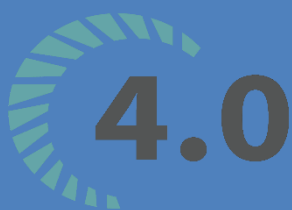


BỘ MÔN HỆ THỐNG THÔNG TIN – KHOA CÔNG NGHỆ THÔNG TIN
ĐẠI HỌC KHOA HỌC TỰ NHIÊN THÀNH PHỐ HỒ CHÍ MINH, ĐẠI HỌC QUỐC GIA TP HCM

MÔN HỆ THỐNG THÔNG TIN PHỤC VỤ TRÍ TUỆ KINH DOANH



Sinh viên thực hiện:
19120680 – Ninh Việt Tiến
19120679 – Nguyễn Văn Tiến
19120693 – Trần Trọng Trí
19120719 – Nguyễn Phước Vinh

GV phụ trách: Hồ Thị Hoàng Vy
ĐỒ ÁN MÔN HỌC - HỆ THỐNG THÔNG TIN PHỤC VỤ TRÍ
TUỆ KINH DOANH

HỌC KỲ I – NĂM HỌC 2022-2023



BẢNG THÔNG TIN CHI TIẾT NHÓM

Mã nhóm:	CQ2019_BI7
Số lượng:	4
MSSV	Họ tên
19120719	Nguyễn Phước Vinh
19120693	Trần Trọng Trí
19120680	Ninh Việt Tiến
19120679	Nguyễn Văn Tiến



Bảng phân công & đánh giá hoàn thành công việc			
Công việc thực hiện	Người thực hiện	Mức độ hoàn thành	Đánh giá của nhóm
Data Mining, Clean Data	Ninh Việt Tiến	90%	9/10
Hỗ trợ ETL, OLAP	Trần Trọng Trí	100%	10/10
Data Visualization, Hỗ trợ ETL	Nguyễn Văn Tiến	90%	9/10
ETL dữ liệu, tổng hợp báo cáo	Nguyễn Phước Vinh	100%	10/10

Phân trăm đóng góp:

STT	Tên thành viên	Phần trăm đóng góp
1	Ninh Việt Tiến	21%
2	Nguyễn Văn Tiến	18%
3	Trần Trọng Trí	29%
4	Nguyễn Phước Vinh	32%



YÊU CẦU ĐỒ ÁN- BÀI TẬP

Loại bài tập	<input type="checkbox"/> Lý thuyết <input checked="" type="checkbox"/> Thực hành <input checked="" type="checkbox"/> Đồ án <input type="checkbox"/> Bài tập
Ngày bắt đầu	05/11/2022
Ngày kết thúc	29/12/2022

MỤC LỤC

A.	Yêu cầu của Đồ án.....	5
B.	Kết quả.....	7
1	Clean Data.....	7
1.1	Lựa chọn dữ liệu.....	7
1.2	Clean data	8
2	Mô tả dữ liệu.....	15
3	Mô tả quá trình ETL	17
3.1	ETL từ Source vào Stage.....	17
3.2	ETL từ Stage vào NDS	26
3.2.1	Phân tích yêu cầu và chuyển đổi dữ liệu.....	26
3.2.2	ETL từ Stage vào NDS	30
3.3	ETL từ NDS vào DDS	41
3.3.1	Phân tích yêu cầu và thiết kế DDS.....	41
3.3.2	ETL từ NDS vào DDS	44
4	OLAP	52
4.1	OLAP Cube:.....	52
5	Report và Visualize:	55
5.1	Thống kê Số ca nhiễm, số ca tử vong, số ca phục hồi của dịch Covid-19 theo từng PHU trong từng năm:	56
5.2	Thống kê số ca tử vong của dịch Covid-19 theo PHU, Mức Độ Nghiêm Trọng và theo các Quý trong từng năm.	57
5.3	Thống kê tổng số người tử vong theo Giới Tính và Nhóm Tuổi theo các năm.	58
5.4	Thống kê số ca nhiễm, tử vong theo Mức Độ Nghiêm Trọng theo Ngày trong Tháng của các Năm.	60



5.5	Thống kê số ca nhiễm, tử vong, số người đã được tiêm vaccin theo Mức Độ Nghiêm Trọng, khu vực (PHU_Group, City), và trong các năm.....	61
5.6	Thống kê số ca nhiễm theo Mức Độ Nghiêm Trọng, loại tiếp xúc của từng khu vực trong các năm.....	62
5.7	Thống kê số ca tử vong, ca nhiễm, số lượng người được chích vắc xin theo nhóm tuổi, City trong các năm.....	63
6	Data mining.....	65
7	Data Visualization	70

A. Yêu cầu của Đồ án

Xây dựng và phân tích dữ liệu về Covid-19 trong các năm 2020 - 2022.

- **Mô tả dữ liệu:** Mô tả ý nghĩa các thuộc tính của các nguồn dữ liệu.
- **Thiết kế kho dữ liệu (KDL), tổng hợp, nạp dữ liệu các nguồn vào KDL và thiết kế, xây dựng Cube:**

Gợi ý:

- Mapping các nguồn dữ liệu trên và đề xuất giải pháp xây dựng Geography dimension với phân cấp: City > PHU_Group > PHU
 - Chuyển đổi dữ liệu ngày tháng sao cho có thể tạo được Date dimension với phân cấp chiều: Year > Quarter > Month > Day
 - Xác định và thiết kế các phân cấp chiều khác để đáp ứng yêu cầu OLAP và report
- **OLAP và Report:**
 1. Thống kê Số ca nhiễm, số ca tử vong, số ca phục hồi của dịch Covid-19 theo từng PHU trong từng năm.
 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 vaccin trong các năm.
 6. Thống kê số ca nhiễm theo Mức Độ Nghiêm Trọng, nhóm bùng phát của từng khu vực trong các năm
 7. Sinh viên tự thiết kế những bảng thống kê khác để có thêm nhiều chiều đánh giá số ca nhiễm và tử vong ở Ontario.
 8. Xây dựng đồ thị/ biểu đồ cho các bảng thống kê ở trên.
 9. [Data Visualization] Dùng regional map để biểu diễn trực quan (bằng màu sắc) số lượng ca nhiễm, số ca tử vong ở các vùng trong năm.

- **Data Mining:**

Gợi ý:



- Sử dụng thuật toán mining để xác định các luật (pattern), ví dụ ở vùng nào, vào thời điểm nào, nhóm tuổi nào, nhóm người nào, ... thường dễ nhiễm, tử vong.
- Sinh viên tự đề xuất các yêu cầu phân tích khác, lựa chọn mô hình phù hợp.



B. Kết quả

1 Clean Data

1.1 Lựa chọn dữ liệu

Dựa trên yêu cầu đề bài cũng như nội dung tập tin Cases Report không cần thiết. Mặc khác data trên bảng Cases Report không thể mapping với bảng Compiled_COVID-19_Case_Deatails (trong một ngày có nhiều đối tượng cùng 1 PHU, 1 Gender, độ tuổi giữa 2 bảng, ví dụ như 2 ảnh bên dưới). Đó đó, ta sẽ bỏ qua dữ liệu của bảng Cases Report.

A	B	C	D	E	F	G	H	I
ObjectId	row_id	date_reported	health_region	age_grc	gender	exposure	case_statu	province
29418	28900	2020/10/09 12:00:00+00	Peel Public Health	20-29	Male	Close Cont	Recovered	Ontario
29505	28987	2020/10/09 12:00:00+00	Peel Public Health	20-29	Male	Close Cont	Recovered	Ontario
30021	31004	2020/10/09 12:00:00+00	Peel Public Health	20-29	Female	Close Cont	Recovered	Ontario
30590	31540	2020/10/09 12:00:00+00	Peel Public Health	20-29	Male	Close Cont	Recovered	Ontario
30910	31860	2020/10/09 12:00:00+00	Peel Public Health	20-29	Female	Close Cont	Recovered	Ontario
31026	31976	2020/10/09 12:00:00+00	Peel Public Health	20-29	Male	Close Cont	Recovered	Ontario
31115	32069	2020/10/09 12:00:00+00	Peel Public Health	20-29	Male	Close Cont	Recovered	Ontario
31385	32339	2020/10/09 12:00:00+00	Peel Public Health	20-29	Male	Close Cont	Recovered	Ontario
31450	32404	2020/10/09 12:00:00+00	Peel Public Health	20-29	Female	Close Cont	Recovered	Ontario
31451	32405	2020/10/09 12:00:00+00	Peel Public Health	20-29	Male	Not Repor	Recovered	Ontario
31515	32469	2020/10/09 12:00:00+00	Peel Public Health	20-29	Male	Close Cont	Recovered	Ontario
31525	32479	2020/10/09 12:00:00+00	Peel Public Health	20-29	Male	Not Repor	Recovered	Ontario
31554	30033	2020/10/09 12:00:00+00	Peel Public Health	20-29	Male	Not Repor	Recovered	Ontario
31595	30074	2020/10/09 12:00:00+00	Peel Public Health	20-29	Male	Close Cont	Recovered	Ontario
31664	30143	2020/10/09 12:00:00+00	Peel Public Health	20-29	Male	Close Cont	Recovered	Ontario
32064	30563	2020/10/09 12:00:00+00	Peel Public Health	20-29	Male	Not Repor	Recovered	Ontario
32205	30704	2020/10/09 12:00:00+00	Peel Public Health	20-29	Male	Not Repor	Recovered	Ontario
32323	30822	2020/10/09 12:00:00+00	Peel Public Health	20-29	Male	Not Repor	Recovered	Ontario
33405	33384	2020/10/09 12:00:00+00	Peel Public Health	20-29	Female	Close Cont	Recovered	Ontario
34625	34624	2020/10/09 12:00:00+00	Peel Public Health	20-29	Male	Close Cont	Recovered	Ontario
499009	499452	2020/10/09 12:00:00+00	Peel Public Health	20-29	Female	Not Repor	Recovered	Ontario
527288	526287	2020/10/09 12:00:00+00	Peel Public Health	20-29	Male	Close Cont	Recovered	Ontario
528545	529520	2020/10/09 12:00:00+00	Peel Public Health	20-29	Male	Travel-Rel.	Recovered	Ontario

Compiled_COVID-19_Case_Details_ (+)

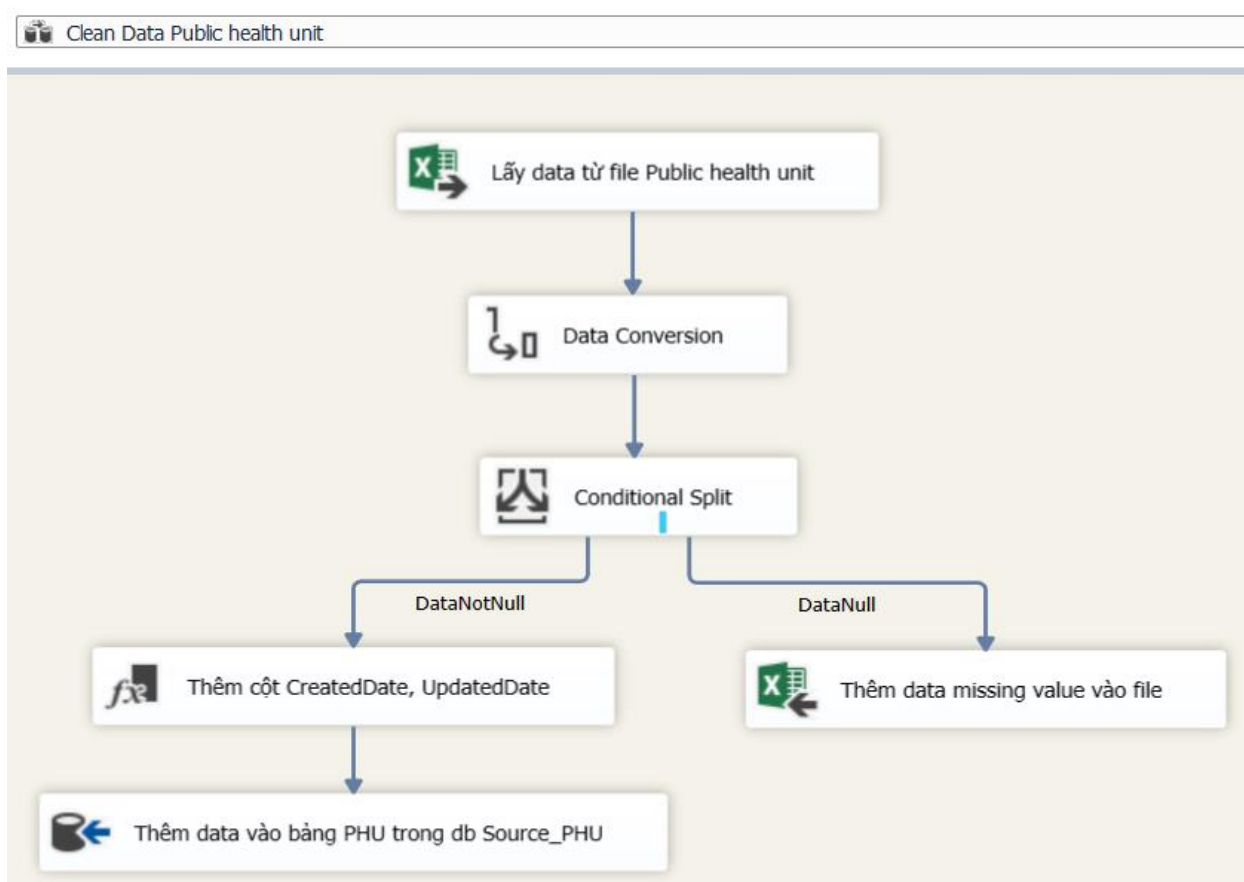
A	B	C	D	E	F	G	H
Outcome	Age	Gender	Reporting PHU	Sex	CaseReported Date	PHUCity	T
Resolved	20s	MALE	Peel Public Health	###	10/9/2020	Mississauga	
Resolved	20s	MALE	Peel Public Health	###	10/9/2020	Mississauga	
Resolved	20s	MALE	Peel Public Health	###	10/9/2020	Mississauga	
Resolved	20s	MALE	Peel Public Health	###	10/9/2020	Mississauga	
Resolved	20s	MALE	Peel Public Health	###	10/9/2020	Mississauga	
Resolved	20s	FEMALE	Peel Public Health	###	10/9/2020	Mississauga	
Resolved	20s	MALE	Peel Public Health	###	10/9/2020	Mississauga	
Resolved	20s	FEMALE	Peel Public Health	###	10/9/2020	Mississauga	
Resolved	20s	FEMALE	Peel Public Health	###	10/9/2020	Mississauga	
Resolved	20s	FEMALE	Peel Public Health	###	10/9/2020	Mississauga	
Resolved	20s	FEMALE	Peel Public Health	###	10/9/2020	Mississauga	
Resolved	20s	FEMALE	Peel Public Health	###	10/9/2020	Mississauga	
Resolved	20s	MALE	Peel Public Health	###	10/9/2020	Mississauga	
Resolved	20s	FEMALE	Peel Public Health	###	10/9/2020	Mississauga	
Resolved	20s	MALE	Peel Public Health	###	10/9/2020	Mississauga	
Resolved	20s	MALE	Peel Public Health	###	10/9/2020	Mississauga	
Resolved	20s	FEMALE	Peel Public Health	###	10/8/2020	Mississauga	
Resolved	20s	MALE	Peel Public Health	###	10/8/2020	Mississauga	
Resolved	20s	MALE	Peel Public Health	###	10/8/2020	Mississauga	
Resolved	20s	MALE	Peel Public Health	###	10/8/2020	Mississauga	
Resolved	20s	MALE	Peel Public Health	###	10/8/2020	Mississauga	

1.2 Clean data

- Làm sạch dữ liệu trước khi tiến hành ETL. (Xử lý các giá trị NULL và dữ liệu không phù hợp)
- Tiến hành thực hiện clean data trên các tập tin Public health unit, Public Health Units GROUP, Vaccines_by_age_phu, Compiled_COVID-19_Case_Deatails, ongoing_outbreaks_phu và insert data các bảng vào source trong SQL.
- Dữ liệu về trường Date_reported trong bảng Compiled_COVID-19_Case_Deatails không thể convert về dạng Datetime trong SQL (lỗi định dạng). Do vậy ta sẽ cắt bỏ chuỗi thời gian của trường Date_reported trước khi clean data. (Sử dụng Python)
- Các bảng còn lại ta tiến hành clean data bằng cách loại bỏ các record trong các bảng nếu chúng là data cần thiết cho phần ETL, OLAP và giá trị của chúng là NULL.
- Các bước tiến hành:
 - Tạo các Control Flow cho 5 nguồn dữ liệu:



- Data Flow (nguồn dữ liệu Public health unit):



- Đọc dữ liệu từ file:



Excel Source Editor

Configure the properties that enable the Data Flow task to obtain data from Excel provider.

Connection Manager
Columns
Error Output

Specify a connection manager, data source, or data source view for the Excel source. Then, select the mode used to access data within the source. After selecting the data access mode, select from among the additional data access options that appear.

Excel connection manager:
Excel Connection Manager

New...

Data access mode:
Table or view

Name of the Excel sheet:
'Public health unit\$'

Preview...

OK Cancel Help

Excel Source Editor

Configure the properties that enable the Data Flow task to obtain data from Excel provider.

Connection Manager
Columns
Error Output

Available External Columns

- ☒ Name
- ☒ PHU_ID
- ☒ Reporting_PHU
- ☒ Reporting_PHU_Address
- ☒ Reporting_PHU_City
- ☒ Reporting_PHU_Postal_Code
- ☒ Reporting_PHU_Website
- ☒ Reporting_PHU_Latitude

External Column	Output Column
PHU_ID	PHU_ID
Reporting_PHU	Reporting_PHU
Reporting_PHU_Address	Reporting_PHU_Address
Reporting_PHU_City	Reporting_PHU_City
Reporting_PHU_Postal_Code	Reporting_PHU_Postal_Code
Reporting_PHU_Website	Reporting_PHU_Website
Reporting_PHU_Latitude	Reporting_PHU_Latitude
Reporting_PHU_Longitude	Reporting_PHU_Longitude

OK Cancel Help

- Chuyển hóa dữ liệu cho phù hợp (Data conversion):

Data Conversion Transformation Editor

Configure the properties used to convert the data type of an input column to a different data type. Depending on the data type to which the column is converted, set the length, precision, scale, and code page of the column.

Available Input Columns

- ☒ Name
- ☐ PHU_ID
- ☒ Reporting_PHU
- ☒ Reporting_PHU_Address
- ☒ Reporting_PHU_City
- ☒ Reporting_PHU_Postal_Code
- ☒ Reporting_PHU_Website

Input Column	Output Alias	Data Type	Length	Precision	Scale	Code Page
Reporting_PHU	Copy of Reporting_P...	string [DT_STR]	255			1252 (ANS
Reporting_PHU_Add...	Copy of Reporting_P...	string [DT_STR]	255			1252 (ANS
Reporting_PHU_City	Copy of Reporting_P...	string [DT_STR]	255			1252 (ANS
Reporting_PHU_Post...	Copy of Reporting_P...	string [DT_STR]	255			1252 (ANS
Reporting_PHU_We...	Copy of Reporting_P...	string [DT_STR]	255			1252 (ANS
Reporting_PHU_Latit...	Copy of Reporting_P...	float [DT_R4]				
Reporting_PHU_Lon...	Copy of Reporting_P...	float [DT_R4]				

Configure Error Output... OK Cancel Help

- Phân luồng dữ liệu (Data Null và Data Not Null):



Conditional Split Transformation Editor

Specify the conditions used to direct input rows to specific outputs. If an input row matches no condition, the row is directed to a default output.

+ Variables and Parameters

+ Columns

+ Mathematical Functions

+ String Functions

+ Date/Time Functions

+ NULL Functions

+ Type Casts

+ Operators

Description:

Order	Output Name	Condition
1	DataNull	ISNULL(PHU_ID) ISNULL(Reporting_PHU) ISNULL(...

Default output name: DataNotNull

Configure Error Output... OK Cancel Help

- Data Not Null (thêm cột CreatedDate, UpdatedDate và insert dữ liệu vào database đã tạo sẵn):



Derived Column Transformation Editor

Specify the expressions used to create new column values, and indicate whether the values update existing columns or populate new columns.

Variables and Parameters

Columns

Mathematical Functions

String Functions

Date/Time Functions

NULL Functions

Type Casts

Operators

Description:

Derived Column Name	Derived Column	Expression	Data Type	Le
CreatedDate	<add as new column>	GETDATE()	database timestamp ...	
UpdatedDate	<add as new column>	GETDATE()	database timestamp ...	

OLE DB Destination Editor

Configure the properties used to insert data into a relational database using an OLE DB provider.

Connection Manager

Mappings

Error Output

Specify an OLE DB connection manager, a data source, or a data source view, and select the data access mode. If using the SQL command access mode, specify the SQL command either by typing the query or by using Query Builder. For fast-load data access, set the table update options.

OLE DB connection manager:
MSI.Source_PHU New...

Data access mode:
Table or view

Name of the table or the view:
[dbo].[PHU] New...

- Data Null (Xuất data ra file excel):

Excel Destination Editor

Configure the properties that enable the insertion of data via an Excel provider.

Connection Manager
Mappings
Error Output

Specify a connection manager, data source, or data source view for the Excel destination. Then, select the mode used to access data within the destination. After selecting the data access mode, select from among the additional data access options that appear.

Excel connection manager:
Excel Connection Manager New...

Data access mode:
Table or view

Name of the Excel sheet:
Public_health_unit_error New...

View Existing

OK Cancel Help

* Đối với các nguồn dữ liệu còn lại làm tương tự.

2 Mô tả dữ liệu

Tên thuộc tính	Mô tả
Date	Ngày báo cáo
PHU ID	Định danh của đơn vị chăm sóc y tế cộng đồng
At least one dose_cumulative	Số người tiêm được ít nhất 1 mũi
Second_dose_cumulative	Số người tiêm được 1 mũi

fully_vaccinated_cumulative	Số người tiêm đủ vaccin. Tiêm đầy đủ nghĩa là: - 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 y 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 (tập con của số người tiêm đủ)
Reporting_PHU	Các PHU được báo cáo
Reporting_PHU_Address	Địa chỉ PHU được báo cáo
Reporting_PHU_City	Thành phố của các PHU được báo cáo.
Reporting_PHU_Postal_Code	Mã bưu điện của PHU được báo cáo
Reporting_PHU_Latitude	Vĩ tuyến PHU
Reporting_PHU_Longitude	Kinh tuyến PHU
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
number_ongoing_outbreaks	Số đợt bùng phát đang diễn ra
row_id	Mã dòng
age_group	Nhóm tuổi, được phân loại gồm: - 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

	<ul style="list-style-type: none"> - 50 to 59 years old - 60 to 69 years old - 70 to 79 years old - 80 years and older - Adults_18plus - Ontario_12plus - Ontario_5plus - Undisclosed_or_missing
gender	Giới tính bệnh nhân
exposure	Phơi nhiễm <ul style="list-style-type: none"> - Outbreak - Bùng phát - Close Contact - Liên hệ chặt chẽ - Not Reported - Không được báo cáo - Travel-Related - Du lịch
case_status	Trạng thái ca nhiễm <ul style="list-style-type: none"> - Recovered - Phục hồi - Deceased - Tử vong - Active - Điều trị tích cực
outcome	Kết quả: <ul style="list-style-type: none"> - Resolved - Điều trị - Fatal - Tử vong
specimenDate	Ngày lấy mẫu
TestReported Date	Ngày trả kết quả
CaseAcquisition info	Thông tin ca nhiễm: <ul style="list-style-type: none"> - CC: dương tính xác định được nguồn lây (closed contact) - No known Epi-link: dương tính không rõ nguồn lây - OB: bùng phát (Outbreak) - Travel
AccurateEpisode Dt	Ngày khởi phát
OutbreakRelated	Có liên quan đến đợt bùng phát

3 Mô tả quá trình ETL

3.1 ETL từ Source vào Stage

Ta thực hiện đồ dữ liệu từ nguồn (data đã được clean):



Các bước thực hiện:

- Tạo control flow cho từng table:



- Set CET bằng ngày giờ ETL trong bảng Data_Flow trong database METADATA_PHU tại record có giá trị trường TenBang là tên các bảng trong nguồn dữ liệu.

Configure the properties required to run SQL statements and stored procedures using the selected connection.

Property	Value
Name	Set CET PHU to Metadata
Description	Execute SQL Task
Timeout	0
CodePage	1252
TypeConversionMode	Allowed
ResultSet	None
ConnectionType	OLE DB
Connection	MSI.METADATA_PHU
SQLSourceType	Direct input
SQLStatement	UPDATE Data_FlowSET CET = GETDATE() WHERE (TenBang = 'PHU')
IsQueryStoredProcedure	False
BypassPrepare	True

SQLStatement
Specifies the query to be run by the task.

Buttons: Browse..., Build Query..., Parse Query, OK, Cancel, Help

- Truncate dữ liệu cũ trong Stage

Execute SQL Task Editor

Configure the properties required to run SQL statements and stored procedures using the selected connection.

General
Parameter Mapping
Result Set
Expressions

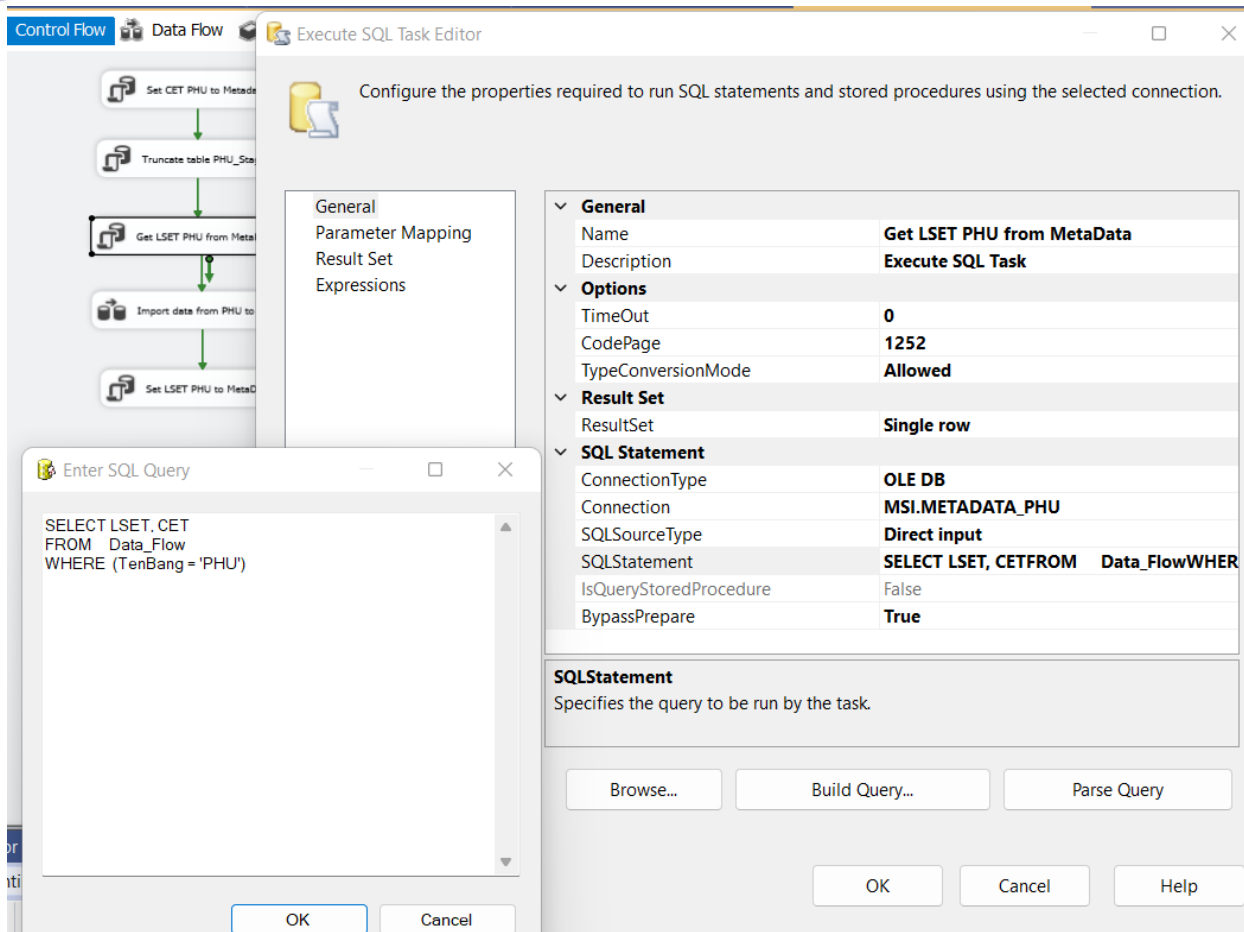
General	
Name	Truncate table PHU_Stage
Description	Execute SQL Task
Options	
TimeOut	0
CodePage	1252
TypeConversionMode	Allowed
Result Set	
ResultSet	None
SQL Statement	
ConnectionType	OLE DB
Connection	MSI.Stage_PHU
SQLSourceType	Direct input
SQLStatement	TRUNCATE TABLE PHU_Stage
IsQueryStoredProcedure	False
BypassPrepare	True

Name
Specifies the name of the task.

Browse... Build Query... Parse Query

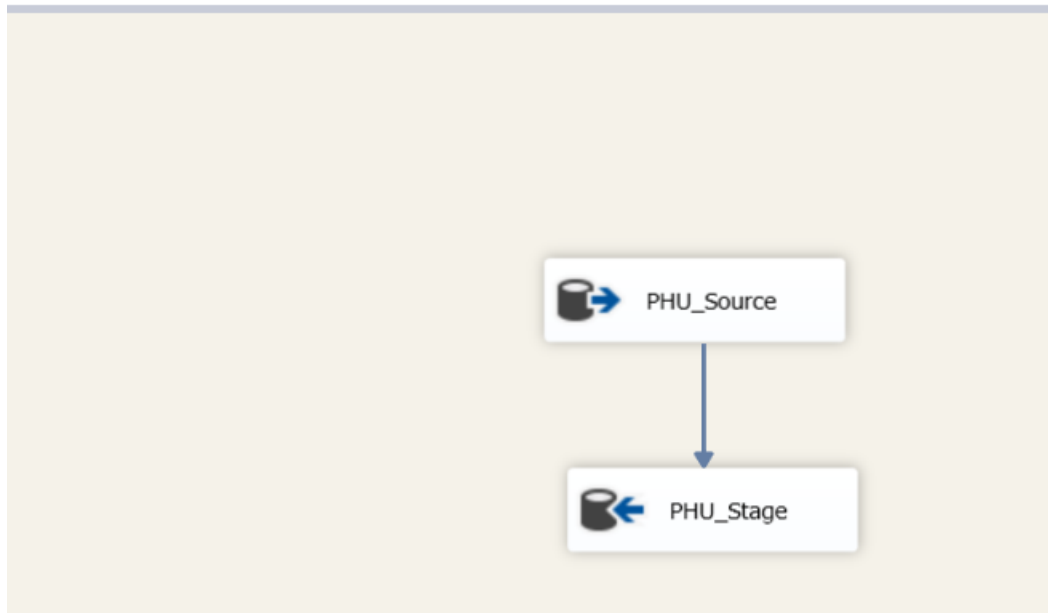
OK Cancel Help

- Lấy giá trị LSET (ngày cập nhật cuối cùng) trong bảng Data_Flow trong database METADATA_PHU tại record có giá trị trường TenBang là tên các bảng trong nguồn dữ liệu.



– Đồ dữ liệu vào Stage

Import data from PHU to Stage



- Load dữ liệu từ nguồn:



OLE DB Source Editor

Configure the properties used by a data flow to obtain data from any OLE DB provider.

Connection Manager
Columns
Error Output

Specify an OLE DB connection manager, a data source, or a data source view, and select the data access mode. If using the SQL command access mode, specify the SQL command either by typing the query or by using Query Builder.

OLE DB connection manager:

MSI.Source_PHU

New...

Data access mode:

SQL command

SQL command text:

```
SELECT PHU_ID, Reporting_PHU, Reporting_PHU_Address,  
Reporting_PHU_City, Reporting_PHU_Postal_Code,  
Reporting_PHU_Website, Reporting_PHU_Latitude,  
Reporting_PHU_Longitude, CreatedDate, UpdatedDate  
FROM PHU  
WHERE (CreatedDate < ?) AND (CreatedDate >= ?)
```

Parameters...

Build Query...

Browse...

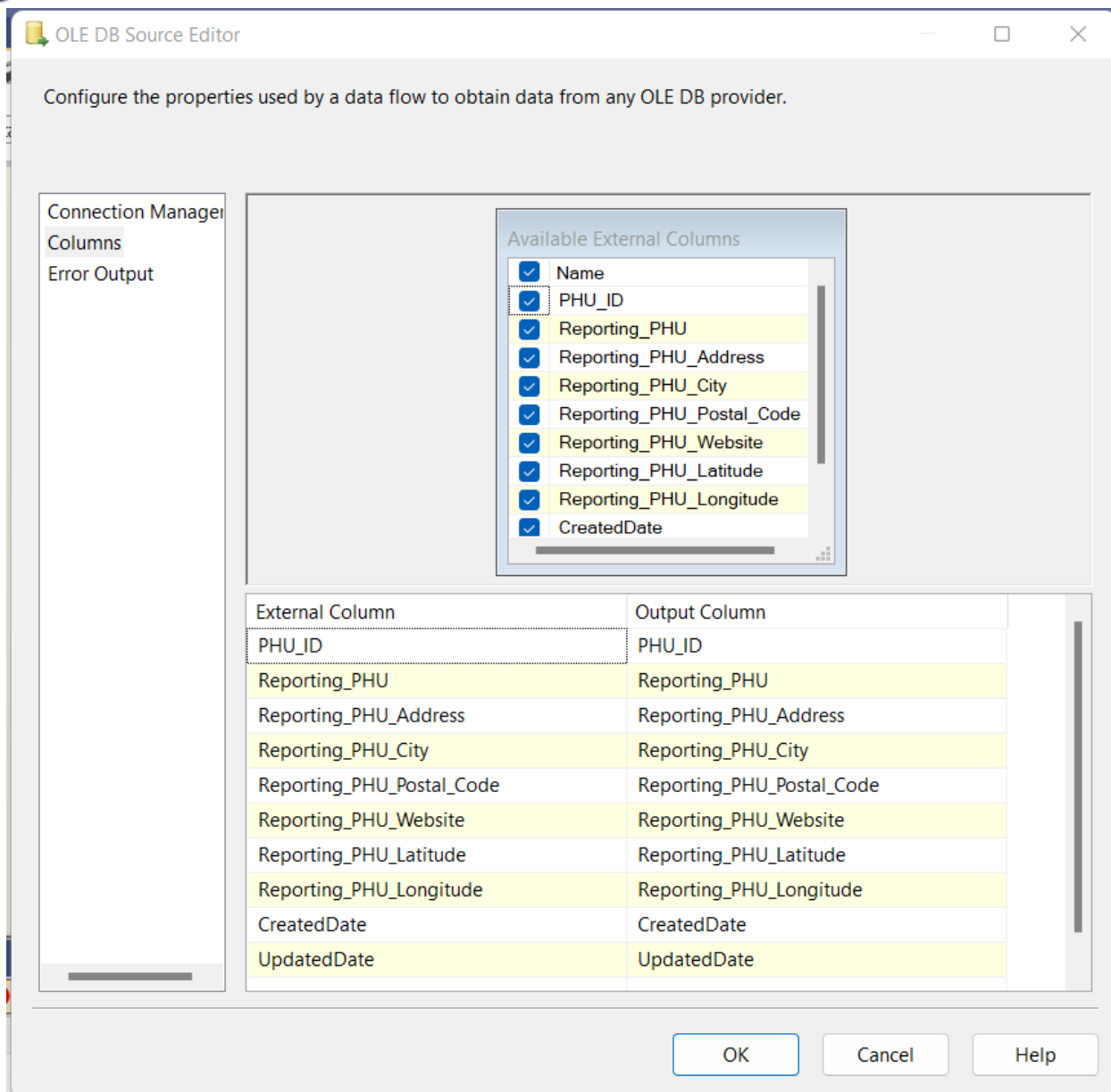
Parse Query

Preview...

OK

Cancel

Help



- Đồ dữ liệu vào stage:



OLE DB Destination Editor

Configure the properties used to insert data into a relational database using an OLE DB provider.

Connection Manager
Mappings
Error Output

Specify an OLE DB connection manager, a data source, or a data source view, and select the data access mode. If using the SQL command access mode, specify the SQL command either by typing the query or by using Query Builder. For fast-load data access, set the table update options.

OLE DB connection manager:
MSI.Stage_PHU New...

Data access mode:
Table or view - fast load

Name of the table or the view:
[dbo].[PHU_Stage] New...

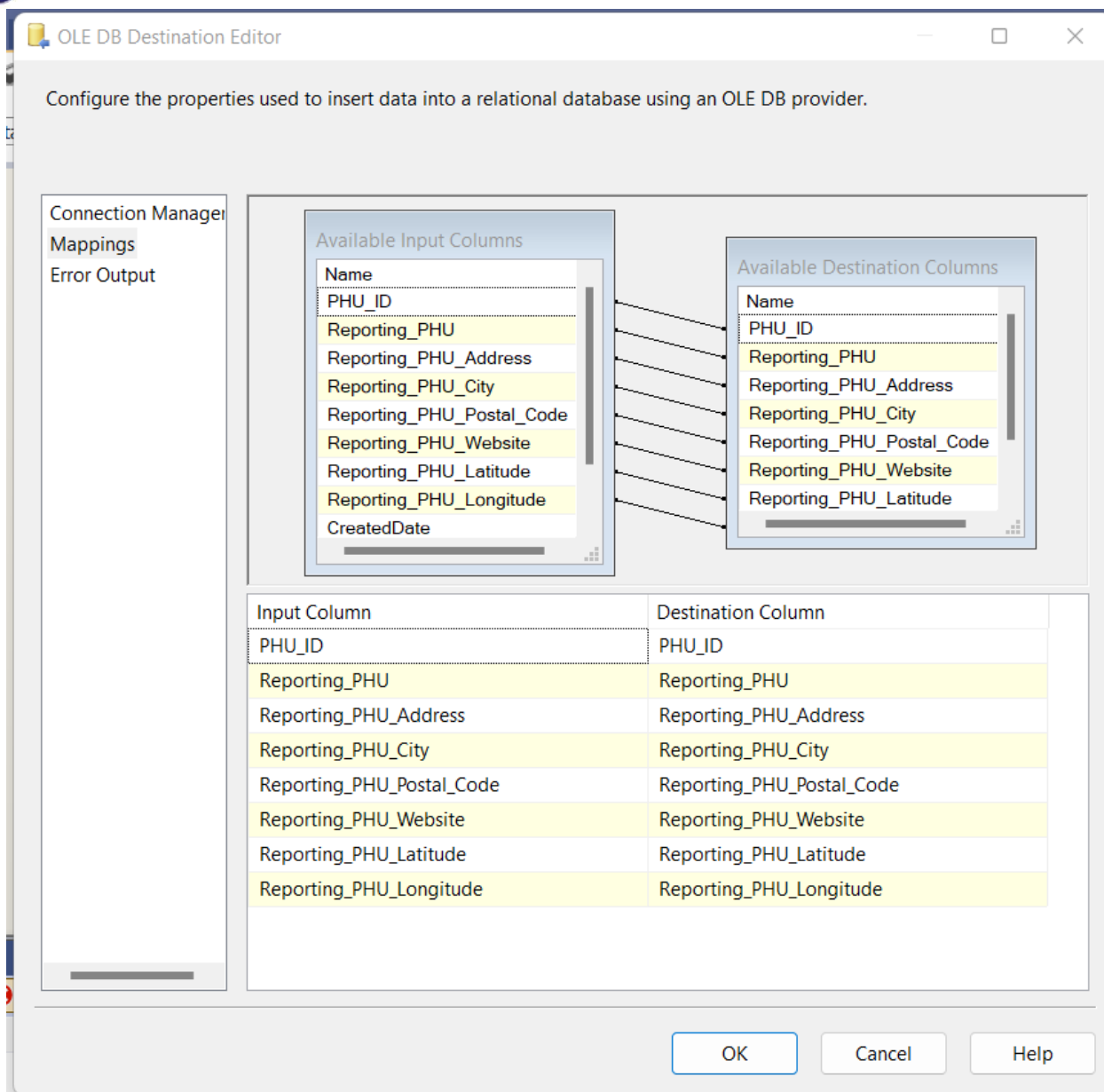
☐ Keep identity ☒ Table lock
☐ Keep nulls ☒ Check constraints

Rows per batch:

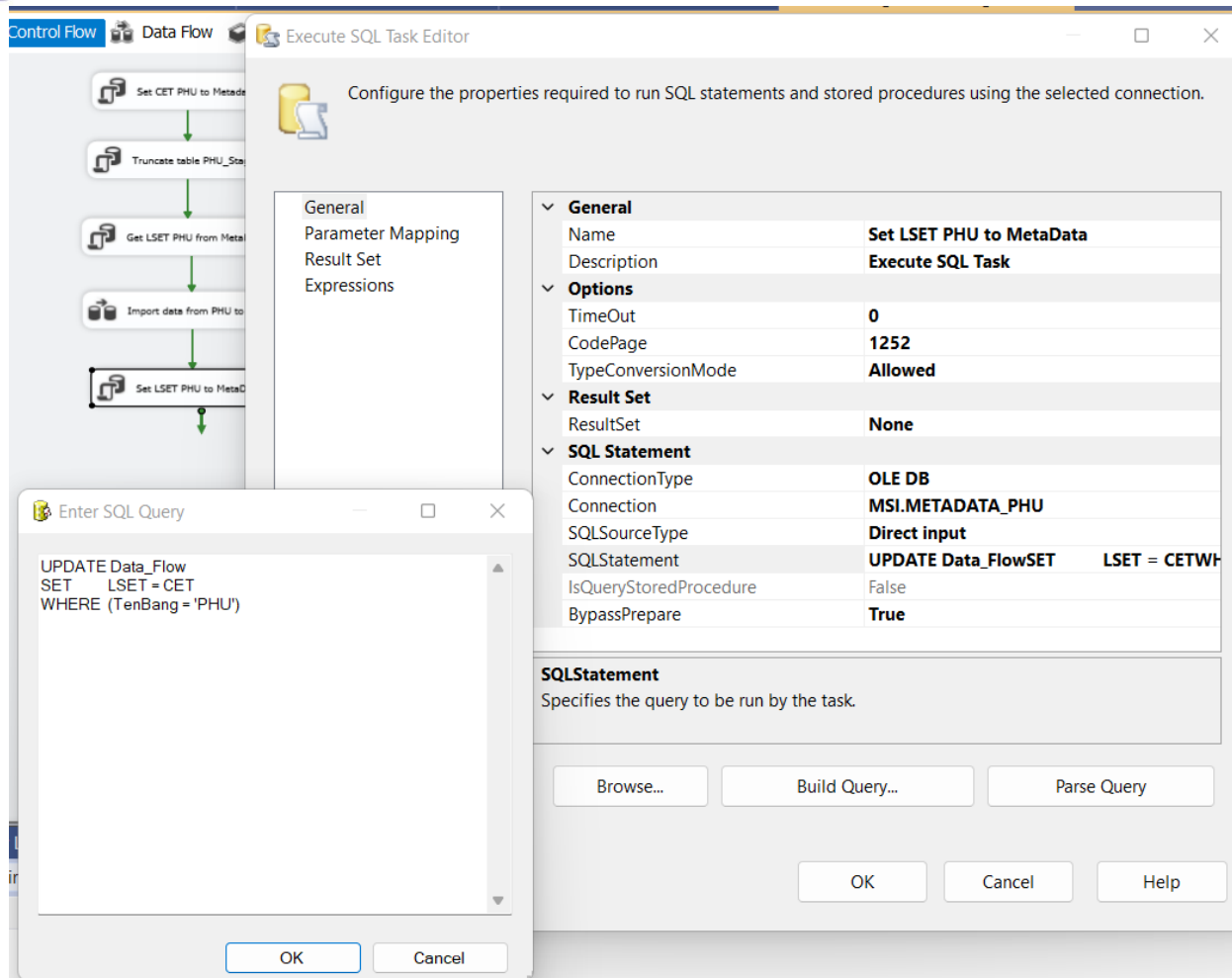
Maximum insert commit size: 2147483647

View Existing

OK Cancel Help



- Set LSET bằng ngày giờ ETL trong bảng Data_Flow trong database METADATA_PHU tại record có giá trị trường TenBang là tên các bảng trong nguồn dữ liệu.

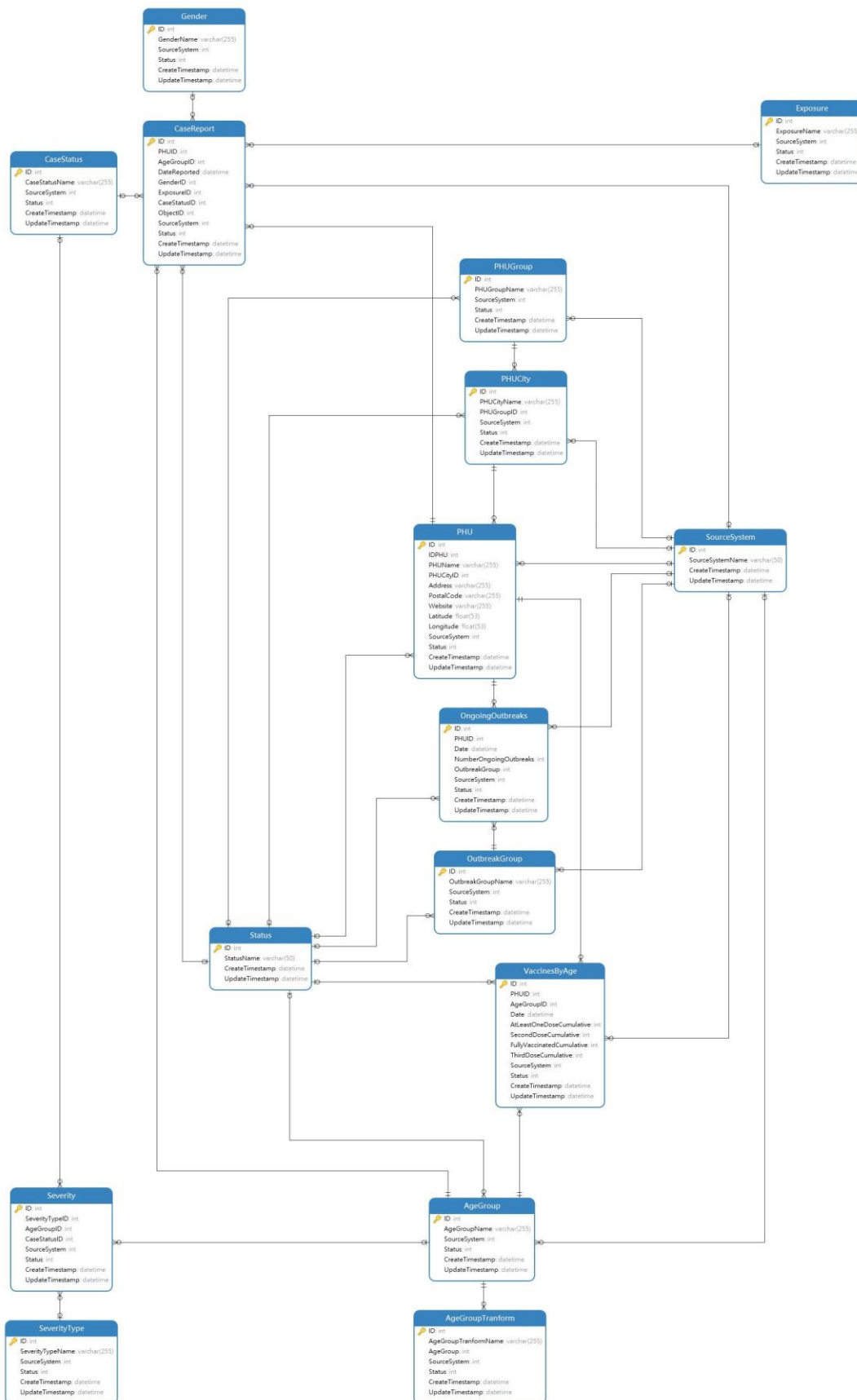


* Đối với các nguồn dữ liệu còn lại làm tương tự.

3.2 ETL từ Stage vào NDS

3.2.1 Phân tích yêu cầu và chuyển đổi dữ liệu

- Yêu cầu report liên quan đến việc thống kê số ca nhiễm, ca tử vong, ca phục hồi theo phân cấp khu vực PHU, PHUCity, PHUGroup, nhóm tuổi, giới tính, mức độ nghiêm trọng, sự phơi nhiễm. Do vậy, ta sẽ tách dữ liệu các tiêu chí trên thành các bảng.



– Giải thích các component:

- Bảng CaseReport: lấy data từ bảng Compiled_COVID_19_Case_Details trong Source
- Bảng CaseStatus: lấy data từ bảng Compiled_COVID_19_Case_Details (trường Case_Status)
- Bảng Gender: lấy data từ bảng Compiled_COVID_19_Case_Details (trường Gender) trong Source
- Bảng PHU: lấy data từ bảng PHU trong Source
- Bảng PHUCity: lấy data từ bảng PHU_Group (trường PHU_City) trong Source
- Bảng PHUGroup: lấy data từ bảng PHU_Group (trường PHU_Group) trong Source
- Bảng Exposure: lấy data từ bảng Compiled_COVID_19_Case_Details (trường Exposure) trong Source
- Bảng VaccinesByAge: lấy data từ bảng Vaccine_By_Age trong Source
- Bảng AgeGroup: lấy data từ bảng Compiled_COVID_19_Case_Details (trường Age_Group) trong Source. Bảng này được lấy làm chuẩn để chuyển đổi các trường tương tự (chi tiết ở mục bên dưới)
- Bảng AgeGroupTranform: lấy data từ bảng Vaccine_By_Age (trường AgeGroup) trong Source để chuyển đổi (nhóm tự định nghĩa)
- Bảng Severity: chi tiết các mức độ nghiêm trọng được nhóm định nghĩa dựa trong dữ liệu trường AgeGroup và CaseStatus trong bảng Compiled_COVID_19_Case_Details trong Source
- Bảng SeverityType: các loại mức độ nghiêm trọng
- Bảng Status: tình trạng của các bảng
- Bảng SourceSystem: bảng nguồn dữ liệu
- Bảng OutbreakGroup: lấy data từ bảng Ongoing_Outbreaks_PHU (trường Outbreak_Group) trong Source
- Bảng OngoingOutBreaks: lấy data từ bảng Ongoing_Outbreaks_PHU trong Source

(Dữ liệu bảng OutbreakGroup, OngoingOutBreaks không được sử dụng do không thể mapping với dữ liệu về ca nhiễm)

– Chuyển đổi dữ liệu:

- Dữ liệu về nhóm tuổi ở các bảng VaccineByAge và Compiled_COVID-19_Case_Deatails khác nhau. Do vậy, ta sử dụng giá trị bảng Compiled_COVID-19_Case_Deatails làm chuẩn và chuyển đổi giá trị AgeGroup về dạng tương ứng.

Dữ liệu chuẩn:

ID	AgeGroupName	SourceSystem	Status	CreateTimestamp	UpdateTimestamp
1	<20	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
2	20-29	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
3	30-39	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
4	40-49	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
5	50-59	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
6	60-69	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
7	70-79	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
8	80+	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
9	Not Reported	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50

Dữ liệu chuyển đổi:

ID	AgeGroupTranformName	AgeGroup	SourceSystem	Status
2	5-11yrs	1	1	(Null)
3	11-17yrs	1	1	(Null)
4	18-29yrs	2	1	(Null)
5	30-39yrs	3	1	(Null)
6	40-49yrs	4	1	(Null)
7	50-59yrs	5	1	(Null)
8	60-69yrs	6	1	(Null)
9	70-79yrs	7	1	(Null)
10	80+	8	1	(Null)
11	Undisclosed_or_missing	9	1	(Null)

- Dữ liệu về mức độ nghiêm trọng: nhóm chọn các tiêu chí nhóm tuổi (AgeGroup), tình trạng (Case_Status) để đánh giá về mức độ nghiêm trọng của các ca nhiễm theo các mức độ sau:

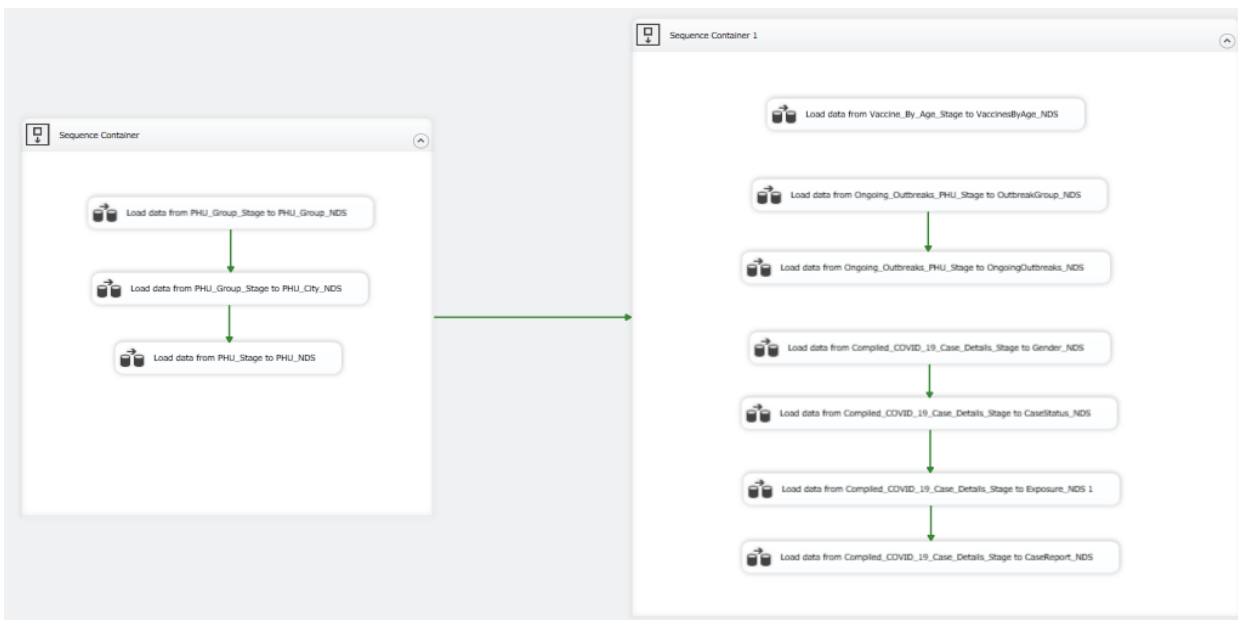


ID	SeverityTypeName	SourceSystem	Status	CreateTimestamp	UpdateTimestamp
1	Low	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
2	Medium	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
3	High	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
4	Very High	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50

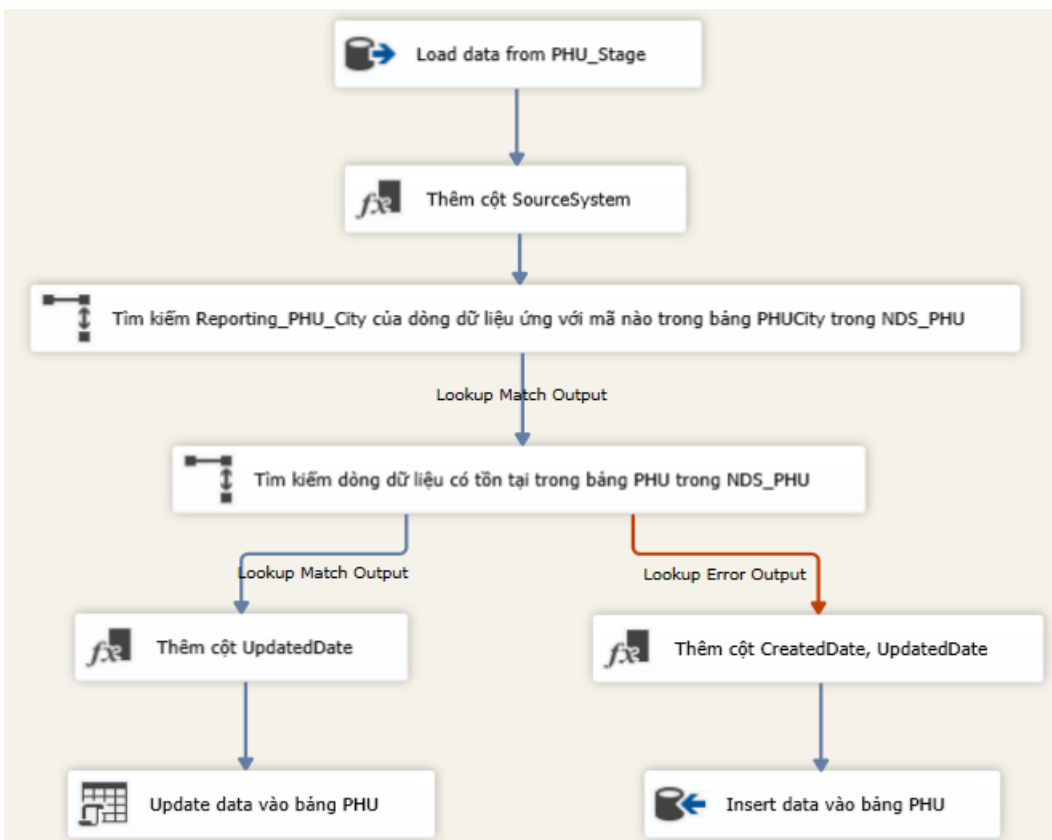
ID	SeverityTypeID	AgeGroupID	CaseStatusID	SourceSystem	Status	CreateTimestamp	UpdateTimestamp
2	2	1	3	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
3	1	2	3	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
4	1	3	3	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
5	1	4	3	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
6	1	5	3	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
7	1	6	3	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
8	2	7	3	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
9	2	8	3	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
10	4	1	1	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
11	3	2	1	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
12	3	3	1	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
13	3	4	1	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
14	3	5	1	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
15	3	6	1	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
16	4	7	1	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
17	4	8	1	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
18	3	1	2	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
19	2	2	2	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
20	2	3	2	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
21	2	4	2	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
22	2	5	2	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
23	2	6	2	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
24	3	7	2	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
25	3	8	2	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
26	1	9	3	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
27	3	9	1	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50
28	2	9	2	1	(Null)	2022-12-24 23:55:47.50	2022-12-24 23:55:47.50

3.2.2 ETL từ Stage vào NDS

- Tạo các Data flow tương ứng theo thứ tự ràng buộc dữ liệu



- Đổ data vào NDS cho từng bảng trong Stage (Minh họa đổ dữ liệu vào bảng PHU):



- Load data từ Stage:



OLE DB Source Editor

Configure the properties used by a data flow to obtain data from any OLE DB provider.

Connection Manager
Columns
Error Output

Specify an OLE DB connection manager, a data source, or a data source view, and select the data access mode. If using the SQL command access mode, specify the SQL command either by typing the query or by using Query Builder.

OLE DB connection manager:
MSI.Stage_PHU

New...

Data access mode:
SQL command

SQL command text:
SELECT DISTINCT PHU_ID, Reporting_PHU,
Reporting_PHU_Address, Reporting_PHU_City,
Reporting_PHU_Postal_Code, Reporting_PHU_Website,
Reporting_PHU_Latitude, Reporting_PHU_Longitude
FROM PHU_Stage

Parameters...

Build Query...

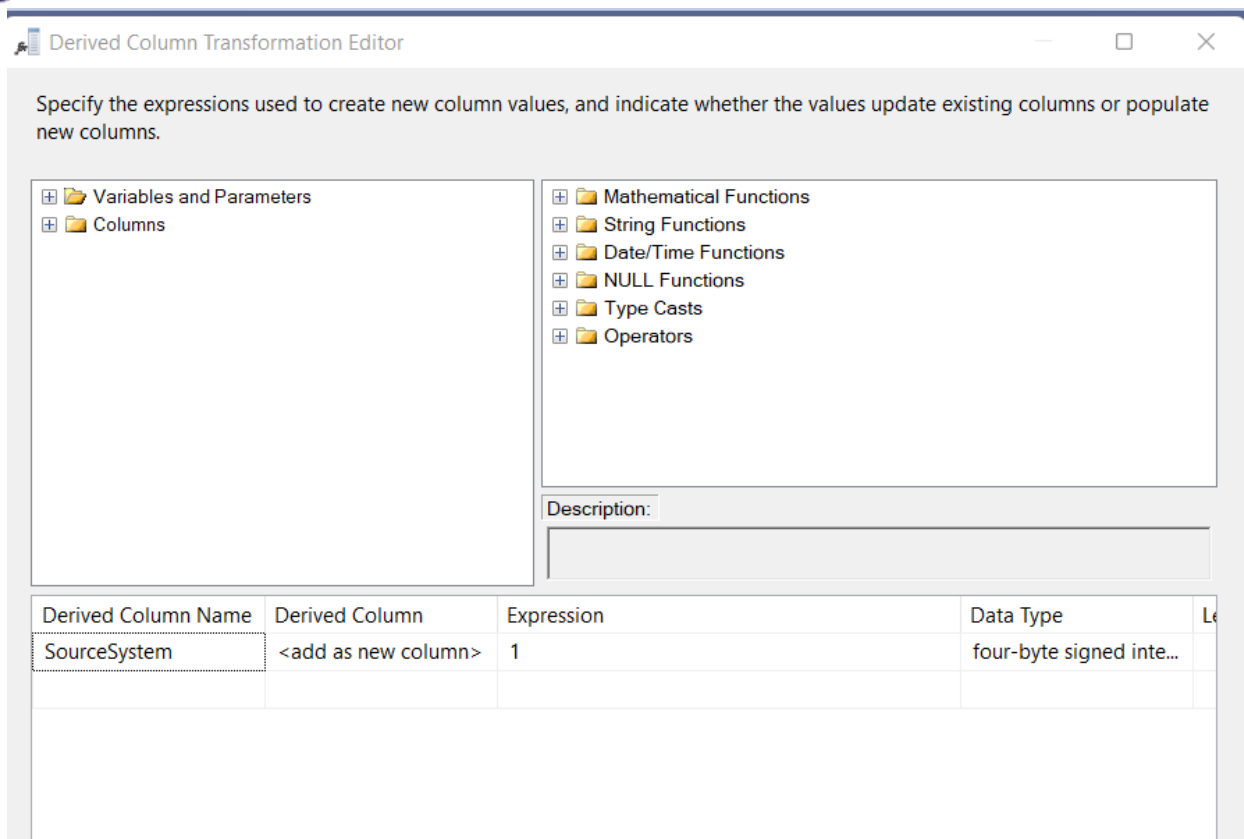
Browse...

Parse Query

Preview...

OK Cancel Help

- Thêm cột SourceSystem:



- Look up giá trị tại trường Reporting_PHU_City của từng dòng dữ liệu có tồn tại trong bảng PHUCity trong NDS hay không.



Lookup Transformation Editor

This transform enables the performance of simple equi-joins between the input and a reference data set.

General
Connection
Columns
Advanced
Error Output

Specify a data source to use. You can select a table in a data source view, a table in a database connection, or the results of an SQL query.

OLE DB connection manager:
MSI.NDS_PHU New...

☐ Use a table or a view:
New...

☒ Use results of an SQL query:
SELECT ID, PHUCityNameFROM PHUCityWHERE
(SourceSystem = 1)
Build Query...
Browse...
Parse Query

Preview...

OK Cancel Help

Lookup Transformation Editor

This transform enables the performance of simple equi-joins between the input and a reference data set.

General
Connection
Columns
Advanced
Error Output

Available Input Columns

Name
PHU_ID
Reporting_PHU
Reporting_PHU_Address
Reporting_PHU_City
Reporting_PHU_Postal_Code
Reporting_PHU_Website
Reporting_PHU_Latitude
Reporting_PHU_Longitude

Available Lookup Columns

Name	I...
<input checked="" type="checkbox"/> ID	
<input type="checkbox"/> PHUCityName	

Lookup Column	Lookup Operation	Output Alias
ID	<add as new column>	PHUCityIDNDS

OK Cancel Help

- Look up dòng dữ liệu có tồn tại trong bảng PHU trong NDS không:
 - Nếu có (Thêm cột UpdatedDate, Update data cũ):



Derived Column Transformation Editor

Specify the expressions used to create new column values, and indicate whether the values update existing columns or populate new columns.

+

 Variables and Parameters

+

 Columns

+

 Mathematical Functions

+

 String Functions

+

 Date/Time Functions

+

 NULL Functions

+

 Type Casts

+

 Operators

Description:

Configure Error Output...

OK

Cancel

Help



Default Start

[Design] CleanData.dtsx [Design] StageToNDS.dtsx

Data Flow Parameters Event Handlers Package Explorer

Load data from PHU_Stage to PHU_NDS

String Value Editor

String value:

```
update phu set phuname=?, phucityid=?,address=?,  
postalcode=?,website=?,  
latitude=?,longitude=?,UpdateTimestamp=? where idphu=? and  
sourcesystem=?
```

OK Cancel

Thêm cột UpdatedDate

Update data vào bảng PHU

0 Errors 0 Warnings 0 Messages

ption

Advanced Editor for Update data vào bảng PHU

The advanced editor provides access to the low-level properties of data flow components. Additionally, the advanced editor can be used to configure components that do not have a custom user interface.

Connection Managers Component Properties Column Mappings Input and Output Properties

Specify advanced properties for the data flow component.

Properties:

Common Properties

ComponentClassID	{4AA3F603-2558-44FD-9CAD-F8005FB6A880}
ContactInfo	OLE DB Command;Microsoft Corporation; Microsoft SQL S
Description	Runs an SQL statement for each row in a data flow. For exa
ID	162
IdentificationString	Update data vào bảng PHU
IsDefaultLocale	True
LocaleID	English (United States)
Name	Update data vào bảng PHU
PipelineVersion	0
UsesDispositions	True
ValidateExternalMetadata	True
Version	2

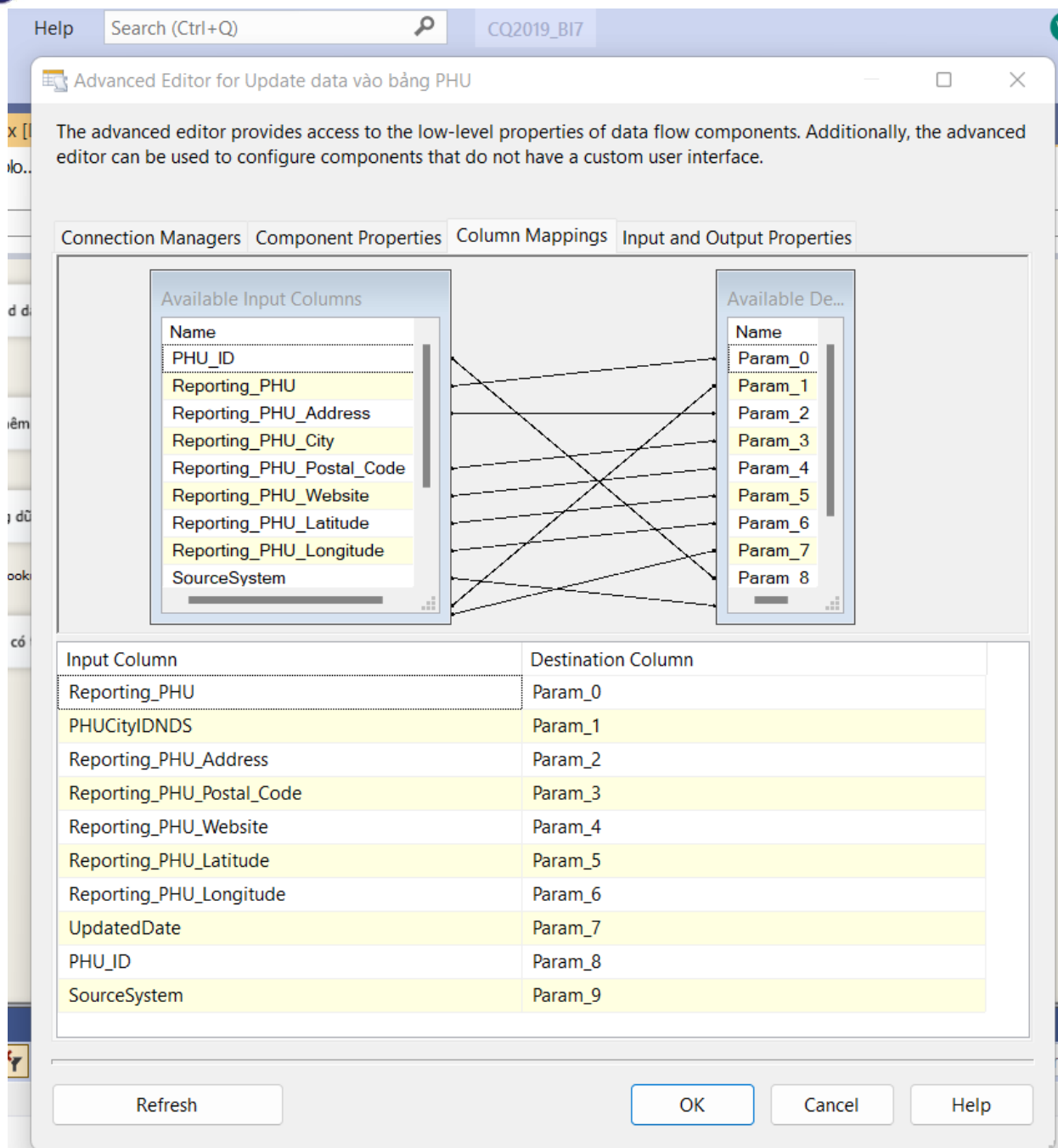
Custom Properties

CommandTimeout	0
DefaultCodePage	1252
SqlCommand	update phu set phuname=?, phucityid=?,address=?, posta

SqlCommand

The SQL command to be executed.

Refresh OK Cancel Help



- Nếu không (Thêm cột CreatedDate, UpdatedDate, insert data mới)



Derived Column Transformation Editor

Specify the expressions used to create new column values, and indicate whether the values update existing columns or populate new columns.

Variables and Parameters

Columns

Mathematical Functions

String Functions

Date/Time Functions

NULL Functions

Type Casts

Operators

Description:

Derived Column Name	Derived Column	Expression	Data Type	Le
CreateDate	<add as new column>	GETDATE()	database timestamp ...	
UpdatedDate	<add as new column>	GETDATE()	database timestamp ...	

Configure Error Output... OK Cancel Help



OLE DB Destination Editor

Configure the properties used to insert data into a relational database using an OLE DB provider.

Connection Manager
Mappings
Error Output

Specify an OLE DB connection manager, a data source, or a data source view, and select the data access mode. If using the SQL command access mode, specify the SQL command either by typing the query or by using Query Builder. For fast-load data access, set the table update options.

OLE DB connection manager:
MSI.NDS_PHU New...

Data access mode:
Table or view - fast load

Name of the table or the view:
[dbo].[PHU] New...

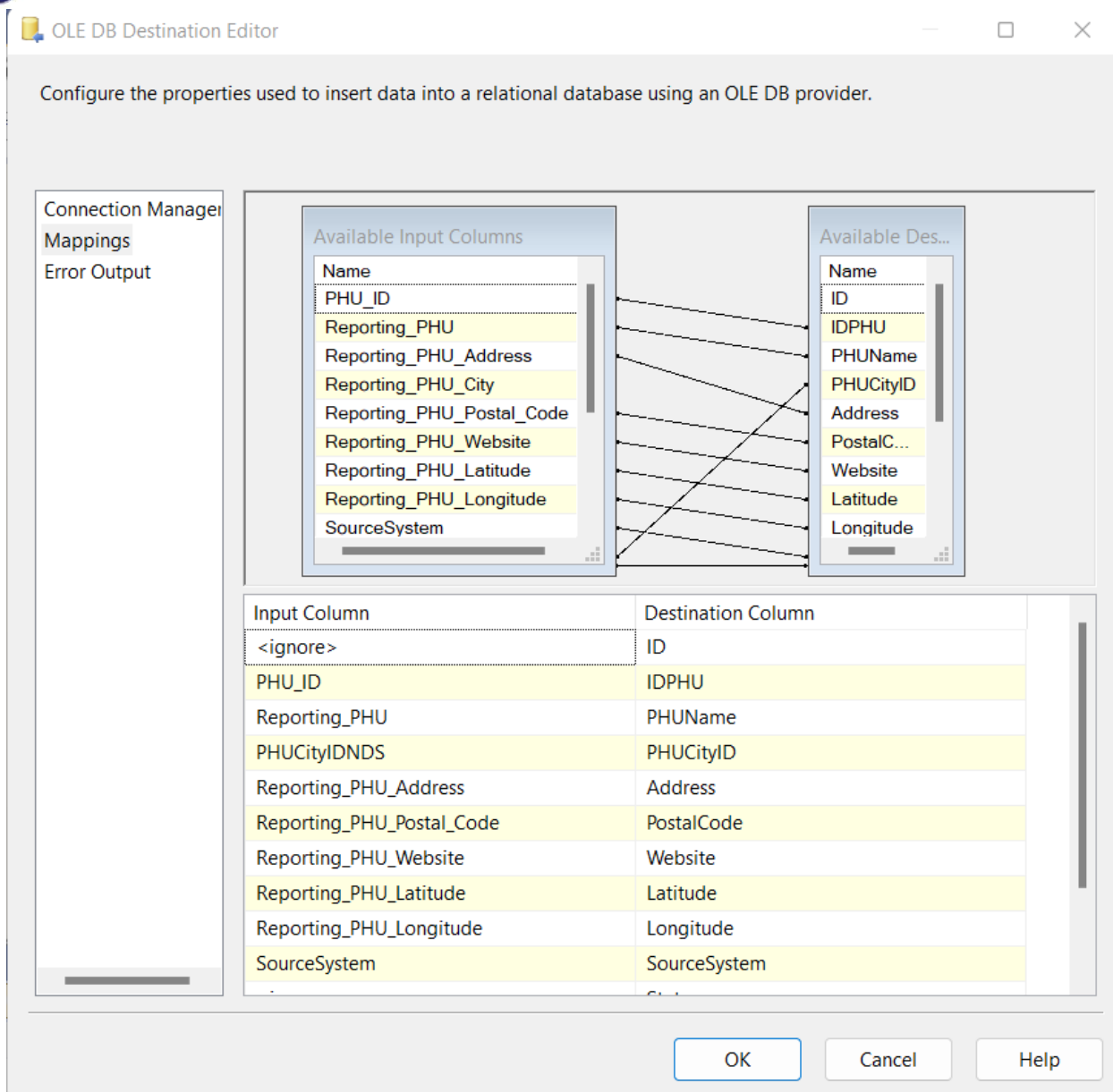
☐ Keep identity ☒ Table lock
☐ Keep nulls ☒ Check constraints

Rows per batch:

Maximum insert commit size: 2147483647

View Existing

OK Cancel Help

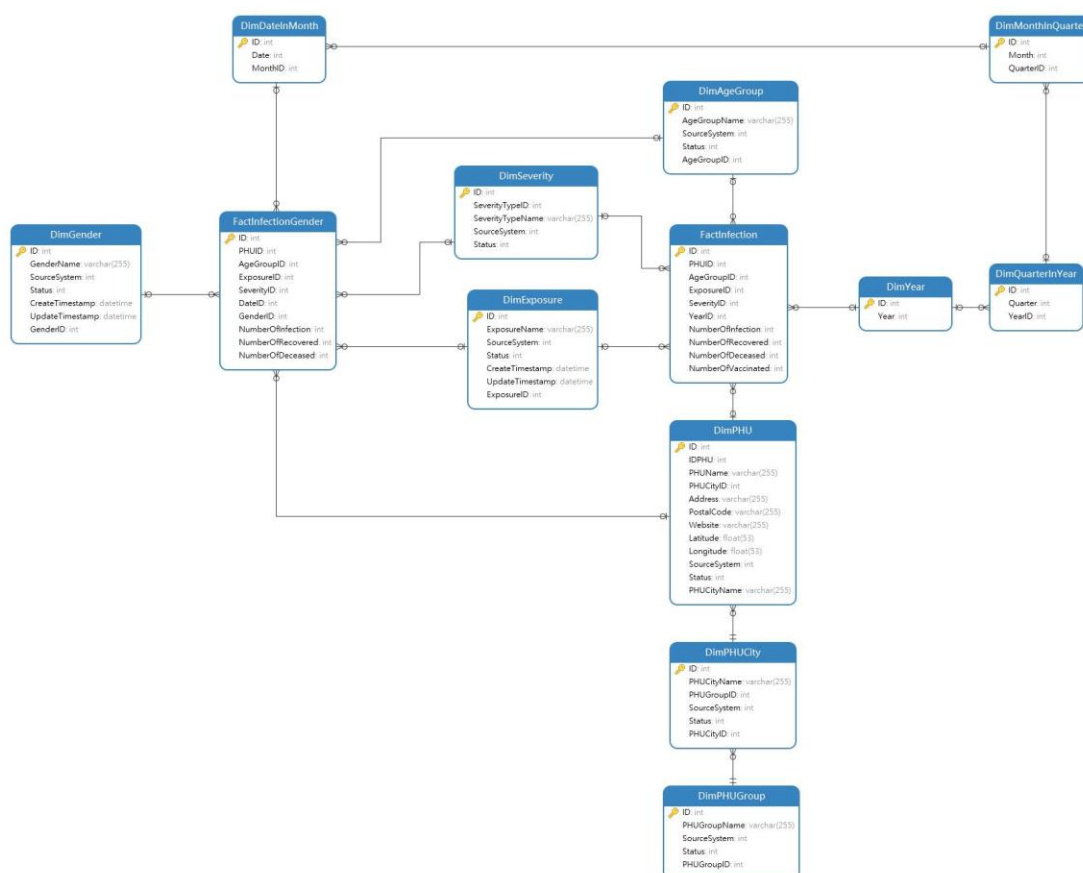


3.3 ETL từ NDS vào DDS

3.3.1 Phân tích yêu cầu và thiết kế DDS

- Yêu cầu report liên quan đến việc thống kê số ca nhiễm, ca tử vong, ca phục hồi theo phân cấp khu vực PHU, PHUCity, PHUGroup, phân cấp chiều thời gian, nhóm tuổi, giới tính, mức độ nghiêm trọng, sự phơi nhiễm. Do vậy, ta sẽ tách dữ liệu các tiêu chí trên thành các bảng Dimension tương ứng và phân cấp chiều tương ứng.

- Ta nhận thấy dữ liệu bảng VaccineByAge không có trường về giới tính so với dữ liệu ca nhiễm CaseReport (dữ liệu từ bảng Compiled_COVID-19_Case_Deatails) và có yêu cầu thống 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) nên ta sẽ tạo 1 bảng FactInfection.
- Các yêu cầu thống kê khác không cần sử dụng đến data bảng VaccineByAge và yêu cầu thống kê theo giới tính. Do đó, ta sẽ tạo 1 bảng FactInfectionGender.
- Thiết kế DDS:



- Giải thích ý nghĩa các component:
 - Bảng FactInfection:
 - Load data từ bảng CaseReport, VaccinesByAge, Severity trong NDS. Trước khi đổ dữ liệu, data đã được group by theo PHU, nhóm tuổi, độ phơi nhiễm (Exposure), mức độ nghiêm trọng (Severity) và năm (Year) để tính tổng các Measures
 - Mức độ chi tiết (độ mịn): mỗi dòng trong bảng tương ứng với dữ liệu về ca nhiễm của từng PHU, nhóm tuổi, độ phơi nhiễm (Exposure), mức độ

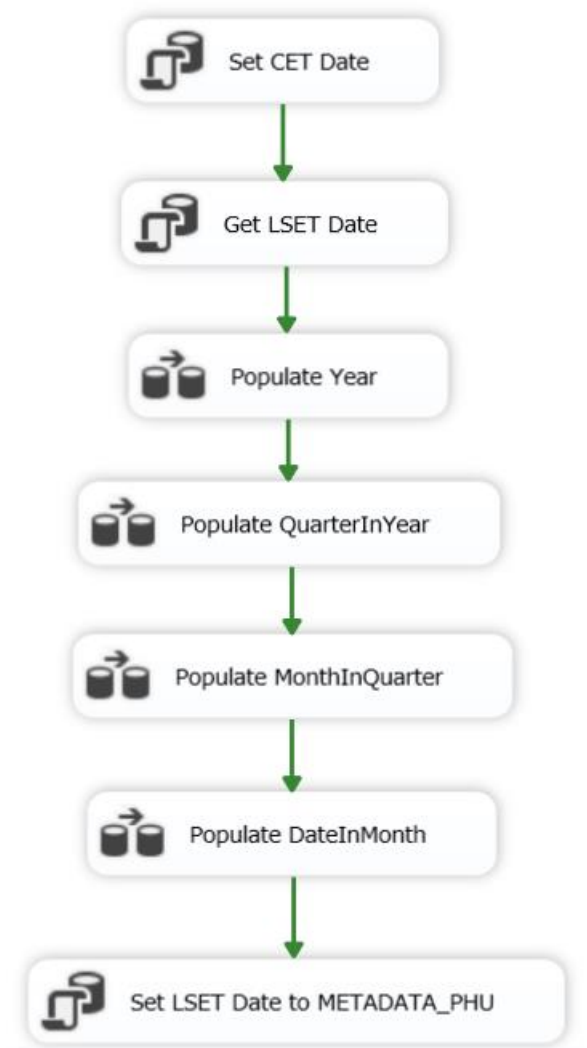
nghiêm trọng (Severity) và theo từng năm (Year) (Theo từng năm là do dữ liệu bảng CaseReport trong NDS có trường Date_Report từ ngày 1/1/2020 đến ngày 2/6/2021, còn bảng VaccinesByAge trong NDS có trường Date từ ngày 26/7/2021 đến ngày 25/8/2022 nên mapping theo Date sẽ không có data)

- Measures: Số lượng ca nhiễm, Số lượng ca phục hồi, Số lượng ca tử vong, Số lượng đã tiêm vaccine
- Bảng FactInfectionGender:
 - Load data từ bảng CaseReport, Severity trong NDS. Trước khi đổ dữ liệu, data đã được group by theo PHU, nhóm tuổi, độ phơi nhiễm (Exposure), mức độ nghiêm trọng (Severity), giới tính (Gender) và ngày (Date) để tính tổng các Measures
 - Mức độ chi tiết: mỗi dòng trong bảng tương ứng với dữ liệu về ca nhiễm của từng PHU, nhóm tuổi, độ phơi nhiễm (Exposure), mức độ nghiêm trọng (Severity), giới tính (Gender) và theo từng ngày (Date)
 - Measures: Số lượng ca nhiễm, Số lượng ca phục hồi, Số lượng ca tử vong
- Phân cấp chiều thời gian (DimDateInMonth, DimMonthInQuarter, DimQuarterInYear, DimYear): load data từ bảng CaseReport (trường DateReport)
- Phân cấp chiều vị trí (DimPHU, DimPHUCity, DimPHUGroup): load data từ bảng PHU, PHUCity, PHUGroup trong NDS
 - Bảng DimPHU lưu lại các giá trị lịch sử tọa độ, PHUName, địa chỉ, postalcode
- Bảng DimExposure (Phơi nhiễm): load data từ bảng Exposure trong NDS
- Bảng DimSeverity (Mức độ nghiêm trọng): load data từ bảng SeverityType trong NDS (mức độ nghiêm trọng dựa theo các trường AgeGroup và CaseStatus đã được định nghĩa sẵn trong bảng Severity trong NDS)
- Bảng DimGender (Giới tính): load data từ bảng Gender trong NDS
- Bảng DimAgeGroup (Độ tuổi): load data từ bảng AgeGroup trong NDS

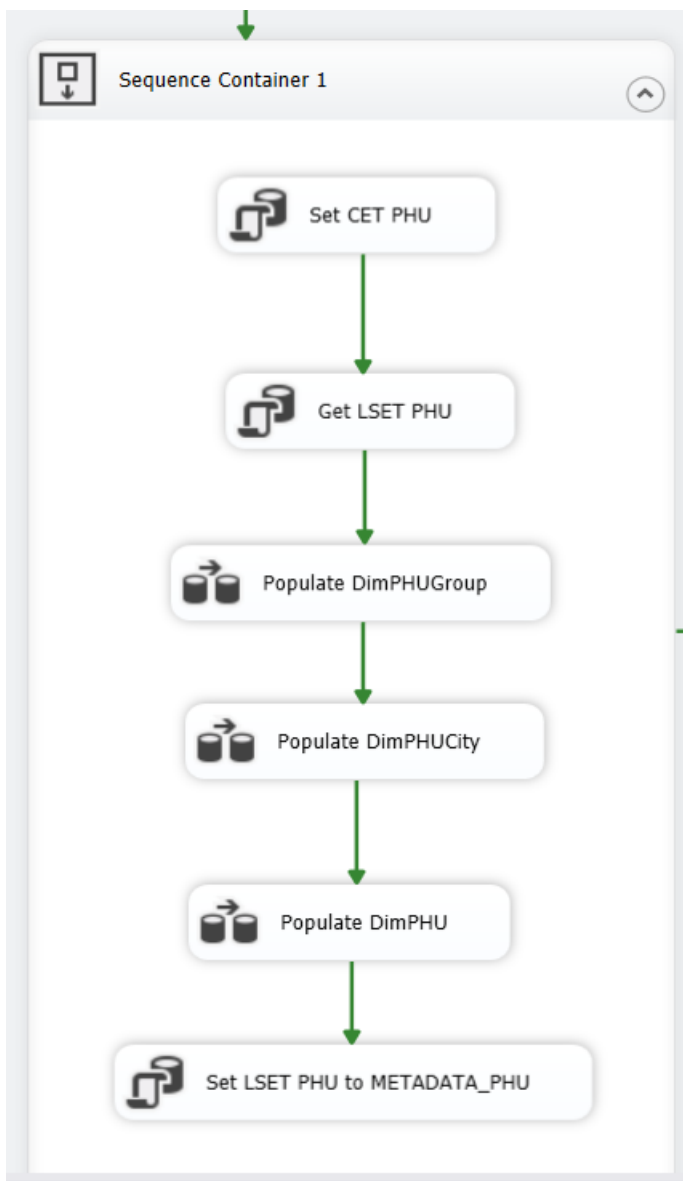
3.3.2 ETL từ NDS vào DDS



- Phân cấp chiều thời gian: Year > Quarter > Month > Day



- Phân cấp chiều theo khu vực: PHU_Group > PHU_City > PHU

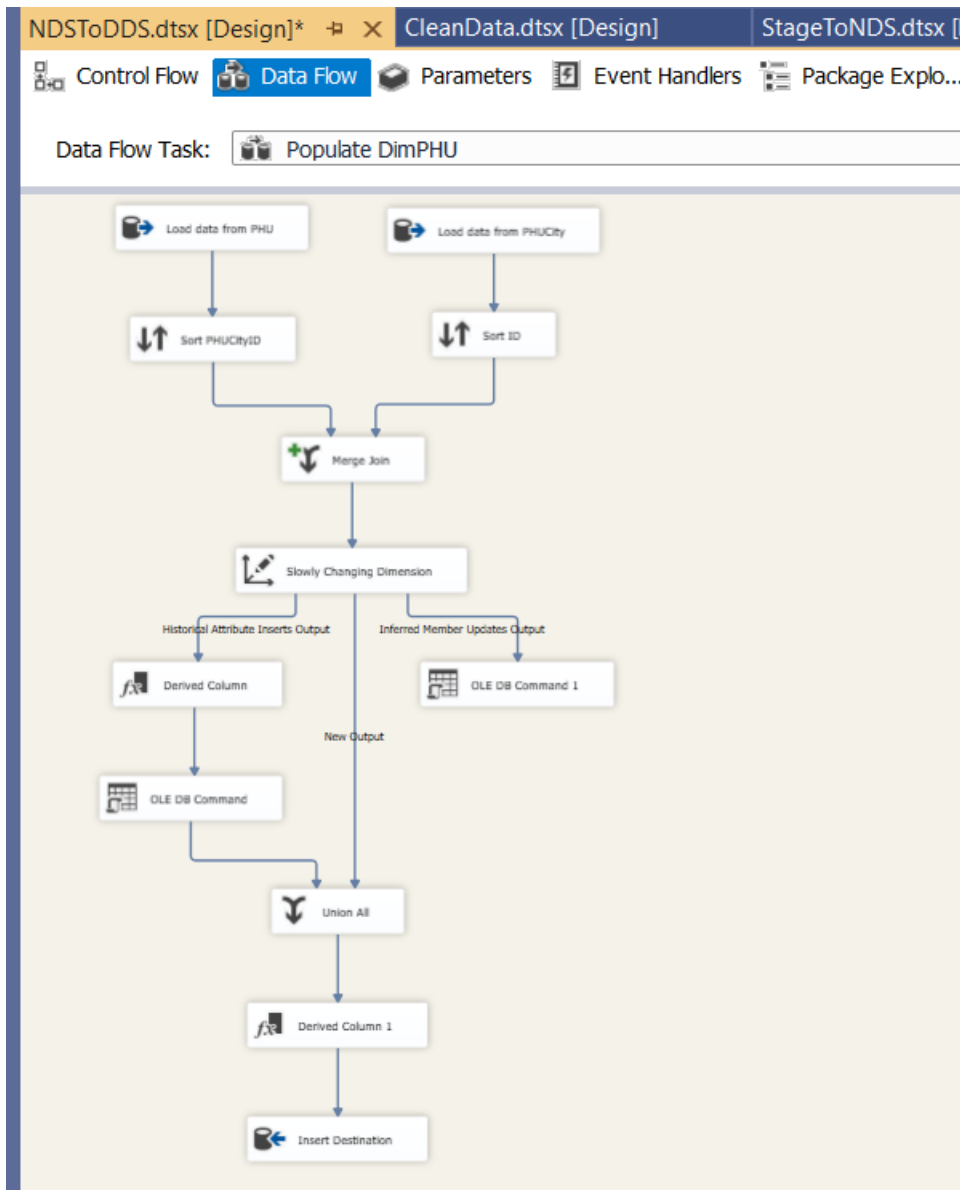


– Đổ dữ liệu vào các bảng Dimension:





- Tạo chiều thay đổi chậm trên bảng PHU:





Slowly Changing Dimension Wizard

Select a Dimension Table and Keys

Select a dimension table to load and map columns in the transformation input to

Connection manager:
MSI.DDS_PHU

Table or view:
[dbo].[DimPHU]

Input Columns	Dimension Columns	Key Type
Address	Address	Not a key column
ID	IDPHU	Business key
Latitude	Latitude	Not a key column
Longitude	Longitude	Not a key column
PHUCityID	PHUCityID	Not a key column
PHUCityName	PHUCityName	Not a key column

Help < Back Next > Finish >>| Cancel

Slowly Changing Dimension Wizard

Slowly Changing Dimension Columns
Manage the changes to column data in your slowly changing dimensions by setting the

Fixed Attribute
Select this type when the value in a column should not change. Changes are treated as errors.

Changing Attribute
Select this type when changed values should overwrite existing values. This is a Type 1 change.

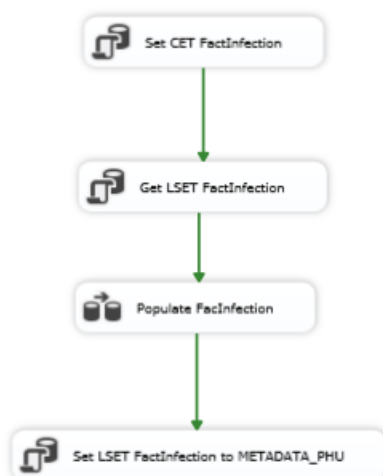
Historical Attribute
Select this type when changes in column values are saved in new

Select a change type for slowly changing dimension columns:

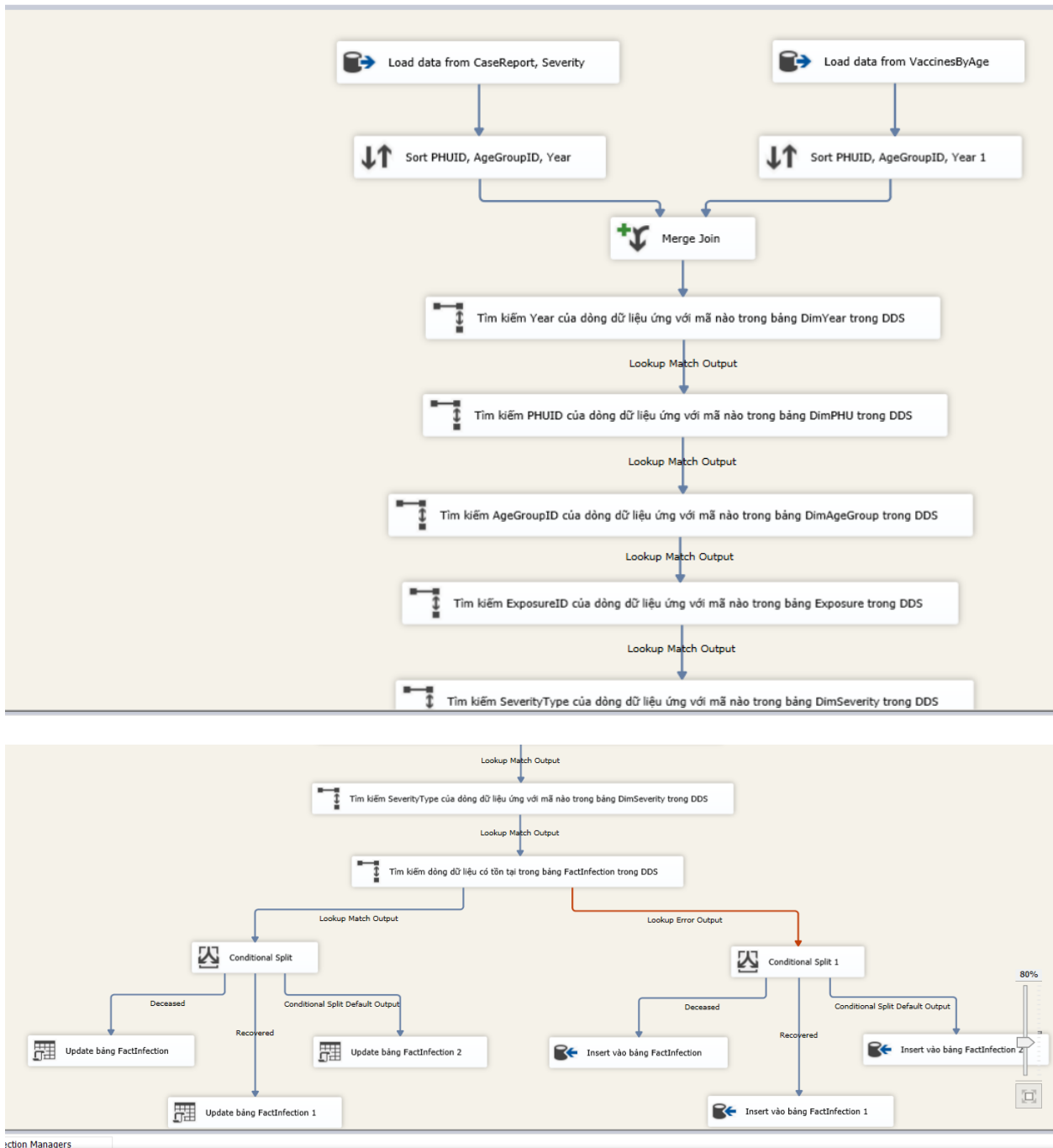
Dimension Columns	Change Type
Address	Historical a...
Latitude	Historical a...
Longitude	Historical a...
PHUName	Historical a...
PostalCode	Historical a...

Help < Back Next > Finish >>| Cancel

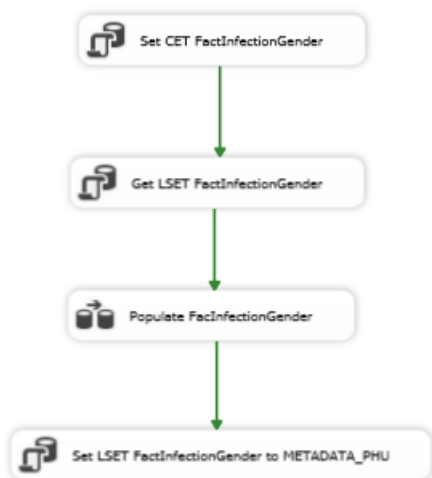
- Đồ dữ liệu vào bảng FactInfection:



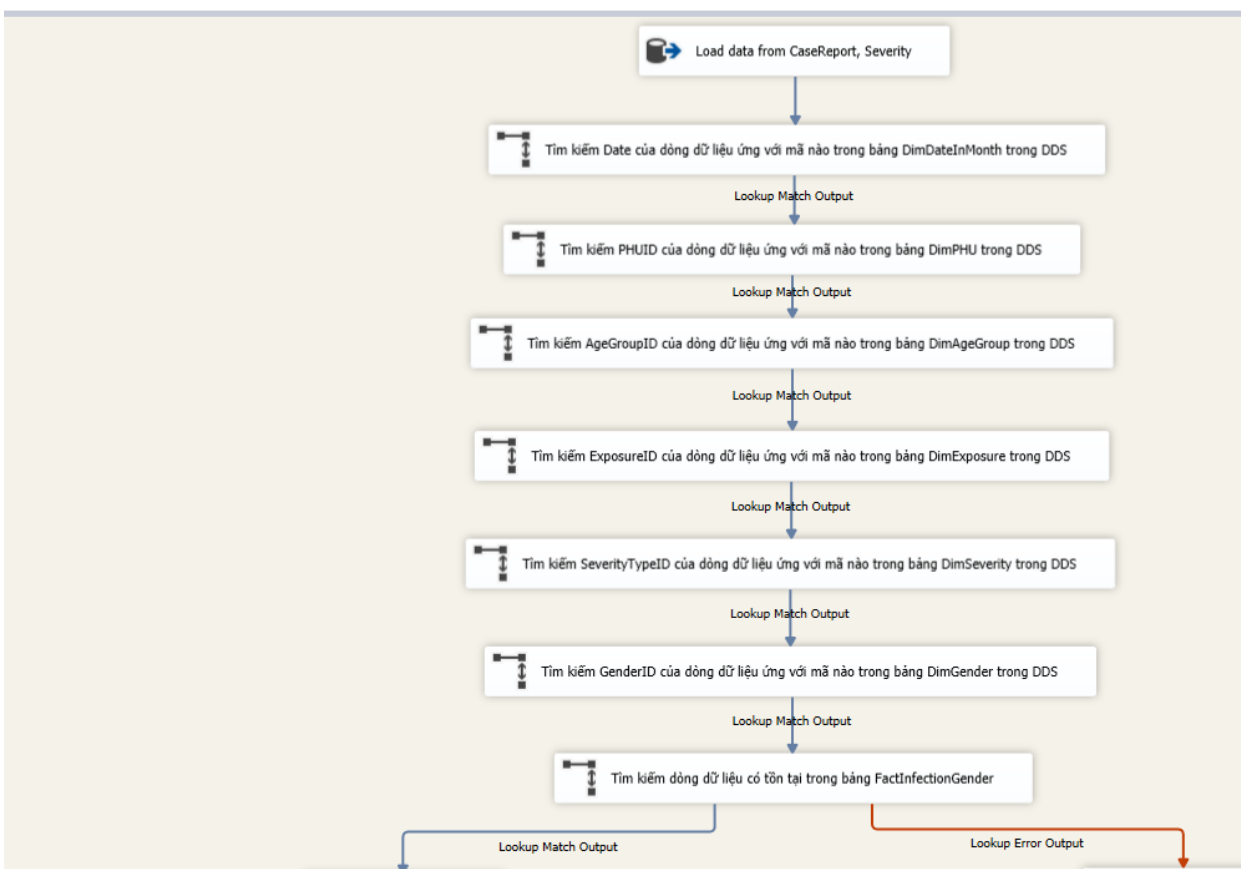
Populate FacInfection

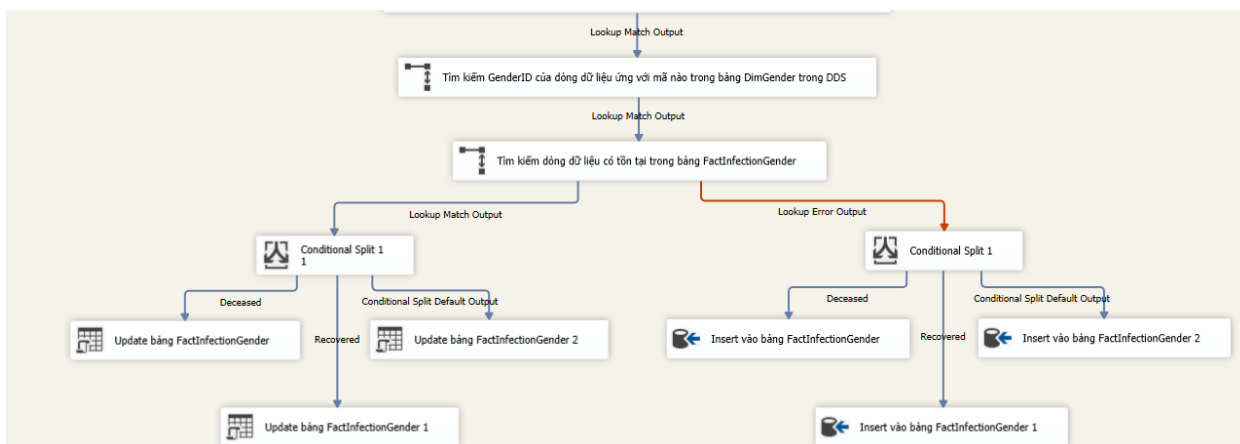


– Đổ dữ liệu vào bảng FactInfectionGender:



Populate FacInfectionGender

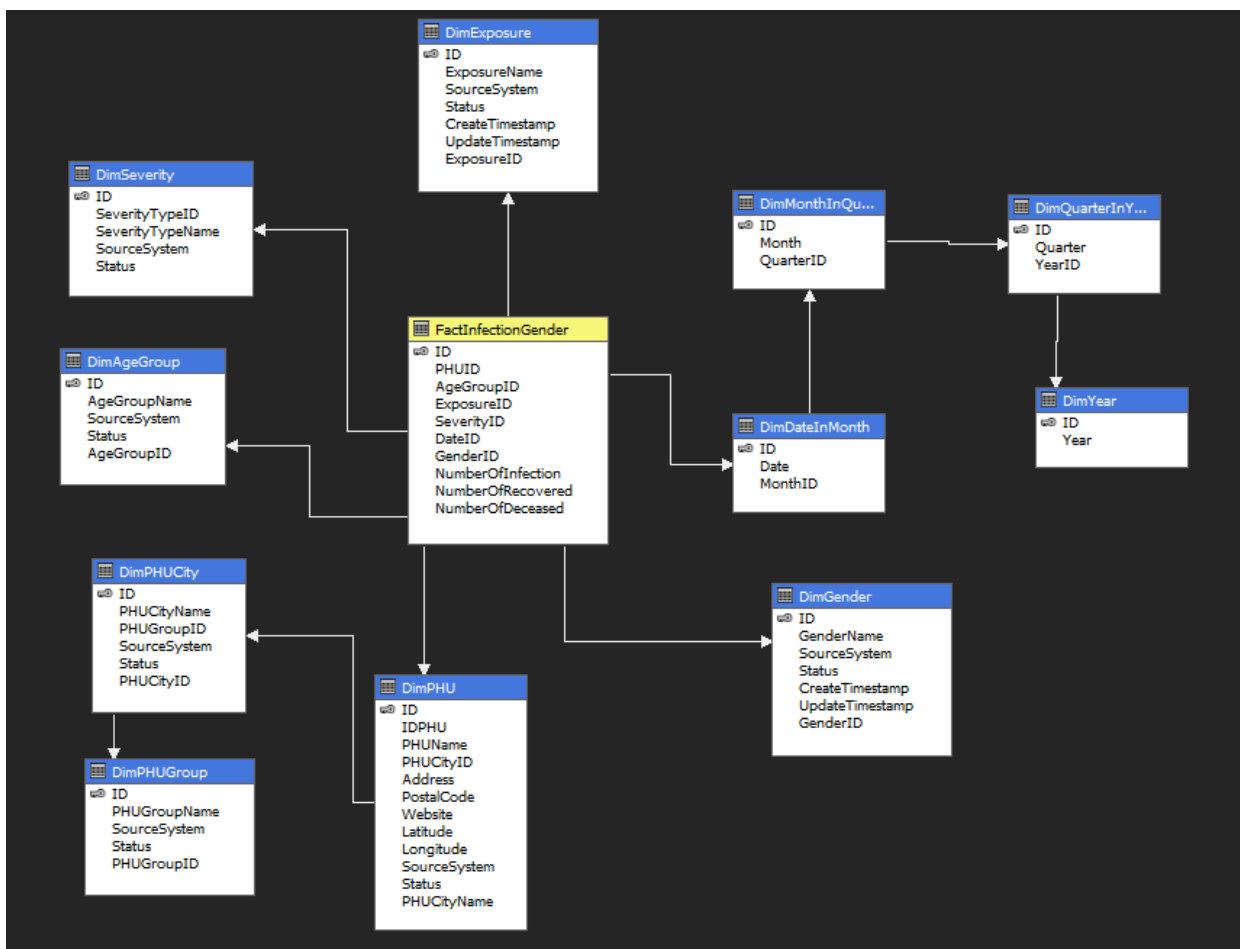




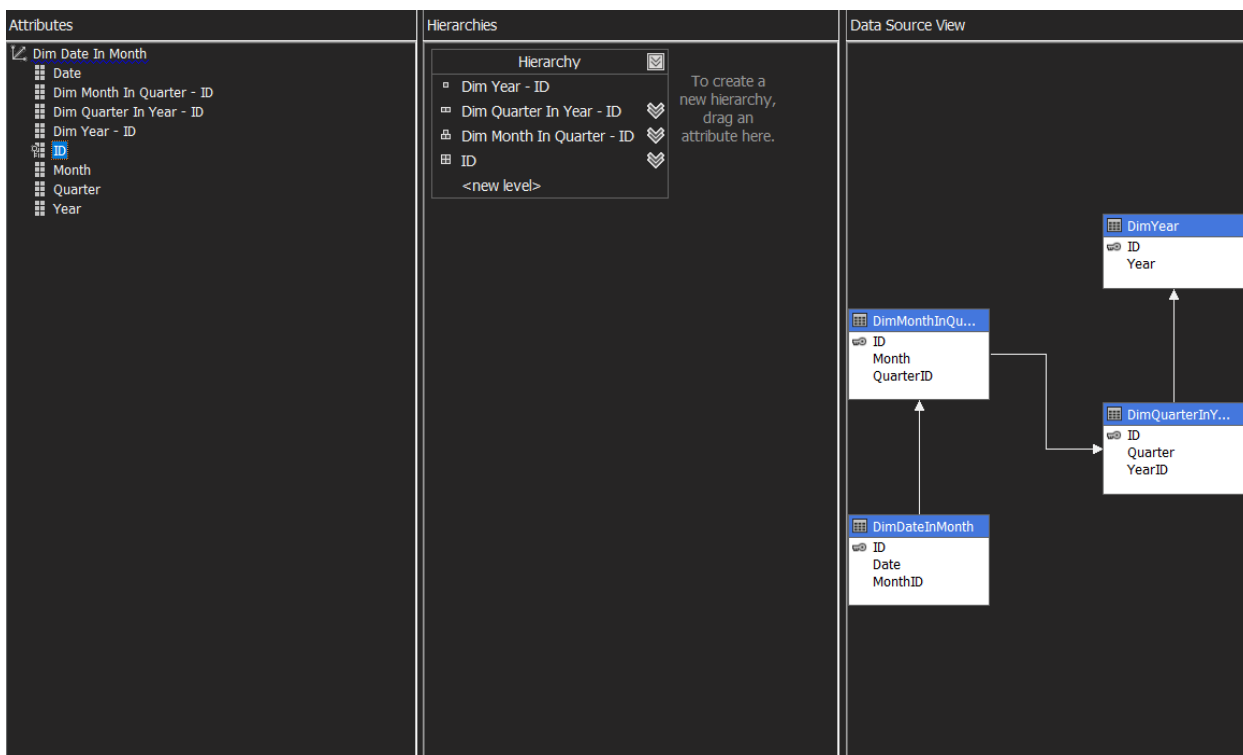
4 OLAP

4.1 OLAP Cube:

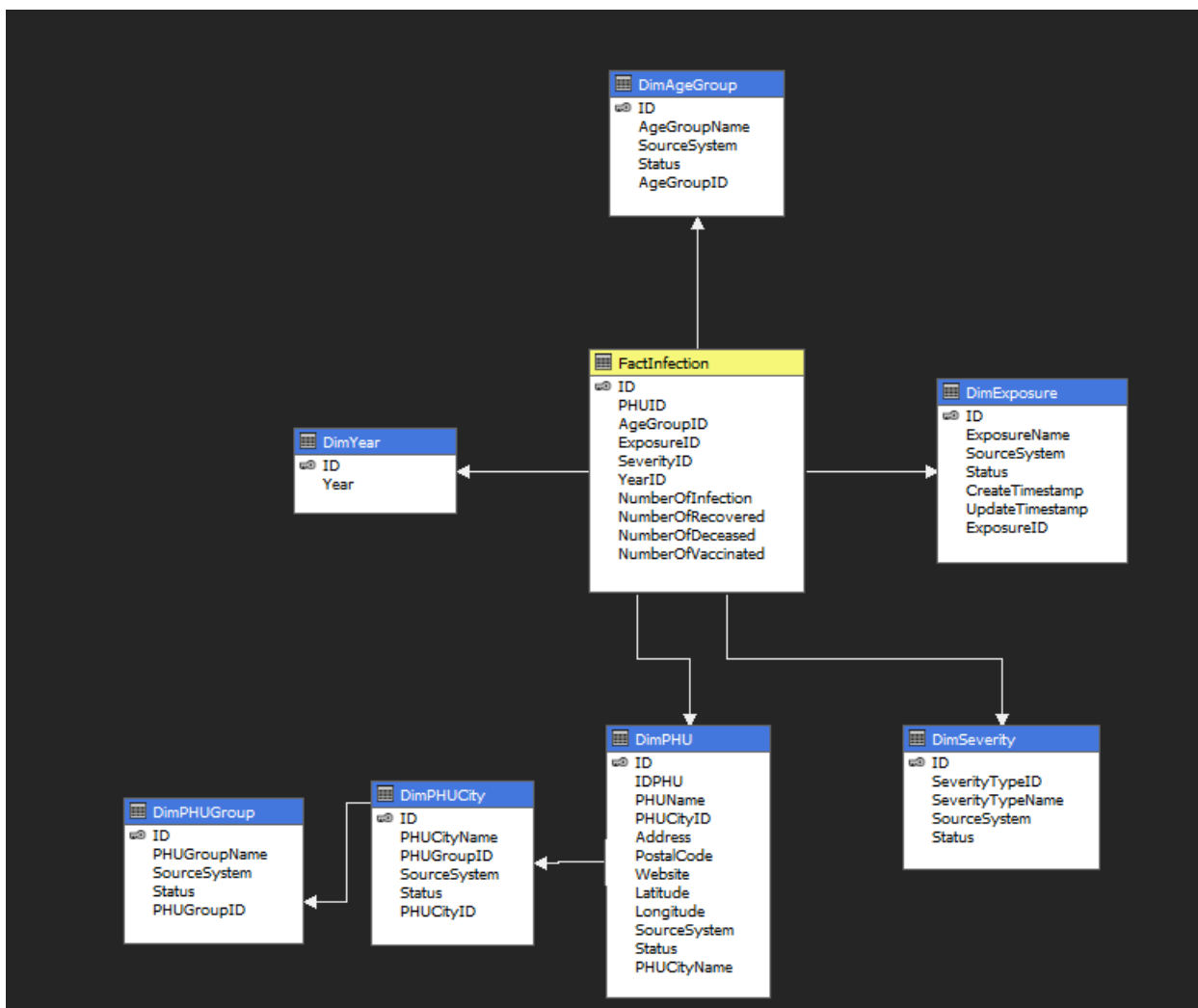
- Sau khi có được dữ liệu DDS, ta tiến hành thực hiện đưa DDS vào trong Cube.
- Một bảng Fact sử dụng chiều thời gian theo phân cấp, còn bảng Fact còn lại chỉ dùng chiều thời gian Year, nhóm em tạo ra hai Cube.
- Cube 1:



- Phân cấp chiều thời gian, các chiều phân cấp còn lại tương tự:



- Cube 2:



- Những Report dựa trên Vaccine trong năm sẽ dựa trên Cube 2.
- Sau khi process các cube, tiến hành report trên excel.

5 Report và Visualize:

- Những kết quả có số dòng ít sẽ được tạo thêm đồ thị thống kê.



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

Filters		Columns		
		Σ Values		
Rows		Σ Values		
Year		Number Of Deceas...		
PHU Name		Number Of Recover...		
		Number Of Infection		
1	Row Labels	Number Of Deceased	Number Of Recovered	Number Of Infection
2	2020			
3	Algoma Public Health Unit	1	87	88
4	Brant County Health Unit	9	1121	1130
5	Chatham-Kent Health Unit	4	822	826
6	Durham Region Health Department	257	8019	8276
7	Eastern Ontario Health Unit	52	1611	1663
8	Grey Bruce Health Unit	1	507	508
9	Haldimand-Norfolk Health Unit	43	969	1012
10	Haliburton, Kawartha, Pine Ridge District Health Unit	27	578	605
11	Halton Region Health Department	143	6136	6279
12	Hamilton Public Health Services	218	6544	6762
13	Hastings and Prince Edward Counties Health Unit	5	301	306
14	Huron Perth District Health Unit	24	786	810
15	Kingston, Frontenac and Lennox & Addington Public Health	0	562	562
37	2021			
38	Algoma Public Health Unit	5	295	306
39	Brant County Health Unit	14	2482	2607
40	Chatham-Kent Health Unit	17	1002	1035
41	Durham Region Health Department	109	15926	16421
42	Eastern Ontario Health Unit	64	2841	2950
43	Grey Bruce Health Unit	6	777	794
44	Haldimand-Norfolk Health Unit	10	1553	1609
45	Haliburton, Kawartha, Pine Ridge District Health Unit	35	1344	1462
46	Halton Region Health Department	80	10279	10738
47	Hamilton Public Health Services	172	13253	13909
67	Timiskaming Health Unit	2	124	126
68	Toronto Public Health	1113	94135	98156
69	Wellington-Dufferin-Guelph Public Health	59	4834	5073
70	Windsor-Essex County Health Unit	162	7117	7445
71	York Region Public Health Services	268	31935	32767
72	Grand Total	8801	514999	533761



5.2 Thống kê số ca tử vong của dịch Covid-19 theo PHU, Mức Độ Nghiêm Trọng và theo các Quý trong từng năm.

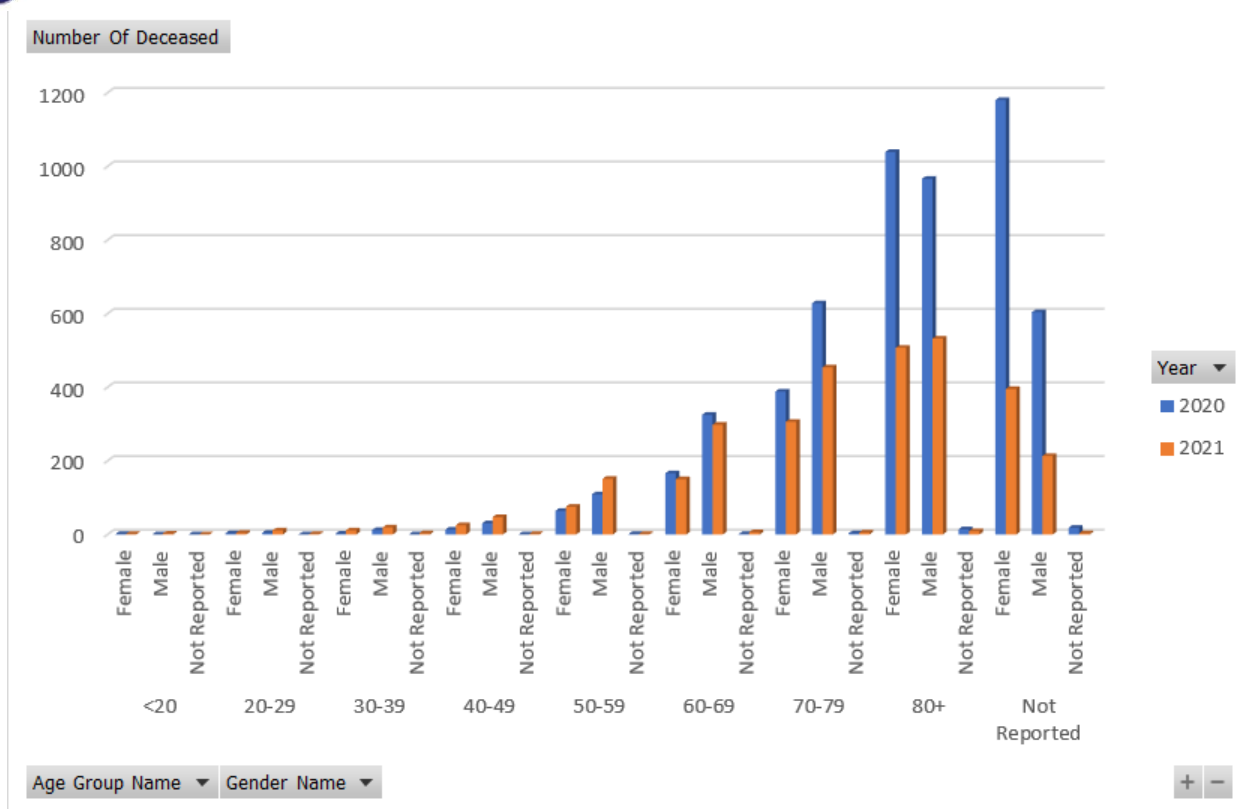
Filters		Columns	
		Year	
		Quarter	
Rows		Values	
PHU Name		Number Of Deceased	
Severity Type Name			

1	Number Of Deceased	Column Labels					
2		2020				2021	
3	Row Labels	1	2	3	4	1	2
4	Algoma Public Health Unit						
5	High					1	1
6	Low	0	0	0	0	0	0
7	Medium	0	0		0	0	0
8	Very High				1	2	1
9	Brant County Health Unit						
10	High	1	1	1	2	4	1
11	Low	0	0	0	0	0	0
12	Medium	0	0	0	0	0	0
13	Very High	1	2		1	3	6
164	Windsor-Essex County Health Unit						
165	High	6	30	3	79	49	11
166	Low	0	0	0	0	0	0
167	Medium	0	0	0	0	0	0
168	Very High	9	25	4	111	95	7
169	York Region Public Health Services						
170	High	18	109	6	69	72	36
171	Low	0	0	0	0	0	0
172	Medium	0	0	0	0	0	0
173	Very High	31	94	13	125	114	46
174	Grand Total	455	2448	174	2494	2230	1000



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

Filters		Columns	
		Year	
Rows		Values	
Age Group Name		Number Of Deceased	
Gender Name			
1	Number Of Deceased	Column Labels	
2	Row Labels	2020	2021 Grand Total
3	<20		
4	Female	1	1 2
5	Male	0	2 2
6	Not Reported	0	0 0
7	20-29		
8	Female	3	4 7
9	Male	4	11 15
10	Not Reported	0	1 1
11	30-39		
12	Female	2	11 13
13	Male	12	19 31
14	Not Reported	0	3 3
31	80+		
32	Female	1039	507 1546
33	Male	966	532 1498
34	Not Reported	14	9 23
35	Not Reported		
36	Female	1180	395 1575
37	Male	604	213 817
38	Not Reported	18	3 21
39	Grand Total	5571	3230 8801



– Nhận xét:

- Chủ yếu từ độ tuổi 40+ trở đi thì số ca tử vong bắt đầu tăng đáng kể so với các nhóm tuổi thấp hơn.
- Ở độ tuổi 40-79 thì số ca tử vong ở giới tính nam luôn nhiều hơn giới tính nữ.
- Ở độ tuổi 80+ có số ca tử vong cao đáng kể so với các nhóm tuổi khác, cả nam và nữ có số ca tử vong gần như nhau.

5.4 Thống kê số ca nhiễm, tử vong theo Mức Độ Nghiêm Trọng theo Ngày trong Tháng của các Năm.

Filters		Columns	
		Σ Values	
Rows		Σ Values	
Year		Number Of Deceased	
Month		Number Of Infection	
Date			
Severity Type Name			
1	Row Labels	Number Of Deceased	Number Of Infection
2	2020		
3	1		
4	1		
5	Low	0	2
6	Medium	0	1
7	10		
8	Low	0	1
9	13		
10	Low	0	1
11	16		
12	Low	0	1
2240	6		
2241	1		
2242	Medium	0	235
2243	High	0	109
2244	2		
2245	Medium	0	6
2246	High	0	2
2247	Grand Total	8801	533761



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

Filters		Columns		
		Σ Values		
Rows		Σ Values		
Year		Number Of Infection		
PHU Group Name		Number Of Deceas...		
PHU City Name		Number Of Vaccina...		
Severity Type Name				
1	Row Labels	Number Of Infection	Number Of Deceased	Number Of Vaccinated
2	2021			
3	Central			
4	Newmarket			
5	High	150	108	2358155
6	Low	24152	0	2941616
7	Medium	2260	0	2891304
8	Very High	160	160	401226
9	Toronto			
10	High	704	501	6055357
11	Low	73598	0	7795633
12	Medium	7905	0	8752994
13	Very High	612	612	979419
14	Central East			
15	Durham Region			
16	High	70	48	1273538
17	Low	11937	0	1783276
178	Waterloo Wellington			
179	Waterloo			
180	High	58	42	698503
181	Low	6905	0	1510922
182	Medium	685	0	1538469
183	Very High	34	34	186516
184	Grand Total	271976	3227	97140308



5.6 Thống kê số ca nhiễm theo Mức Độ Nghiêm Trọng, loại tiếp xúc của từng khu vực trong các năm.

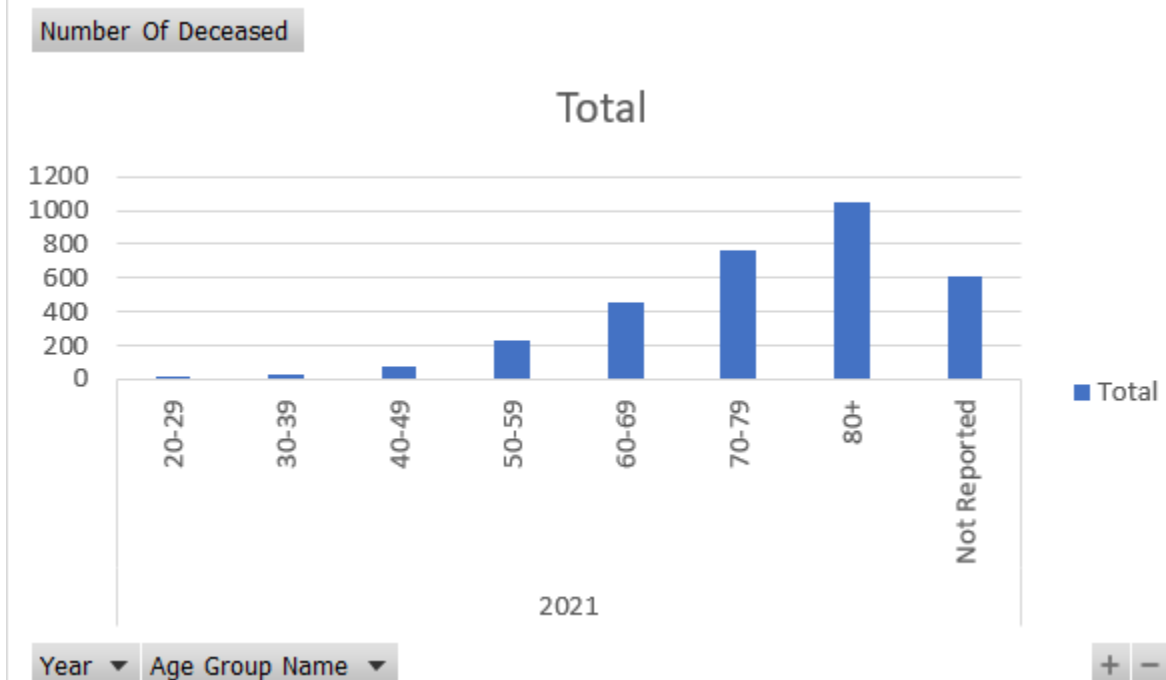
Filters		Columns	
		Year	
		Exposure Name	
Rows		Values	
PHU Group Name		Number Of Infection	
PHU City Name			
Severity Type Name			

1	Number Of Infection	Column Labels							
2		2020				2021			
3	Row Labels	Close Contact	Not Reported	Outbreak	Travel-Related	Close Contact	Not Reported	Outbreak	Travel-Related
4	Central								
5	Newmarket								
6	High	24	27	150	3	131	89	36	3
7	Low	7793	4681	1799	270	11406	10561	1967	218
8	Medium	2729	759	739	53	5039	2363	760	34
9	Very High	38	50	171	4	63	42	55	
10	Toronto								
11	High	69	173	814	5	352	686	179	3
12	Low	11001	29699	8057	739	12227	53496	7286	589
13	Medium	4566	4412	3651	91	10964	9588	2035	139
14	Very High	86	256	829	21	81	382	149	
15	Central East								
16	Durham Region								
17	High	3	13	110		81	41	18	4
18	Low	2901	1962	1361	121	6142	4421	1205	169
19	Medium	968	261	425	20	2982	774	492	31
20	Very High	8	12	106	5	25	14	21	1
21	Peterborough								
22	High			4	1	13	1	3	
23	Low	169	63	55	35	509	149	97	12
24	Medium	39	11	13	10	197	24	91	2
25	Very High					3	3	4	
181	Waterloo Wellington								
182	Waterloo								
183	High	7	17	46		55	24	30	
184	Low	2172	1772	1043	113	3303	2240	1179	183
185	Medium	809	321	286	18	1479	499	269	28
186	Very High	6	25	78	1	3	8	23	
187	Grand Total	87429	70148	41299	4178	149680	136834	39823	4370

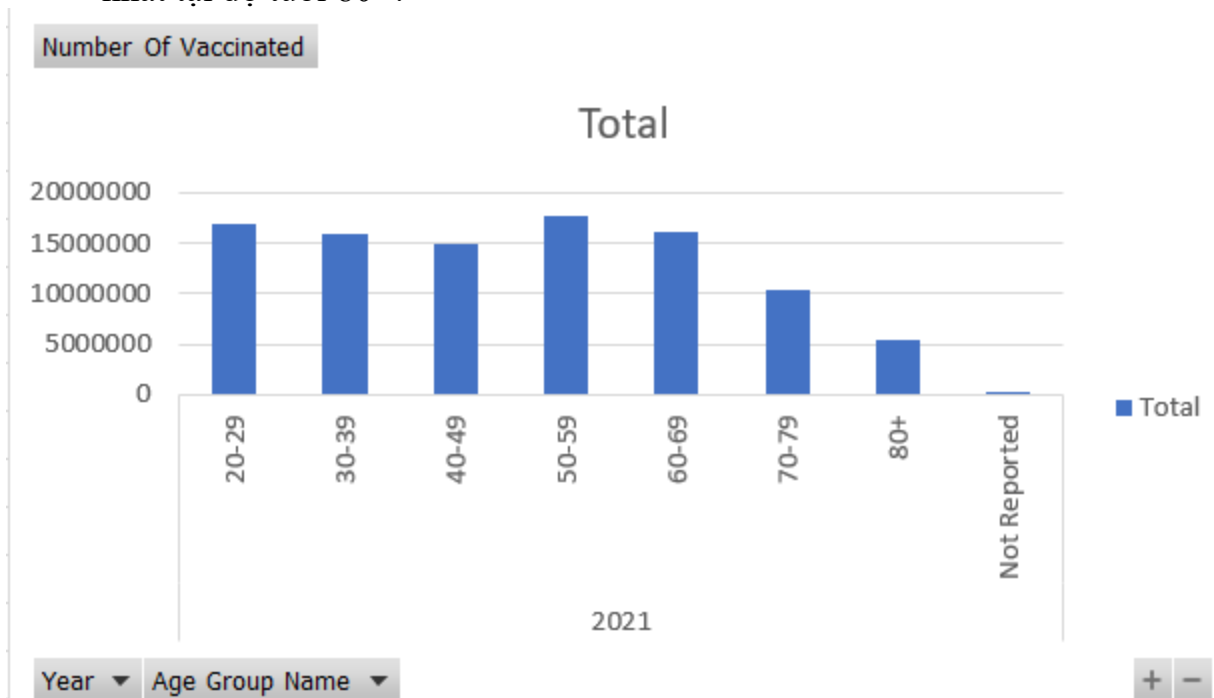
5.7 Thống kê số ca tử vong, ca nhiễm, số lượng người được chích vắc xin theo nhóm tuổi, City trong các năm.

Filters		Columns		
		Σ Values		
Rows		Σ Values		
Year		Number Of Infection		
Age Group Name		Number Of Deceas...		
		Number Of Vaccina...		
1	Row Labels	Number Of Infection	Number Of Deceased	Number Of Vaccinated
2	2021			
3	20-29	70979	16	16884896
4	30-39	54916	33	15862335
5	40-49	47740	73	14879113
6	50-59	46180	227	17582071
7	60-69	28268	454	16008669
8	70-79	13445	765	10417226
9	80+	7375	1048	5468546
10	Not Reported	3073	611	37452
11	Grand Total	271976	3227	97140308

- Vì số liệu chênh lệch ở các cột nên nhóm em tách thành 3 biểu đồ cho 3 cột.

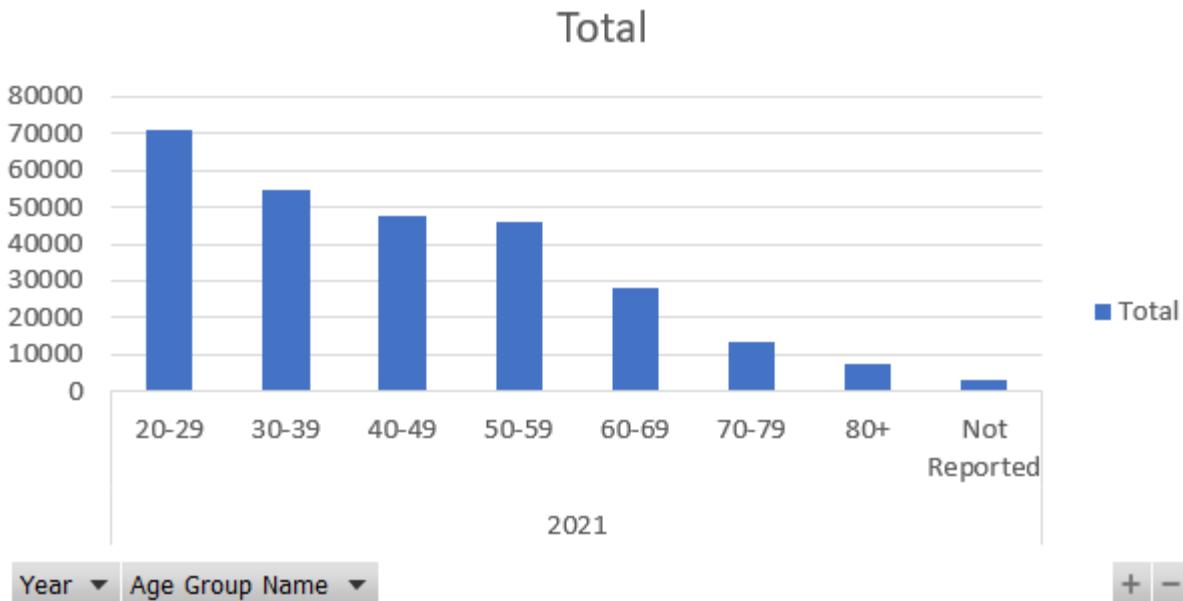


- Nhận xét: nhìn chung, số ca tử vong nhiều tỉ lệ thuận theo nhóm tuổi và cao nhất tại độ tuổi 80+.



- Nhận xét: nhìn chung ở độ tuổi 20-69, số lượng người tiêm vacxin được khá cao, khoảng 15000000 người ở các nhóm tuổi trong đó. Tuy nhiên, số lượng đó giảm dần kể từ độ tuổi 70.

Number Of Infection



- Nhận xét: số ca nhiễm cao nhất là từ 20-29, sau đó giảm dần theo độ tuổi tăng dần.
- Nhận xét chung:
 - Độ tuổi cao (từ 60+ trở đi) có số ca nhiễm thấp nhưng có số người tử vong cao nhất, nguyên nhân 1 phần do số lượng người tiêm vaccine thấp.
 - Độ tuổi thấp (20-29) thì ngược lại, có số ca nhiễm cao và có số người tử vong thấp nhất bởi một phần do số lượng người tiêm vaccine cao.

6 Data mining

- Mục tiêu: Xác định khu vực nào, độ tuổi nào, giới tính nào có khả năng tử vong cao và dễ bị nhiễm do bùng phát.
- Dữ liệu: sử dụng bảng Compiled_COVID_19_Case_Details_Stage trong Stage.
- Trường Exposure, Case Status thực hiện mining qua Decision Tree model trên các trường Age group, Gender, Province.
- Đầu tiên là dự đoán tỉ lệ tử vong, với đầu vào là Age_Group, Gender, Province và đầu ra là Case_Status:

Data Mining Wizard

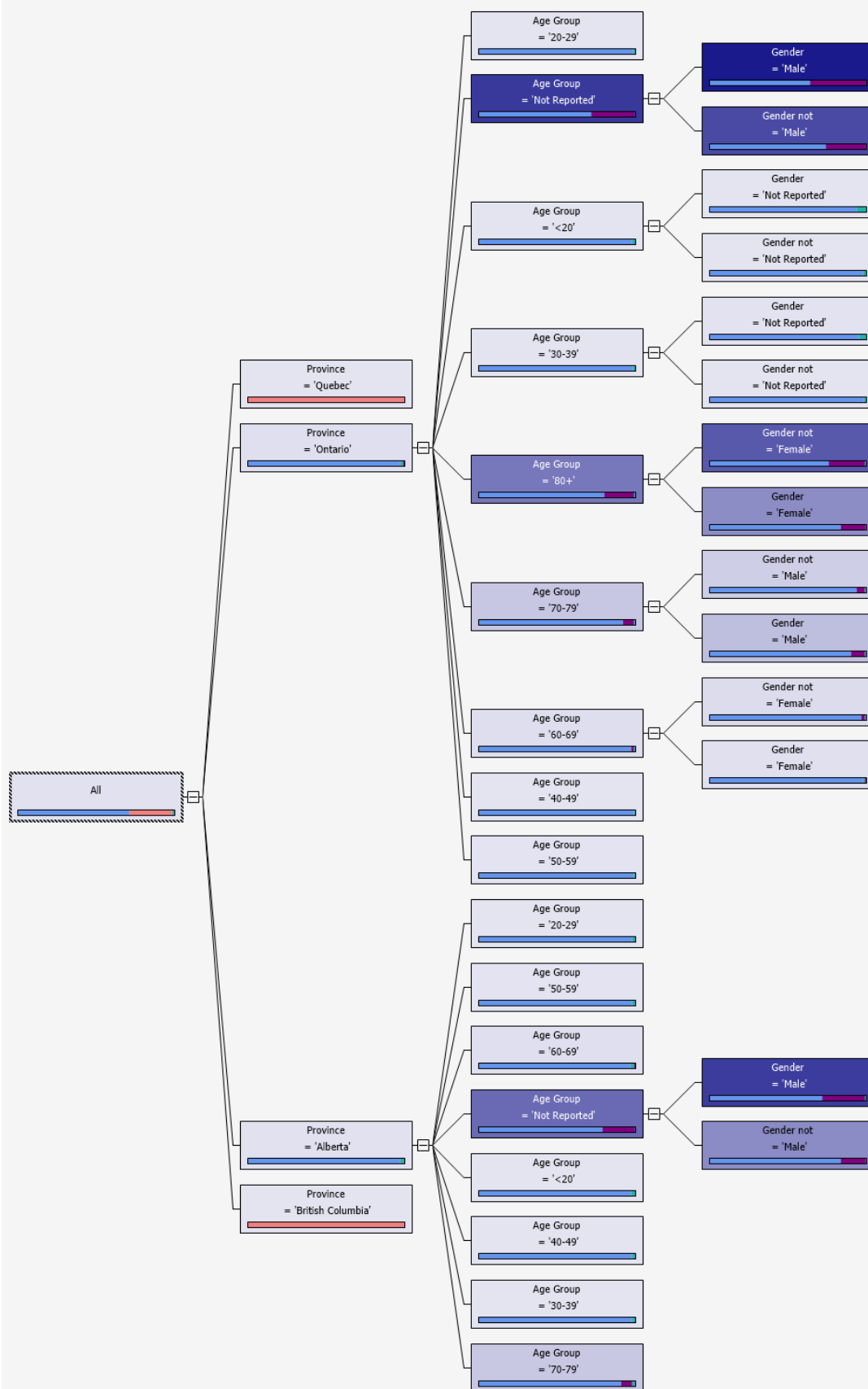
Specify the Training Data
Specify the columns used in your analysis.

Mining model structure:

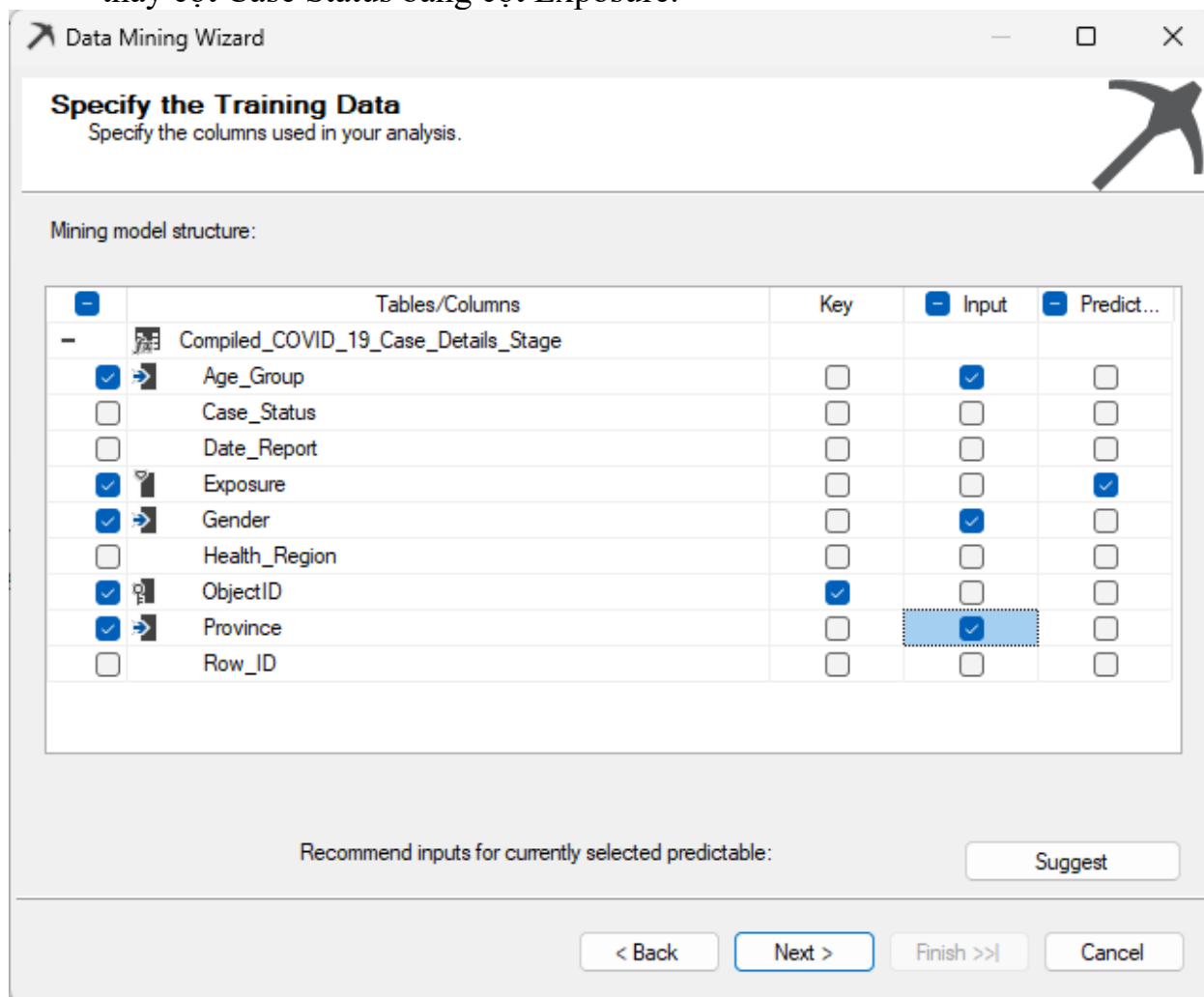
	Tables/Columns	Key	Input	Predict...
-	Compiled_COVID_19_Case_Details_Stage			
<input checked="" type="checkbox"/>	Age_Group	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Case_Status	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	Date_Report	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	Exposure	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Gender	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	Health_Region	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	ObjectID	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Province	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	Row_ID	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Recommend inputs for currently selected predictable:

- Sau khi thực hiện mining trên công cụ Data mining của Visual Studio, ta chọn giá trị Case_Status là Deceased ta được cây quyết định dự đoán mức độ tử vong:



- Dựa vào mô hình ta thấy:
 - Tại vùng Ontario, AgeGroup = "Not Reported", ở cả 2 giới tính Nam và Nữ thì có tỉ lệ tử vong (Deceased) cao nhất, tiếp đến là độ tuổi = "80+", giới tính Nam có tỉ lệ tử vong cao hơn Nữ
 - Tại Alberta, độ tuổi AgeGroup="Not Reported" có tỉ lệ tử vong cao nhất, Nam giới có tỉ lệ tử vong cao hơn Nữ giới
- Khả năng bùng phát: làm tương tự bên trên, ở bước tạo Mining Structures, thay cột Case Status bằng cột Exposure.



Data Mining Wizard

Specify the Training Data
Specify the columns used in your analysis.

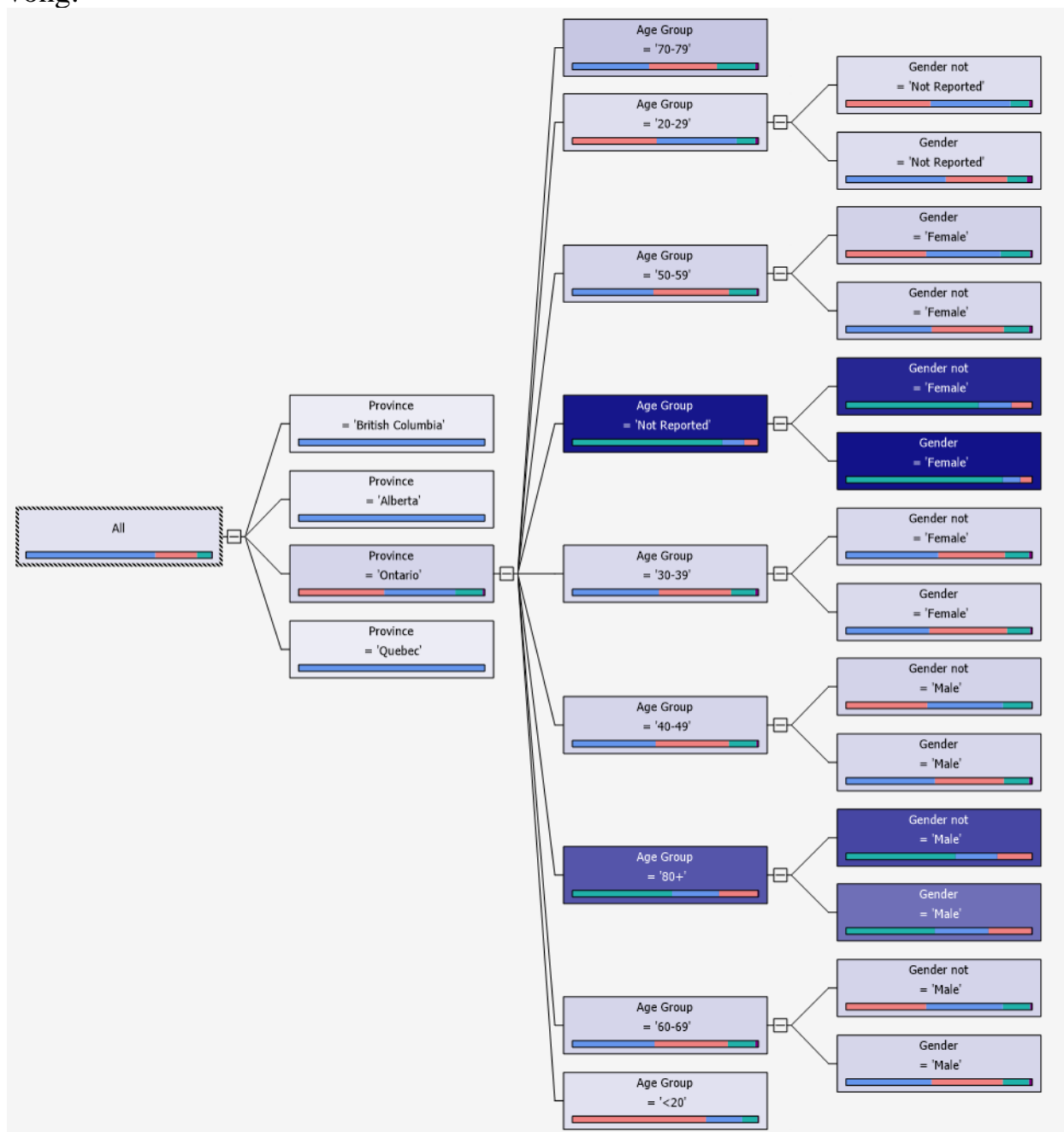
Mining model structure:

	Tables/Columns	Key	Input	Predict...
-	Compiled_COVID_19_Case_Details_Stage			
<input checked="" type="checkbox"/>	Age_Group	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	Case_Status	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	Date_Report	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Exposure	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	Gender	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	Health_Region	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	ObjectID	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Province	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	Row_ID	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Recommend inputs for currently selected predictable:

- Kết quả khi chọn giá trị Exposure là Outbreak ta được cây quyết định dự đoán tử

vong:



– Nhận xét:

- Dựa vào mô hình, ở Ontario, độ tuổi AgeGroup = "Not Reported" có tỉ lệ nhiễm do bùng phát cao nhất và phân đều ở cả các giới tính.
- Tiếp đến là độ tuổi > 80 cũng có tỉ lệ nhiễm do bùng phát cao, với tỉ lệ giới tính không phải Nam dễ bị nhiễm nhất.

7 Data Visualization

- Dùng region map để biểu diễn trực quan (bằng màu sắc) số lượng ca nhiễm ở các vùng trong năm 2019
- Dữ liệu:
 - Lấy từ bảng Case Report để truy xuất dữ liệu PHU_LATITUDE và PHU_LONGITUDE của các PHU_CITY
 - Và lấy kết quả truy xuất số ca nhiễm trong từng PHU qua từng năm.

CITY	VALUE	LATITUDE	LONGITUDE
Barrie	3971	44.41071258	-79.68630597
Belleville	306	44.18667362	-77.39144554
Brantford	1130	43.151811	-80.27437415
Brockville	709	44.61584261	-75.70283308
Chatham	826	42.403861	-82.208561
Cornwall	1663	45.02915233	-74.73629779
Hamilton	6762	43.2576311	-79.87134089
Kenora	199	49.76961482	-94.48825435
Kingston	562	44.2278735	-76.5252108
London	3945	42.98146842	-81.25401572
Mississauga	42680	43.6474713	-79.7088933
New Liskeard	79	47.5092835	-79.681632
Newmarket	19290	44.048023	-79.480239
North Bay	153	46.31320706	-79.4678405
Oakville	6279	43.41399692	-79.74479581
Ottawa	10625	45.3456651	-75.7639122
Owen Sound	508	44.57619612	-80.94097993
Pembroke	267	45.799406	-77.118727
Peterborough	400	44.30163229	-78.32134748
Point Edward	1187	42.98641646	-82.40480836
Port Hope	605	43.96817279	-78.28579239
Sault Ste. Marie	88	46.5323728	-84.3148358
Simcoe	1012	42.84782526	-80.30381491
St. Thomas	1642	42.77780366	-81.15115646
Stratford	810	43.3686615	-81.00191283
Sudbury	292	46.46609195	-80.99805884
Thorold	4773	43.1165366	-79.2412197

- Sử dụng DataWrapper (<https://app.datawrapper.de/map/CWOt4/basemap>) để visualize dữ liệu region map



- Bước 1: Chọn map

This map is in My archive

1 Select your map 2 Add your data 3 Visualize 4 Publish & Embed

Choropleth map [Proceed](#)

What type of map do you want to create?

[or Upload Map](#)

- ☐ Canada » Alberta » Census Subdivisions (2022)
- ☐ Canada » Alberta » Census Subdivisions
- ☐ Canada » Atlantic » Provincial Electoral Districts
- ☐ Canada » Atlantic » Provinces
- ☐ Canada » British Columbia » Census Subdivisions (2022)
- ☐ Canada » British Columbia » Census Subdivisions
- ☐ Canada » Alberta » Provincial Electoral Districts
- ☐ Canada » British Columbia » Health Authority Boundaries
- ☐ Canada » British Columbia » Health Districts
- ☐ Canada » Calgary » Wards
- ☐ Canada » Calgary » Neighbourhoods

- Bước 2: Import file data đã chuẩn bị

1 Select your map 2 Add your data 3 Visualize 4 Publish & Embed

Now your map needs data!

We prefilled the table with Name keys.
You can start **adding your values** or **upload your own file**.

[Upload](#) [Match](#) [Check](#)

Upload a file (CSV or Excel)
You can also simply drop it here [Upload file](#)

OR

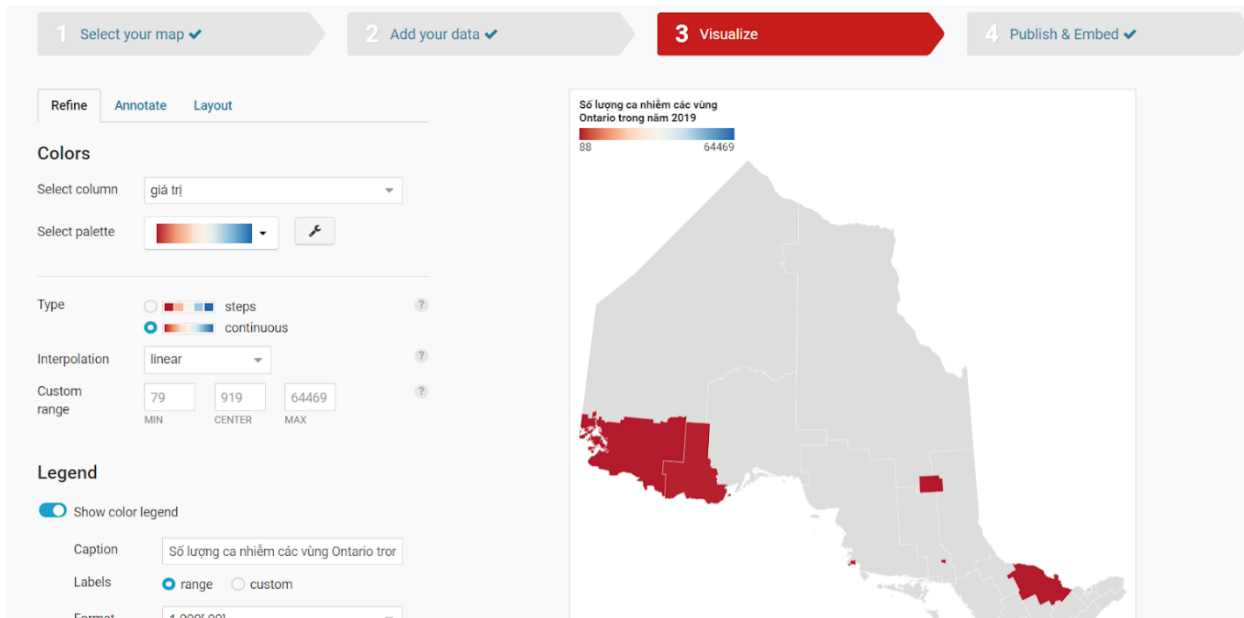
Copy & paste your data (including header row/column):

☒ First row as label

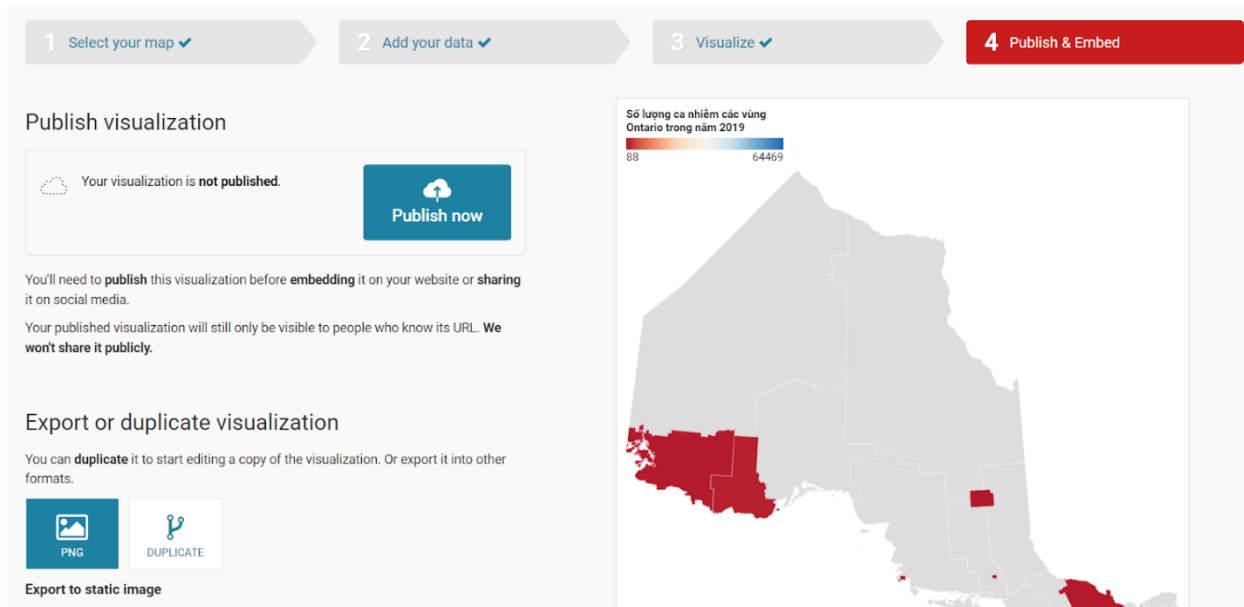
	A	Name	B	Values	C	D
1	city	south		giá trị	Latitude	Longitude
21		Parkdale High Park		1,187	42.986416	-82.484888
22		Oxford		685	43.968173	-78.285792
23		Sault Ste. Marie		88	46.532373	-84.314836
24		Simcoe North		1,012	42.847825	-80.383815
25		St. Catharines		1,642	42.777884	-81.151156
26		Scarborough Centre		810	43.368662	-81.081913
27		Sudbury		292	46.466092	-80.998059
28		Thornhill		4,773	43.116537	-79.24122
29		Thunder Bay Atikokan		613	48.480572	-89.25885
30		Timmins		137	48.47251	-81.32875
31		Toronto Centre		64,469	43.656591	-79.379358
32		Waterloo		6,714	43.462876	-80.520913
33		Windsor West		9,157	42.388797	-83.03367
		Ajax		-	-	-
		Algoma Manitoulin		-	-	-
		Aurora Oak Ridges Richmond Hill		-	-	-

- Bước 3: Thực hiện visualize

Chọn màu giá trị từ thấp đến cao

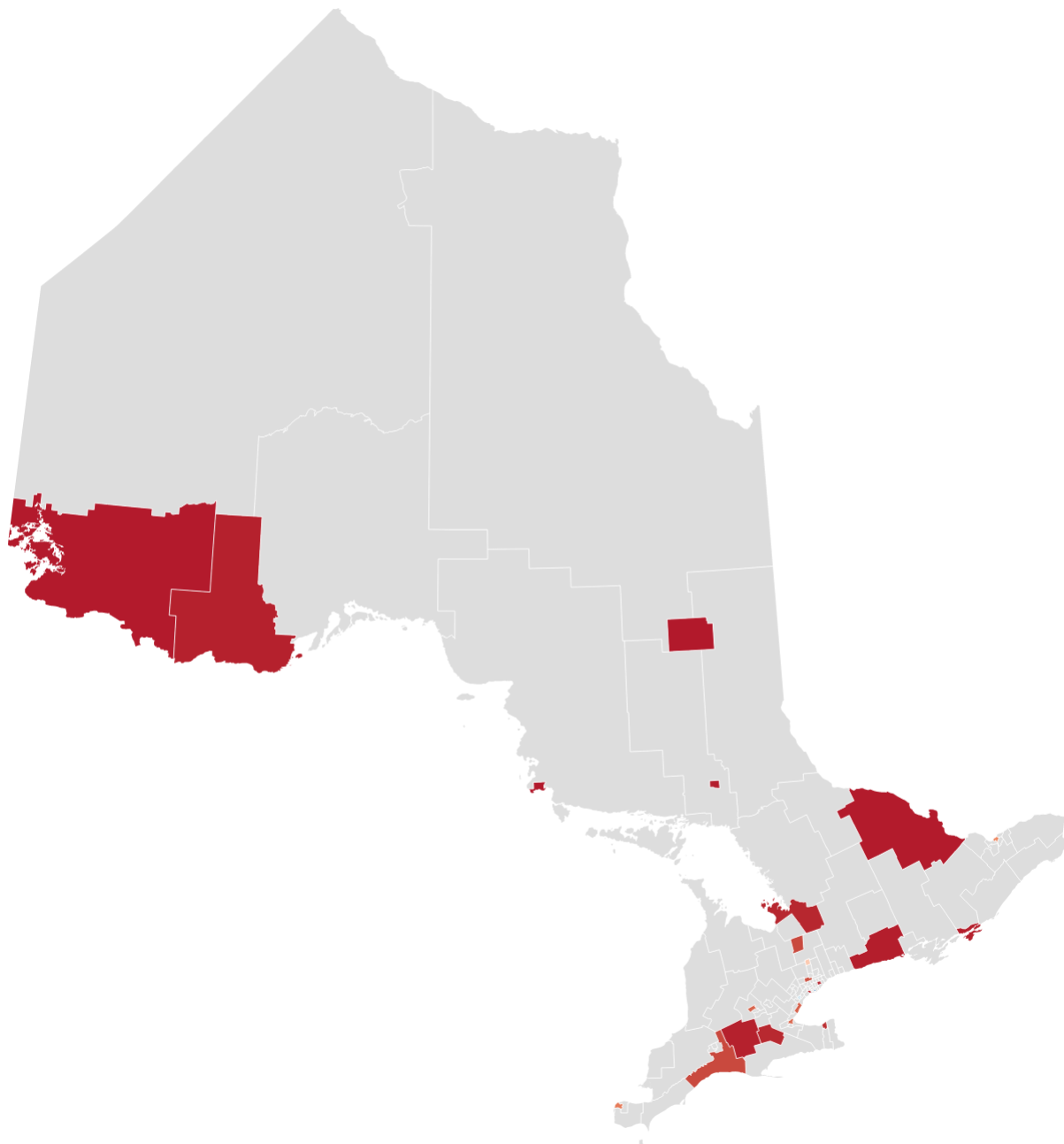


- Bước 4: Tải region map về
Xuất file với định dạng .png



– KẾT QUẢ

Số lượng ca nhiễm các vùng
Ontario trong năm 2019



Created with Datawrapper