CITS5504

Data Warehousing

Project 1

Building a Data Warehouse

02/04/2021

Table of Contents

Table of Tables

Table of Figures

Introduction

On 31 December 2019, the World Health Organisation (WHO) China Country Office was informed that some cases of pneumonia unknown etiology detected in Wuhan, Hubei Province of China (WHO, 2020a). On 12 March 2020, WHO announced that the global COVID-19 outbreak is a controllable pandemic (WHO, 2020b). As of 14 March 2021, there are 119218587 cumulative cases and 2642673 cumulative deaths worldwide (WHO, 2020c). In order to develop therapeutic solutions and understand COVID-19 biology, many countries have collected and analysed data related to COVID-19, therefore, developing a new tool for organisations to use their data to make decisions is important.

This project aims to use data from four Comma Separated Value (CSV) files to build a data warehouse, which will be used to answer four business queries. In order to achieve this goal, this report will be divided into eight parts. Firstly, part one provided two assumptions for this project. Subsequently, part two will present a big picture of these CSV files. The third part will introduce the four business queries and a related StarNet diagram briefly. Next, part four is important in this report since data warehouse modelling, and design of dimension tables and fact table will be explained in this part. Part five is related to data process, it includes data extraction, transformation and cleaning, and loading. Then, OLAP cubes will be designed in part six. The results for business queries will be presented and visualised in part seven. Finally, a further business query and a galaxy schema will be provided in part eight; this part will use five Comma Separated Value files.

This project includes two parts: four business queries and task 8. Two databases will be built in this project, the first database, Project_1_Covid19, will be used to answer four business queries, and the second databases Project_1_Covid19_2, will be used to answer Task 8 (further research section).

Assumptions

1) In this project, sovereign states' data are used to build the data warehouse and be analysed. Other regions or entities will be removed.

2) This project only focuses on local cases (confirmed, deaths and recovered); all dependent territories and their data will be removed.

Data Overview

In this project, a total of five CSV files will be used. See Table 1.

CSV files	Overview	Main Contents
acaps_covid19_government_measures_dataset.csv	18 columns and 23923 rows (excluding column header)	The measures implemented by governments worldwide in response to the COVID-19 pandemic.
owid-covid-data.csv	50 columns and 62201 rows (excluding column header)	The basic information of some countries.
time_series_covid19_deaths_global.csv (Snapshot table)	409 columns and 274 rows (excluding column header)	The number of deaths from 1/23/2020 to 1/3/2021 for some countries.
time_series_covid19_confirmed_global.csv (Snapshot table)	409 columns and 274 rows (excluding column header)	The number of confirmed cases from 1/23/2020 to 1/3/2021 for some countries.
time_series_covid19_recovered_global.csv (Snapshot table)	409 columns and 274 rows (excluding column header)	The number of recovered cases from 1/23/2020 to 1/3/2021 for some countries.

Table 1: Data overview

Requirements Analysis and StarNet

In this report, four business queries need to be answered.

The potential dimensions and measurements can be identified (See Table 2).

Queries	Keywords	Potential dimensions	Measurements
Query 1	Month, Quarter, Year and Australia	Time (Month, Quarter and Year) Country (Australia)	Confirmed cases
Query 2	September, Americas, United State, Canada and Mexico	Time (September) Region (Americas) Country (the United States, Canada and Mexico)	Recovered cases
Query 3	Population	Population (Country size)	Deaths
Query 4	Life expectancy	Health condition (life expectancy)	Recovered cases and confirmed cases

Table 2: Business queries analysis

According to the requirements analysis, the StarNet should have four dimensions: time, country location, population, and health condition. A benefit of StarNet is that checking those dimensions could help reduce the time for data processing in the project's model (Kao, Hung & Hsu, 2008).

For the time dimension, it should include month, quarter, and year. Business queries 1, 2 and 3 are related to year (2020), month (September) and quarters. However, all deaths, confirmed cases and recovered cases information is stored in three time series covid19 data files in units of days. In order to use this data warehouse to collect and analyse further data and update the data warehouse every day, "day" will be added into the time dimension. Therefore, the time dimension includes all, year, quarter, month and day.

For the country location dimension, it includes all, region and country. Since a region (Americas) and some countries, such as Australia, the United States, Canada and Mexico, are involved in business queries 1 and 2.

The population dimension includes all and country size. Although there are three country sizes, they are not in a sequence of mappings from a set of low-level concepts to high-level. From this point of view, in addition to "all", only country size is included in the population dimension.

Finally, the health condition dimension includes all and life expectancy. Same as population dimension, greater than 75 and less than or equal to 75 are not in a sequence of mappings from a set of low-level concepts to high-level. The StarNet see Figure 1.

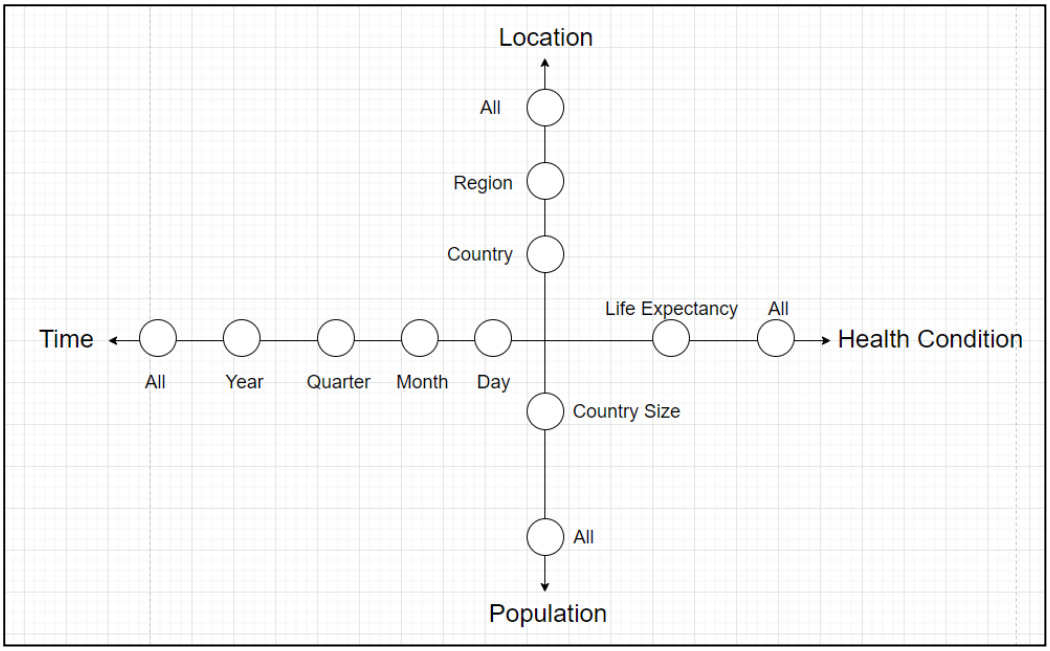


Figure 1: StarNet

Business query 1 focuses on the number of confirmed cases in Australia in 2020, each quarter of 2020 and each month of 2020. Therefore, the footprint for business query 1 should include time (Year is for sub-query 1, Quarter is used to sub-query 2, and Month can be applied for sub-query 3) and location dimensions (Country) (For population and Health Condition, “All” is suitable). See Figure 2, 3 and 4.

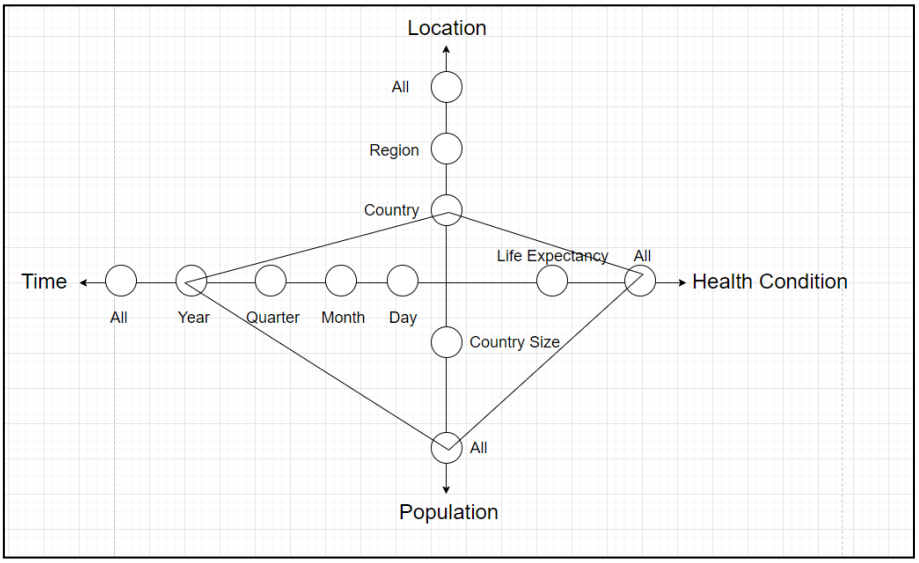


Figure 2: The footprint of the number of confirmed cases in Australia in 2020

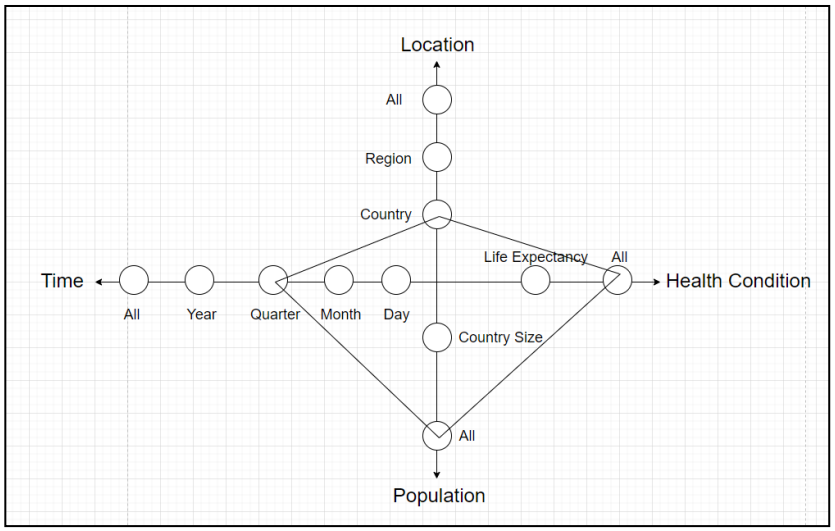


Figure 3: The footprint of the number of confirmed cases in each quarter of 2020 in Australia

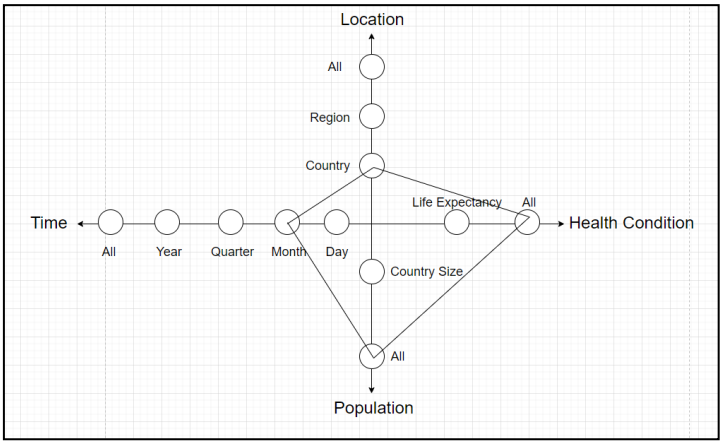


Figure 4: The footprint of the number of confirmed cases in each month of 2020 in Australia

Business query 2 pays attention to the Americas and three countries (the United States, Canada and Mexico). Besides, this business query only asks for the number of recovered cases in September 2020. Hence, the related dimensions are location (Region is applied for sub-query 1 and Country is used to sub-query 2) and time (Month) (For population and Health Condition, “All” is appropriate). See Figure 5 and 6.

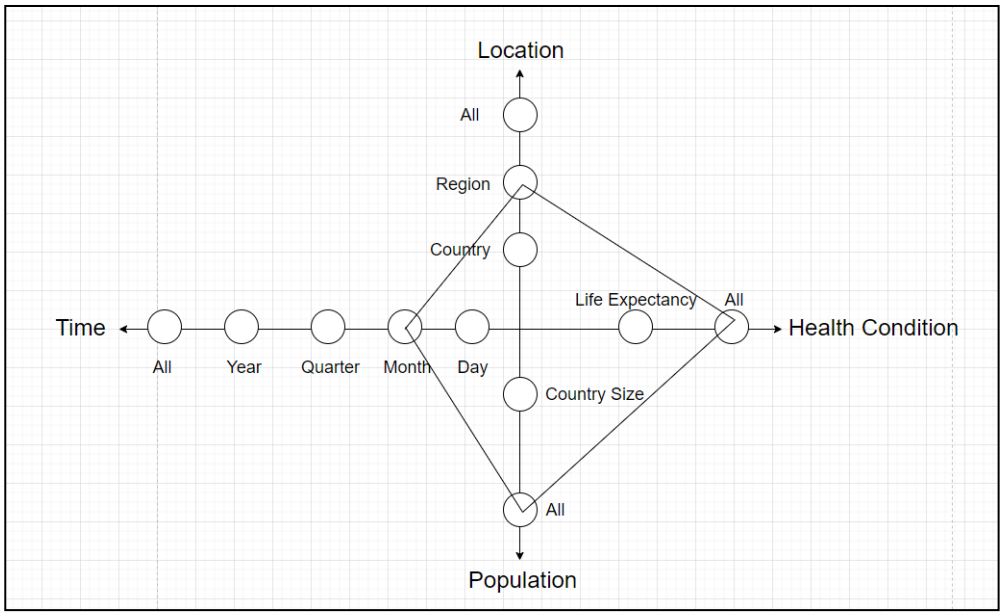


Figure 5: The footprint of the number of recovered cases in the Americas in September 2020

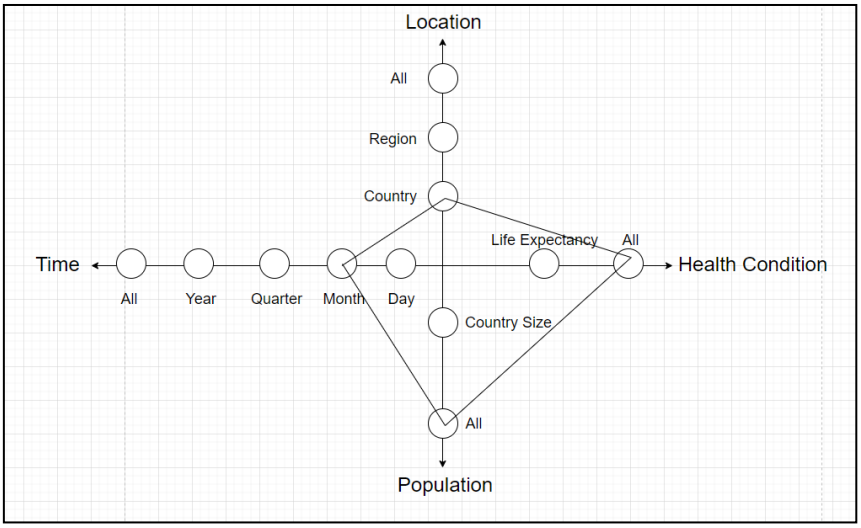


Figure 6: The footprint of the number of recovered cases in the United States, Canada and Mexico in September 2020

Business query 3 asks for presenting the total deaths worldwide in 2020. Thus, the time dimension (Year) and all countries should be applied to sub-query 1 and 2. Meanwhile, all countries were divided into large, medium and small groups based on their population (sub-query 2 only). Therefore, the population dimension (Country Size) also should be involved in sub-query 2 (For Health Condition, “All” is appropriate). See Figure 7 and 8.

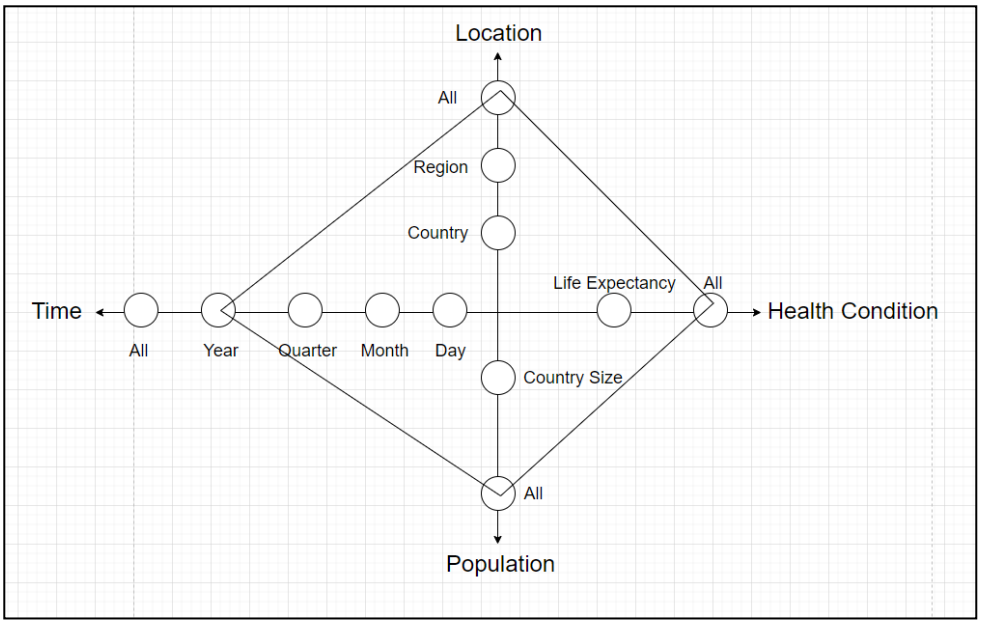


Figure 7: The footprint of the total number of covid deaths worldwide in 2020

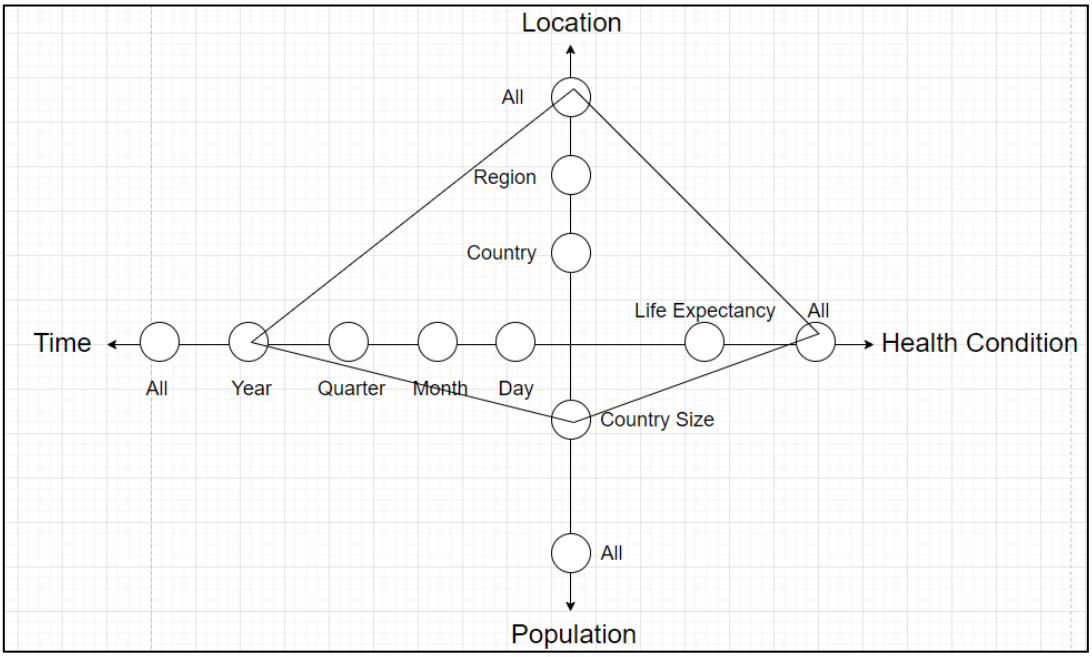


Figure 8: The footprint of the total number of covid deaths in large countries, medium countries and small countries in 2020

The final business query is related to life expectancy and recovery rate for all countries in 2020 and 2021; therefore, the health condition dimension (Life Expectancy) will be used for this business query. For other dimensions, “All” is appropriate. See Figure 9.

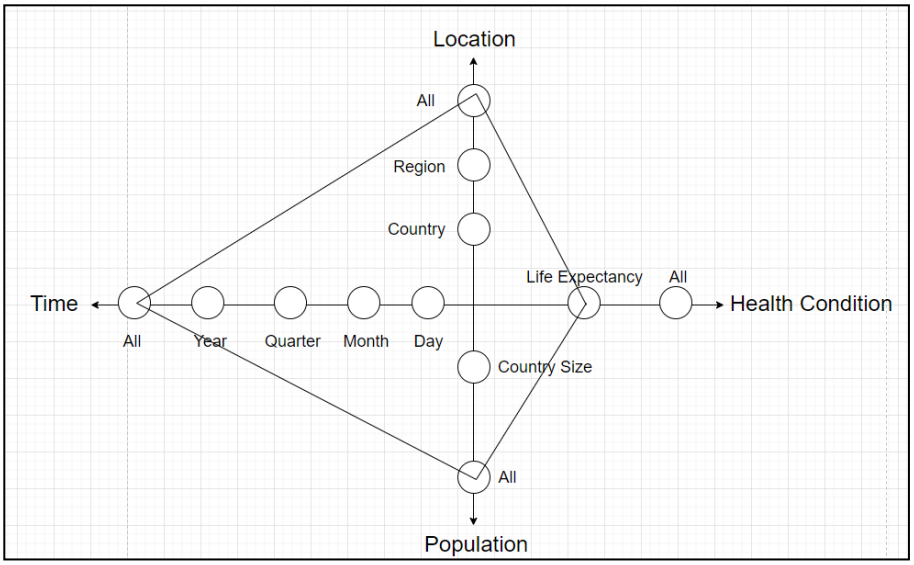


Figure 9: The footprint of the relationship between life expectancy and recovery rate

Dimensional Modelling and Tables Design

In a data warehouse, dimensional modelling is an important part since good dimensional modelling could help users minimise the query execution time and save the storage memories (Kimball and Ross, 2011).

Star Schema and Snowflake Schema

In a star schema, all dimensions data is saved in dimension tables; there are various dimension tables; however, for each dimension table, it only includes a group of dimensions, and these dimension tables connect to centralised fact tables. The fact tables include the measurements and foreign keys, which refers to the dimension tables. (Sidi et al., 2016). An example of a star schema. See Figure 10.

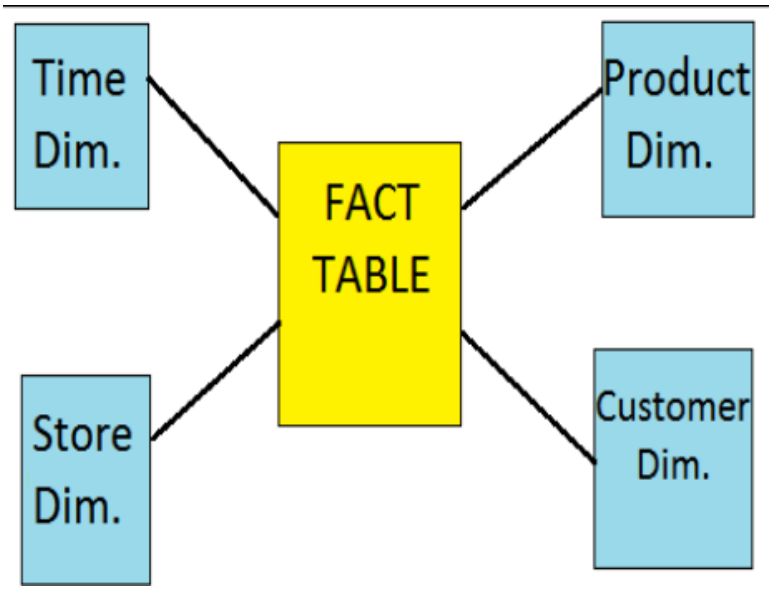


Figure 10: An example of star schema

Similar to a star schema, there are some dimension tables connected to centralised fact tables in a snowflake schema. However, in a snowflake schema, each dimension table is split into different hierarchies. For each hierarchy, it represents a single table. See Figure 11.

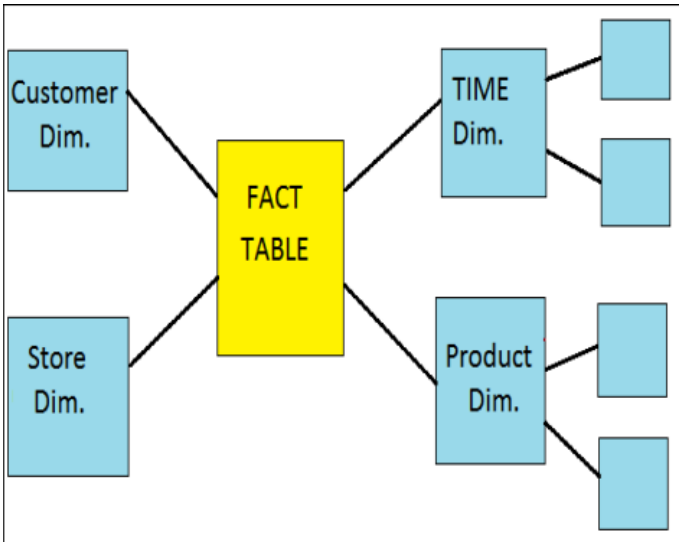


Figure 11: An example of snowflake schema

The Suitable Schema for This Project

Snowflake schema and star schema are used for different projects. They have their benefits. For example, using snowflake schema could use less disk

space since data is normalised and minimise data redundancy (Levene and Loizou, 2003). However, the snowflake schema is challenging to design. Mohammed (2019) points out that star schema is better than snowflake schema. The reasons are that the queries in star schema are easy to understand; the queries performance is more efficient in star schema; and in a star schema, the number of foreign keys is less than snowflake schema. From these points of view, the star schema was chosen for this project. Based on StarNet and analysis of business queries, there are four dimension tables and one fact table. The fact table connects to four dimension tables through four foreign keys. The star schema sees Figure 12. The relations and tables were created by SQL in SQL Server Studio, the SQL code please see the submitted SQL file (Create Tables.sql).

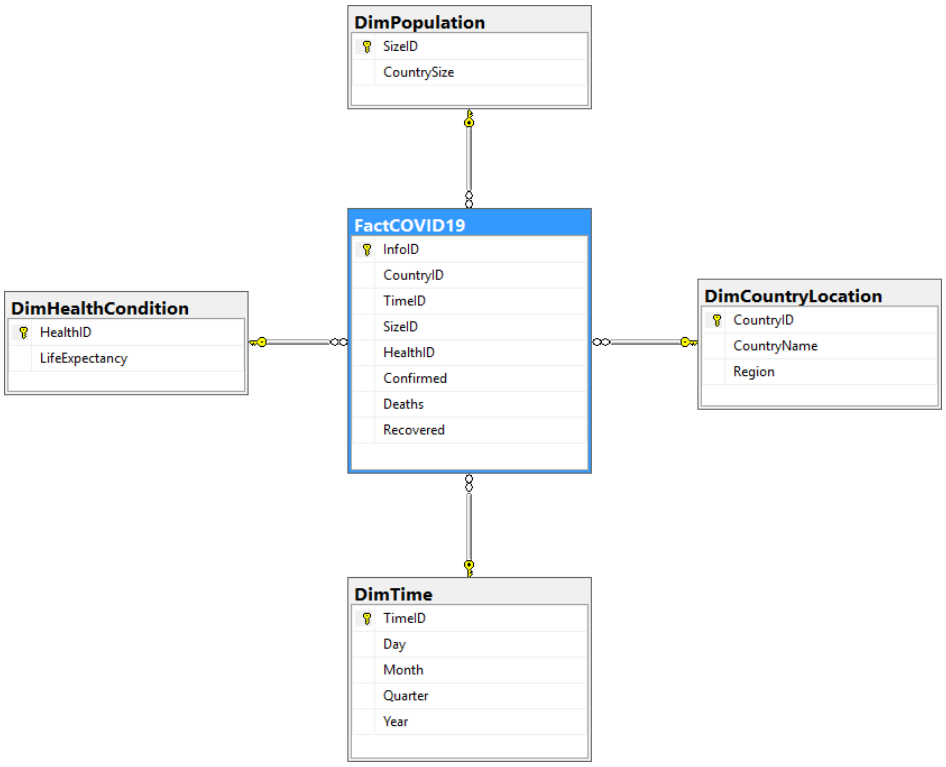


Figure 12: Star schema in this project

Dimension Tables Design

Refer to the StarNet and dimensional modelling; there are four dimension

tables in this project. See Table 3.

Dim_Country_Location	CountryID (1-188) Primary key.
	CountryName (the name of 188 countries)
	Region (5 Regions, Americas, Asia, Europe, Africa, Oceania)
Dim_Time	TimeID (1-405, total 405 days from 1/22/2020 to 3/1/2021). Primary key.
	Day (1/22/2020 to 3/1/2020)
	Month (1/2020 to 3/2021)
	Quarter (2020 Q1, 2020 Q2, 2020 Q3, 2020 Q4 and 2021 Q1)
	Year (2020 and 2021)
Dim_Population	SizeID (1, 2 and 3). Primary key.
	CountrySize (Large, Medium, and Small)
Dim_Health_Condition	HealthID (1 and 2). Primary key.
	LifeExpectancy (Greater than 75, and Less than or equal to 75)

Table 3: Dimension tables in this project

LocationID, TimeID, SizeId and HealthID are surrogate keys, not natural keys; hence, the data warehouse could insulate from changes to operational systems, and it easy to integrate data from multiple systems. Moreover, using surrogate keys is better handling of exceptional cases.

Fact Table Design

There are three measures (confirmed cases, deaths and recovered cases) in the fact table. Four dimension tables are connected to the fact table through foreign keys. And the fact table has its own primary key (InfoID) See Table 4.

Primary key	InfoID
Foreign keys	CountryID
	TimeID
	SizeID
	HealthID
Measures	Confirmed
	Deaths
	Recovered

Table 4: Fact table in this project

Data Process

After created the StarNet and footprints, the data from these four given CSV files should be processed. In data warehouses, the process is extraction-transformation-loading (ETL). El-Sappagh et al. (2011) point out that first, data should be extracted from different data sources, such as Excel files, a relational database, or a web site. Subsequently, the extracted data is propagated to the

data staging area (DSA). In the DSA, data will be transformed and cleansed. Finally, the clean data can be loaded to the data warehouse. See Figure 13. In this project, the principles of data process are accuracy and reasonability. Since for a data warehouse, the majority objects are querying and adding, high accuracy for a data warehouse is important.

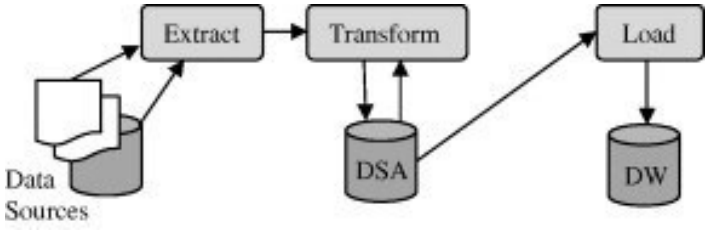


Figure 13: A general framework for ETL processes

Data Extraction

Extraction aims to extract data from some source systems. Usually, the extraction process will use ODBC/JDBC drivers to connect to some database resources. Then, understand the data structure of resource. Finally, take some measures to handle the sources with different nature (El-Sappagh et al., 2011). In this project, five CSV files were extracted from three different data sources. For example, acaps_covid19_government_measures_dataset.csv was extracted from the ACAPS website, an independent information provider. owid-covid-data.csv was extracted from the Our World in Data website, which is an online scientific publication. time_series_covid19_confirmed_gaobal.csv, time_series_covid19_deaths_global.csv and time_series_covid19_recovered_global.csv were extracted from the same data source, Johns Hopkins University Centre for System Science and Engineering (JHU CCSE). However, the method of data extraction was not provided by the CITS5504 teaching team. Therefore, this report cannot explain the method of data extraction for this project.

Data Transformation and Data Cleansing

After data extraction, data need to be transformed. In order to gain accurate data that is correct, complete, consistent, and unambiguous, the extracted data needs to be cleaned, transformed and integrated. Besides, fact tables and dimension tables should be generated in data transformation process.

There are various methods that can be used for data cleansing; the two most popular methods are using coding-based approaches (such as Python and R) and a software approach (such as Excel). In this project, Excel was chosen for cleaning data for this project, since for business analytics students, they are familiar with Excel.

In this project, four datasets need to be cleaned, which are related to four business queries. These four CSV files see Table 5.

Files' name	owid-covid-data.csv
	time_series_covid19_confirmed_global.csv
	time_series_covid19_deaths_global.csv
	time_series_covid19_recovered_global.csv

Table 5: Related CSV files for four business queries

Country names issue

The data cleansing process for county names includes a clean owid covid dataset and a clean three time series covid19 dataset. See Figure 14.

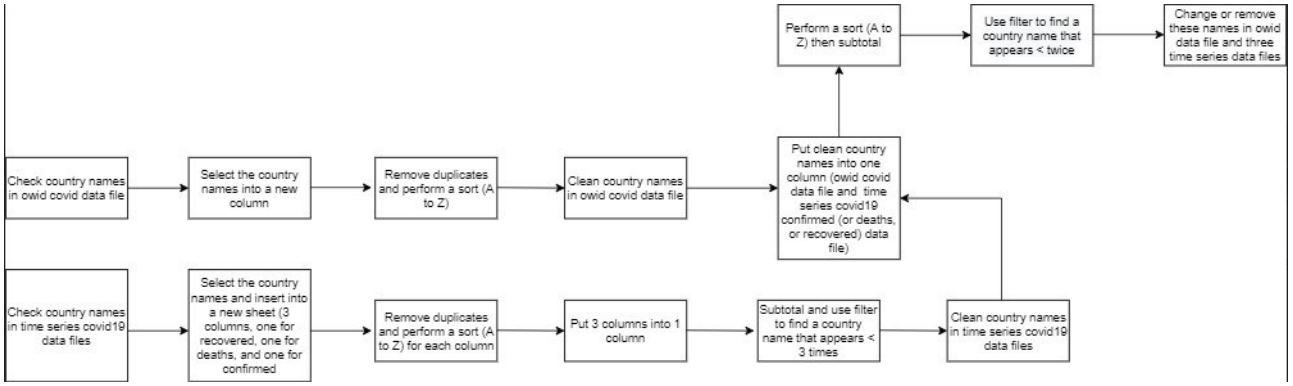


Figure 14: Cleansing process for country names

The countries in the owid covid data file have many duplicates; therefore, removing the duplicates is necessary. After duplicates removing, there are 192 unique countries (or regions). See Figure 15 and 16.

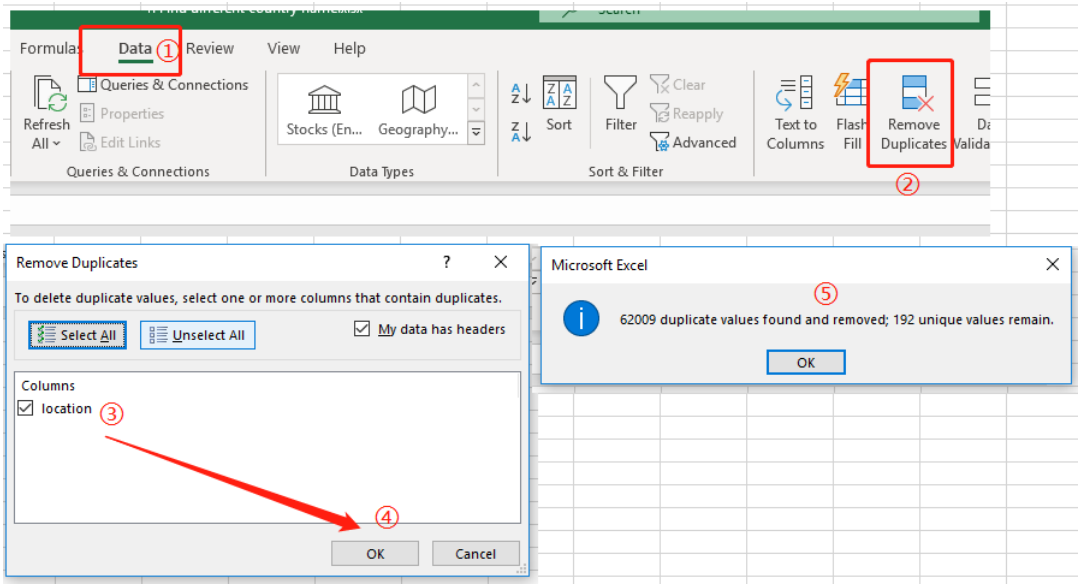


Figure 15: Remove duplicates for names in owid covid data file

Figure 16: Names in owid covid data file without duplicates

In this project, three time series covid19 data files need to be cleaned; the first step is that find the different country names in these files. The step is the same as the owid covid data cleaning process. However, there are three

columns in one sheet; therefore, the duplicates should be removed column by column. See Figure 17 and 18.

	A	B	C
1	Country/Region	Country/Region	Country/Region
2	Afghanistan	Afghanistan	Afghanistan
3	Albania	Albania	Albania
4	Algeria	Algeria	Algeria
5	Andorra	Andorra	Andorra
6	Angola	Angola	Angola
7	Antigua and Barbuda	Antigua and Barbuda	Antigua and Barbuda
8	Argentina	Argentina	Argentina
9	Armenia	Armenia	Armenia
10	Australia	Australia	Australia
11	Australia	Australia	Australia
12	Australia	Australia	Australia
13	Australia	Australia	Australia
14	Australia	Australia	Australia
15	Australia	Australia	Australia
16	Australia	Australia	Australia

In this sheet, all country names were from time_series_covid19 files.
Column A is country/region in confirmed cases file.
Column B is country/region in deaths file.
Column C is country/region in recovered file.

Figure 17: Names in three times series covid19 data files

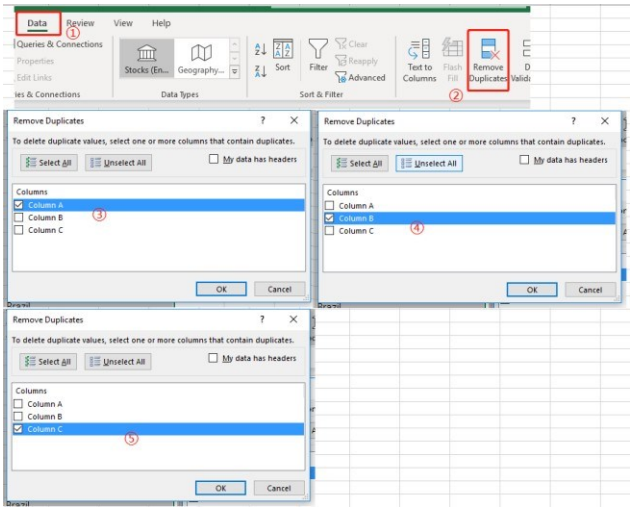


Figure 18: Remove duplicates for names in three time series covid19 data files

Then, put all country names into one column; in order to identify that which file the country name belongs to, colours were applied. Subsequently, perform an ascending sort by the cell values of country/region column. See Figure 19. Finally, use “subtotal” to count the number of each country, and use filter to find countries that appear less than 3 times. See Figure 20. As a result, all countries appear three times; therefore, the country names in three time series covid19 data files are same.

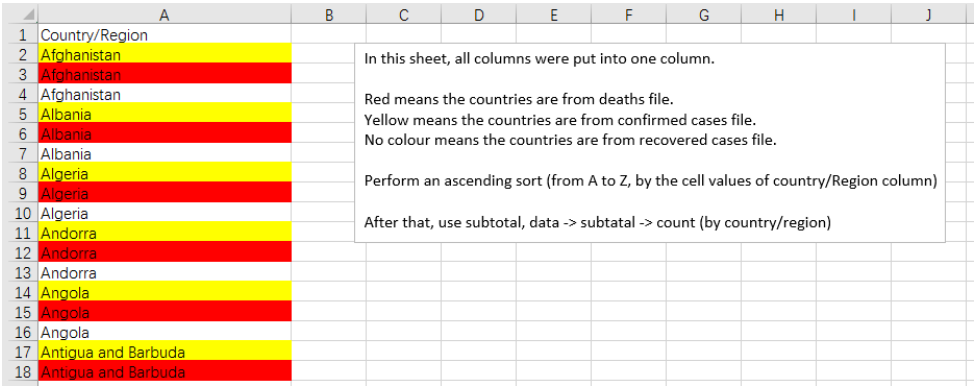


Figure 19: Preparation for finding different names in three time series covid19 data files

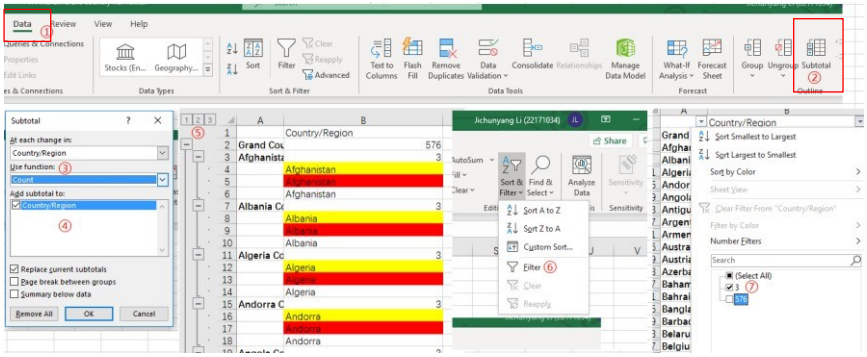


Figure 20: Looking up the different names in three time series covid19 data files

The country names have been cleaned in the owid covid data file and three time series covid19 data files. Then,1) put country names from owid covid data file and time series confirmed (or deaths or recovered) data file into one column (yellow means the country names were extracted from confirmed cases file, and no colour means the country names were extracted from owid covid data file). See Figure 21. 2) Perform an ascending sort by the cell values of country/region column. 3) Use “subtotal” and filter to find the country names that appear once. See Figure 22. Finally, the country names that need to be changed or removed can be found. See Figure 23.

B
Country/region
Afghanistan
Afghanistan
Albania
Albania
Algeria
Algeria
Andorra
Andorra
Angola
Angola
Antigua and Barbuda
Antigua and Barbuda
Argentina
Argentina
Armenia
Armenia
Australia
Australia
Austria
Austria
Azerbaijan
Azerbaijan
Bahamas
Bahamas
Bahrain
Bahrain

Figure 21: Preparation for finding different names in three time series covid19 data files and owid covid data file

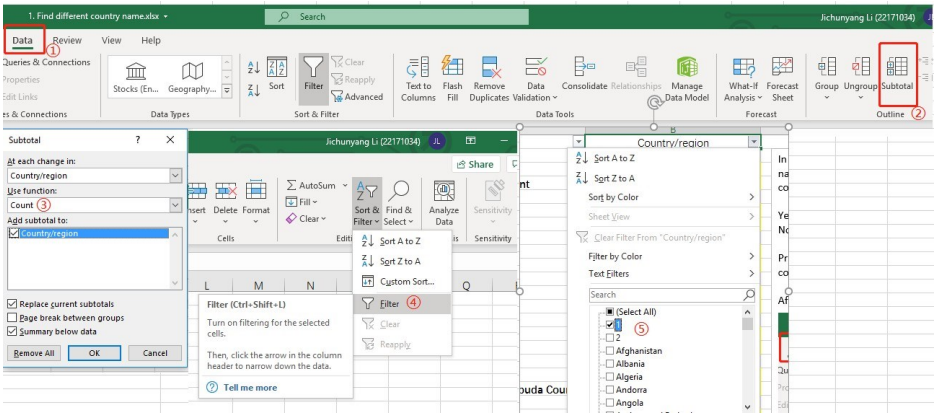


Figure 22: Looking up the different names in three time series covid19 data files and owid covid data file

Burma
Cabo Verde
Cape Verde
Congo
Congo (Brazzaville)
Congo (Kinshasa)
Democratic Republic of Congo
Diamond Princess
Holy See
Hong Kong
International
Korea, South
Micronesia
MS Zaandam
Myanmar
Palestine
South Korea
Taiwan
Taiwan*
Timor
Timor-Leste
United States
US
Vatican
West Bank and Gaza
World

Figure 23: Country names that need to be changed or removed

However, changing or removing these names relies on geographic knowledge and common political senses. See Table 6.

Names in owid covid data file	Names in time series covid19 data files	Final name	Reason
Myanmar	Burma	Myanmar	One country has two names; in 1989, the ruling military junta changed Burma to Myanmar.
Cape Verde	Cabo Verde	Cape Verde	One country has two names. Anyone is acceptable.
South Korea	Korea, South	South Korea	One country has two names. Anyone is acceptable.
United States	US	United States	US is the abbreviation of United States.
Palestine	West Bank and Gaze	Palestine	One country has two names. Anyone is acceptable.
Timor	Timor-Leste	Timor-Leste	One country has two names. Anyone is acceptable.
Vatican	Holy See	Holy See	One country has two names. Anyone is acceptable.
Democratic Republic of Congo	Congo (Kinshasa)	Congo (Kinshasa)	One country has two names. Anyone is acceptable.
Congo	Congo (Brazzaville)	Congo (Brazzaville)	One country has two names. Anyone is acceptable.
N/A	Diamond Princess	N/A	This was removed because this is a cruise ship rather than a country.
N/A	MS Zaandam	N/A	This was removed because this is a cruise ship rather than a country.
N/A	Micronesia	N/A	This country was removed because no deaths in this country, only one confirmed case, and one recovered case.
International	N/A	N/A	Removed.
World	N/A	N/A	Removed.
Taiwan	N/A	N/A	Removed. Taiwan is a province of China. The number of confirmed (deaths or recovered) cases will be summed into China.
Hong Kong	N/A	N/A	Removed. Hong Kong is a Special Administrative Region of the People's Republic of China.

Table 6: Operations for changing or removing country names

Provinces/States issue

There is not a total number for a country, such as Australia and China in three time series covid19 data files. For these countries, they only have the data of each Province (or State). See Table 7.

Provinces/States	Country	Operations
Australian Capital Territory	Australia	Create a new row named Australia, then sum the number of cases in all provinces/states. Finally, remove these provinces/states.
New South Wales		
Northern Territory		
Queensland		
South Australia		
Tasmania		
Victoria		
Western Australia		
Alberta	Canada	Create a new row named Canada, then sum the number of cases in all provinces/states. Finally, remove these provinces/states. (Only for confirmed cases and deaths files.)
British Columbia		
Manitoba		
New Brunswick		
Newfoundland and Labrador		
Northwest Territories		
Nova Scotia		
Nunavut		
Ontario		
Prince Edward Island		
Quebec		
Saskatchewan		
Yukon		
Diamond Princess	Canada	Removed. Because Diamond Princess and Grand Princess are cruise ships, repatriated travelers are not local cases. (Only for confirmed cases and deaths files.)
Grand Princess		
Repatriated Travellers		
Anhui	China	Create a new row named China, then sum the number of cases in all provinces/states. Finally, remove these provinces/states.
Beijing		
Chongqing		
Fujian		
Gansu		
Guangdong		
Guangxi		
Guizhou		
Hainan		
Hebei		
Heilongjiang		
Henan		
Hong Kong		
Hubei		
Hunan		
Inner Mongolia		
Jiangsu		
Jiangxi		
Jilin		
Liaoning		
Macau		
Ningxia		
Qinghai		
Shaanxi		
Shandong		
Shanghai		
Shanxi		
Sichuan		
Tianjin		
Tibet		
Xinjiang		
Yunnan		
Zhejiang		
Taiwan		

Table 7: Province/States issues (1)

Some counties, such as Denmark, United Kingdom and France, have

some dependent territories. based on assumption 2, these Provinces/States need to be removed or re-calculated. See Table 8.

Dependent Territory	Country	Operation
Anguilla	United Kingdom	Remove
Bermuda		
British Virgin Islands		
Cayman Islands		
Channel Islands		
Falkland Islands (Malvinas)		
Gibraltar		
Isle of Man		
Montserrat		
Saint Helena, Ascension and Tristan da Cunha		
Turks and Caicos Islands		
Aruba	Netherlands	
Bonaire, Sint Eustatius and Saba		
Curacao		
Sint Maarten		
French Guiana	France	
French Polynesia		
Guadeloupe		
Martinique		
Mayotte		
New Caledonia		
Reunion		
Saint Barthelemy		
Saint Pierre and Miquelon		
St Martin		
Wallis and Futuna		
Faroe Islands	Denmark	
Greenland		

Table 8: Province/States issues (2)

Dropping columns

In owid covid data file and three time series covid19 data files, not all columns are useful for this project; therefore, some columns need to be dropped. See Figure 24.

In three time series covid19 data files, Lat column and Long column need to be dropped since they are not used in this project.

Country/Region	22/01/2020	23/01/2020	24/01/2020	25/01/2020	26/01/2020	27/01/2020	28/01/202
Afghanistan	0	0	0	0	0	0	
Albania	0	0	0	0	0	0	
Algeria	0	0	0	0	0	0	
Andorra	0	0	0	0	0	0	
Angola	0	0	0	0	0	0	
Antigua and Barbuda	0	0				0	
Argentina	0	0				0	
Armenia	0	0				0	
Australia	0	0				5	
Austria	0	0				0	
Azerbaijan	0	0				0	
Bahamas	0	0	0	0	0	0	
Bahrain	0	0	0	0	0	0	
Bangladesh	0	0	0	0	0	0	
Barbados	0	0	0	0	0	0	
Belarus	0	0	0	0	0	0	
Belgium	0	0	0	0	0	0	
Belize	0	0	0	0	0	0	
Benin	0	0	0	0	0	0	

Remove:
The Lat column and Long column

Figure 24: Dropping columns in three time series covid19 data files

In the owid covid data file, only the region column (changed from the continent, South America and North America change to Americas), location column, population column and life_expectancy column are necessary. Other columns were dropped. See Figure 25.

Region	location	population	life_expectancy
Asia	Afghanistan	38,928,341.00	64.83
Europe	Albania	2,877,800.00	78.57
Africa	Algeria	43,851,043.00	76.88
Europe	Andorra	77,265.00	83.73
Africa	Angola	32,866,268.00	61.15
Americas	Antigua and Barbuda	97,928.00	77.02
Americas	Argentina	45,195,777.00	76.67
Asia	Armenia	2,963,234.00	75.09
Oceania	Australia	25,499,881.00	83.44
Europe	Austria	9,006,400.00	81.54
Asia	Azerbaijan	10,139,175.00	73
Americas	Bahamas	393,248.00	73.92
Asia	Bahrain	1,701,583.00	77.29
Asia	Bangladesh	164,689,383.00	72.59
Americas	Barbados	287,371.00	79.19
Europe	Belarus	9,449,321.00	74.79
Europe	Belgium	11,589,616.00	81.63
Americas	Belize	397,621.00	74.62
Africa	Benin	12,123,198.00	61.77
Asia	Bhutan	771,612.00	71.78
Americas	Bolivia	11,673,029.00	71.51
Europe	Bosnia and Herzegovina	3,280,815.00	77.4
Africa	Botswana	2,351,625.00	69.59
Americas	Brazil	212,559,409.00	75.88
Asia	Brunei	437,483.00	75.86
Europe	Bulgaria	6,948,445.00	75.05
Africa	Burkina Faso	20,903,278.00	61.58
Africa	Burundi	11,890,781.00	61.58

Figure 25: Dropping columns in owid covid data file

Missing values

In Excel, using “Find & Select” to find the missing values. In the owid covid data file, only one missing value, the life expectancy for Kosovo. See Figure 26.

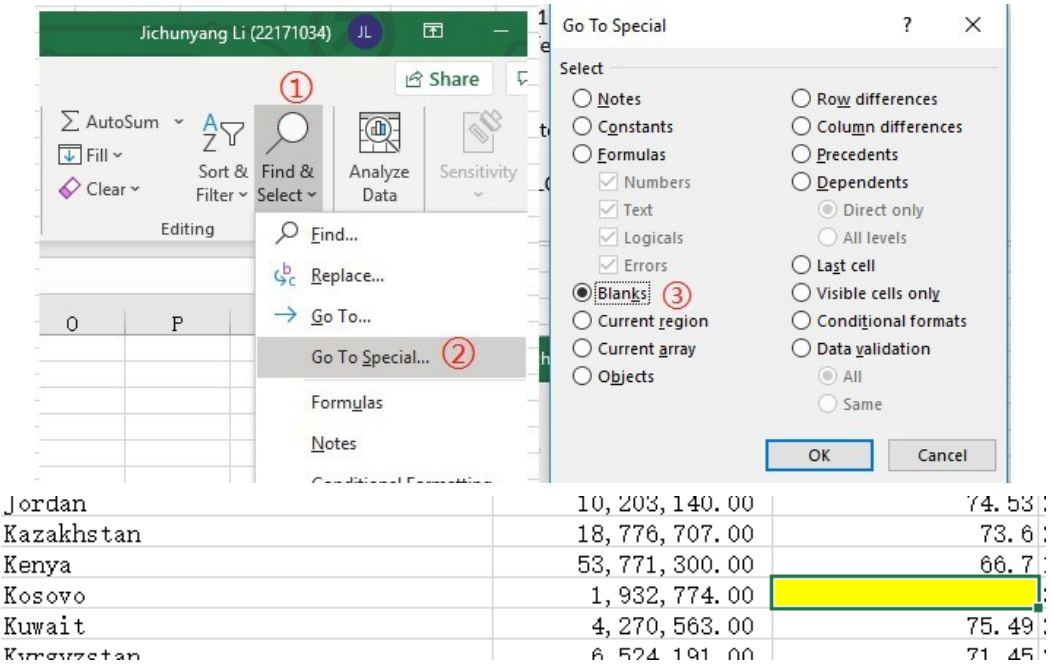


Figure 26: Missing values in owid covid data file

In order to solve this issue, the life expectancy for Kosovo has been researched; in 2018, the life expectancy for Kosovo is 72.2 (World Bank, 2018); therefore, fill 72.2 into the cell.

For three time series covid19 data files, the above method also is suitable. After checking, there are no missing values in the time series covid19 data files. See Figure 27.

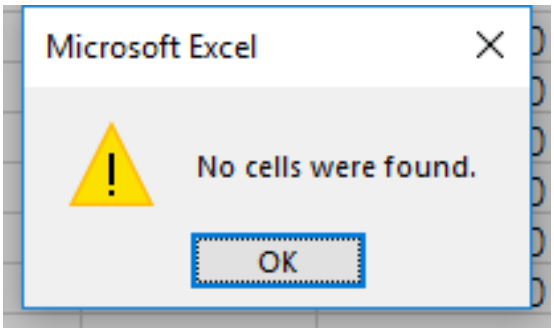
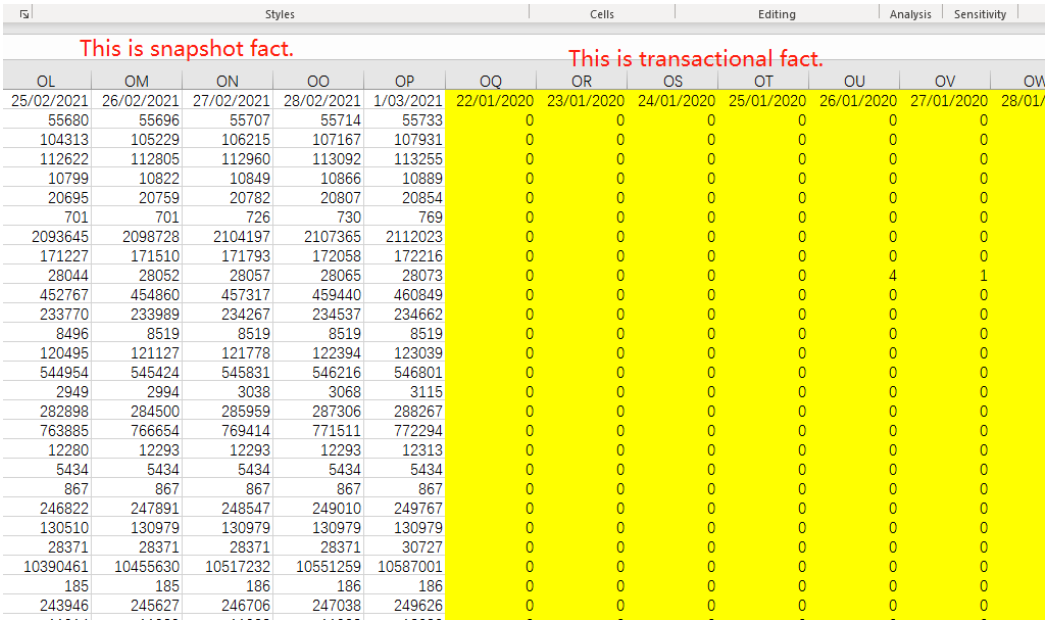


Figure 27: Missing values in three time series covid19 data files

Transactional facts generation

The three time series covid19 files are snapshot facts since they used cumulative cases from 1/22/2020 to 3/1/2021. In order to handle further data easier, and update the data in the data warehouse every day, using

transactional facts is more convenient than snapshot facts (directly insert daily cases into tables). For example, on 1/22/2020, there are no deaths in Afghanistan, there is no operation for 1/22/2020 (=B2, in a new cell). Then, using the number of 1/23/2020 minus the number of 1/22/2020 (=C2-B2, in a new cell), the result is the number of deaths on 1/23/2020. Repeat this process until the final day (3/1/2021). Finally, we can gain the number of cases each day; after that, using the new daily data overwrites the original snapshot facts. See Figure 28 and 29.



OL	OM	ON	OO	OP	OQ	OR	OS	OT	OU	OV	OW
25/02/2021	26/02/2021	27/02/2021	28/02/2021	1/03/2021	22/01/2020	23/01/2020	24/01/2020	25/01/2020	26/01/2020	27/01/2020	28/01/
55680	55696	55707	55714	55733	0	0	0	0	0	0	
104313	105229	106215	107167	107931	0	0	0	0	0	0	
112622	112805	112960	113092	113255	0	0	0	0	0	0	
10799	10822	10849	10866	10889	0	0	0	0	0	0	
20695	20759	20782	20807	20854	0	0	0	0	0	0	
701	701	726	730	769	0	0	0	0	0	0	
2093645	2098728	2104197	2107365	2112023	0	0	0	0	0	0	
171227	171510	171793	172058	172216	0	0	0	0	0	0	
28044	28052	28057	28065	28073	0	0	0	0	4	1	
452767	454860	457317	459440	460849	0	0	0	0	0	0	
233770	233989	234267	234537	234662	0	0	0	0	0	0	
8496	8519	8519	8519	8519	0	0	0	0	0	0	
120495	121127	121778	122394	123039	0	0	0	0	0	0	
544954	545424	545831	546216	546801	0	0	0	0	0	0	
2949	2994	3038	3068	3115	0	0	0	0	0	0	
282898	284500	285959	287306	288267	0	0	0	0	0	0	
763885	766654	769414	771511	772294	0	0	0	0	0	0	
12280	12293	12293	12293	12313	0	0	0	0	0	0	
5434	5434	5434	5434	5434	0	0	0	0	0	0	
867	867	867	867	867	0	0	0	0	0	0	
246822	247891	248547	249010	249767	0	0	0	0	0	0	
130510	130979	130979	130979	130979	0	0	0	0	0	0	
28371	28371	28371	28371	30727	0	0	0	0	0	0	
10390461	10455630	10517232	10551259	10587001	0	0	0	0	0	0	
185	185	186	186	186	0	0	0	0	0	0	
243946	245627	246706	247038	249626	0	0	0	0	0	0	

Figure 28: Transactional fact (1)

	A	B	C	D	E	F	G	H	I	J
1	Country/Region	22/01/2020	23/01/2020	24/01/2020	25/01/2020	26/01/2020	27/01/2020	28/01/2020	29/01/2020	30/01/2020
2	Afghanistan	0	0	0	0	0	0	0	0	0
3	Albania	0	0	0	0	0	0	0	0	0
4	Algeria	0	0	0	0	0	0	0	0	0
5	Andorra	0	0	0	0	0	0	0	0	0
6	Angola	0	0	0	0	0	0	0	0	0
7	Antigua and Barbuda	0	0	0	0	0	0	0	0	0
8	Argentina	0	0	0	0	0	0	0	0	0
9	Armenia	0	0	0	0	0	0	0	0	0
10	Australia	0	0	0	0	4	1	0	1	3
11	Austria	0	0	0	0	0	0	0	0	0
12	Azerbaijan	0	0	0	0	0	0	0	0	0
13	Bahamas	0	0	0	0	0	0	0	0	0
14	Bahrain	0	0	0	0	0	0	0	0	0
15	Bangladesh	0	0	0	0	0	0	0	0	0
16	Barbados	0	0	0	0	0	0	0	0	0
17	Belarus	0	0	0	0	0	0	0	0	0
18	Belgium	0	0	0	0	0	0	0	0	0
19	Belize	0	0	0	0	0	0	0	0	0
20	Benin	0	0	0	0	0	0	0	0	0
21	Bhutan	0	0	0	0	0	0	0	0	0
22	Bolivia	0	0	0	0	0	0	0	0	0
23	Bosnia and Herzegovina	0	0	0	0	0	0	0	0	0
24	Botswana	0	0	0	0	0	0	0	0	0
25	Brazil	0	0	0	0	0	0	0	0	0
26	Brunei	0	0	0	0	0	0	0	0	0
27	Bulgaria	0	0	0	0	0	0	0	0	0

Figure 29: Transactional fact (2)

Negative values issue

In three time series covid19 data files, there are some negative values. On 4/11/2020 in France, the number of confirmed cases decreased by 47301. However, this issue is common in a pandemic. For example, the Western Australia Department of Health announced that historical cases would be removed from the State total since these cases are no longer infectious and present no risk to the community on 31 July 2020 (Government of Western Australia Department of Health, 2020). For deaths, on 14 February 2020, 108 deaths were removed in Hubei Province since these deaths were counted repeatedly (National Heath Commission of the People's Republic of China, 2020). Besides, some positive RT-PCR test results occurred in patients recovered from Covid-19 (Lan et al., 2020). In other words, some recovered patients were re-infected by Covid-19. From these points of view, the negative values could appear in three time series covid19 data files. Thus, the negative values that appear in three time series covid19 data files are not outliers. From this point of view, they cannot be changed or removed. However, in the recovered cases files, some countries should be focused on, such as Belgium, the United States and the United Kingdom. For these countries, the number of recovered cases suddenly reduced to 0 on one day; currently, no evidence explains this issue, the reasons for this issue in these countries will be researched in the future.

Generation for Fact Table and Dimension Tables

After all cleaning processes, the selecting data need to be filled into dimension tables and fact table. Before that, the order of countries in four CSV files need to be sorted, the cell value is country names, and the order is A to Z. Therefore, the order of countries is the same in all CSV files. In Excel, "VLOOKUP" is a useful and powerful function that can be used to complete this step. First, in the fact table, we need to assign the CountryID and TimeID; for

this step, we can use:

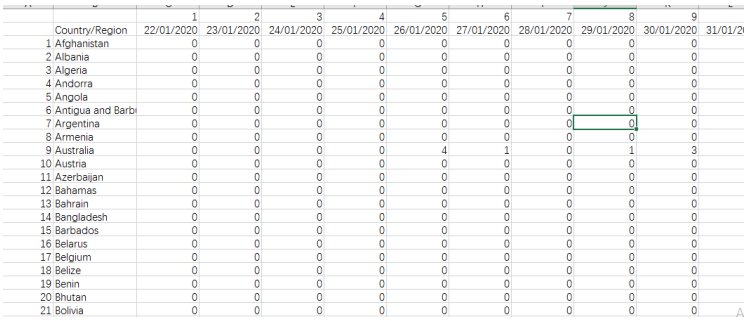
$=\mathrm{INT}(\mathrm{ROW}(A405)/405)$

(Set CountryID)

$=\mathrm{MOD}(\mathrm{ROW}(A405),405)+1$

(Set TimeID)

Since each country, its records have 405 days. Using this function, the CountryID can be filled into a column (76140 cells). Secondly, set the ID for each country (1-188) and date (1-405) in time series covid19 data files (for owid covid data file, only need to set the CountryID). This step aims to provide a target for the VLOOKUP function. See Figure 30.



		1	2	3	4	5	6	7	8	9	
	Country/Region	22/01/2020	23/01/2020	24/01/2020	25/01/2020	26/01/2020	27/01/2020	28/01/2020	29/01/2020	30/01/2020	31/01/2
1	Afghanistan	0	0	0	0	0	0	0	0	0	
2	Albania	0	0	0	0	0	0	0	0	0	
3	Algeria	0	0	0	0	0	0	0	0	0	
4	Andorra	0	0	0	0	0	0	0	0	0	
5	Angola	0	0	0	0	0	0	0	0	0	
6	Antigua and Barb	0	0	0	0	0	0	0	0	0	
7	Argentina	0	0	0	0	0	0	0	0	0	
8	Armenia	0	0	0	0	0	0	0	0	0	
9	Australia	0	0	0	0	4	1	0	1	3	
10	Austria	0	0	0	0	0	0	0	0	0	
11	Azerbaijan	0	0	0	0	0	0	0	0	0	
12	Bahamas	0	0	0	0	0	0	0	0	0	
13	Bahrain	0	0	0	0	0	0	0	0	0	
14	Bangladesh	0	0	0	0	0	0	0	0	0	
15	Barbados	0	0	0	0	0	0	0	0	0	
16	Belarus	0	0	0	0	0	0	0	0	0	
17	Belgium	0	0	0	0	0	0	0	0	0	
18	Belize	0	0	0	0	0	0	0	0	0	
19	Benin	0	0	0	0	0	0	0	0	0	
20	Bhutan	0	0	0	0	0	0	0	0	0	
21	Bolivia	0	0	0	0	0	0	0	0	0	

Figure 30: Set CountryID and TimeID for VLOOKUP function (an example of confirmed cases)

After that, we need to use the VLOOKUP function to find and insert data automatically.

$=\mathrm{VLOOKUP}(A1,\text{time\_series\_covid19\_confirmed\_global.csv!\$A:\$OQ},B1+2,0)$
(look up the values from confirmed cases CSV)

$=\mathrm{VLOOKUP}(A1,\text{time\_series\_covid19\_deaths\_global.csv!\$A:\$OQ},B1+2,0)$ (look up the values from death CSV)

$=\mathrm{VLOOKUP}(A1,\text{time\_series\_covid19\_recovered\_global.csv!\$A:\$OQ},B1+2,0)$
(look up the values from recovered cases CSV)

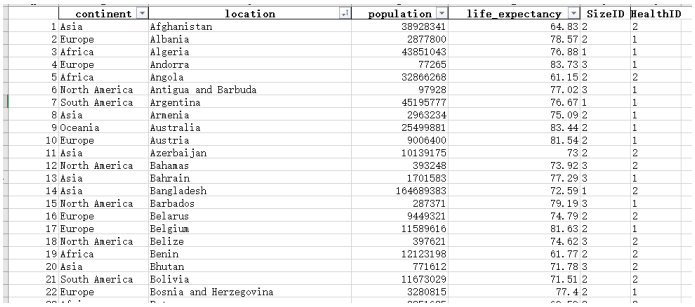
Same as time series covid19 data files, we also use VLOOKUP to lookup values from owid covid data and then insert this data into the fact table; however, we need to use IF function first, since we need to generate the SizeID and HealthID, which depend on the population and life expectancy. This project

uses nested IF statements for the population since there are 3 types of counties (1 = Large, 2 =Medium and 3 = Small). See Figure 31.

=IF(D2>=40000000,"1",IF(AND(D2>2000000,D2<40000000),"2",IF(D2<=2000000,"3")))

For HealthID, the simple IF statement is enough (1 = Greater than 75, and 2 = Less than or equal to 75).

=IF(E2>75,"1","2")



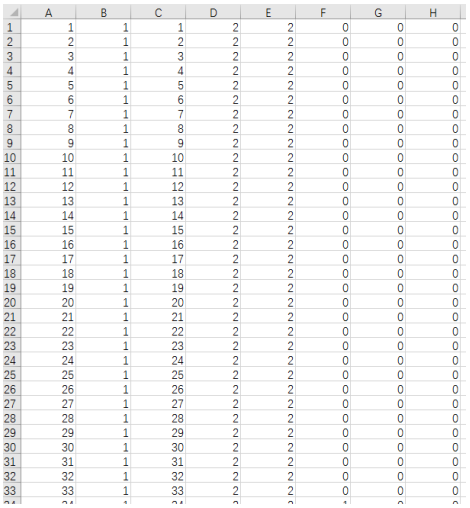
	continent	location	population	life_expectancy	SizeID	HealthID
1	Asia	Afghanistan	38928341	64.83	2	2
2	Europe	Albania	2877800	78.57	2	1
3	Africa	Algeria	43851043	76.88	1	1
4	Europe	Andorra	77265	83.73	3	1
5	Africa	Angola	32866268	61.15	2	2
6	North America	Antigua and Barbuda	97928	77.02	3	1
7	South America	Argentina	45195777	76.67	1	1
8	Asia	Armenia	2963234	75.09	2	1
9	Oceania	Australia	25499881	83.44	2	1
10	Europe	Austria	9006400	81.54	2	1
11	Asia	Azerbaijan	10139175	73.2		2
12	North America	Bahamas	393248	73.92	3	2
13	Asia	Bahrain	1701583	77.29	3	1
14	Asia	Bangladesh	164689383	72.59	1	2
15	North America	Barbados	287371	79.19	3	1
16	Europe	Belarus	9449321	74.79	2	2
17	Europe	Belgium	11589616	81.63	2	1
18	North America	Belize	397621	74.62	3	2
19	Africa	Benin	12123198	61.77	2	2
20	Asia	Bhutan	771612	71.78	3	2
21	South America	Bolivia	11673029	71.51	2	2
22	Europe	Bosnia and Herzegovina	3280815	77.4	2	1

Figure 31:SizeID and HealthID in owid covid data

Then we can use the VOOLUP function:

=VLOOKUP(A1,owid-covid-data.xlsx!$A:$G, 6,0) (lookup the SizeID from owid covid data)
=VLOOKUP(A1,owid-covid-data.xlsx!$A:$G, 7,0) (lookup the HealthID from owid covid data)

Finally, insert the primary key (InfoID) into the fact table. The final fact table. See Figure 32.



	A	B	C	D	E	F	G	H
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
19	19	1	19	2	2	0	0	0
20	20	1	20	2	2	0	0	0
21	21	1	21	2	2	0	0	0
22	22	1	22	2	2	0	0	0
23	23	1	23	2	2	0	0	0
24	24	1	24	2	2	0	0	0
25	25	1	25	2	2	0	0	0
26	26	1	26	2	2	0	0	0
27	27	1	27	2	2	0	0	0
28	28	1	28	2	2	0	0	0
29	29	1	29	2	2	0	0	0
30	30	1	30	2	2	0	0	0
31	31	1	31	2	2	0	0	0
32	32	1	32	2	2	0	0	0
33	33	1	33	2	2	0	0	0

Figure 32:Final fact table (part of)

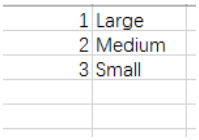
For Dim_Population and Dim_Health_Condition, we can fill them by manual method. Since only two columns and four rows in Dim_Population and two columns and three rows in Dim_Health_Condition. For Dim_Country_Location, the data can be directly copied from the owid covid data file. For Dim_Time we can use the “automatically fill” and “Month” function to automatically produce 405 days, 15 months and 2 years. See Figure 33, 34, 35 and 36.

1	Afghanistan	Asia
2	Albania	Europe
3	Algeria	Africa
4	Andorra	Europe
5	Angola	Africa
6	Antigua and Barbuda	Americas
7	Argentina	Americas
8	Armenia	Asia
9	Australia	Oceania
10	Austria	Europe
11	Azerbaijan	Asia
12	Bahamas	Americas
13	Bahrain	Asia
14	Bangladesh	Asia
15	Barbados	Americas
16	Belarus	Europe
17	Belgium	Europe
18	Belize	Americas
19	Benin	Africa
20	Bhutan	Asia
21	Bolivia	Americas
22	Bosnia and Herzegovina	Europe
23	Botswana	Africa
24	Brazil	Americas

Figure 33: Dim_Country_Location

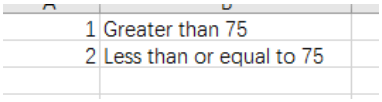
1	22/01/2020	2020-01	2020 Q1	2020
2	23/01/2020	2020-01	2020 Q1	2020
3	24/01/2020	2020-01	2020 Q1	2020
4	25/01/2020	2020-01	2020 Q1	2020
5	26/01/2020	2020-01	2020 Q1	2020
6	27/01/2020	2020-01	2020 Q1	2020
7	28/01/2020	2020-01	2020 Q1	2020
8	29/01/2020	2020-01	2020 Q1	2020
9	30/01/2020	2020-01	2020 Q1	2020
10	31/01/2020	2020-01	2020 Q1	2020
11	1/02/2020	2020-02	2020 Q1	2020
12	2/02/2020	2020-02	2020 Q1	2020
13	3/02/2020	2020-02	2020 Q1	2020
14	4/02/2020	2020-02	2020 Q1	2020
15	5/02/2020	2020-02	2020 Q1	2020
16	6/02/2020	2020-02	2020 Q1	2020
17	7/02/2020	2020-02	2020 Q1	2020
18	8/02/2020	2020-02	2020 Q1	2020
19	9/02/2020	2020-02	2020 Q1	2020
20	10/02/2020	2020-02	2020 Q1	2020
21	11/02/2020	2020-02	2020 Q1	2020
22	12/02/2020	2020-02	2020 Q1	2020
23	13/02/2020	2020-02	2020 Q1	2020
24	14/02/2020	2020-02	2020 Q1	2020

Figure 34: Dim_Time



1	Large
2	Medium
3	Small

Figure 35: Dim_Population



1	Greater than 75
2	Less than or equal to 75

Figure 36: Dim_Health_Condition

Data Loading

After data extraction and transformation, the data needs to be loaded into the Microsoft SQL Server Management Studio (SSMS) by SQL codes.

Insert the data into the created tables

In this part, the main steps include: 1) set the environment variable to data path and use the targeted database; 2) insert data into tables. In this section, in order to prevent SSMS from assigning identity values while bulk importing data rows into a table, “KEEPIDENTITY” was used. As a result, SSMS uses the identity values in the data file, rather than assigning an identity value. For the full SQL script, please refer to the submitted SQL file (Insert Data.sql).

OLAP Cube Design

In order to answer the four business queries, a new multi-dimensional project was created in Microsoft Visual Studio. Then a cube was created; they are four dimensions: Health Condition.dim, Country Location.dim, Time.dim and Population.dim. These dimensions need to be populated with unique values of each dimension. Finally, the concept hierarchies were created. See Table 9,10, 11 and 12.

Hierarchy
Life Expectancy

Table 9: Hierarchy for Health Condition.dim

Hierarchy
Region
Country Name

Table 10: Hierarchy for Country Location.dim

Hierarchy
Year
Quarter
Month
Day

Table 11: Hierarchy for Time.dim

Hierarchy
Country Size

Table 12: Hierarchy for Population.dim

After that, the data can be visualised in Power BI.

Results, Visualisation and Discussion

For business query 1, in 2020, a total of 28425 confirmed cases in Australia. See Figure 37.

Figure 37: 28425 confirmed cases in Australia in 2020

The number of confirmed cases in each quarter of 2020 in Australia see Figure 38.

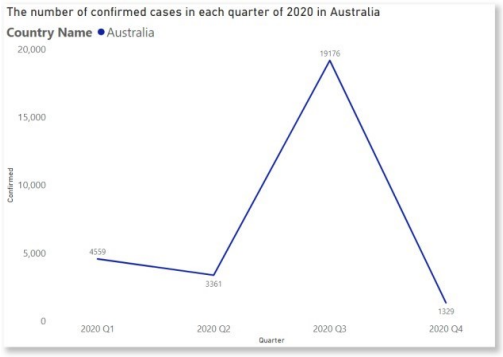


Figure 38: The number of confirmed cases in each quarter of 2020 in Australia

For each month, there are new confirmed cases in Australia. More than half of the confirmed cases occurred in July and August. From March to September 2020, the confirmed cases dramatically increased; after September 2020, the curve was flattened. See Figure 39. Based on Figure 38 and 39, the Covid-19 has been initially controlled in quarter 4 of 2020 in Australia.

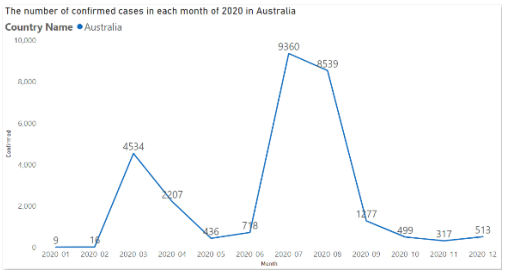


Figure 39: The number of confirmed cases in each month of 2020 in Australia

In the Americas, a total of 2874960 recovered cases in September 2020. See Figure 40.

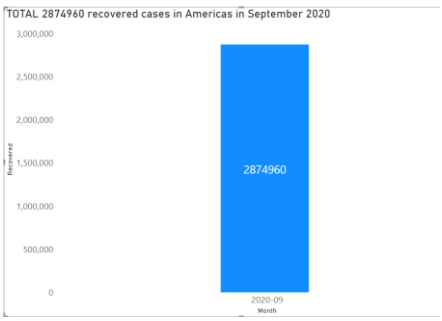


Figure 40: Total 2874960 recovered cases in the Americas in September 2020

In three major countries, the United States, Canada and Mexico, in North America, September 2020, United States had the largest recovered cases (655863). Mexico had the second largest recovered cases (131785), and Canada had the third largest recovered cases (21161). However, that does not mean the United States did well in the pandemic; since the largest recovered cases could be caused by the largest confirmed cases. See Figure 41.

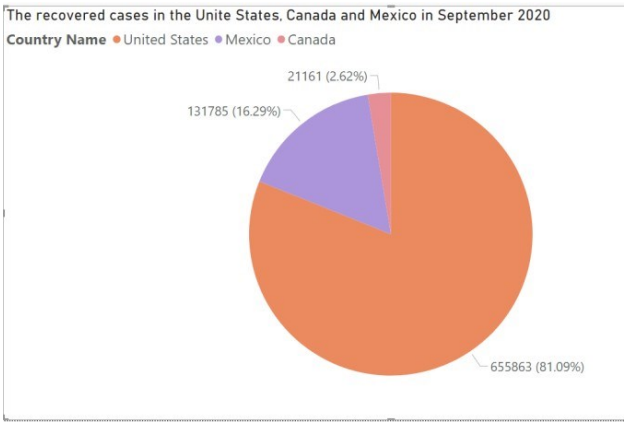


Figure 41: The recovered cases in the United States, Canada and Mexico September 2020

In 2020, a total of 1825020 deaths worldwide because of Covid-19. See Figure 42.

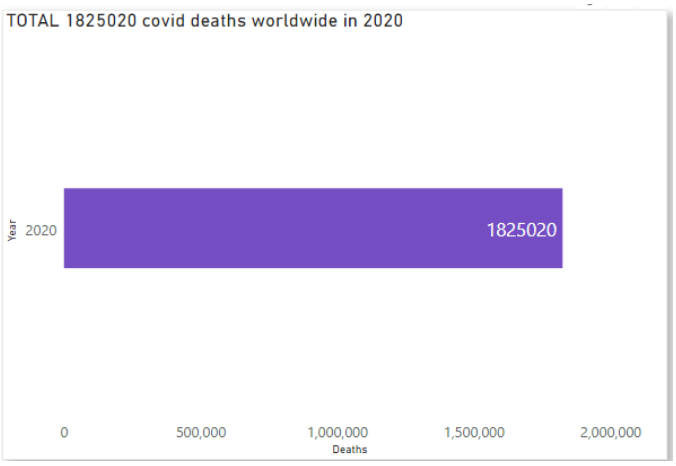


Figure 42: Total 1825020 covid deaths worldwide in 2020

Large countries have the largest number of deaths in three different country sizes since these countries have a large population base. See Figure 43.

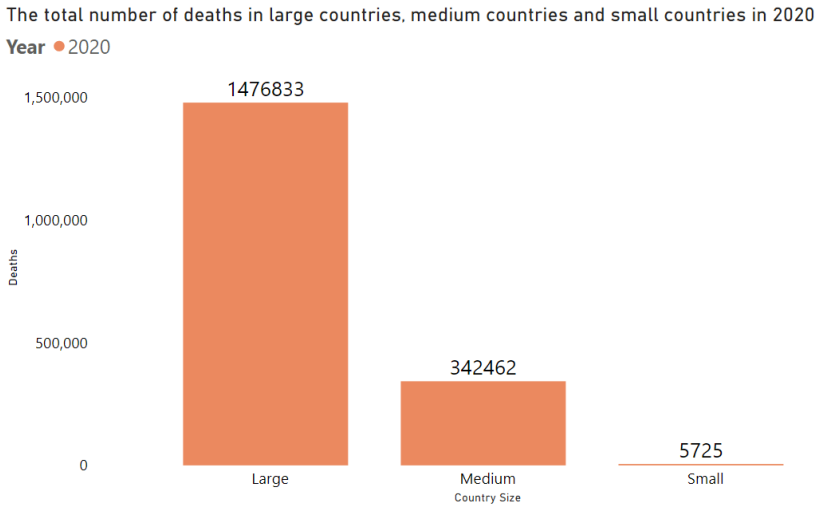


Figure 43: The total number of deaths in large counties, medium counties and small counties in 2020

Consider the recovery rate and health condition for countries (in 2020 and 2021); the countries with a life expectancy greater than 75 (39334516/87087107=45.17%) are poor than the countries with life expectancy less than or equal to 75 (25237994/27242949=92.64%). However, the reasons for this issue are various, for example, the countries with poor health condition could not record all confirmed cases, or they could not test people who have some symptoms. See Figure 44.

Recovery rate and life expectancy

Life Expectancy	Recovered	Confirmed
Greater than 75	39,334,516.00	87,087,107.00
Less than or equal to 75	25,237,994.00	27,242,949.00

Figure 44: Recovery rate and life expectancy

Further Research

In addition to the four business queries, task 8 requires the project to design a galaxy schema that includes two fact tables, one fact table is for the number of confirmed, recovered, and death cases and the other fact table is for government measures. Therefore, the acaps covid19 government measures need to be transformed. Since this task did not ask us to insert the data, selecting the useful columns is enough, data cleansing will be not applied for

this task.

For a galaxy schema, multiple fact tables share dimension tables; in other words, these fact tables are connected further with multiple dimension tables that are normalised (Pedamkar, 2020). The galaxy schema. See Figure 45.

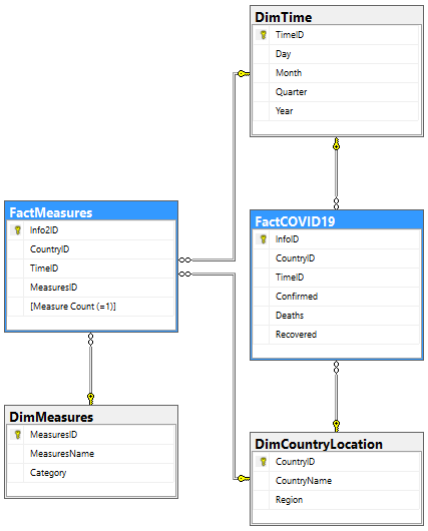


Figure 45: Galaxy schema

In the acsps covid19 government measures file, country, region, category, measure and date implemented are acceptable for task 8. This acsps coivd19 government measures table is an event factless fact table. Since in this table, the events recorded a set of dimensional entities (measures). Facless fact tables include factless fact table for events and factless fact table for conditions (Kimball Group, 2016). The measure for event factless fact table usually is count (=1). Dimension tables should include dim_country, dim_time and dim_measures. See Table 13.

Table name	Dimensions	Primary key
Dimcountry	Region and Country name	CountryID
DimTime	Year, Quarter, Month and Day	TimeID
DimMeasures	Category and measures	MeasuresID

Table 13: Dimension tables for task 8

Fact tables include FactCovid19 and Fact_Measures. See Table 14.

Table name	Foreign keys	Measurements	Primary key
FactCovid19	CountryID and TimeID	Confirmed, Deaths, and Recovered	InfoID
FactMeasures	MeasureID, CountryID, TimeID	Measure Count (=1)	Info2ID

Table 14: Fact tables for task 8

In SSMS, a new database was created for task 8. In this part, the important steps include: 1) clean the environment; 2) create a database named Project_1_Covid19_2 and use it and create two fact tables and three dimension tables. At the same time, the data types for each column in these tables, since all measures are integers, the data type for confirmed, deaths and recovered are int; 3) add relations between fact tables and dimension tables (through foreign keys). The SQL script please see submitted file (Task_8.sql).

Based on the schema and tables, the possible business query is that

In Europe in August 2020, was the number of deaths in the country with the largest number of measures implemented less than the country with the second-largest number of measures implemented?

Conclusion

In conclusion, this report uses some data from five CSV files to build a data warehouse to answer four business queries. The schema for this data warehouse is star schema, and data was cleaned by Excel. Data was loaded into SSMS by SQL scripts and then produced cubes and their hierarchies in Visual Studio. Finally, the results were visualised by Power BI.

In the further research section, a galaxy schema was provided. In this galaxy schema, two fact tables and three dimension tables were produced. The further researched business query is: In Europe in August 2020, was the number of deaths in the country with the largest number of measures implemented less than the country with the second-largest number of measuresimplemented?

References

Agapito, G., Zucco, C., & Cannataro, M. (2020). COVID-WAREHOUSE: A Data Warehouse of Italian COVID-19, Pollution, and Climate Data. *International Journal of Environmental Research and Public Health*, 17(15), 5596–5617.

El-Sappagh, S. H. A., Hendawi, A. M. A., & El Bastawissy, A. H. (2011). A proposed model for data warehouse ETL processes. *Journal of King Saud University-Computer and Information Sciences*, 23(2), 91-104.

Government of Western Australia Department of Health. (2020). *COVID-19 update – 31 July 2020*. Government of Western Australia Department of Health. https://ww2.health.wa.gov.au/Media-releases/2020/COVID19- update-31-July-2020

Kao, W., Hung, J., & Hsu, V. (2008). Using Data Mining in MURA Graphic Problems. *Journal of Software*, 3(8), 73-79.

Kimball Group. (2016). *Factless fact tables*. Kimball Group. https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dimensional-modeling-techniques/factless- fact-table/

Kimball, R., & Ross, M. (2011). *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons.

Lan, L., Xu, D., Ye, G., Xia, C., Wang, S., Li, Y., & Xu, H. (2020). Positive RT-PCR test results in patients recovered from COVID-19. *Jama*, 323(15), 1502-1503.

Levene, M., & Loizou, G. (2003). Why is the snowflake schema a good data warehouse design?. *Information Systems*, 28(3), 225-240.

Mohammed, K. (2019). Data Warehouse Design and Implementation Based on Star Schema vs. Snowflake Schema. *International Journal of Academic Research in Business and Social Sciences*, 9(14), 25-38.

National Heath Commission of the People's Republic of China. (2020). *Covid- 19 situation report by 13 February 2020*. National Heath Commission of the People's Republic of China. http://www.nhc.gov.cn/xcs/yqtb/202002/553ff43ca29d4fe88f3837d49d6b6ef1.shtml

Pedamkar, P. (2020). *Galaxy schema*. EDUCBA. https://www.educba.com/galaxy-schema/

Sidi, E., El Merouani, M., Amin, E., & Abdelouarit, A. (2016). Star schema advantages on data warehouse: using bitmap index and partitioned fact tables. *International Journal of Computer Applications*, 134(13), 11-13.

World Bank. (2018). *Life expectancy at birth, total (years) -Kosovo*. Word Bank.

https://data.worldbank.org/indicator/SP.DYN.LE00.IN?locations=XK

World Health Organisation. (2020a). *Novel Coronavirus (2019-nCoV) Situation Report -1*. World Health Organisation. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200121-sitrep-1-2019-ncov.pdf?sfvrsn=20a99c10_4

World Health Organisation. (2020b). *Novel Coronavirus (2019-nCoV) Situation Report -52*. World Health Organisation. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200312-sitrep-52-covid-19.pdf?sfvrsn=e2bfc9c0_4

World Health Organisation. (2020c). *COVID-19 Weekly Epidemiological Update*. World Health Organisation. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20210316_weekly_epi_update_31.pdf?sfvrsn=c94717c2_17&download=true