòng dữ liệu |
| UpdatedDate | datetime | Ngày cập nhật dòng dữ liệu |
| **PublicHealthUnitsGroup** | | |
| GroupID | int | Mã định danh của nhóm đơn vị y tế |
| PHUGroup | [nvarchar](255) | Tên khu vực đơn vị y tế |
| CreatedDate | datetime | Ngày tạo dòng dữ liệu |
| Status | int | Trạng thái |
| UpdatedDate | datetime | Ngày cập nhật dòng dữ liệu |
| **PublicHealthUnitsCity** | | |
| CityID | int | Mã định danh thành phố |
| GroupID | int | Mã định danh của nhóm đơn vị y tế |
| PHUCity | [nvarchar](255) | Tên thành phố |
| CreatedDate | datetime | Ngày tạo dòng dữ liệu |
| UpdatedDate | datetime | Ngày cập nhật dòng dữ liệu |
| **AgeGroup** | | |
| AgeID | int | Mã định danh của nhóm tuổi |
| AgeGroup | [nvarchar](255) | Nhóm tuổi |
| CreatedDate | datetime | Ngày tạo dòng dữ liệu |
| UpdatedDate | datetime | Ngày cập nhật dòng dữ liệu |
| **CaseStatus** | | |
| CaseStatusID | int | Mã định danh trạng thái ca nhiễm |
| CaseStatusName | [nvarchar](255) | Tên trạng thái |
| CreatedDate | datetime | Ngày tạo dòng dữ liệu |
| UpdatedDate | datetime | Ngày cập nhật dòng dữ liệu |
| **Exposure** | | |
| ExposureID | int | Mã định danh thông tin về ca nhiễm |
| ExposureName | [nvarchar](255) | Tên thông tin ca nhiễm |
| CreatedDate | datetime | Ngày tạo dòng dữ liệu |
| UpdatedDate | datetime | Ngày cập nhật dòng dữ liệu |
| **Gender** | | |
| GenderID | int | Mã định danh giới tính |
| Gender | [nvarchar](50) | Tên giới tính |
| CreatedDate | datetime | Ngày tạo dòng dữ liệu |
| UpdatedDate | datetime | Ngày cập nhật dòng dữ liệu |
| **Outbreak** | | |
| OutbreakID | int | Mã định danh nhóm bùng phát |
| OutbreakGroup | [nvarchar](255) | Tên nhóm bùng phát |
| CreatedDate | datetime | Ngày tạo dòng dữ liệu |
| UpdatedDate | datetime | Ngày cập nhật dòng dữ liệu |
| **DataSource** | | |
| ID | int | Mã định danh của bảng |
| DataSource | [nvarchar](255) | Tên bảng dữ liệu |

* ***Thiết kế cơ sở dữ liệu NDS:***

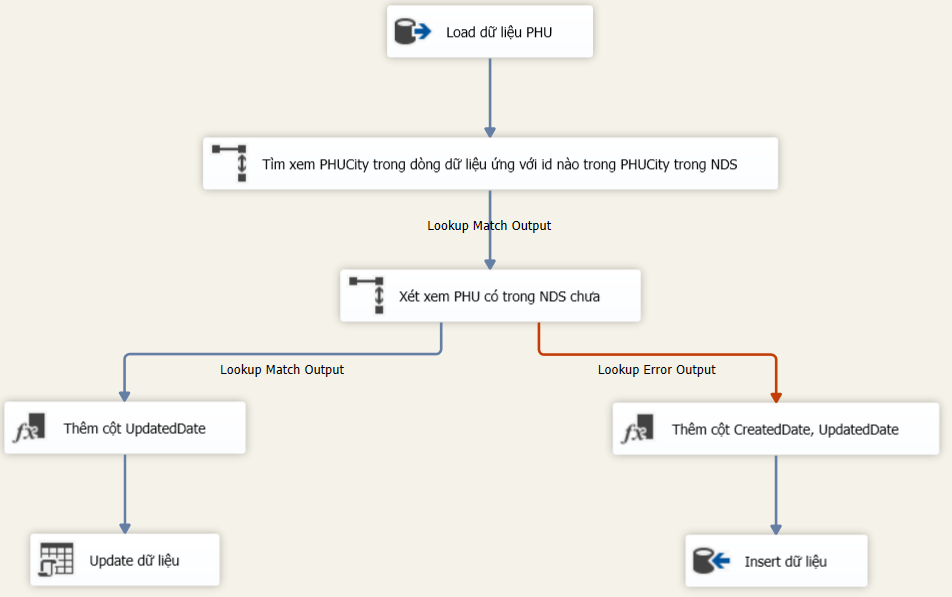


* ***Quy trình đổ dữ liệu từ Stage vào NDS:***
* Dữ liệu sẽ được đổ theo thứ tự khóa ngoại như trong hình.



***Với mỗi Data Flow Task:***

1. Load dữ liệu từ Stage.
2. Chuẩn hóa dòng dữ liệu bằng cách tìm xem dữ liệu đó ứng với ID nào trong NDS. (Bảng có bao nhiêu khóa ngoại thì LookUp ID của khóa ngoại bấy nhiêu lần).
3. Tìm xem dòng dữ liệu có nằm trong NDS hay chưa.
4. Nếu rồi thì thêm cột UpdatedDate và cập nhật dữ liệu trong NDS.
5. Nếu chưa thì thêm cột CreatedDate và UpdatedDate và insert dữ liệu vào NDS.



*(Data Flow Task trong NDS)*

## **Populating DDS**

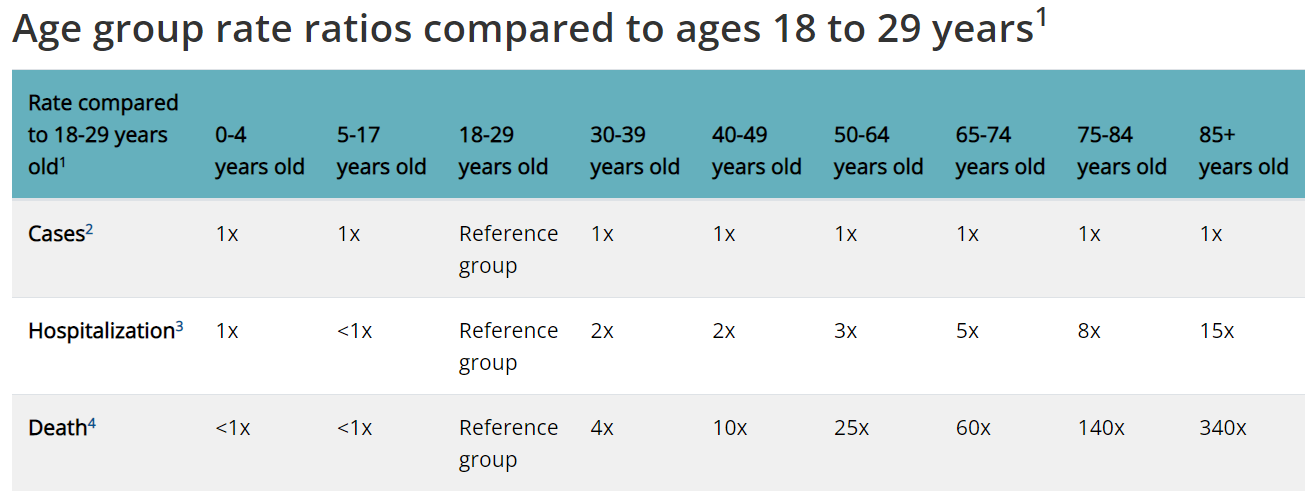
1. **Xác định yêu cầu nghiệp vụ:**

* Thống kê Số ca nhiễm, số ca tử vong, số ca phục hồi của dịch Covid-19 theo từng PHU trong từng năm.
* Thống kê Mức Độ Nghiêm Trọng 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)
* Thống kê số người đã được tiêm vacxin trong các năm.
* Thống kê số ca nhiễm theo Mức Độ Nghiêm Trọng, nhóm bùng phát của từng khu vực trong các năm
* ***Xác định mức độ nghiêm trọng:***

Mức độ nghiêm trọng được đo lường dựa trên tình trạng ca nhiễm và độ tuổi của bệnh nhân.

Độ tuổi bệnh nhân được chia thành 5 mức độ dựa theo khảo sát trên toàn cầu:

<https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-age.html>



* Mỗi tình trạng và mỗi độ tuổi sẽ tương ứng với số điểm như bảng sau:

|  | **0** | **1** | **2** | **3** | **4** |
| --- | --- | --- | --- | --- | --- |
| **Outcome** | recovered | active |  |  | death |
| **Age** | <30 | 30-49 | 49-69 | >=70 | >=80 |

* Mức độ nghiêm trọng của ca nhiễm sẽ được tính bằng công thức outcome + age như một số ví dụ dưới đây:
* Ca nhiễm đã phục hồi và có độ tuổi dưới 30: 0 + 0 = 0
  + Ca nhiễm đang điều trị và có độ tuổi 49-69 : 1 + 2 = 3
  + Ca nhiễm đã mất và có độ tuổi >= 70 và < 80 : 4 + 3 = 7

**Từ các số điểm trên sẽ xếp hạng mức độ nghiêm trọng theo bảng dưới đây**:

| **Low** | **Moderate** | **High** | **Critical** |
| --- | --- | --- | --- |
| **0, 1** | **2** | **3** | **>=4** |

1. **Xác định độ mịn của dữ liệu:**

* Số ca theo loại ca nhiễm (đang nhiễm, tử vong, phục hồi), theo PHU, giới tính, nhóm tuổi, nhóm bùng phát, mức độ nghiêm trọng của từng ngày.
* Số người tiêm vaccine theo PHU, theo tuổi trong từng năm.

1. **Xác định các Dimension**

Sau khi xác định yêu cầu nghiệp vụ và độ mịn dữ liệu, thiết kế và phân cấp trong các dimension như sau:

| **Tên thuộc tính** | **Kiểu dữ liệu** | **Mô tả** |
| --- | --- | --- |
| **DimPHU ( Group > City > PHU )** | | |
| PHUID\_SK | int | Surrogate Key của bảng DimPHU |
| PHUID | int | ID của đơn vị y tế cộng động  trong NDS |
| PHUName | [nvarchar](255) | Tên của đơn vị y tế cộng động |
| PHUAddress | [nvarchar](255) | Địa chỉ của đơn vị y tế cộng động |
| PHUWebsite | [nvarchar](255) | Trang web của đơn vị y tế cộng động |
| PHULatitude | [nvarchar](255) | Vĩ độ của đơn vị y tế cộng động |
| PHULongitude | [nvarchar](255) | Kinh độ của đơn vị y tế cộng động |
| CityID | int | ID của thành phố |
| CityName | [nvarchar](255) | Tên của thành phố |
| GroupID | int | ID của vùng miền |
| GroupName | [nvarchar](255) | Tên của vùng miền |
| Status | int | Trạng thái để hỗ trợ xóa dữ liệu |
| EffectiveDate | datetime | Ngày có hiệu lực (phục vụ cho SCD loại 2) |
| ExpiryDate | datetime | Ngày hết hiệu lực (phục vụ cho SCD loại 2) |
| **DimGender** | | |
| GenderID | int | ID của giới tính |
| GenderName | [nvarchar](255) | Giới tính |
| **DimAge** | | |
| AgeID | int | ID của nhóm tuổi |
| Age | [nvarchar](255) | Nhóm tuổi |
| **DimExposure** | | |
| ExposureID | int | ID của thông tin về ca nhiễm |
| ExposureName | [nvarchar](255) | Thông tin về ca nhiễm |
| **DimOutcome** | | |
| OutcomeID | int | ID của trạng thái ca nhiễm |
| Outcome | [nvarchar](255) | Trạng thái ca nhiễm |
| **DimSeverity** | | |
| SeverityID | int | ID mức độ nghiêm trọng của ca nhiễm |
| Severity | [nvarchar](255) | Mức độ nghiêm trọng của ca nhiễm (Low, Moderate, High, Critical) |
| **DimOutbreak** | | |
| OutbreakID | int | ID nhóm bùng phát |
| OutbreakGroup | [nvarchar](255) | Nhóm bùng phát |
| **DimDate (Year > Quarter > Month > Date)** | | |
| DateID | int | ID Ngày |
| Date | datetime | Ngày tháng năm |
| Year | int | Năm |
| Quarter | int | Quý trong năm |
| Month | int | Tháng trong năm |
| MonthName | [nvarchar](255) | Tên của tháng |
| Day | int | Ngày trong tháng |

1. **Xác định bảng Fact:**

**Measure:**

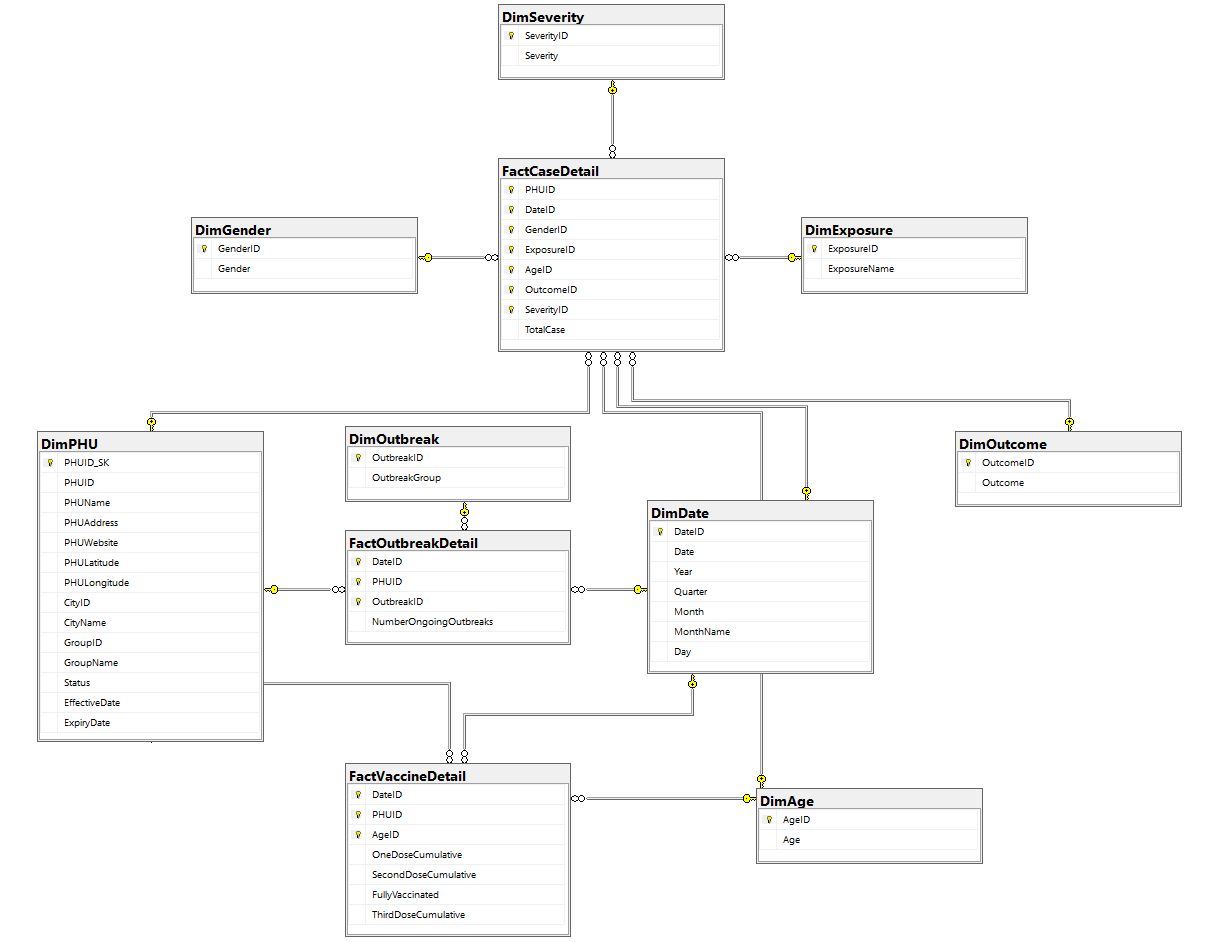
* Tổng số ca nhiễm
* Số người được tiêm vaccine.

**Dữ liệu đổ vào bảng Fact phải thỏa điều kiện:**

* Thông tin của PHU còn hiệu lực (Status = 1).
* Thông tin của dòng dữ liệu là dòng mới nhất (ExpiryDate = NULL)
* Dòng dữ liệu chưa tồn tại trong bảng Fact.

| **FactCaseDetail** | | |
| --- | --- | --- |
| PHUID | int | ID của cơ sở y tế |
| DateID | int | ID ngày |
| GenderID | int | ID của giới tính |
| ExposureID | int | ID của thông tin về ca nhiễm |
| AgeID | int | ID nhóm tuổi |
| OutcomeID | int | ID của trạng thái ca nhiễm |
| SeverityID | int | ID mức độ nghiêm trọng |
| TotalCase | int | Tổng số ca nhiễm |
| **FactVaccineDetail** | | |
| DateID | int | ID ngày |
| PHUID | int | ID của cơ sở y tế |
| AgeID | int | ID nhóm tuổi |
| OneDoseCumulative | int | Số lượng cá nhân đã tiêm được ít nhất một liều. |
| SecondDoseCumulative | int | Số lượng cá nhân đã tiêm được hai liều. |
| FullyVaccinated | int | Tổng số người được tiêm phòng đầy đủ. |
| ThirdDoseCumulative | int | Tổng số người được tiêm phòng ba liều. |
| **FactOutbreakDetail** | | |
| DateID | int | ID ngày |
| PHUID | int | ID của cơ sở y tế |
| OutbreakID | int | ID nhóm bùng phát |
| NumberOngoingOutbreaks | int | Số lượng bùng phát |

* ***Quy trình thực hiện chuyển đổi dữ liệu từ NDS sang DDS:***
* **Thiết kế database DDS:**

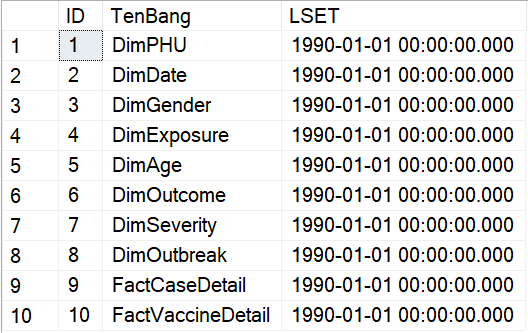


* **Tương tự như stage, dữ liệu đổ vào DDS cũng sẽ thỏa điều kiện:**

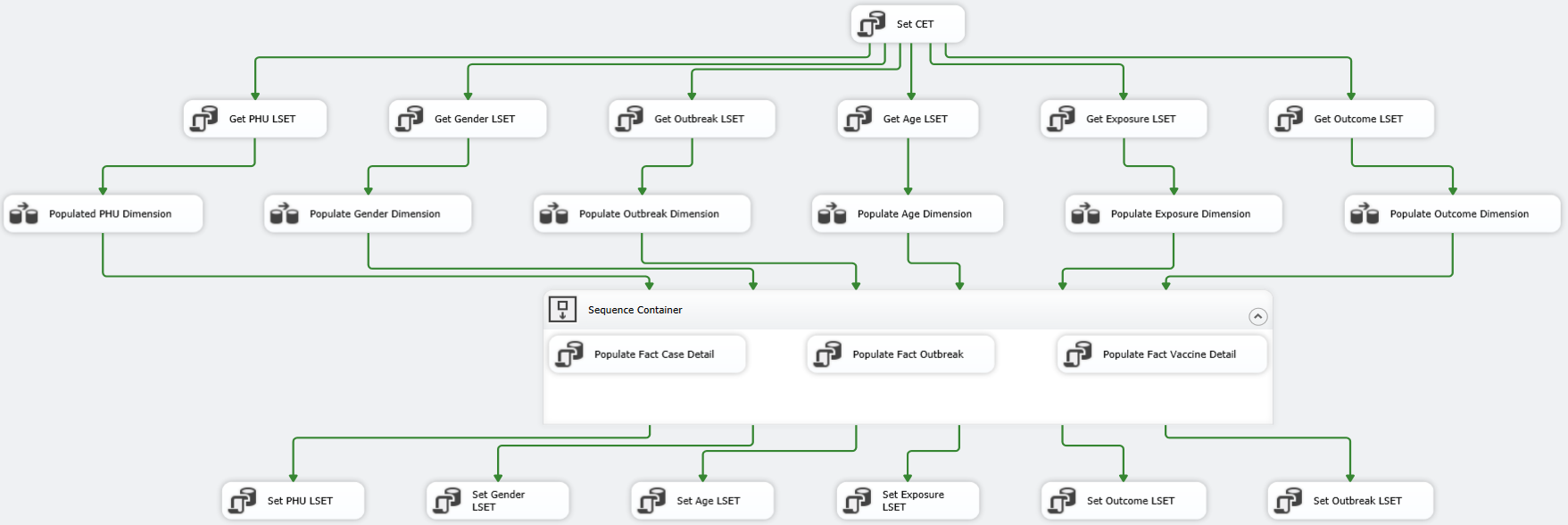
(CreatedDate < CET AND CreatedDate >= LSET)

OR ( UpdatedDate < CET AND UpdatedDate >= LSET)

***LSET*** sẽ được lấy và cập nhật từ bảng ***DDS*** trong ***METADATA***:

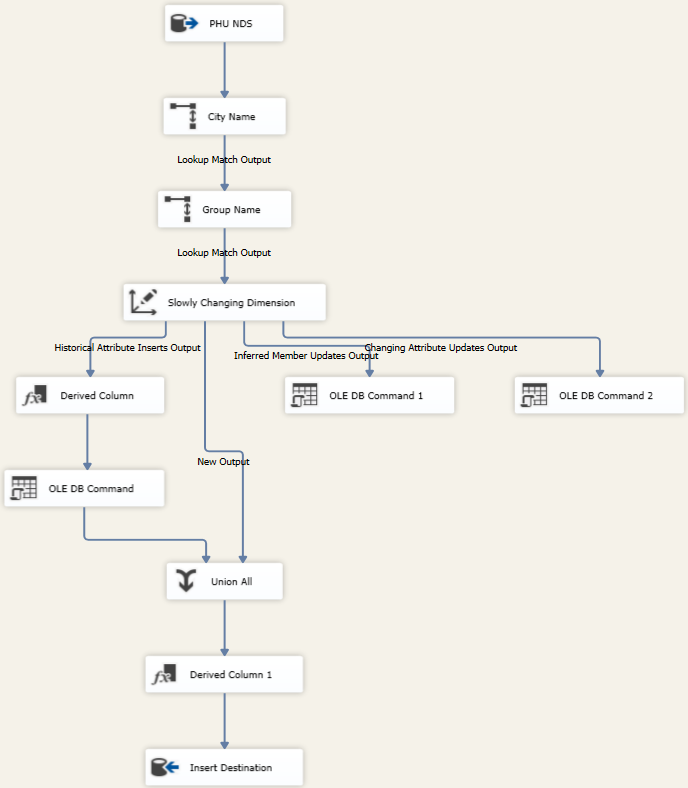


* **Quy trình đổ dữ liệu từ NDS sang DDS:**



**DataFlow ở dimension DimPHU (các dimension khác được cài đặt tương tự):**

1. Load dữ liệu ở NDS.
2. Lấy ra CityName và GroupName để cho vào DimPHU.
3. Sử dụng Slowly Changing để lưu trữ các dữ liệu thay đổi trong lịch sử.
4. Insert dữ liệu vào bảng dimension.



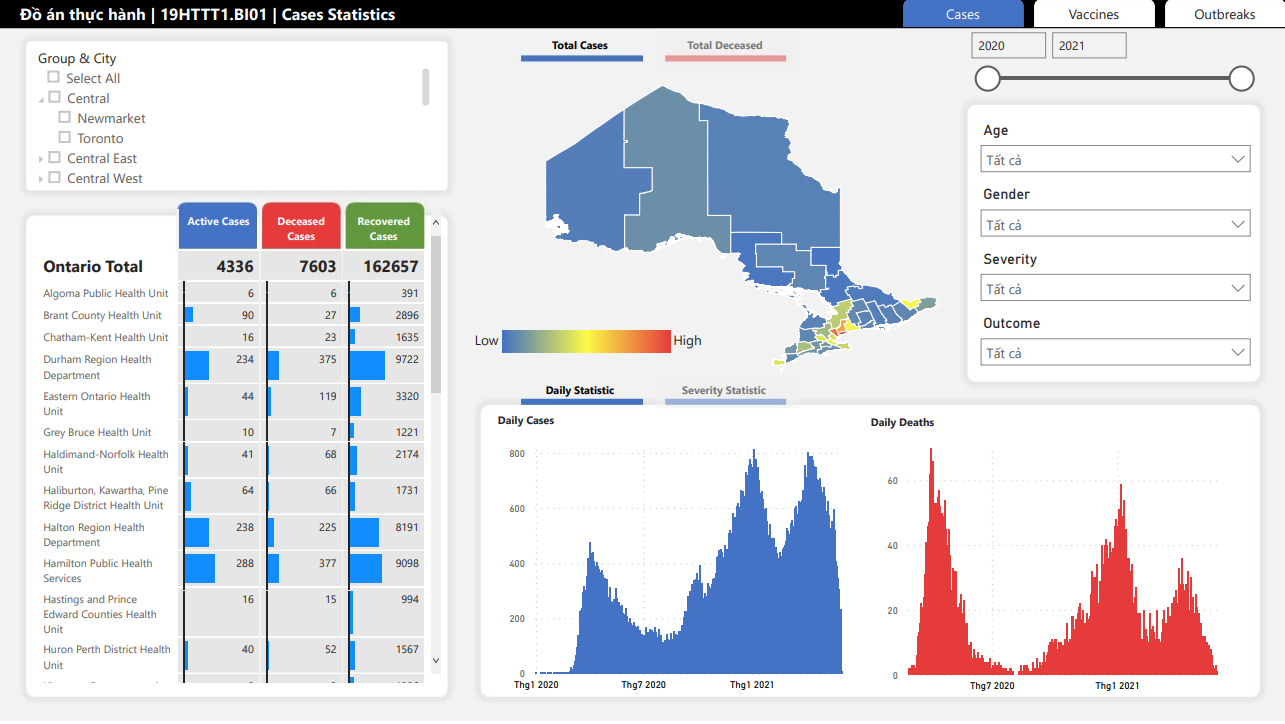
# OLAP và Report:

## OLAP Cube

Nhóm triển khai OLAP Cube và phân lớp các dimension sau:

* **DimDate:** Year > Quarter > Month > Day
* **DimPHU:** PHU Group > PHU City > PHU

## Report



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

* Đối với các PHU ở trung tâm, thì số lượng các ca nhiễm, ca tử vong, ca phục hồi ở mức cao. Các vùng còn lại ví dụ như vùng phía Đông Bắc có lượng ca nhiễm ở mức thấp.
* Cơ sở y tế Toronto ở trung tâm có số lượng ca nhiễm nhiều nhất, trong khi đơn vị y tế Timiskaming ở phía Đông Bắc có lượng ca nhiễm thấp nhất.

**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.**

* Các mức độ nghiêm trọng: Low > Moderate > High > Critical dựa vào nhóm tuổi và loại ca nhiễm.
* Trong quý 1 năm 2020 có lượng ca nhiễm thấp nhất, và quý 1 năm 2021 có lượng ca nhiễm cao nhất. Trong vòng 1 năm từ quý 1 - 2020 đến quý 1 - 2021, tuy có giảm nhẹ ở Q3-2020 nhưng sau đó đã tăng trở lại. Đến ngày mới nhất của Q2 - 2021 thì số lượng ca nhiễm có giảm nhẹ ở tất cả các mức độ.
* Số ca nhiễm ở mức Low là cao nhất, mức Critical là thấp nhất, trong đó tập trung chủ yếu ở vùng Central và xung quanh.
* Các ca nhiễm theo từng mức độ vẫn tập trung chủ yếu ở trung tâm và rải rác ở các vùng còn lại.

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

Vào năm 2020:

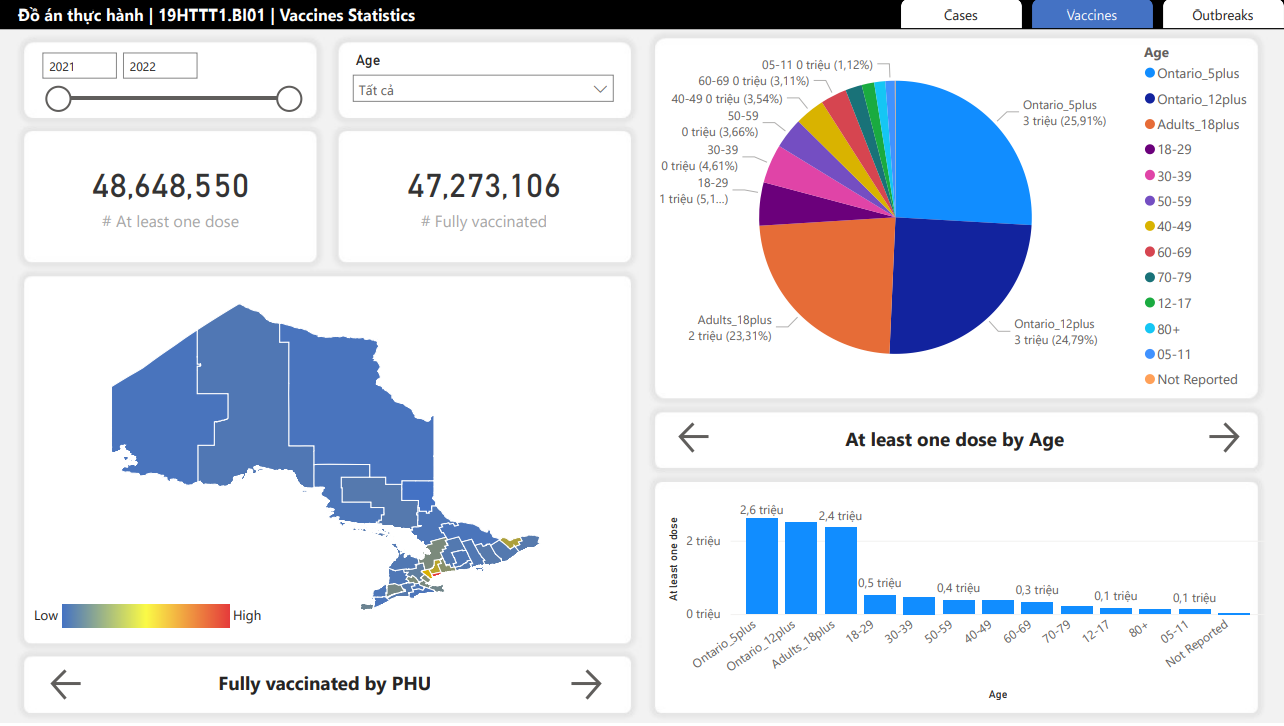
* Số lượng tử vong tỉ lệ thuận với độ tuổi.
* Số lượng tử vong ở các nhóm tuổi như 60-69, 70-79, 80+ và các ca không được khai báo tuổi ở mức cao, đặc biệt là vào đầu năm và cuối năm.
* Các nhóm tuổi chỉ ở mức thấp trong năm.
* Không có sự khác biệt đáng kể ở giới tính.

Vào năm 2021:

* Số lượng tử vong ở giới tính nam vào năm này cao hơn giới tính nữ.
* Số lượng tử vong vẫn tỉ lệ thuận với độ tuổi.
* Đầu năm thì các ca tử vong khá cao, tuy nhiên vào giữa năm cho đến ngày dữ liệu mới nhất thì bắt đầu giảm dần.

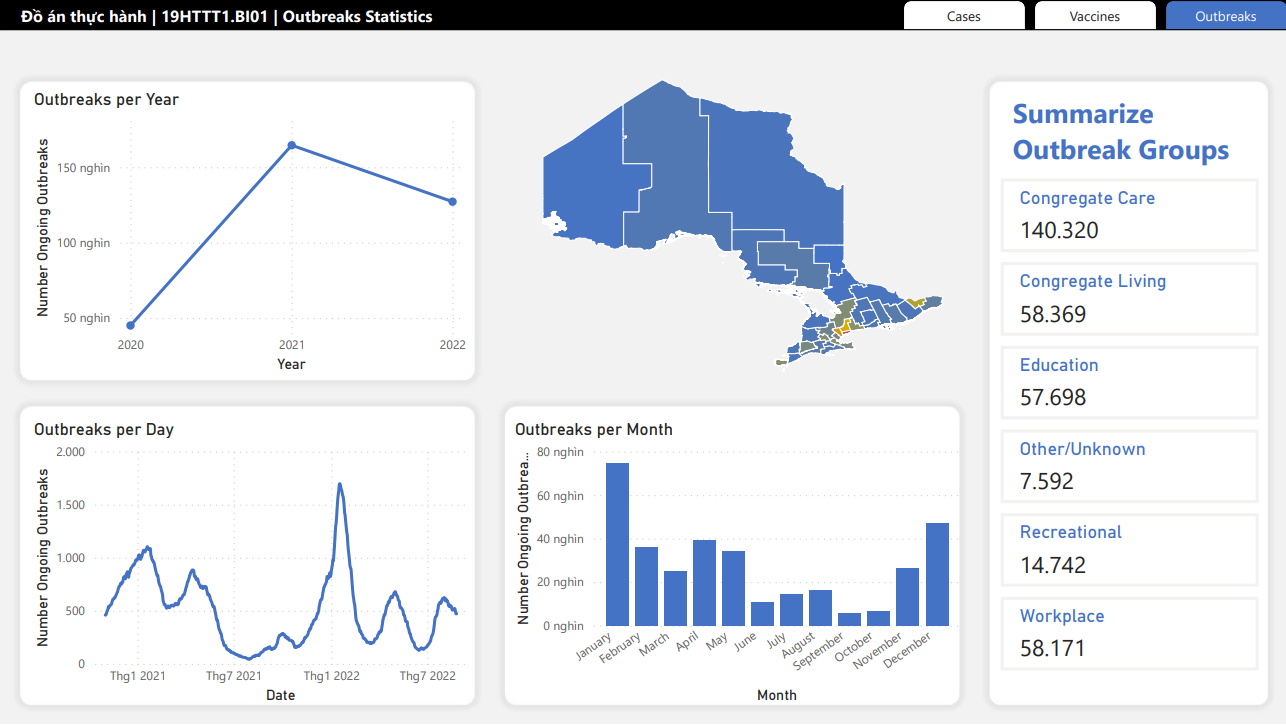
**Thống kê số ca nhiễm, tử vong theo Mức Độ Nghiêm Trọng theo Ngày Trong Tháng của các năm.**

* Vào năm 2020, các ngày của tháng 3 có sự tăng vọt trong số ca nhiễm một cách đáng kể. Sau đó, giảm dần cho đến đầu tháng 9 và có xu hướng tăng đến đầu tháng 4 - 2021.
* Số lượng ca nhiễm theo mức độ nghiêm trọng giảm dần theo thứ tự: Low -> Moderate -> Critical -> High.
* Các ca tử vong được tính là mức độ Critical, có sự tăng vọt vào đầu tháng 4 - 2020 và giảm đến tháng 7 - 2020 thì tăng trở lại. Số ca tử vong cao nhất là vào tháng 4 - 2020 với 66 ca.



**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.**

* Các ca nhiễm và ca tử vong vẫn tập trung vào các thành phố của các vùng Central.
* Số người được tiêm ít nhất một mũi cho tới năm 2022 là khoảng 48 triệu người.
* Số người được tiêm đầy đủ cho tới năm 2022 là khoảng 47 triệu người.



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

* Số lượng các ca bùng phát tập trung chủ yếu ở các vùng trung tâm.
* Nhóm bùng phát Congregate Care có số lượng bùng phát cao nhất, trong khi nhóm Other/Unknown có số lượng thấp nhất.

# Data mining:

Nhóm triển khai thuật toán gom cụm (clustering) giúp phân dữ liệu thành các nhóm dựa trên các đặc tính của dữ liệu.

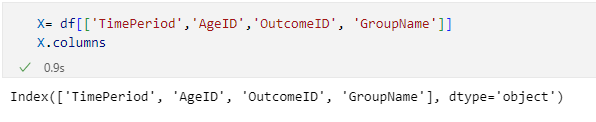
## Chuẩn bị dữ liệu

Tạo bảng dữ liệu miningCaseDetail bao gồm thông tin của các ca nhiễm và cột Label để lưu lại nhãn của cụm sau khi chạy thuật toán.

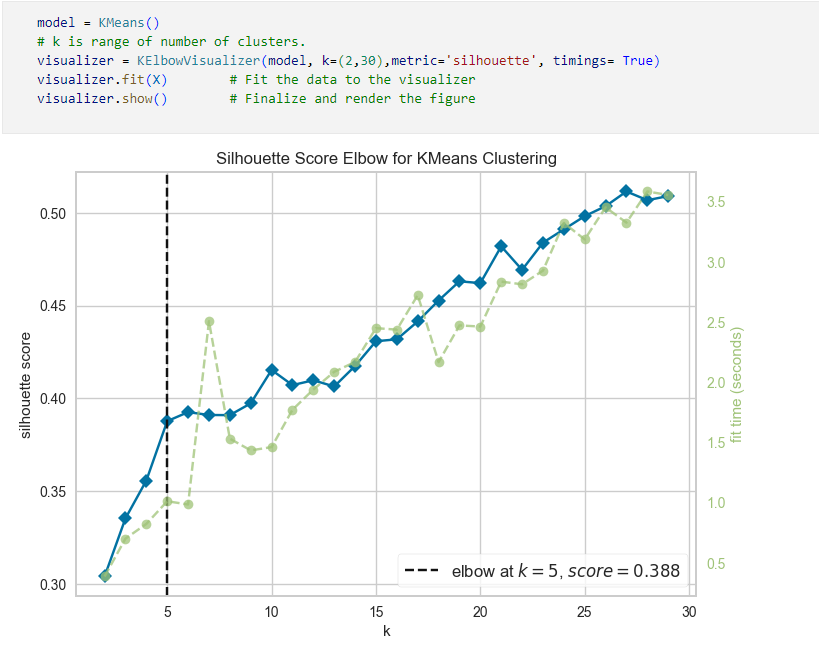
| **miningCaseDetail** | | |
| --- | --- | --- |
| ID | bigint | ID của dòng dữ liệu |
| PHUName | nvarchar(255) | Tên của cơ sở y tế cộng đồng |
| CityName | nvarchar(255) | Tên thành phố mà đơn vị y tế thuộc về |
| GroupName | nvarchar(255) | Tên khu vực mà đơn vị y tế thuộc về |
| Date | nvarchar(255) | Ngày ca nhiễm được ghi nhận |
| TimePeriod | varchar(62) | Quý trong năm |
| Year | int | Năm |
| GenderID | int | ID của giới tính |
| Gender | nvarchar(255) | Giới tính |
| ExposureID | int | Mã định danh thông tin về ca nhiễm |
| ExposureName | nvarchar(255) | Tên thông tin ca nhiễm |
| AgeID | int | Mã định danh của nhóm tuổi |
| Age | nvarchar(255) | Nhóm tuổi |
| OutcomeID | int | ID của trạng thái ca nhiễm |
| Outcome | nvarchar(255) | Trạng thái ca nhiễm |
| SeverityID | int | ID mức độ nghiêm trọng của ca nhiễm |
| Severity | nvarchar(255) | Mức độ nghiêm trọng của ca nhiễm (Low, Moderate, High, Critical) |
| Label | varchar(50) | Nhãn của cụm mà dòng dữ liệu thuộc về |

## Thực hiện phân cụm với mô hình K-means bằng Python

* Các cột dữ liệu được chọn cho việc chạy phân cụm bao gồm: **AgeID, OutcomeID, GroupName, TimePeriod.**



* Sử dụng hệ số silhouette để tìm ra số k tối ưu.

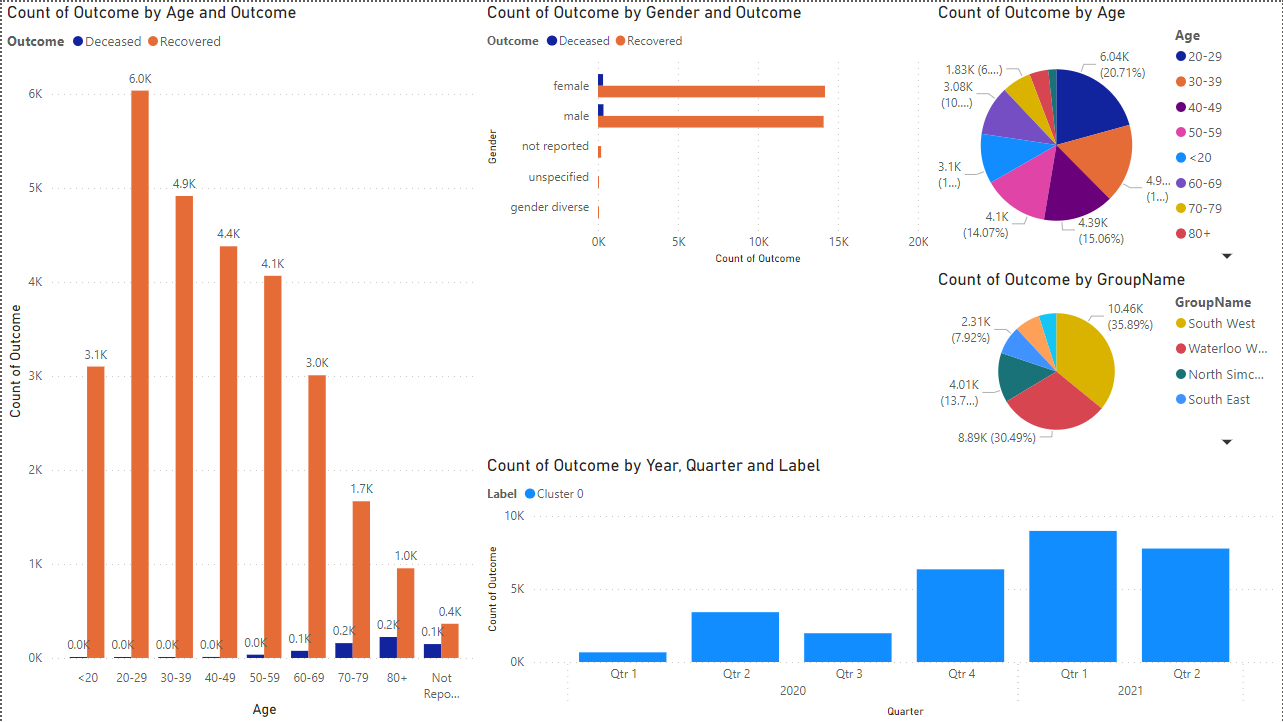


* Thực hiện phân cụm với k = 5. Và gán giá trị nhãn của cụm vào cột Label của bảng miningCaseDetail.



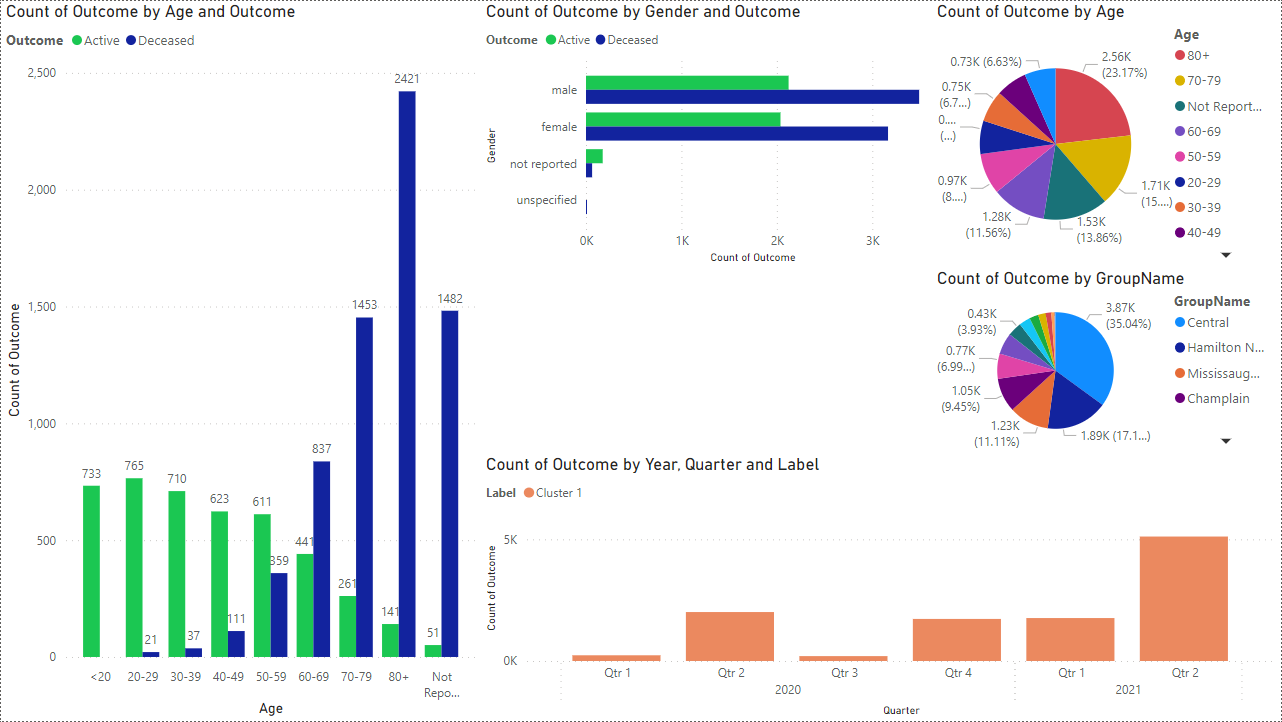
## Nhận xét

**Cluster 0:**



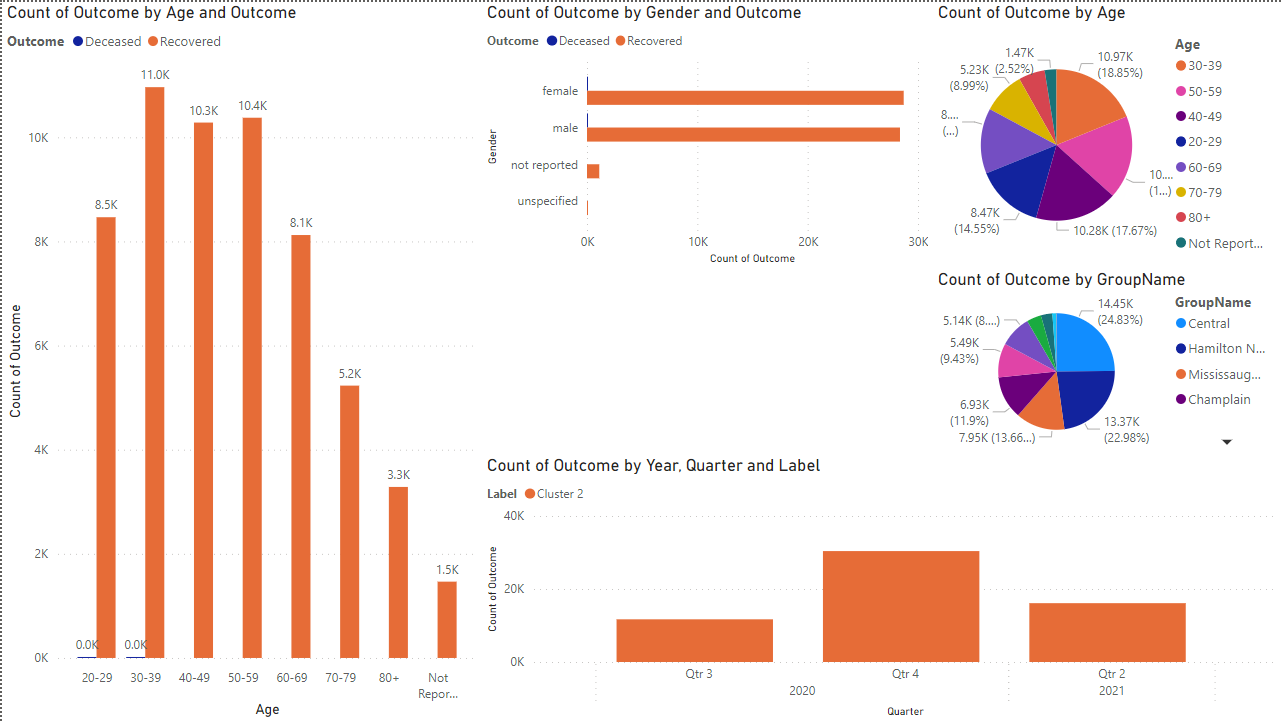
* Xuyên suốt từ đầu năm 2020 đến giữa năm 2021, đỉnh điểm là khoảng thời gian từ quý 4/ 2020 tới quý 2/2021 tại vùng các vùng South West, Waterloo Wellington nhóm tuổi từ 20-29 có số lượng ca nhiễm cao nhất.
* Tuy nhiên số lượng hồi phục của nhóm tuổi này cũng là cao nhất, trong khi nhóm tuổi trên 80 có số lượng ca nhiễm thấp nhất nhưng lại có số lượng ca tử vong lớn nhất.

**Cluster 1:**



* Xuyên suốt từ đầu năm 2020 đến giữa năm 2021, đỉnh điểm là khoảng thời gian từ quý 4/2020 tới quý 2/2021 tại vùng các vùng Central, Hamilton Niagara Haldimand Brant nhóm tuổi trên 80 có số lượng ca nhiễm cao nhất.
* Mặc dù có ca nhiễm ở hầu hết các độ tuổi, số lượng ca tử vong đối với những bệnh nhân cao hơn 60 tuổi cao hơn số lượng các ca nhiễm đang điều trị tích cực. Số lượng ca tử vong lớn nhất và số ca nhiễm điều trị tích cực nhỏ nhất ở nhóm tuổi trên 80.

**Cluster 2:**



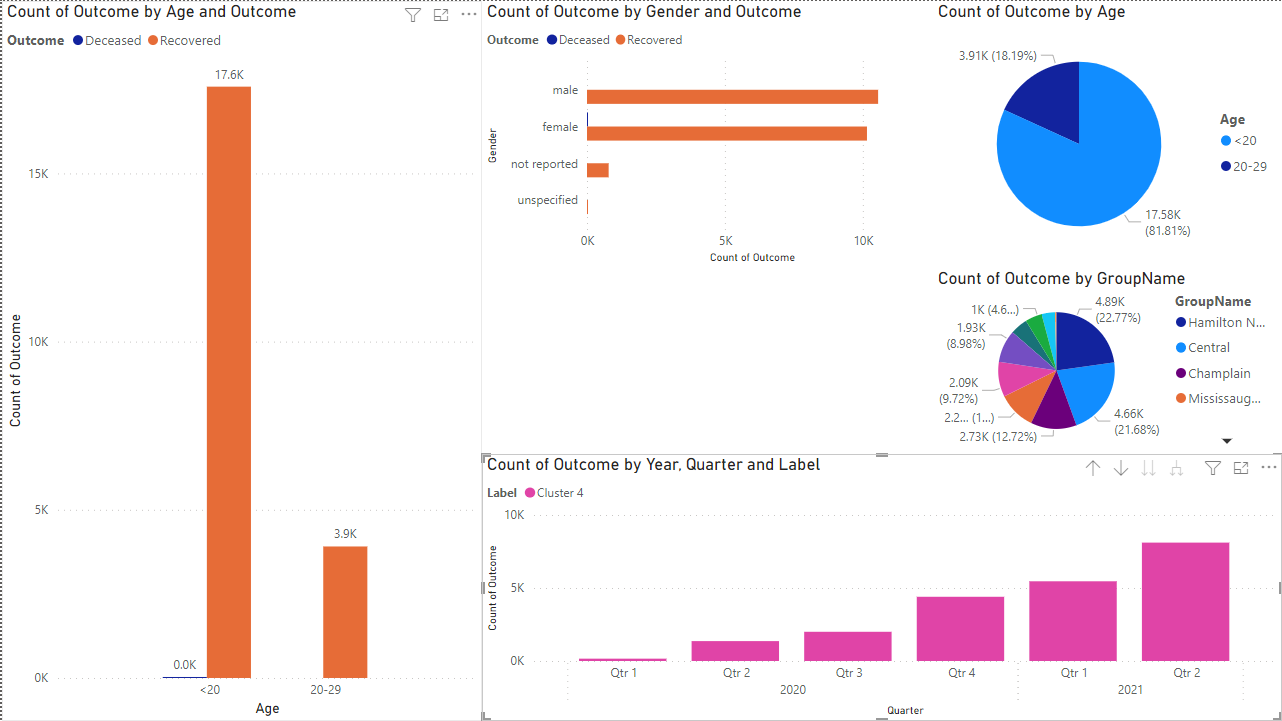
* Trong khoảng thời gian quý 3, 4/2020 và quý 2/2021 tại các vùng Central, Hamilton Niagara Haldimand Brant, Mississauga Halton các ca nhiễm xuất hiện nhiều ở độ tuổi từ 20 đến 69, đỉnh điểm là độ tuổi từ 30 đến 39.
* Hầu hết các ca nhiễm đều ở trạng thái đã hồi phục.

**Cluster 3:**



* Trong khoảng thời gian quý 1, 2/2020 và quý 1/2021 tại các vùng Central, Hamilton Niagara Haldimand Brant, Mississauga Halton các ca nhiễm xuất hiện nhiều ở độ tuổi từ 20 đến 69, đỉnh điểm là độ tuổi từ 20 đến 29 và 50 đến 59.
* Hầu hết các ca nhiễm đều ở trạng thái đã hồi phục.

**Cluster 4:**



* Xuyên suốt từ đầu năm 2020 đến giữa năm 2021, đỉnh điểm là khoảng thời gian từ quý 4/2020 đến quý 2/2021 tại các vùng Hamilton Niagara Haldimand Brant, Central, Champlain các ca nhiễm xuất hiện nhiều ở độ tuổi dưới 29, nhiều nhất là ở độ tuổi dưới 20.
* Hầu hết các ca nhiễm đều ở trạng thái đã hồi phục.

## Kết luận

* Số lượng các ca nhiễm xuất hiện từ đầu năm 2020 với 6 ngàn ca nhiễm ở quý 1 và tăng dần rồi đạt đỉnh điểm vào quý 1 năm 2021 với 43 ngàn ca.
* Các ca nhiễm hầu hết rơi vào độ tuổi từ 20 đến 69. Số lượng ca tử vong nhiều nhất ở các bệnh nhân trên 80 tuổi. Có thể nhận xét các bệnh nhân càng lớn tuổi có tỉ lệ tử vong càng cao.
* Ba vùng có số lượng ca nhiễm cao nhất là Central, Hamilton Niagara Haldimand Brant, Mississauga Halton.

# **Link Youtube:**

<https://youtu.be/_eRn3RbrGWA>

# **Tham khảo:**

<https://www.medintensiva.org/es-covid-19-severity-index-a-predictive-articulo-S021056912030396X>

[Confirmed positive cases of COVID-19 in Ontario - Cases reported in 2020 - Ontario Data Catalogue](https://data.ontario.ca/en/dataset/confirmed-positive-cases-of-covid-19-in-ontario/resource/5cad2b06-e131-44e7-ad46-757ac67e041d)

[COVID-19 Vaccine Data in Ontario - COVID-19 Vaccine data by Public Health Unit (PHU) and by age - Ontario Data Catalogue](https://data.ontario.ca/en/dataset/covid-19-vaccine-data-in-ontario/resource/2a362139-b782-43b1-b3cb-078a2ef19524)

[Compiled COVID-19 Case Details (Canada) | Compiled COVID-19 Case Details (Canada) | COVID-19 Canada (arcgis.com)](https://resources-covid19canada.hub.arcgis.com/datasets/covid19canada::compiled-covid-19-case-details-canada/explore)