CITS5504 Data Warehousing Project 1 Report

Chang Su (22993116)

Introduction

The project is to make a data warehouse based on relevant COVID data from around the world. After the data is filtered, the basic mining, classification and visual presentation are performed. The data tables used in the project are from three statistical agencies, and the basic data standards are different due to the statistical methods, cultural background and other reasons. All forms need to be processed in order to complete the project.

According to the project's business queries, the important data to be used are country name and quantity, the daily number of new confirmed, deaths and rehabilitation, region, date, the country size, population and life expectancy.

Section A Create StarNet and Dimensions

Based on the project requirements, determine the creation of four dimensions: location, time, population, and life span. The Location dimension contains Region and Country. The Time dimension contains the date, month, quarter, and year. Population dimension includes country size. The Life span dimension includes Life expectancy. From this, create the basic StarNet as Figure 1.

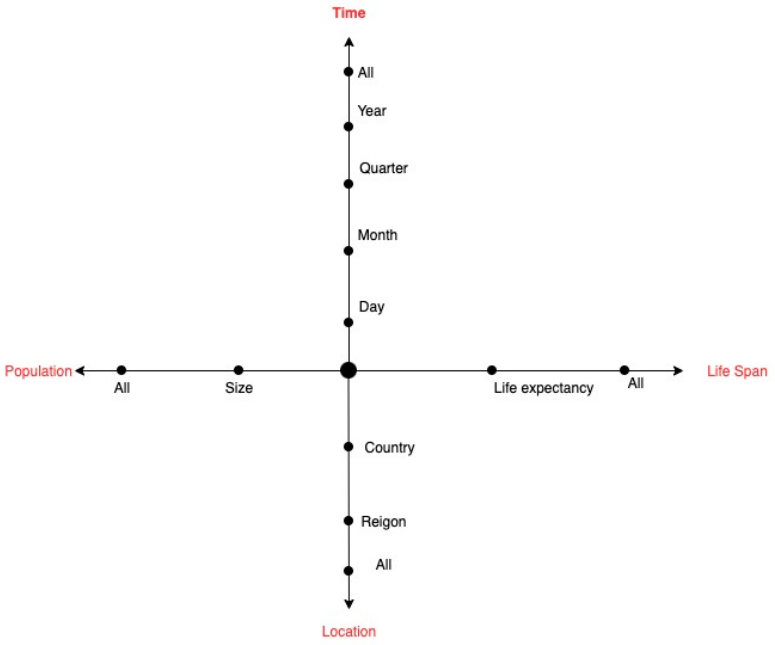


Figure 1

Section B ETL

According to the data provided by the project, TIME_SERIES_COVID19 (hereinafter referred to as time_series) confirmed, recovered, and deaths data are from the same institution. The statistical items, country names, and country number are identical. owid-covid-data (hereinafter referred to as owid) and acaps-covid-19-government-measures (hereinafter referred to as acaps) are not unified in country names and

quantities due to different data sources. The three tables all have the same data statistics problem, which is based on regional statistics and national statistics. For example, the data of each state or province of Australia and China are listed separately in time_series confirmed table. This issue should be merged or deleted the problem-related rows according to the current international rules. Besides, because all the tables contained a large amount of other data that could be deemed invalid based on project requirements, columns containing invalid data in Excel were removed.
The data process was completed using MS Excel.

Step 1 Country name and quantity statistics

In the TIME_SERIES_CONFIRMATION form, the data of Australia's 7 states and territories are counted separately as shown in Figure 2, and all rows need to be summed up using the Excel summation formula $= SUM\ (A\$: D\$)$ to finally merge into the total data of Australia which represented in Figure 3. Similarly, the death table and the recovery table perform similar operations.

	Australia	0	0	0	0	4	5
Australian Capital Territory	Australia	0	0	0	0	0	0
New South Wales	Australia	0	0	0	0	3	4
Northern Territory	Australia	0	0	0	0	0	0
Queensland	Australia	0	0	0	0	0	0
South Australia	Australia	0	0	0	0	0	0
Tasmania	Australia	0	0	0	0	0	0
Victoria	Australia	0	0	0	0	1	1
Western Australia	Australia	0	0	0	0	0	0

Figure 2

9	Armenia	0	0	0	0	0
10	Australia	0	0	0	0	0
11	Austria	0	0	0	0	0

Figure 3

Data consolidation in the following countries:

- Australia
- Canada
- China
- French
- Denmark
- Netherlands

China is a representative country in the data sheet, with more than 30 provinces in China counted separately, all performing the same operations as Australia. Moreover, according to common sense in international diplomacy, Hong Kong and Taiwan are part of China, so Hong Kong and Taiwan's data are merged with China's and are not

counted separately.There are two kinds statistical results for France in time_series, which are calculated by region and calculated by country. Since the three tables are from the same statistical institution, the differences may be caused by the geographical location, so there is no repeat statistics by default. Lines 120 to 131 of the original data in Excel are unified and merged into France shows in Figure 4.

120	French Guia	France	3.9339	-53.1258	0
121	French Poly	France	-17.6797	149.4068	0
122	Guadeloupe	France	16.265	-61.551	0
123	Martinique	France	14.6415	-61.0242	0
124	Mayotte	France	-12.8275	45.166244	0
125	New Caledo	France	-20.904305	165.618042	0
126	Reunion	France	-21.1151	55.5364	0
127	Saint Barthe	France	17.9	-62.8333	0
128	Saint Pierre	France	46.8852	-56.3159	0
129	St Martin	France	18.0708	-63.0501	0
130	Wallis and F	France	-14.2938	-178.1165	0
131		France	46.2276	2.2137	0

Figure 4

Greenland and Faroe Islands belong to Denmark's overseas territories, and territorial ownership and securities jurisdiction still belong to Denmark. Therefore, the data of these two Islands and the local data of Denmark are combined for calculation.

Similarly, overseas territory data for the Netherlands and the United Kingdom are combined with local data.

The data of confirmed，deaths and recovered people on Diamond Princess and Ms. Zaandam Cruises from different countries and regions, have not been identified by nationality, and data from Diamond Princess and Ms. Zaandam have been deleted to ensure consistency and validity of data.

Step 2 Delete duplicate country

Because the data source and statistical format are not statistical, OWID tables and ACASP tables have a large number of duplicate country names. Preserve unique values for all country names that appear multiple times using the Remove Duplicates function in Excel. Delete duplicate country name Sort all data by country name alphabetical order. Do the same for all tables of data.

Step 3 Unify country name

After that, create a new Excel sheet and paste the countries' names in each of the three sheets into the new Excel sheet. In some countries where the names appear only once or in different spelling languages, use the Excel function to get unique values. The three columns are compared using the ‘if nested judgment’. Tables with differences are automatically highlighted in Figure 5. Changes are made to the names of countries in

all tables following the current international practice, resulting in a consistent name and alphabetical order of country names for all tables, the result is shown in Figure 6.

8	Burma	Burundi	Burundi
9	Burundi	Cambodia	CÃ´te d'Ivoire
0	Cabo Verde	Cameroon	Cambodia
1	Cambodia	Canada	Cameroon
2	Cameroon	Cape Verde	Canada
3	Canada	Central African Republic	Cape Verde
4	Central African Republic	Chad	CAR
5	Chad	Chile	Chad
6	Chile	China	Chile
7	China	Colombia	China
8	Colombia	Comoros	Colombia
9	Comoros	Congo	Comoros
0	Congo (Brazzaville)	Costa Rica	Congo
1	Congo (Kinshasa)	Cote d'Ivoire	Costa Rica
2	Costa Rica	Croatia	Croatia
3	Cote d'Ivoire	Cuba	Cuba
4	Croatia	Cyprus	Cyprus
5	Cuba	Czechia	Czech Republic
6	Cyprus	Democratic Republic of Congo	Denmark
7	Czechia	Denmark	Djibouti
8	Denmark	Djibouti	Dominica
9	Diamond Princess	Dominica	Dominican Republic
0	Djibouti	Dominican Republic	DPRK
1	Dominica	Ecuador	DRC

Figure 5

K26 fx

	A	B	C	D	E
1	Afghanistan	Afghanistan	Afghanistan		
2	Albania	Albania	Afghanistan		
3	Algeria	Algeria	Albania		
4	Andorra	Andorra	Algeria		
5	Angola	Angola	Angola		
6	Antigua and Barbuda	Antigua and Barbuda	Antigua and Barbuda		
7	Argentina	Argentina	Argentina		
8	Armenia	Armenia	Armenia		
9	Australia	Australia	Australia		
10	Austria	Austria	Austria		
11	Azerbaijan	Azerbaijan	Azerbaijan		
12	Bahamas	Bahamas	Bahamas		
13	Bahrain	Bahrain	Bahrain		
14	Bangladesh	Bangladesh	Bangladesh		
15	Barbados	Barbados	Barbados		
16	Belarus	Belarus	Belarus		
17	Belgium	Belgium	Belgium		
18	Belize	Belize	Belize		
19	Benin	Benin	Benin		
20	Bhutan	Bhutan	Bhutan		
21	Bolivia	Bolivia	Bolivia		
22	Bosnia and Herzegovina	Bosnia and Herzegovina	Bosnia and Herzegovina		
23	Botswana	Botswana	Botswana		
24	Brazil	Brazil	Brazil		
25	Brunei	Brunei	Brunei		
26	Bulgaria	Bulgaria	Bulgaria		
27	Burkina Faso	Burkina Faso	Burkina Faso		
28	Burma	Burundi	Burundi		
29	Burundi	Cambodia	CÃ´te d'Ivoire		
30	Cabo Verde	Cameroon	Cambodia		
31	Cambodia	Canada	Cameroon		
32	Cameroon	Cape Verde	Canada		
33	Canada	Central African Republic	Cape Verde		
34	Central African Republic	Chad	CAR		
35	Chad	Chile	Chad		
36	Chile	China	Chile		
37	China	Colombia	China		
38	Colombia	Comoros	Colombia		
39	Comoros	Congo	Comoros		
40	Congo (Brazzaville)	Costa Rica	Congo		
41	Congo (Kinshasa)	Cote d'Ivoire	Costa Rica		
42	Costa Rica	Croatia	Croatia		
43	Cote d'Ivoire	Cuba	Cuba		
44	Croatia	Cyprus	Cyprus		

Figure 6

The names of the following countries have been changed：

- Burma——Myanmar
- Timor Leste——Timor-Leste
- Cabo Verde——Cape Verde
- Czech Republic——Czechia
- St. Kitts and Nevis——Saint Kitts and Nevis
- St. Lucia——Saint Lucia
- St. Vincent and the Grenadines——Saint Vincent and the Grenadines
- Timor Leste——Timor-Leste
- CÃ´te d'Ivoire——Cote d'Ivoire

It is important to note that when the country name is changed, the original document needs to be rearranged alphabetically to ensure that the countries are listed in the same order in all tables. Write the input function $=IF(A1=B1,"same","not")$ in the last column to determine if the country names and order are consistent across all the final tables.

After removing duplications and unifying the names of countries, the total number is182.

Step 4 Calculate the daily increment

The new confirmed, dead and recovered cases provided in Time_series are cumulative new cases. To meet the query requirements of projects, the accumulated increase data needs to be converted into daily increase data. By subtracting all the data in the column after the date in Excel from the data before the date to get a new day, a total of 405 columns. After the Excel function completes the calculation, all the data need to be copied in the original position and only pasted the value, so as to avoid all the data changes when a cell in the Excel is changed. This step is required for all Excel operations during the data cleansing process and will not be repeated in the report.

Because of the need to query the national population size and life expectancy, according to the size of population and life expectancy, using the if function in owid table at new columns automatically fill in the corresponding provisions of national. If population is less than two million fill in 1, between two million and forty million to 2, more than forty million to 3, fill out 2 life is greater than or equal to 75 years old, and less than 1, 75 years old.

America stopped updating the recovered number since 14th Dec 2020, so the data on the day need to be changed to zero instead of the original large negative number.

Step 5 Create Fact Table and Dimension Table in Excel

According to the previously designed Starnet, four dimensions Excel need to be created, namely Location, Time, LifeSpan, and Population dimension. DimLocation Excel contains locationID, Region, Country, Region, and Country correspondence from ACASP table. DimTime contains the Time ID, Day, Month, Quarter, and Year. The dates provided in the time_series table are 405 days. And according to common sense, all the dates are divided into 6 quarters, which contain three months in each quarter.

LifeSpan includes LifeSpanID and Life Expectancy. Population Dimension contains PopulationID and Size.

According to the project query requirements analysis, facttable should contain locationID TimeID, LifeSpanID, PopulationID, and measures. Measures formed, deaths, recovered.

Based on the information in the time_sereis table, 405 days of data is required. There are 405 rows for each country in facttable. The LocationID for each country needs to be filled in 405 times, so the formula $= INT\ (ROW\ (A405)/405)$ is automatically filled in in the LocationID column. Similarly, TImeID needs to be repeated 182 times from 1-405 because 182 countries are automatically populated using Excel's built-in Formula $= Mod\ (ROW\ (B1) - 1)\ , 405)\ +\ 1$.

Filled data of LifeSpanID, PopulationID, Conformed, Deformed, and dissolved are respectively derived from three forms, OWID and TIME_Series. Another built-in function of Excel, VLOOKUP function, is needed to query the required data across the table and fill in the corresponding cells of factable Excel.

$$= VLOOKUP(A1,'target\ excel\ path\ /[target.xlsx]target\ shell\ name'!\$A{:}\$OQ, C1 + 2,0)$$

The total number of Factt able rows created is 73710.

Step 6 Identify the surrogate key

After all data collation is completed, start from the first line of data in each Excel table and add surrogate keys to each line.

Step 7 Export the CSV file

After completing all the procedures, export Excel as a CSV file and wait for the import into the database.

Section C Create database in SEMM

Step 1 Create database and tables

- Create the database CITS5504_Project1

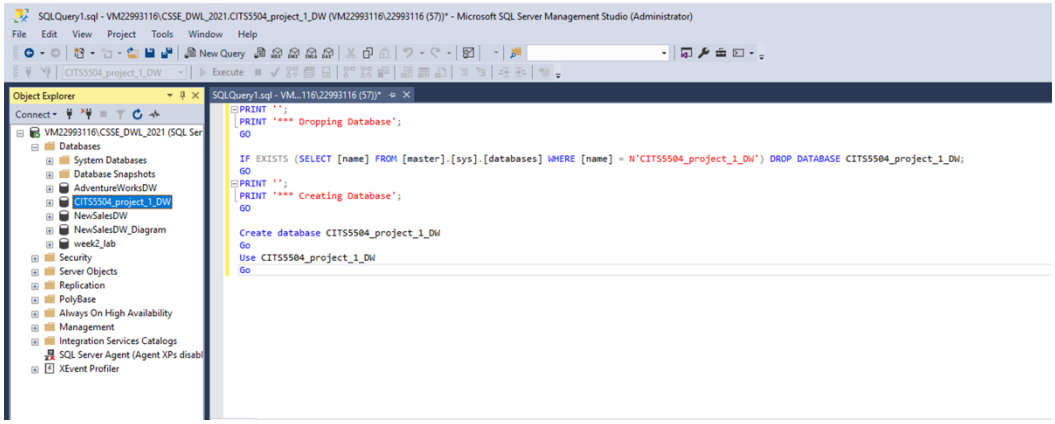


Figure 7

- Create DimLifespan table

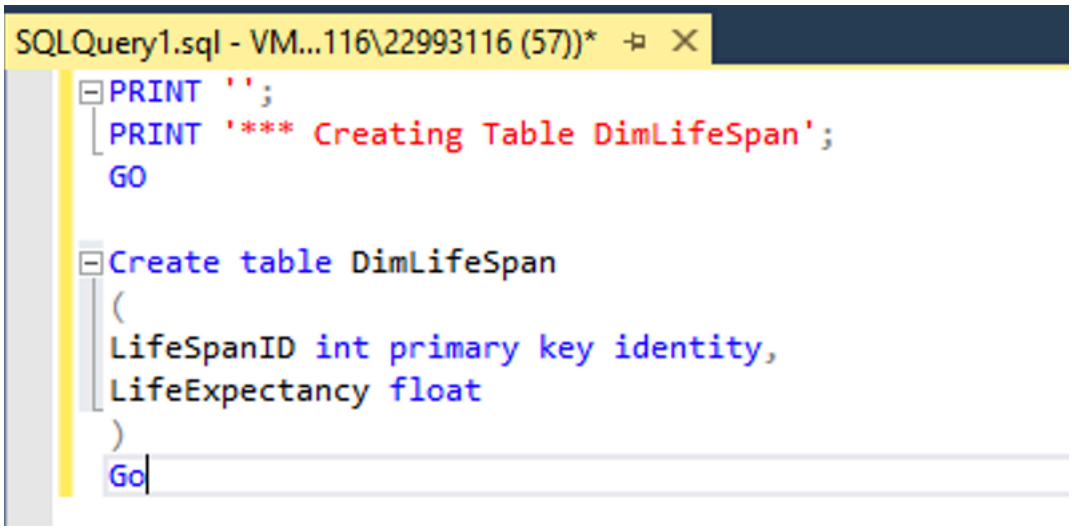


Figure 8

Similarly, create Dimlocation, DimPopulation, DimTime and FactCOVID tables.

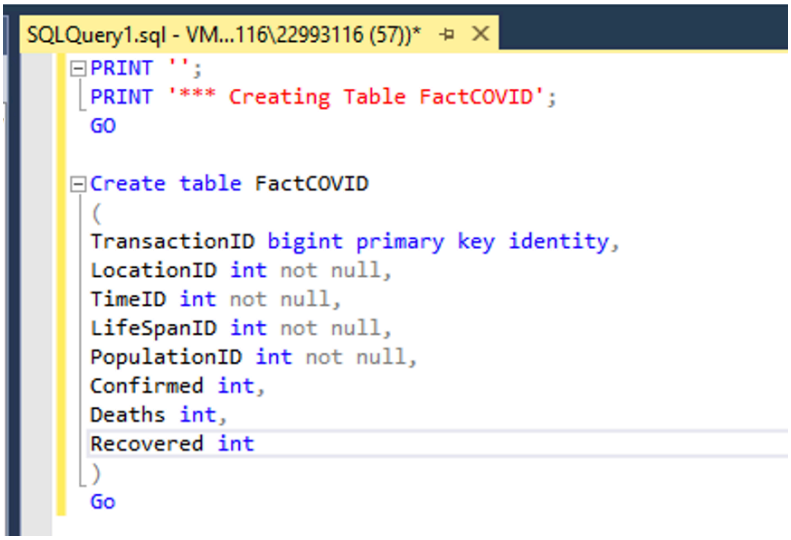


Figure 9

- Create foreign key relationships

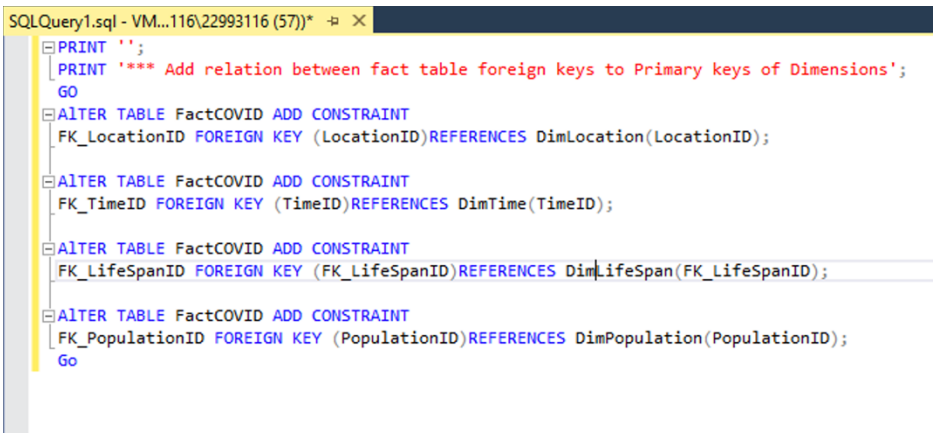


Figure 10

Step 2 Create diagram

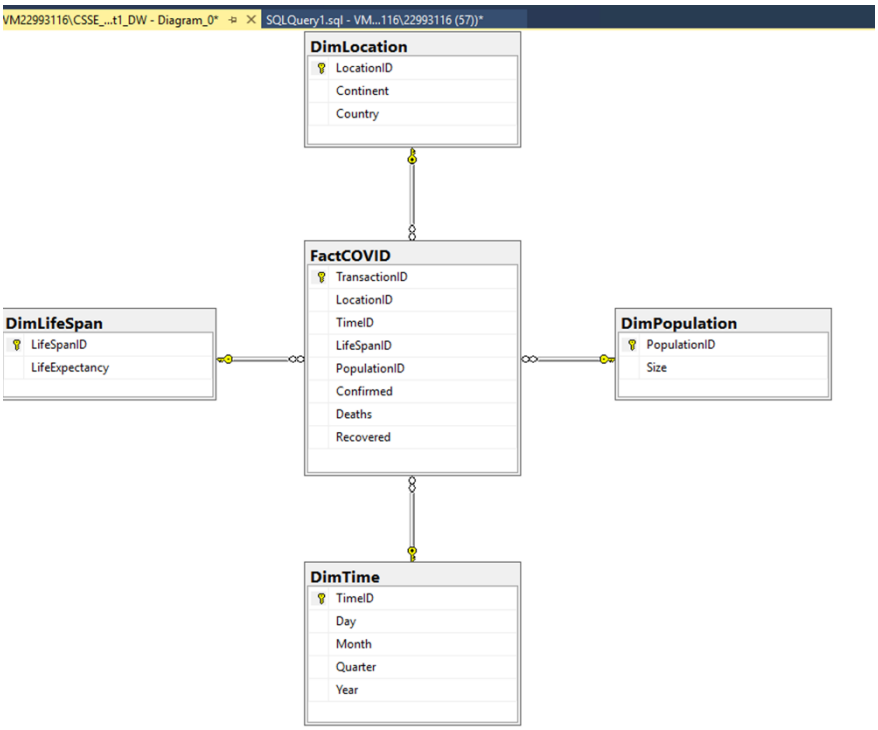


Figure 11

Step 3 Import data

- Import data to LifeSpan table.

```
:setvar SqlSamplesSourceDataPath "C:\Users\22993116\Desktop\Project\dimension table csv\"
USE CITS5504_Project1_DW
BULK INSERT [dbo].[DimLifeSpan] FROM '$(SqlSamplesSourceDataPath)DimLifeSpan.csv'
WITH (
    CHECK_CONSTRAINTS,
    --CODEPAGE='ACP',
    DATAFILETYPE='char',
    FIELDTERMINATOR=',',
    ROWTERMINATOR='\n',
    KEEPIDENTITY,
    TABLOCK
);
```

100 %

Messages

```
(2 rows affected)

Completion time: 2021-03-31T13:57:23.3501901+00:00
```

Figure 12

- Similarly, import data to FactCOVID and other tables.

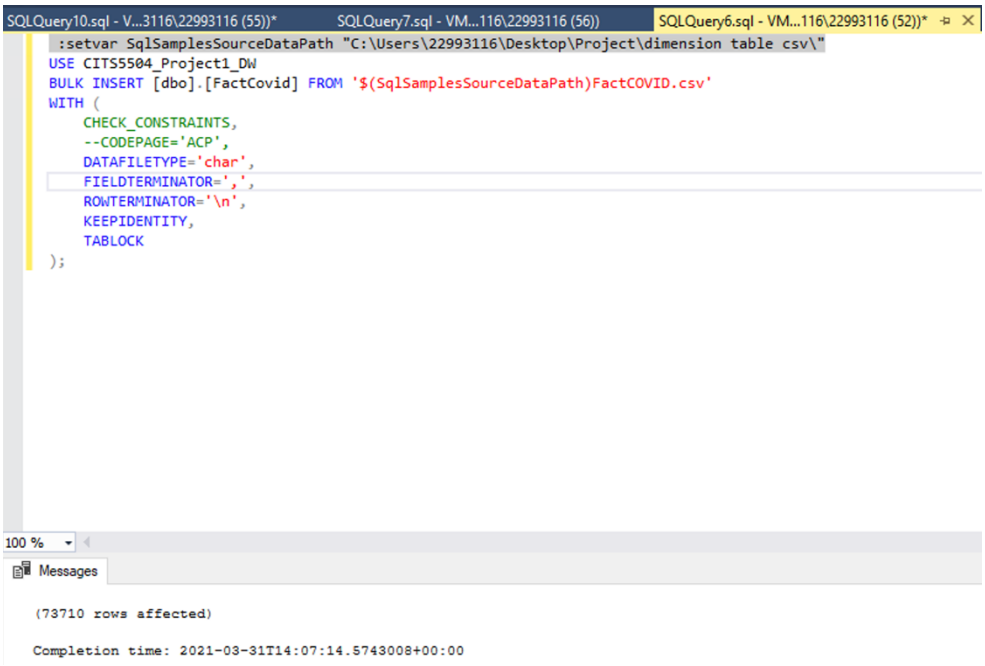
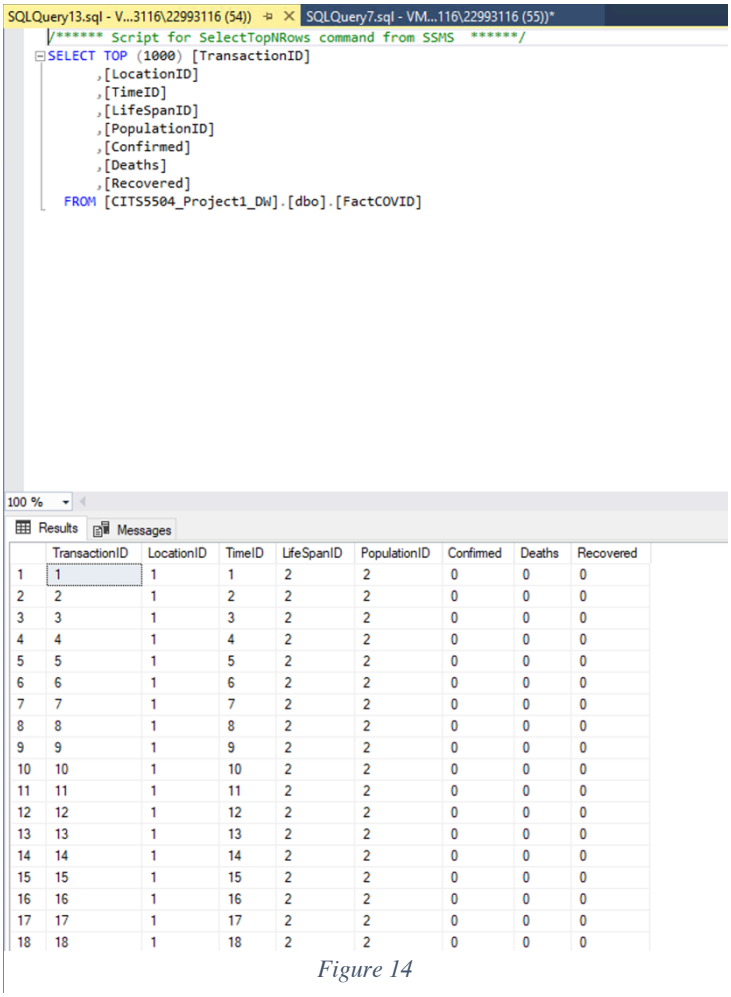


Figure 13

- Checking database contents. For example view first 1000 rows data in FactCOVID.



	TransactionID	LocationID	TimeID	LifeSpanID	PopulationID	Confirmed	Deaths	Recovered
1	1	1	1	2	2	0	0	0
2	2	1	2	2	2	0	0	0
3	3	1	3	2	2	0	0	0
4	4	1	4	2	2	0	0	0
5	5	1	5	2	2	0	0	0
6	6	1	6	2	2	0	0	0
7	7	1	7	2	2	0	0	0
8	8	1	8	2	2	0	0	0
9	9	1	9	2	2	0	0	0
10	10	1	10	2	2	0	0	0
11	11	1	11	2	2	0	0	0
12	12	1	12	2	2	0	0	0
13	13	1	13	2	2	0	0	0
14	14	1	14	2	2	0	0	0
15	15	1	15	2	2	0	0	0
16	16	1	16	2	2	0	0	0
17	17	1	17	2	2	0	0	0
18	18	1	18	2	2	0	0	0

Figure 14

Step 4 Carried out the design in Microsoft Visual Studio

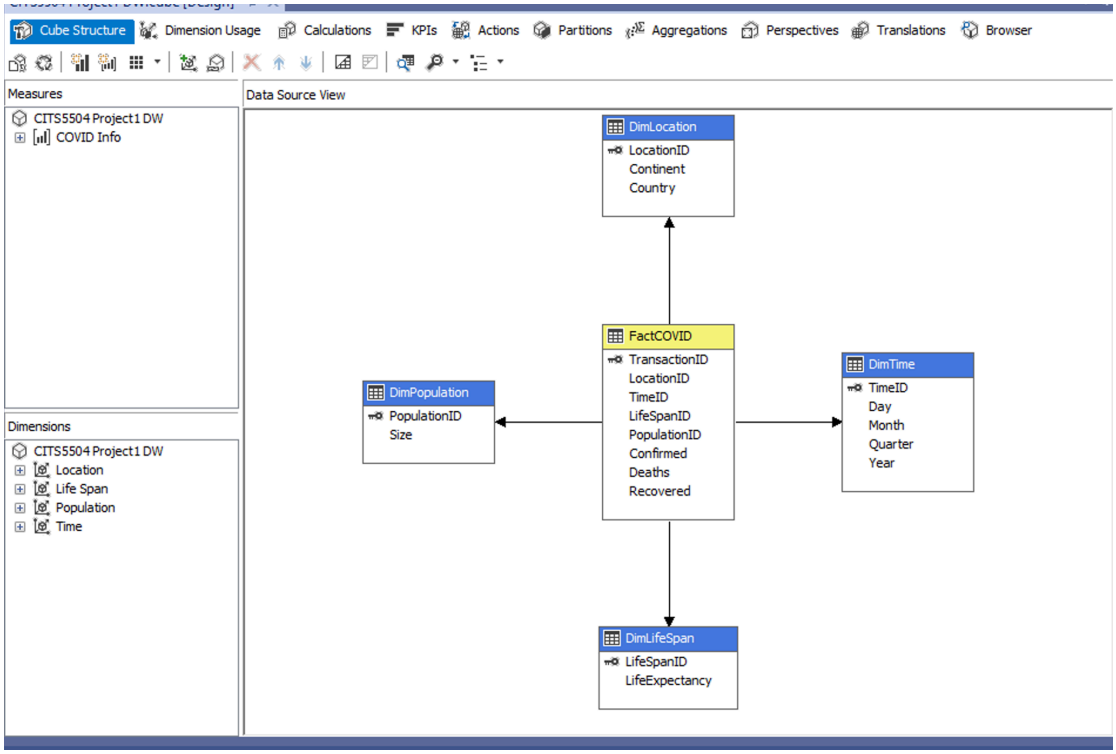


Figure 15

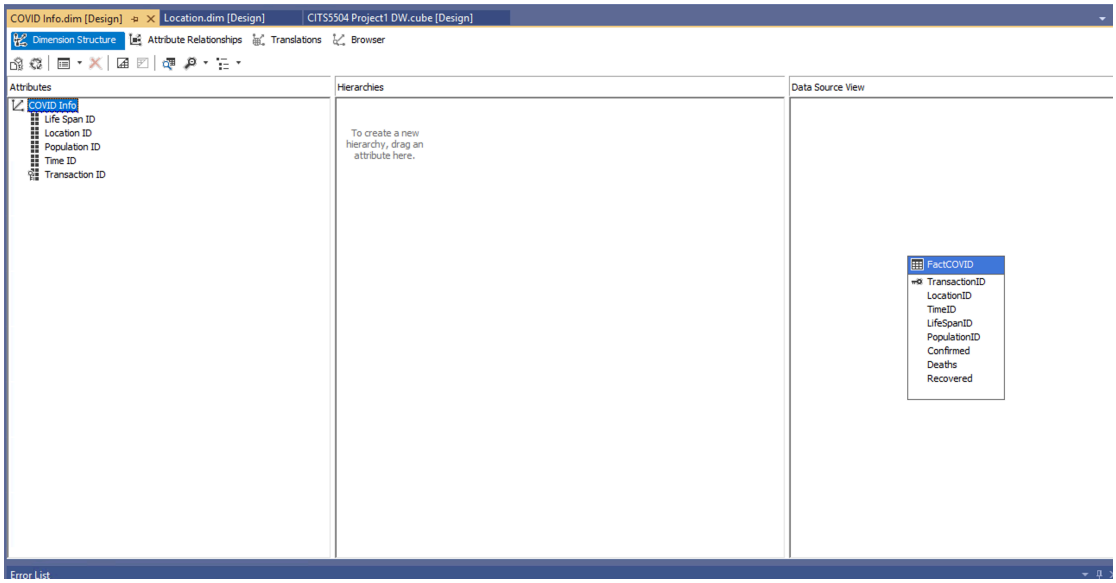


Figure 16

- There are 4 cubes created, LifeSpan, Location, Population and Time.
  There are four hierarchies in Time cube, which are all, year, quarter, month and day from large to small.
  LifeSpan and Population only have one hierarchy.

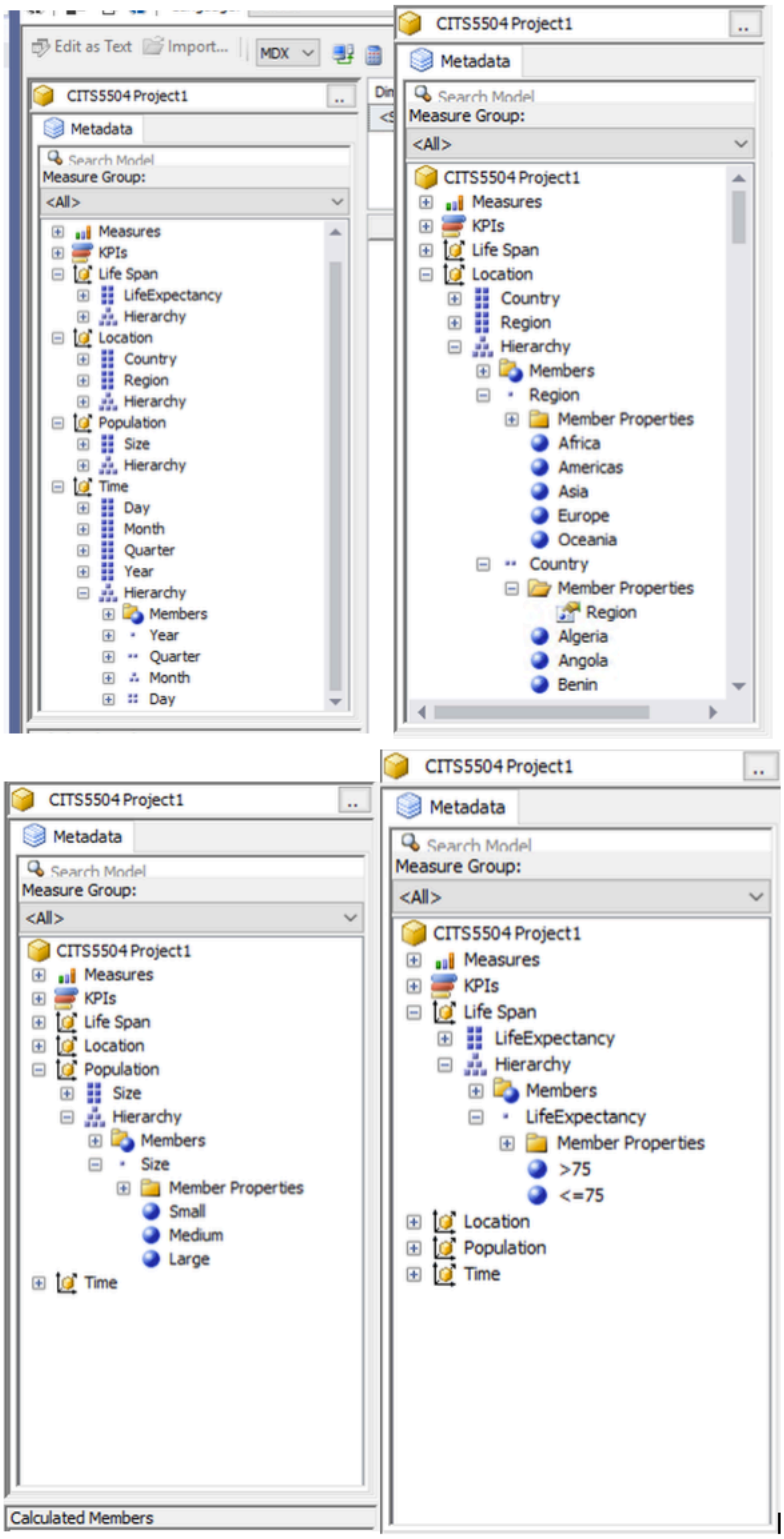


Figure 17

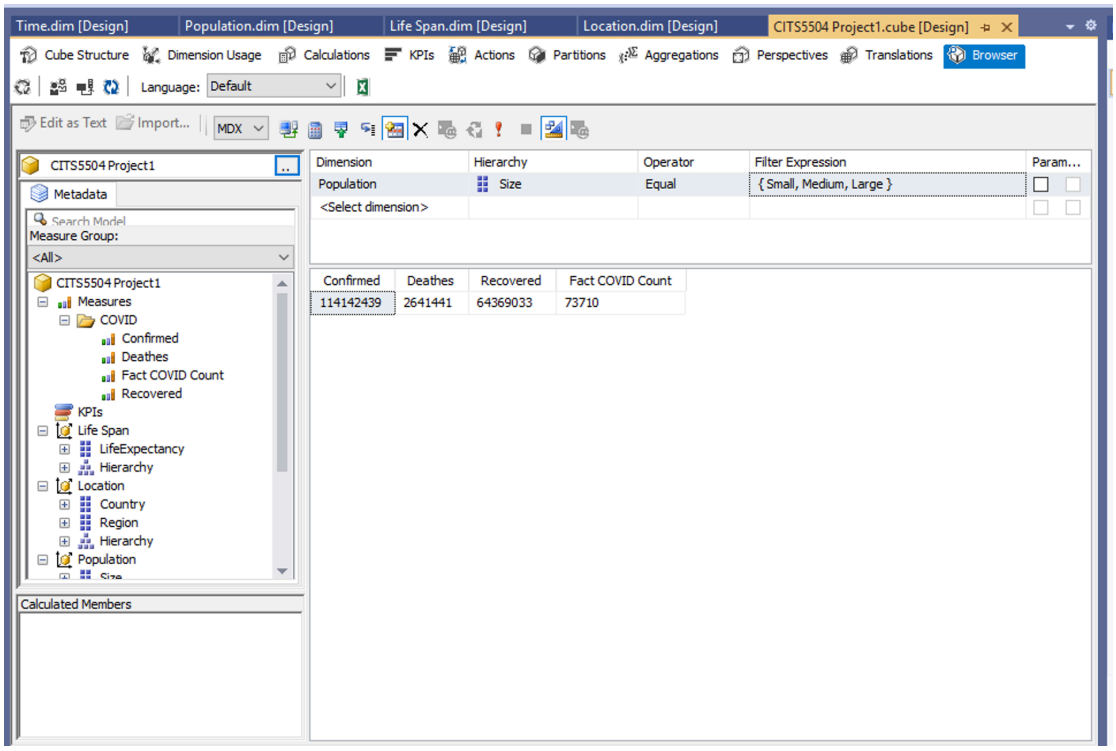


Figure 18

Section D 4 business queries

Q1.1 What is the total number of confirmed cases in Australia in 2020?

This query has involved in Time dimension year hierarchy and Location dimension country hierarchy. The time dimension is 2020, and the place dimension is Australia. Here is the Starnet footpath and visualize result, the total confirmed number of Australia in 2020 is 28425.

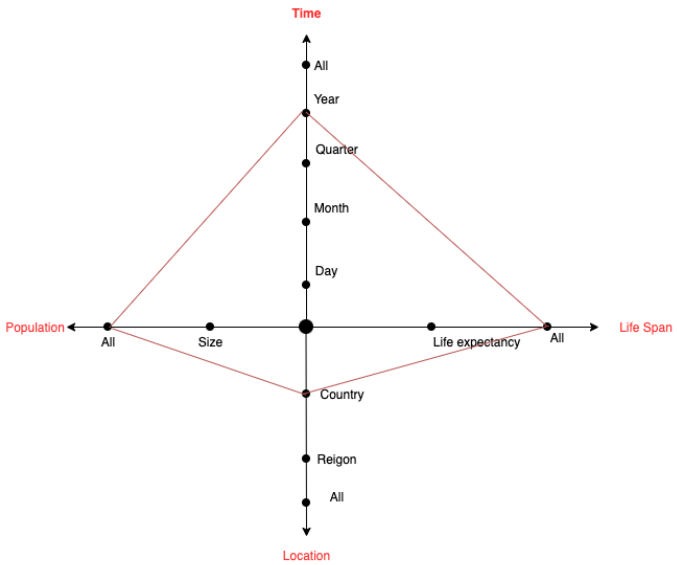


Figure 19

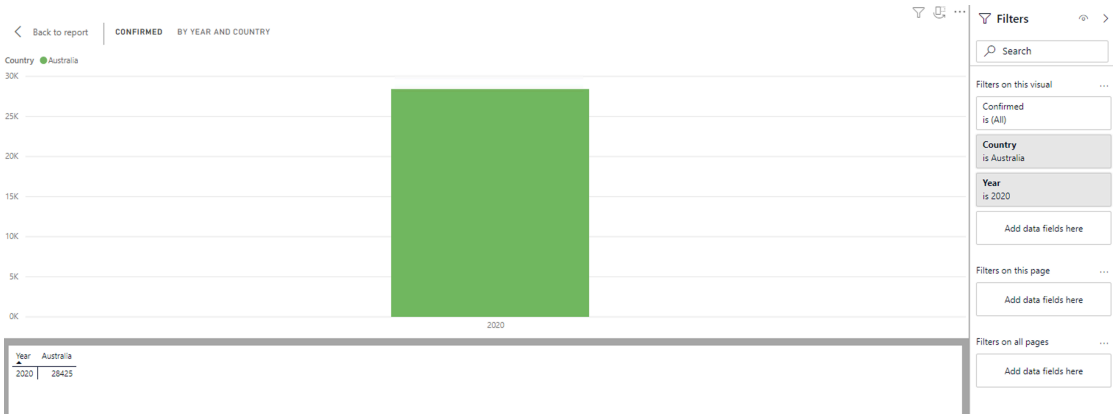


Figure 20

Q1.2 What is the number of confirmed cases in each quarter of 2020 in Australia? This query has involved in Time dimension quarter hierarchy and Location dimension country hierarchy. The time dimension are quarter one to four in 2020, and the location dimension is Australia. Here is the Starnet footpath and visualize result. The number of people confirmed in the first to the fourth quarter of 2020 in Australia is respectively are 4559, 3361, 19176 and 1329.

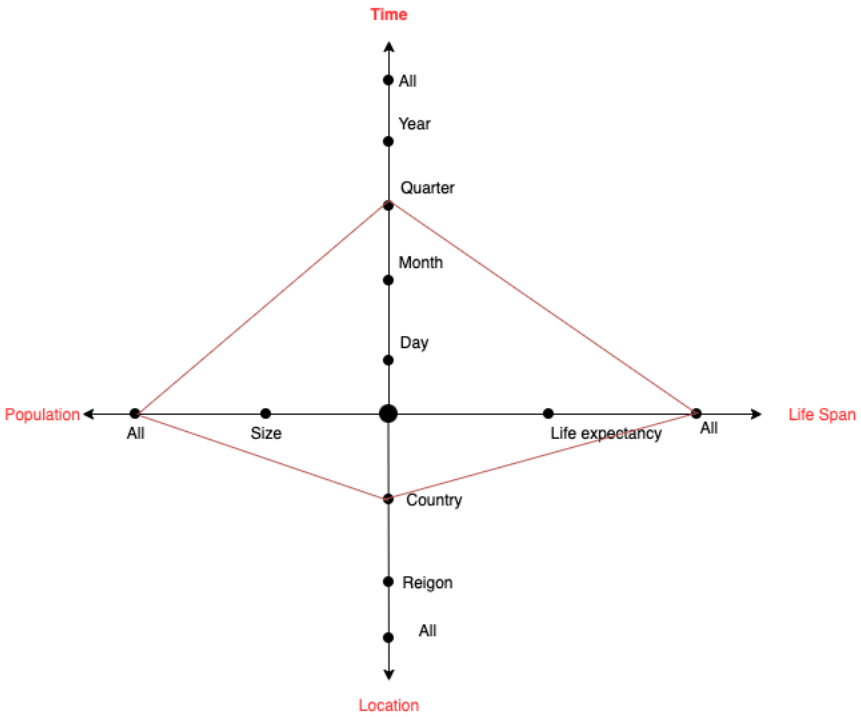


Figure 21

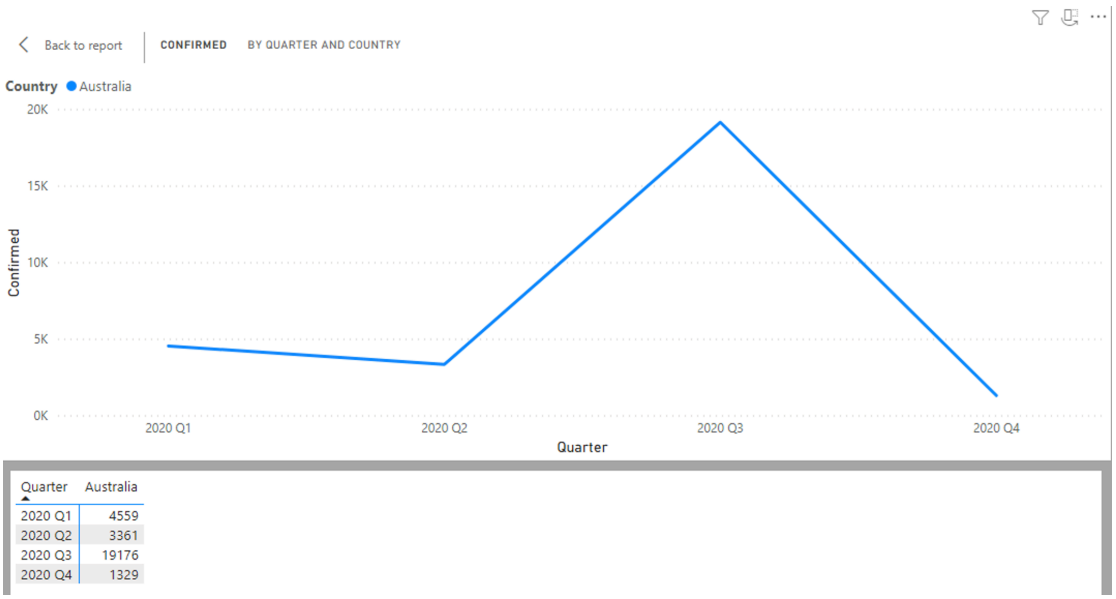


Figure 22

Q1.3 What is the number of confirmed cases in each month of 2020 in Australia? This query has involved in Time dimension month hierarchy and Location dimension country hierarchy. The time dimension is all month in 2020, and the location dimension is Australia. Here is the Starnet footpath and visualize result.

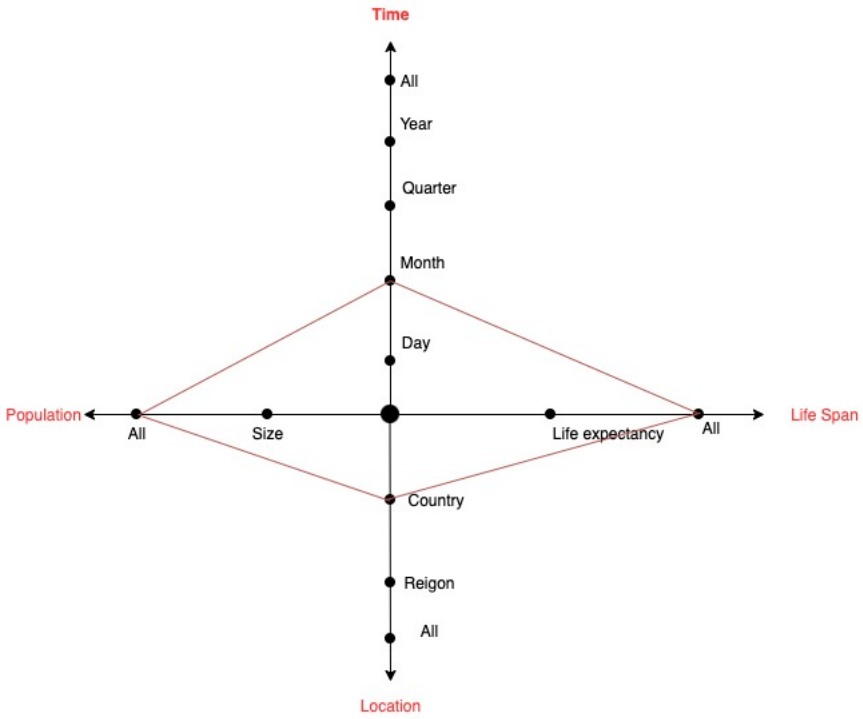


Figure 23

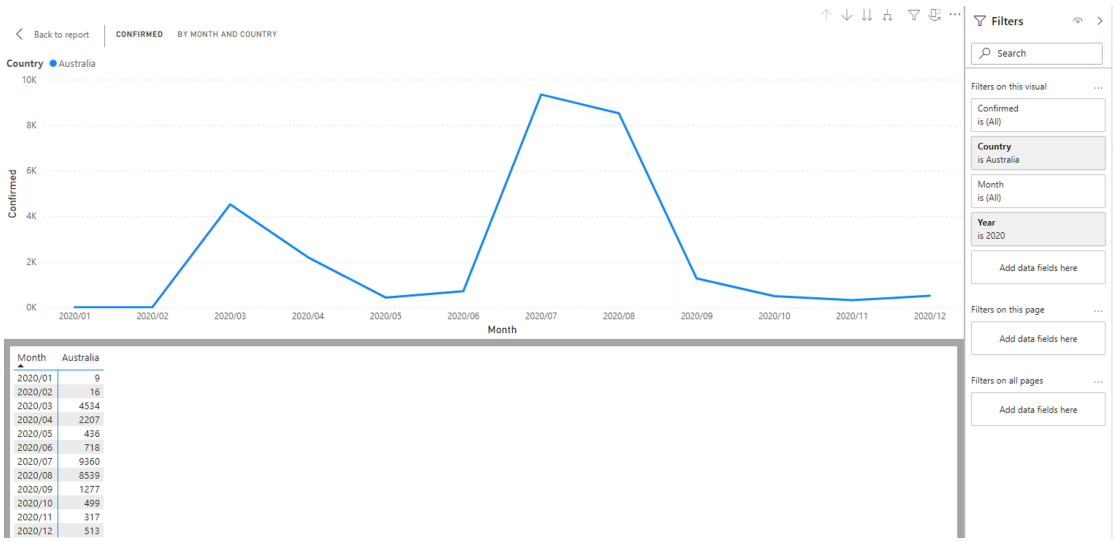


Figure 24

Q2.1 In Sept 2020, how many recovered cases are there in the region of the Americas?

This query has involved in Time dimension year hierarchy and Location dimension region hierarchy. The time dimension is 2020, and the location dimension is Americas region. Here is the Starnet footpath and visualize result. There are 20454370 recovered cases in Americans.

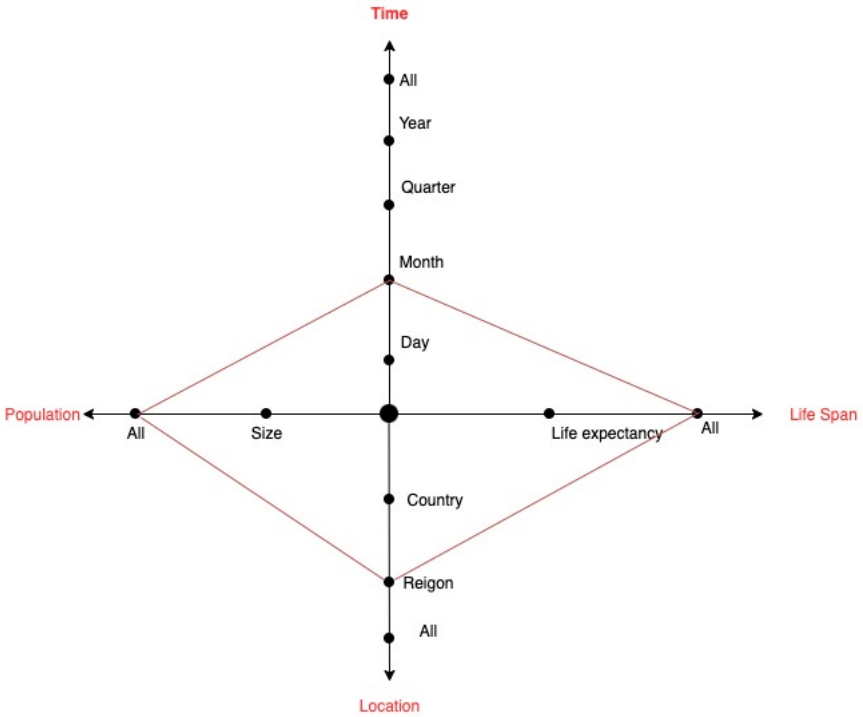


Figure 25

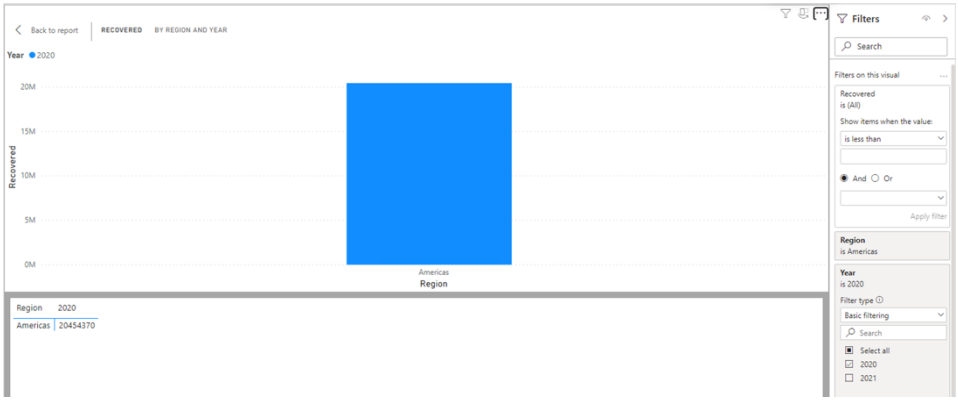


Figure 26

Q2.2 How many recovered cases in the United States, Canada and Mexico, respectively, in Sep 2020?

This query has involved in Time dimension year hierarchy and Location dimension country hierarchy. The time dimension is 2020, and the location dimension is United States, Canada and Mexico region. Here is the Starnet footpath and visualize result. There are 21161 cases in Canada, 131785 cases in Mexico and 655863 cases in United States.

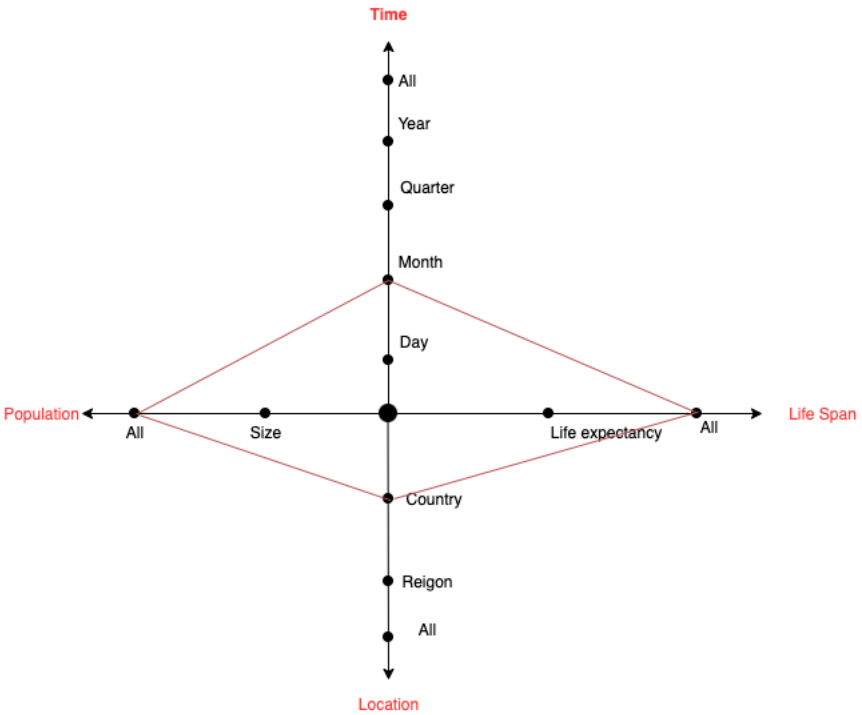


Figure 27

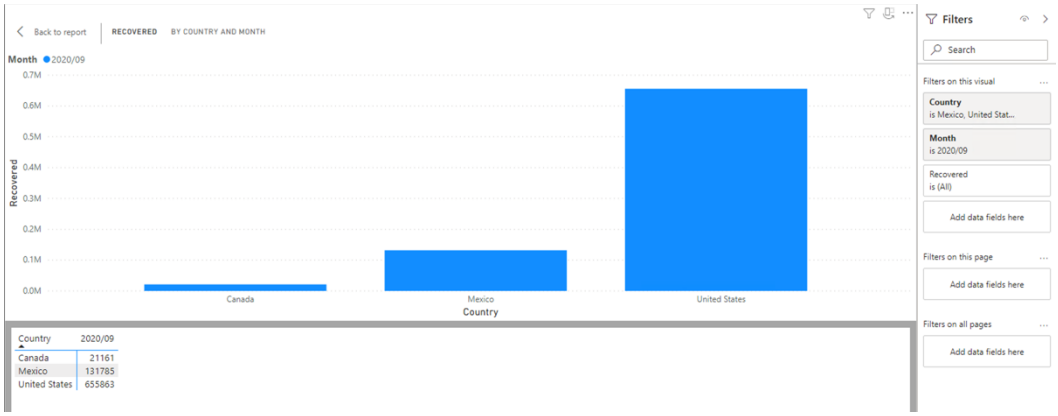


Figure 28

Q3.1 What is the total number of covid deaths worldwide in 2020?

This query has involved in Time dimension year hierarchy and Location dimension country hierarchy. The time dimension is 2020, and the location dimension is all countries. Here is the Starnet footpath and visualize result.

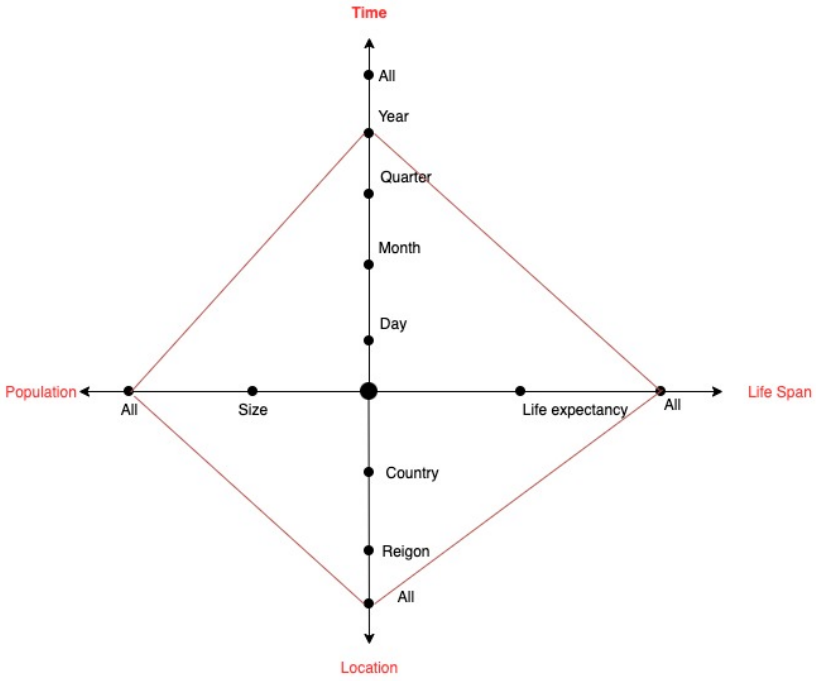


Figure 29

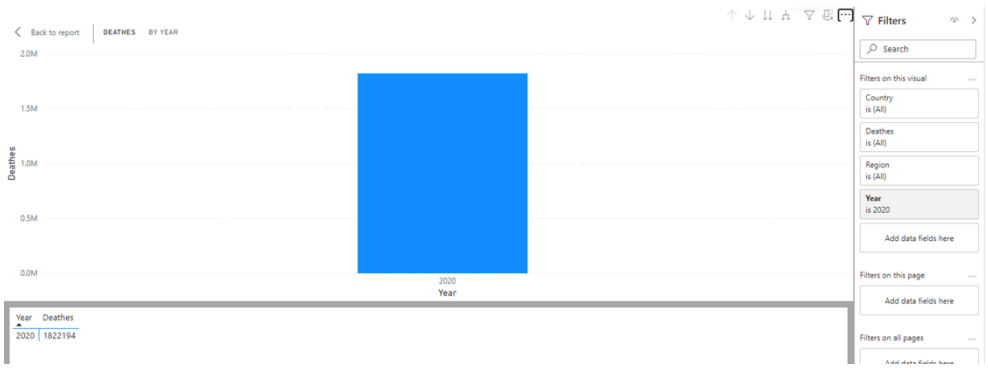


Figure 30

Q3.2 What is the total number of covid deaths in large countries, medium countries and small countries, respectively, in 2020?

This query has involved in Population dimension year hierarchy and Location dimension country hierarchy. The time dimension is 2020, and the population dimension is all size. Here is the Starnet footpath and visualize result. The the total number of covid deaths in large countries, medium countries and small countries, respectively, in 2020 are 1479837, 341048 and 4309.

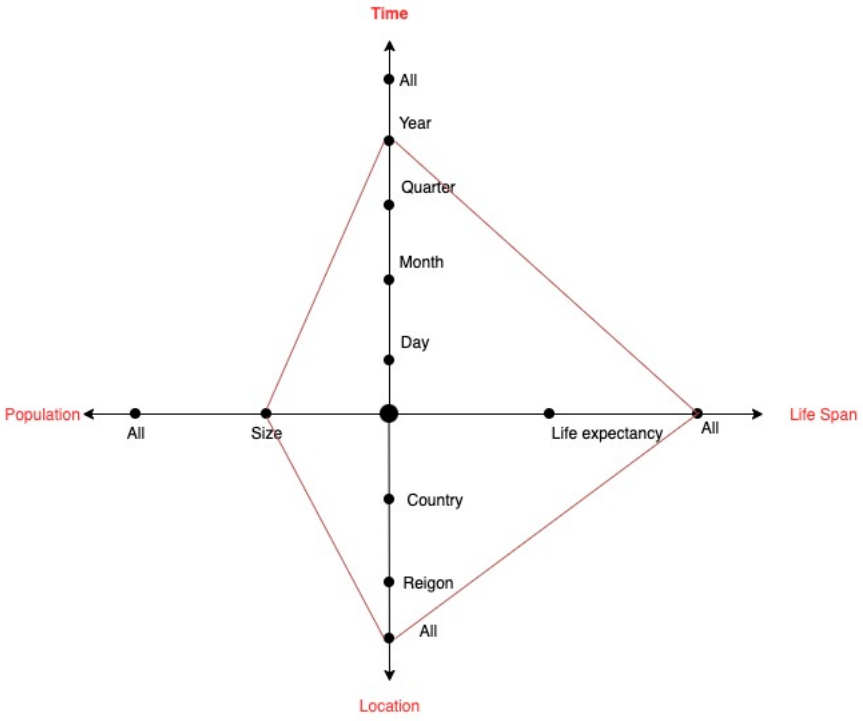


Figure 32

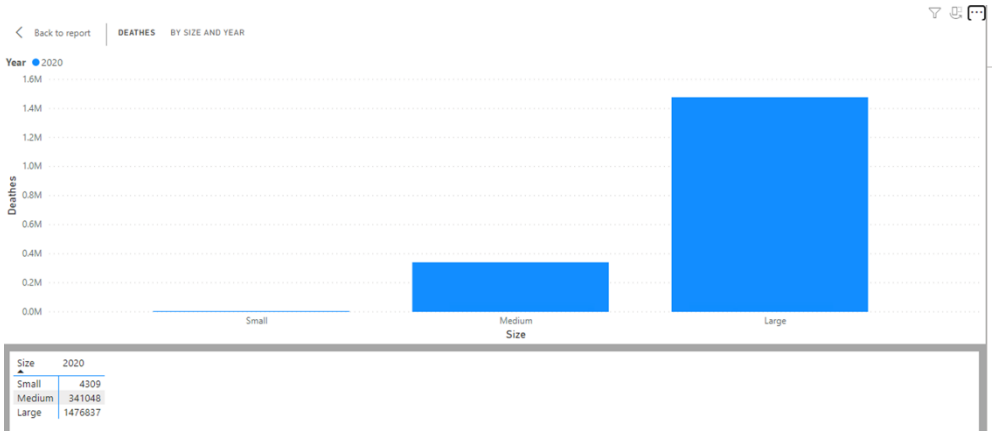


Figure 31

Q4 Do countries with a life expectancy greater than 75 have a higher recovery rate?

This query has involved in LifeSpan dimension life expectancy hierarchy and Location dimension country hierarchy. The time dimension are 2020 and 2021. Here is the Starnet footpath and visualize result.

Because $recovery\ rate = \frac{recovered}{confirmed}$, according to the visualized graph, people over the age of 75 are recovered less than people under the age of 75.

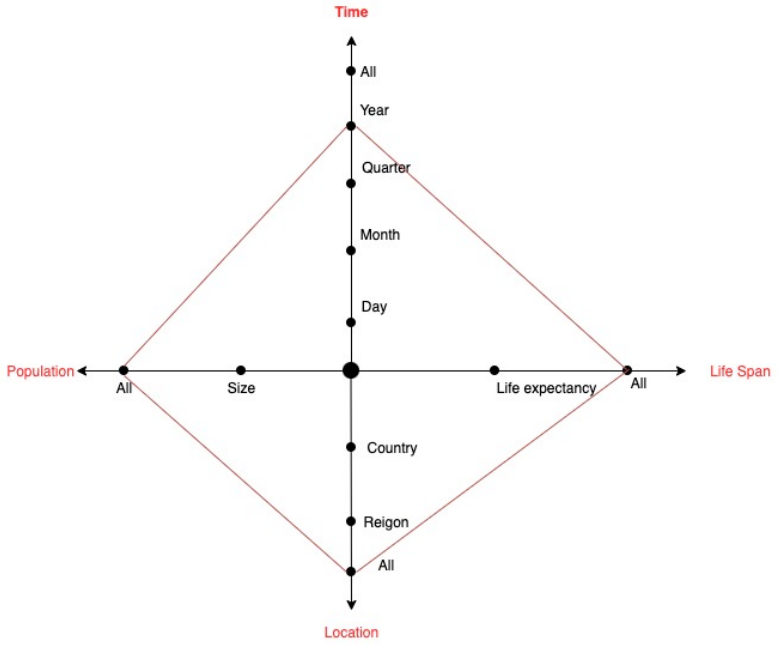


Figure 33

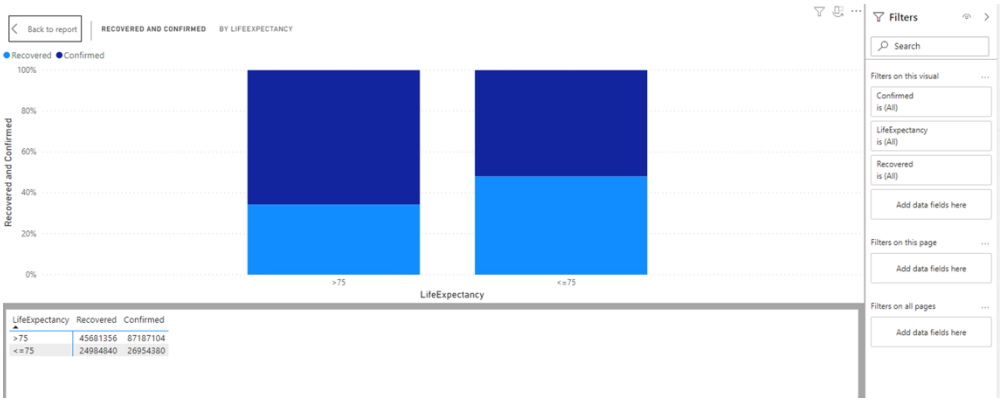


Figure 34

Task 8

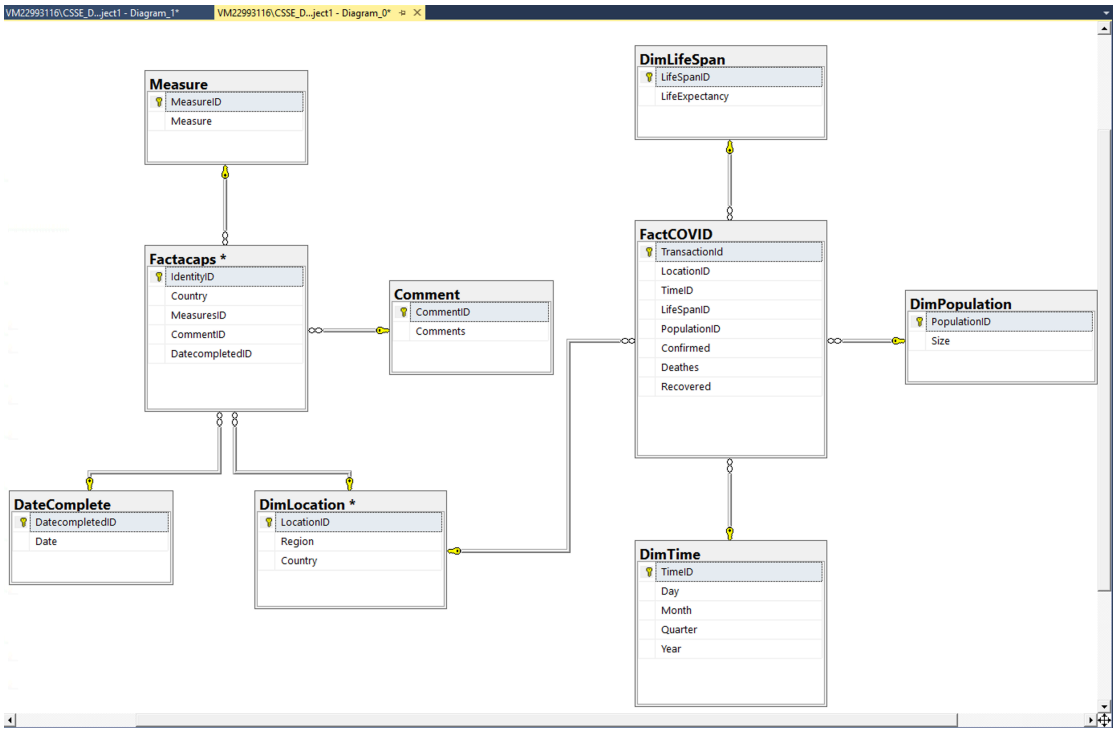


Figure 35

Question:

What is the number of confirmed cases after the adoption of border blockade measures in Australia in 2020?