# Report for Datawarehouse Designing of COVID-19

1. Observation data:

The 4 data tables are involved in the World Bank COVID-19 dataset, which we need to use for answering the business queries. Three csv tables have the data that describe the accumulated number of confirmed, death and recovered people from each country on the world. One large table describe the country information such as population and life expectancy.

From the table, we can have a roughly thinking of data warehouse designing. To answering the queries, we need the information for Time, Location, confirmed numbers, death numbers, recovery numbers, population of country and life expectancy. In addition, the smallest specific time that the queries asked is month. The location that queries asked accurate to the country.

Therefore, to answer the queries, firstly, the information in different tables need be combined. Secondly, the countries name in large table are different with the country names in the csv files. Thirdly, the cases of recovery, death and confirmed cases are recorded accumulated. Lastly, the province information needs to be merged as countries and the date need to be merged as month.

2. StarNet designing

There are four dimensions in my StarNet (Figure 1). Because there are no significant hierarchy relations in life expectancy and population, I show the different types together in those two dimensions. For time dimension, there are three hierarchy which are Year, Quarter and Month. For location dimension, there are two hierarchy. One is continent following by country.

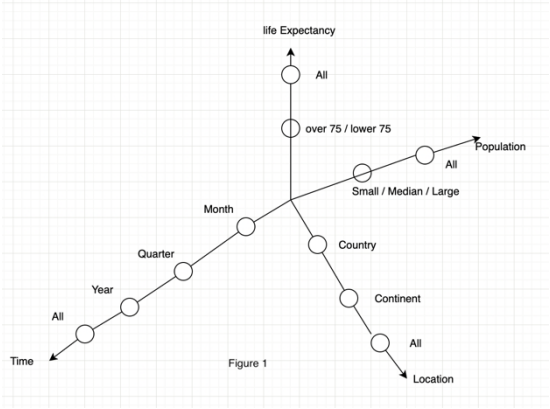


Figure 1

3. Footprints for answering questions

The query one can be answered by green Footprints (Figure 2). The total number of confirmed cases in Australia in each month, each quarter and total year of 2020 are described by green footprints in different hierarchy of time dimension from month to year.

The question two can be answered by red line (Figure 2). The region of the Americas is defined by continent hierarchy of Location dimension. The Sept 2020 can be defined by month hierarchy by Time dimension. The different country of United States, Canada and Mexico can be constrained by country hierarchy of location dimension.

The query three can be answered by red line and blue line of footprints (Figure 3). The red line illustrates the total number of covid deaths worldwide in 2020, which constrain a specific year hierarchy of time dimension and all for country dimension. The total number of covid deaths in large countries, medium countries and small countries, respectively, in 2020 are shown on blue line. There are two specific constrains in time dimension and population dimension and the country are labeled by population's types.

The query four can be answered by green line of footprint (Figure 3). There are only one constrain dimension of life expectancy. The labels of life expectancy can separate data by over 75 and lower 75.

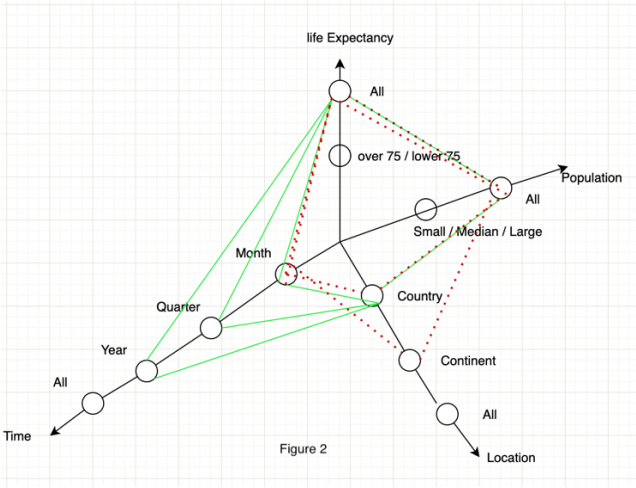


Figure 2

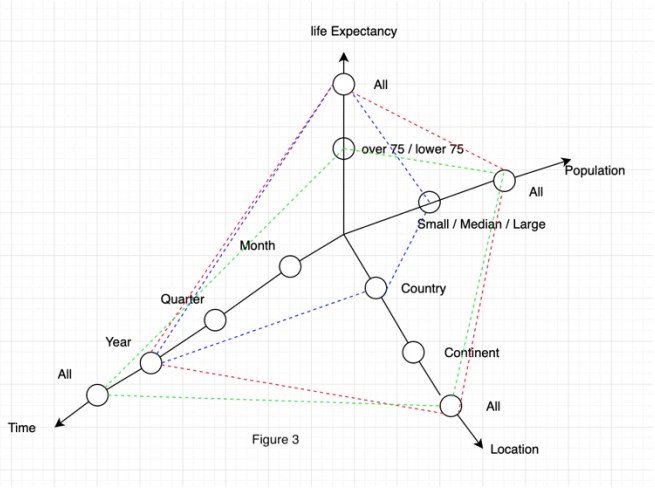


Figure 3

4. Produce a star schema

According to the StarNet designed before, using SQL Server Management Studio (SSMS) implement a star schema (Figure 4). Each dimension and fact table have only one primary key. The values of primary key are assigned by SQL system automatically.

For the life expectancy table, it has two columns which are life_expectancyID (primary key) and LifeAltID. In column of LifeAltID, it going to insert three values which are life expectancy lower 75, life expectancy over 75 and Unknown.

For the population table, it has two columns as well, which are PopulationID(primary key) and PopulationAltID. In column of PopulationAltID, four values will be inserted, which are small country, median country, large country and unknown country.

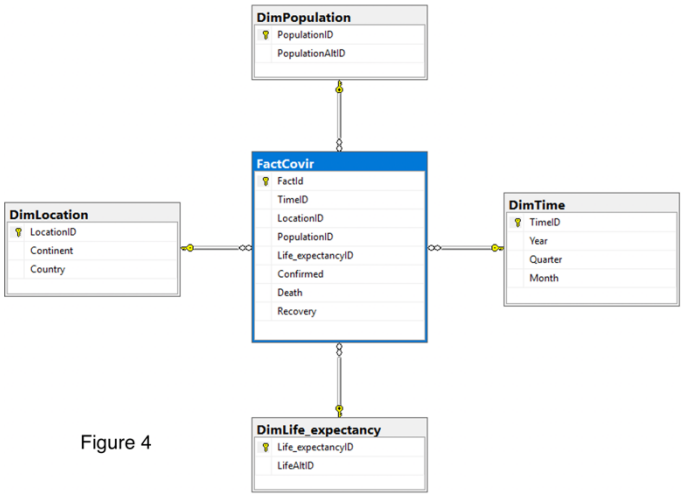


Figure 4

There are four columns in table DimTime shown on the Figure 4, with 15 rows from January 2020 to the March 2021.

The DimLocation table involves three columns (Figure 4). There are 192 number of countries in this table, but only the Americas continents are labeled. Because only the countries in the region of Americas are asked in the business queries.

The table of FactCovir has eight columns (Figure 4). The FactID is the primary key of this table. Its values are given automatically by SQL server system. There are four foreign keys in the table, which reference to the primary keys of the four dimension tables respectively. In addition, the three measures are appended in the table, which record the number of confirm, death and recovery people in each month from 2020 to 2021.

5. Data Cleaning, Integration and ETL processes

1) Selecting useful information and simplify datasheets

- Datasheets of recovery, confirm and death

Firstly, I deleted the columns of longitude and latitude because they are useless. Secondly, I summed the provinces data of the countries that have provinces. Thirdly, I change number of people in everyday to number of people in every month for three data sheets, which simplify datasheets from more than 400 columns to just 15 columns.

- Datasheet of owid-covid-data

Firstly, I removed the useless columns and only keep the useful columns, which are Location, Population and Life_expectancy. Secondly, I removed the duplicate locations and only keep each one of them. There are only 193 locations left after cleaning. Then, I find there are locations called 'international' and 'world', which is meaningless for answering queries. So, I delete these two rows as well.

2). Dealing with the unequal countries name in datasheets.

By observation datasheets, I found some countries name in recovery datasheet are not involved in the owid-covid-data, which means that some countries in recovery datasheet will not have values of population and life expectancy in owid-covid-data. Therefore, I set the value 'unknown' in the values of population and life expectancy in the lost countries. I set the unknown value in both dimension tables of population and life expectancy in case the countries are lost.

3). Creating the dimension and fact tables.

Designing dimension and fact tables according to the Star schema (figure 4). The DimTime table have four columns which are columns of blank, year, quarter and month. The blank column is used to insert values by SQL system automatically. The year column involve year from 2020 to 2021. The quarter column involves four quarters and the month column involve monthly data from January 2020 to March 2021. The total rows of DimTime table are 15.

There are three columns in DimLocation table which are blank, continent and country. The country names used in the country column are same with the country names in confirm datasheet with total 192 countries. In continent column, only the countries in the region of the Americas are marked.

The DimLife_ecpectancy table have two column which are blank column and lifeAltID. The values in blank columns are assigned by SQL server system automatically. The values in lifeAltID are over_75, lower_75 and Unknown

The DimPopulation table have two columns as well, blank column and PopulationAltID. There are four values inserted into the column popultionALtID which are small, median, large and unknown.

The FactCovir table have eight columns, which is an integration of the primary key in dimension tables. There are 2880 rows by assign each country (192) to each month (15). Assign the population and life_expectancy into the table by id of country. Assign each month number of confirme, death and recovery people into the FactCovir table.

6. Building a Multi-dimensional analysis service solution

Designing a cube based the fact and dimension tables created before. The structure of the cube shown below (Figure 5). Change the ID in the four dimension tables by its meaningful attributes names, which make people more easier to understand.

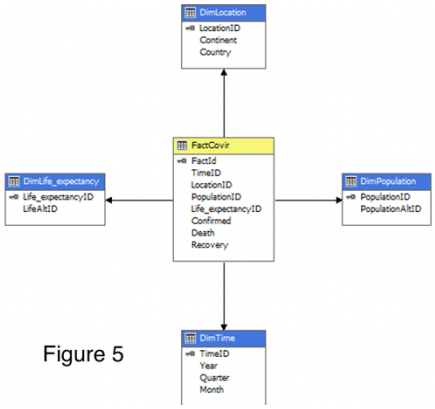


Figure 5

According to the designed cube, we can answer the business queries by dragging relevant dimensions and measures into the query area (the cube queries for answering questions are appended in appendix).

1) Question 1:

In Australia, there is total 28425 number of confirmed cases in 2020. There are 4559, 3361, 19176 and 1329 number of confirmed cases in the first, the second, the third and the last quarter of 2020 respectively. In each month of 2020 in Australia, the confirmed cases are shown below.

| Australia 2020 coronavirus confirmed cases | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Month | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Confirmed cases | 9 | 16 | 4534 | 2207 | 436 | 718 | 9360 | 8539 | 1277 | 499 | 317 | 513 |

2) Question 2:

In Sept 2020, there are 2874960 recovered cases in the region of the Americas.

The number of recovered cases in the United States, Canada and Mexico, respectively, in Sep 2020 are shown below.

| Recovered cases in Sep, 2020 | | | |
|---|---|---|---|
| Locations | United States | Canada | Mexico |
| Recovered cases | 655863 | 21161 | 131785 |

3) Question 3:

The total number of covid deaths worldwide in 2020 is 1825731.

The total number of covid deaths in large countries, medium countries and small countries, respectively, in 2020 are shown below.

| Covid deaths in 2020 | | | |
|---|---|---|---|
| Country population | large | median | small |
| Death cases | 1474746 | 341163 | 4280 |

4) Question 4:

The number of recovery people with a life expectancy greater than 75 is 39394668. The number of recovery people with a life expectancy lower than 75 is 24864448. The number of confirmed people with a life expectancy greater than 75 is 83462558. The number of confirmed people with a life expectancy lower than 75 is 46831814.

| Recovery rate | | |
|---|---|---|
| Life expectancy | Greater than 75 | Lower than 75 |
| Recovery cases | 39394668 | 24864448 |
| Confirmed cases | 83462558 | 46831814 |
| Recovery rate | 47.2% | 53.1% |

The recovery rate in countries with life expectancy greater than 75 is 47.2%, which is lower than the recovery rate in countries with life expectancy lower than 75.

7. Using power BI to visualize business queries

1) Question 1:

The figure 6 shows the total number of confirmed cases in Australia in 2020. The figure 7 and figure 8 illustrate the number of confirmed cases in Australia in each month and each quarter in 2020. The results are same with the Multi-dimensional analysis solution before.

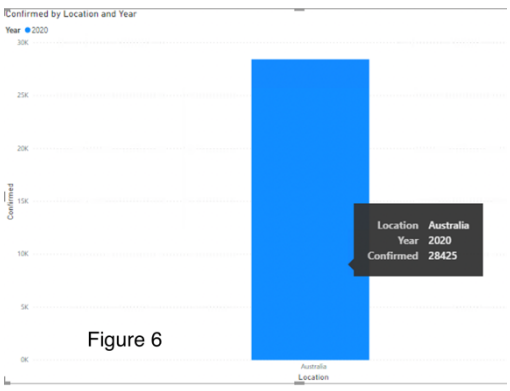



Figure 6

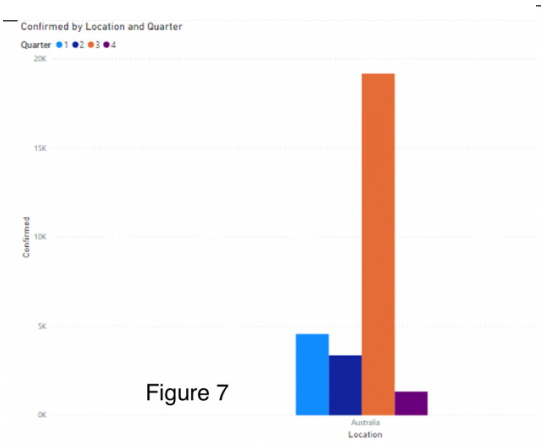



Figure 7

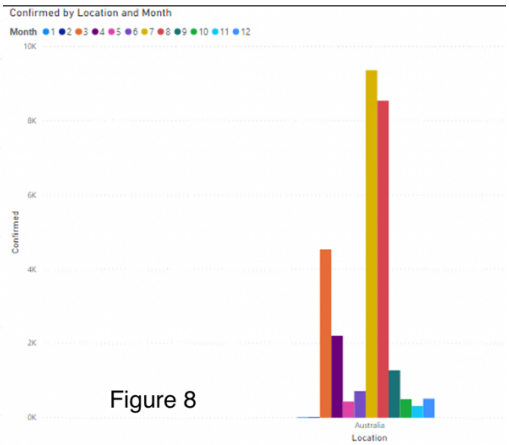



Figure 8

2) Question 2:

The figure 9 illustrate that there total 2874960 number of recovery cases in the region of Americas in Sept 2020. The figure 10 show that the number of recovery cases in Canada, Mexico and US, which are 21161, 131785 and 655863 respectively.

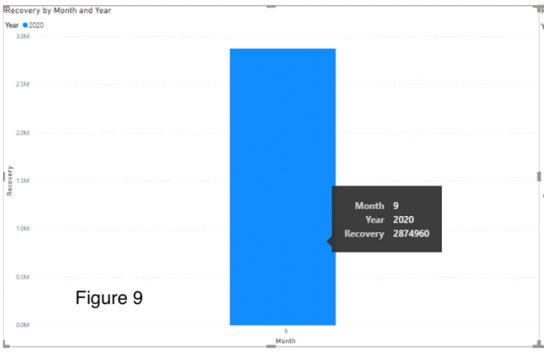



Figure 9

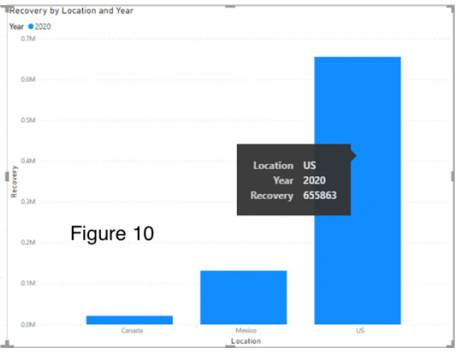



Figure 10

3) Question 3:

The figure 13 shows that the total number of covid deaths worldwide in 2020 is 1825731. The figure 14 shows the number of covid death in small, median and large countries, which are 4280, 341163 and 1474746.

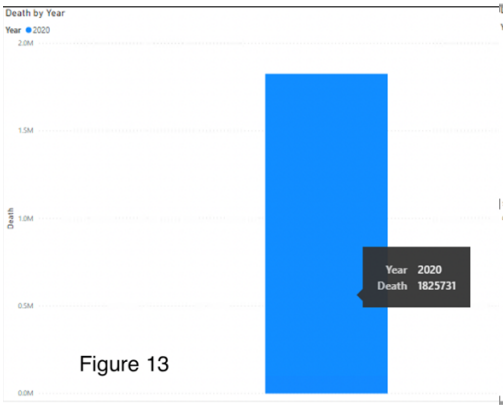



Figure 13

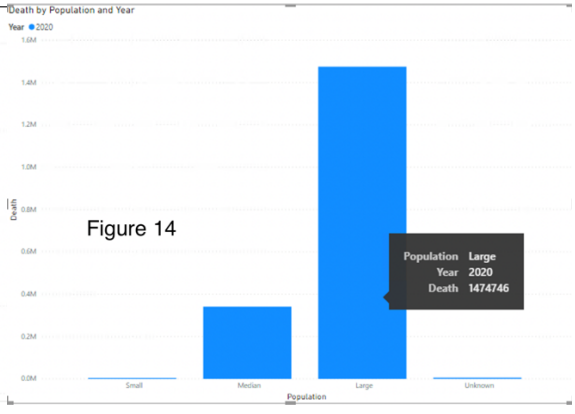



Figure 14

4) Question 4:

The figure 15 illustrates the number of confirmed cases and recovery cases in the countries with life expectancy lower and over 75. The recovery rate can be calculated by number of recovery cases divide number of confirmed cases.

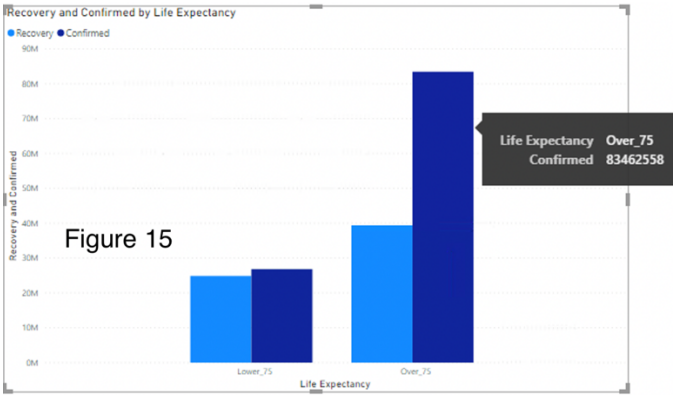
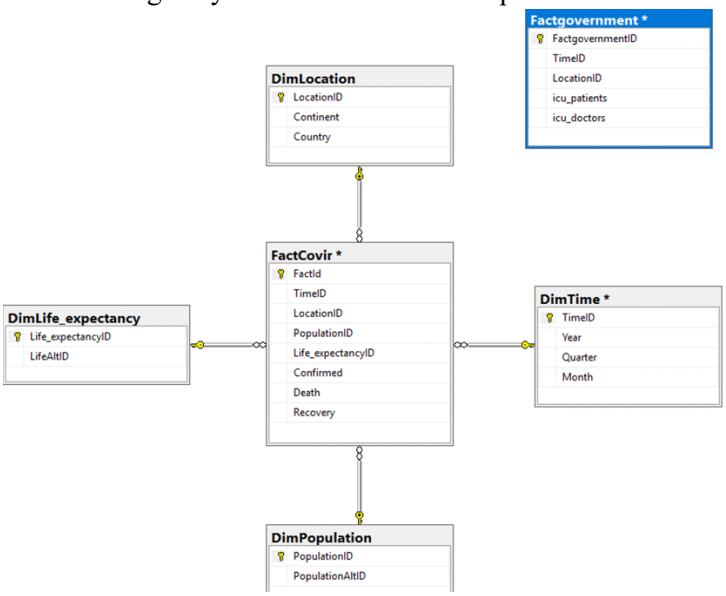



Figure 15

8. Implement a galaxy schema

I try to create a galaxy schema by SSMS but there is a constrain for primary key. Because the primary key cannot connect with the second fact table. However, if my imagine is work, there is a relation between DimLocation and Factgovernment table and a relation between Location and Factgovernment table. Then we can use this galaxy schema to answer the question of How many people dead in icu during 2020 in US.



# Appendix

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Location | Hierarchy | Equal | { Australia } |
| Time | Hierarchy | Equal | { 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, ... |
| <Select dimension> | | | |

| Confirmed |
|---|
| 28425 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Location | Hierarchy | Equal | { Australia } |
| Time | Hierarchy | Equal | { 1, 1, 1 } |
| <Select dimension> | | | |

| Confirmed |
|---|
| 4559 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Location | Hierarchy | Equal | { Australia } |
| Time | Hierarchy | Equal | { 2, 2, 2 } |
| <Select dimension> | | | |

| Confirmed |
|---|
| 3361 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Location | Hierarchy | Equal | { Australia } |
| Time | Hierarchy | Equal | { 3, 3, 3 } |
| <Select dimension> | | | |

| Confirmed |
|---|
| 19176 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Location | Hierarchy | Equal | { Australia } |
| Time | Hierarchy | Equal | { 4, 4, 4 } |
| <Select dimension> | | | |

| Confirmed |
|---|
| 1329 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Location | Hierarchy | Equal | { Australia } |
| Time | Hierarchy | Equal | { 1 } |
| <Select dimension> | | | |

| Confirmed |
|---|
| 9 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Location | Hierarchy | Equal | { Australia } |
| Time | Hierarchy | Equal | { 2 } |
| <Select dimension> | | | |

| Confirmed |
|---|
| 16 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Location | Hierarchy | Equal | { Australia } |
| Time | Hierarchy | Equal | { 3 } |
| <Select dimension> | | | |

| Confirmed |
|---|
| 4534 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Location | Hierarchy | Equal | { Australia } |
| Time | Hierarchy | Equal | { 4 } |
| <Select dimension> | | | |

| Confirmed |
|---|
| 2207 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Location | Hierarchy | Equal | { Australia } |
| Time | Hierarchy | Equal | { 5 } |
| <Select dimension> | | | |

| Confirmed |
|---|
| 436 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Location | Hierarchy | Equal | { Australia } |
| Time | Hierarchy | Equal | { 6 } |
| <Select dimension> | | | |

| Confirmed |
|---|
| 718 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Location | Hierarchy | Equal | { Australia } |
| Time | Hierarchy | Equal | { 7 } |
| <Select dimension> | | | |

| Confirmed |
|---|
| 9360 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Location | Hierarchy | Equal | { Australia } |
| Time | Hierarchy | Equal | { 8 } |
| <Select dimension> | | | |

| Confirmed |
|---|
| 8539 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Location | Hierarchy | Equal | { Australia } |
| Time | Hierarchy | Equal | { 9 } |
| <Select dimension> | | | |

| Confirmed |
|---|
| 1277 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Location | Hierarchy | Equal | { Australia } |
| Time | Hierarchy | Equal | { 10 } |
| <Select dimension> | | | |

| Confirmed |
|---|
| 499 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Location | Hierarchy | Equal | { Australia } |
| Time | Hierarchy | Equal | { 11 } |
| <Select dimension> | | | |

| Confirmed |
|---|
| 317 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Location | Hierarchy | Equal | { Australia } |
| Time | Hierarchy | Equal | { 12 } |
| <Select dimension> | | | |

| Confirmed |
|---|
| 513 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Location | Hierarchy | Equal | {Americas, Americas, America |
| Time | Hierarchy | Equal | {9, 2020} |
| <Select dimension> | | | |

| Recovery |
|---|
| 2874960 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Location | Hierarchy | Equal | {US} |
| Time | Hierarchy | Equal | {9, 2020} |
| <Select dimension> | | | |

| Recovery |
|---|
| 655863 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Location | Hierarchy | Equal | {Canada} |
| Time | Hierarchy | Equal | {9, 2020} |
| <Select dimension> | | | |

| Recovery |
|---|
| 21161 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Location | Hierarchy | Equal | {Mexico} |
| Time | Hierarchy | Equal | {9, 2020} |
| <Select dimension> | | | |

| Recovery |
|---|
| 131785 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Location | Hierarchy | Equal | {All} |
| Time | Hierarchy | Equal | {2020, 2020, 2020, 2020, 202 |
| <Select dimension> | | | |

| Death |
|---|
| 1825731 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Location | Hierarchy | Equal | |
| Time | Hierarchy | Equal | {2020, 2020, 2020, 202 |
| Population | Hierarchy | Equal | {Large} |

| Death |
|---|
| 1474746 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Time | Hierarchy | Equal | {All} |
| Population | Hierarchy | Equal | |
| Life Expectancy | Hierarchy | Equal | {Over_75} |

| Recovery |
|---|
| 39394668 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Time | Hierarchy | Equal | {All} |
| Population | Hierarchy | Equal | |
| Life Expectancy | Hierarchy | Equal | {Over_75} |

| Confirmed |
|---|
| 83462558 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Time | Hierarchy | Equal | {All} |
| Population | Hierarchy | Equal | |
| Life Expectancy | Hierarchy | Equal | {Lower_75} |

| Confirmed |
|---|
| 26831814 |

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Time | Hierarchy | Equal | {All} |
| Population | Hierarchy | Equal | |
| Life Expectancy | Hierarchy | Equal | {Lower_75} |

| Recovery |
|---|
| 24864448 |