111-1 Introduction to Database Systems - HW2

Governments are taking a wide range of measures to tackle the COVID-19 outbreak. Systematic information on which measures governments take, and when, can help us understand the responses in a consistent way, aiding efforts to fight the pandemic.

Oxford COVID-19 Government Response Tracker aims to track and compare worldwide government responses to the coronavirus rigorously and consistently. They collect information on common policy responses, score the stringency of such measures, and aggregate these into a Stringency Index.

Please use the data here. The codebook and coding interpretation can be downloaded from their official website (mainly on their GitHub repo), the files that should be included in the homework are:
(Note: Please use the data here, not on their official website because the data on their website is updated every day.)

1. OxCGRT_nat_latest.csv
   reports country/territory- and state-level data presented in "country/territory-day" format (or "state-day" as the case may be), with a list of all indicators for each country/territory as a single row each day.
2. country-and-continent-codes-list-csv.csv Country code and continents mapping table

You might notice that importing these files into the database directly is not a good idea, with redundant attributes and unnormalized forms (UNF). Please re-design (decompose) the original data schema to fit the request for good Relational Designs (at least 3NF, using BCNF or 4NF if it is possible) by understanding the data sources, designing the ER model for it, drawing the ER diagram, and translating it to relation schemas. Then create the whole database accordingly in AWS RDS.

HW2 requests

1. Draw the ER diagram to show **your design** of ER model for the data (2 csv files) mentioned above (**30**pts in total)
   a. Including entity sets (10pts) and relationship sets (10pts), with or without attributes
   b. Add constraints for the model (10pts)
   c. Avoid redundancy
   d. Please redesign the structure of these tables, do not use the original format directly
   e. You should keep the name consistent throughout the databases, so please modify the country name in **“country-and-continent-codes-list-csv.csv”**
   f. Some relations might have too many attributes and is hard to include all of them in the diagram. If so, you can use “a range of attributes”, such as C1 to C10, or the other ways, to list attributes.

2. Launch a database service with Amazon RDS with PostgreSQL engines (using **Learner Lab** (or new link https://www.awsacademy.com/vforcesite/LMS_Login), we have step-by-step the instruction here, please check it if needed), and create the "oxcgrt" database (**10pts**)
3. Translating your ER model to relations. Include all keys, foreign keys, and uniqueness constraints. Also, please include functional dependency F (**40pts**)
    a. If your table is in UNF, 1NF, or 2NF, please decompose it. (10pts)
    b. Among your tables, what are the normal forms of these tables? Test the normal form using the algorithms (with functional dependencies) introduced in Ch7 (10pts)
    c. List function dependency (listing the simplest set only is ok) in each relation (10pts)
    d. Create the tables you designed in the "oxcgrt" database on Amazon RDS. (10pts)
4. Try to write queries in Query Tool to (**20pts**)
    a. List the country names, continents, and date of the countries with the highest Stringency Index on 2022/06/01, 2021/06/01, and 2020/06/01 in each continent. If there are multiple Stringency Indexes in one country per day, please aggregate them with a reasonable approach and explain why you choose the method (if there is only one Stringency Index in one country per day, then you can use StringencyIndex_Average_ForDisplay directly). Please use StringencyIndex_Average_ForDisplay column in this Query. (10pts)
    b. We define an indicator to show the "over Stringency index" as ((maybe aggregated) Stringency Index)/(7-day moving average of daily confirmed new cases). Please list the country names, continents, "over Stringency index", and date with the highest "over Stringency index" on 2022/06/01, 2021/06/01, and 2020/06/01 in each continent. Please use StringencyIndex_Average_ForDisplay column in this Query. (10pts)
5. (Bonus-20 points) After executing the query above, write a series functions with AWS Lambda to **update the tables** from official OxCGRT GitHub sources (References **may** help, but the reference is using MySQL with different tasks and tables). If you can only finish part of the request, please also include the part you have done in your submission. Include the code and the results after execution.
6. **Please include your work in HW2 format below**, including ER diagram (if it is hard to include in the .pdf file, you can submit the diagram separately), relational schema with normal form test, and function dependency in a PDF file. Please also submit the SQL commands for data retrieval (submit .sql file separately)
7. submit the PDF and SQL files to the E3 system HW2 section by **2022/12/08 23:59**

HW2 format:

1. ER diagram with entity sets and relationship sets, with or without attributes. Add constraints if needed. (30pts) (if it is hard to include your ER diagram in the .pdf file, you can submit the diagram separately)
   Ans:

2. Provide print screens of the 1) AWS RDS launch page, and 2) the way you connect to the AWS RDS (PostgreSQL console tool, pgAdmin, or other IDE's connection page, with the same IP or URL with your AWS RDS) (10pts)
   Ans:

3. Please provide the schema after decomposition, of each table, and a print screen to show that the tables have been created in your database on AWS RDS. (10+10pts)
   Ans:

4. Clearly indicate the level of normal form, test the level of normal form for each table (10pts)
   Ans:

5. List the functional dependency of each table. (10pts)
   Ans:

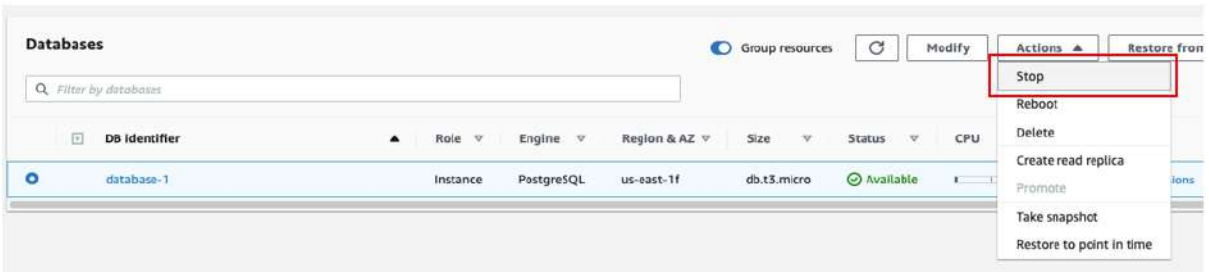
6. The SQL statements (in .sql file) and output results of 4a (10pts)
   Ans:

7. The SQL statements (in .sql file) and output results of 4b (10pts)
   Ans:

8. Database auto-update (from the data provider's GitHub repo) strategy and implementation (bonus 20pts)
   Ans:

Important!

Please turn off (stop) the RDS service and End the Learner Lab when you are not working with the database to save money. You can always re-start the database afterward, and the data is still there.



Q&A:

Q: ER first normalization later? or not

A:
The process of designing a database:

1. Understand the requirements (like what you have from the coding book, then you might know functional dependency here)
2. Design ER-diagram
3. Transfer the ER diagram to the database schema
4. Check if these relations fit BCNF or 3NF, if not, perform decomposition
5. Fine-tune the schema for a specific application.

In this HW, the different part is that you have a huge table on hand, so some of you might think you can do the decomposition first. That is totally fine, but you still need to identify the functional dependencies in the data to perform the decomposition.
在這個作業中，因為你們已經拿到一個大表格，有些人就想要想做normalization再畫ER-diagram，這也是可以的，只是你還是要從觀察資料/coding book 去推敲可能的functional dependency，才能接著做decomposition.

Q: What is the meaning of C1M or C1V or the others?
A:
As mentioned in the coding book, C1M means School closing for the majority of people. The "flag" means its geographic scope. If the status is for general people, then the flag will be one. The meaning of differentiation is here: the 'Majority' value reflects either the policy for everyone (E), or the policy applying to the majority of people fully vaccinated in a country, using vaccination rate data to determine if this is the vaccinated or non-vaccinated part of the population. If there are differentiated policies in place, we report the NV value if vaccination rates are under 50% in that jurisdiction, and V value if vaccination rates are above 50%, hence the majority of the population.

Q: What is the meaning of V2D?
A:
V2D is V2D_Medically/ clinically vulnerable (Non-elderly)
Translation from the coding book: 針對臨床易受傷害族群（老人除外）的疫苗策略，從以下這些類別中，如果有1~2類有打，就紀錄為1。三類以上有打，就紀錄為2。如果沒特別紀錄，但是16歲到80歲都能打的話，也紀錄為2（就是大家都有打了，當然易受傷害族群也有打）
V2_Clinically vulnerable/chronic illness/significant underlying health condition (excluding elderly and disabled)
V2_Disabled people
V2_Pregnant people
V2_People living with a vulnerable/shielding person or other priority group

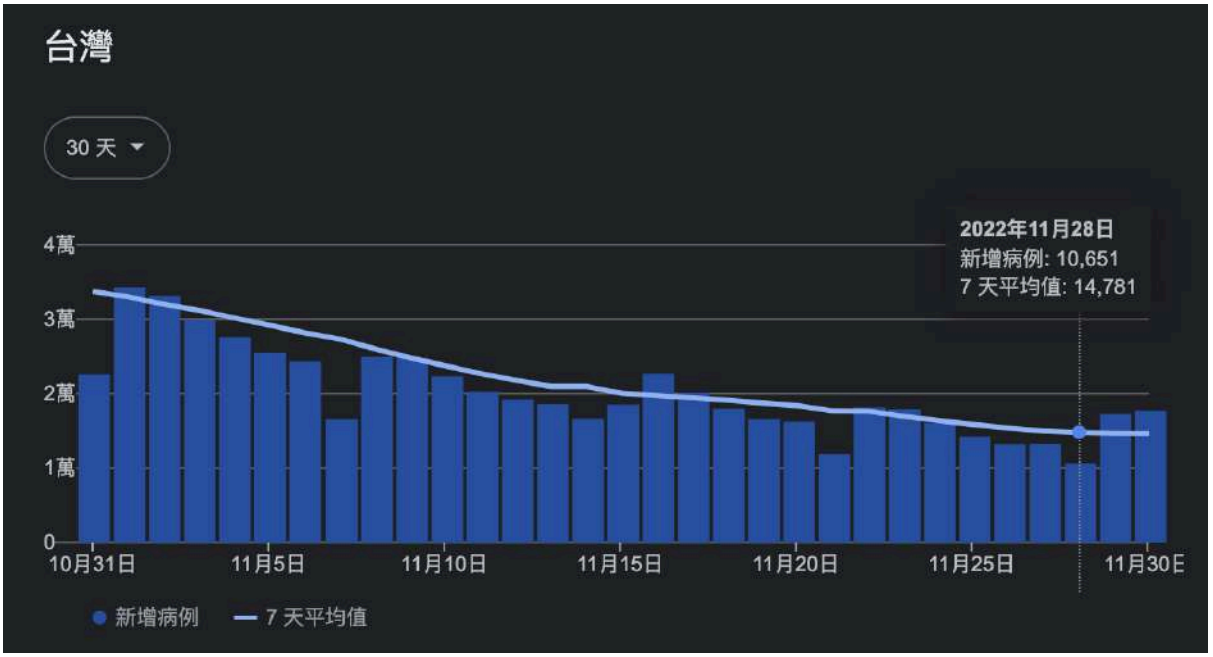
Q: What is the difference between StringencyIndex_Average and StringencyIndex_Average_ForDisplay?
A:
The former is raw index and the latter is the index of "display" version. The display version index is the smooth result of the raw index. You can read the Index Methodology for more details about index calculation.

Q: What is "over Stringency index"? Why do we need a "7 days moving average"?
A:

over Stringency index means a kind of government reaction to COVID-19. The idea of over Stringency index is showing the unreasonable regulation of the government. If the case number is 0, the government should not apply any regulation to society.
over Stringency index的意義是政府控管是否成比例的指標，計算方式是政府控管與COVID-19案例數的比例。照理來說案例數較多的國家，管控通常會越嚴格，若案例數少的國家，管控就應該不會太嚴格。所以用管控嚴格程度/新個案數，如果得到很大的值，就是該政府反應過度（明明只有一點案例卻管東管西），太小的值則可能是該政府已經放飛（明明有一堆案例但是卻什麼都不管）

$$\text{“}over\ Stringency\ index\text{”} = \frac{StringencyIndex\_Average\_ForDisplay}{7\text{-}day\ moving\ average\ of\ daily\ confirmed\ new\ cases}$$

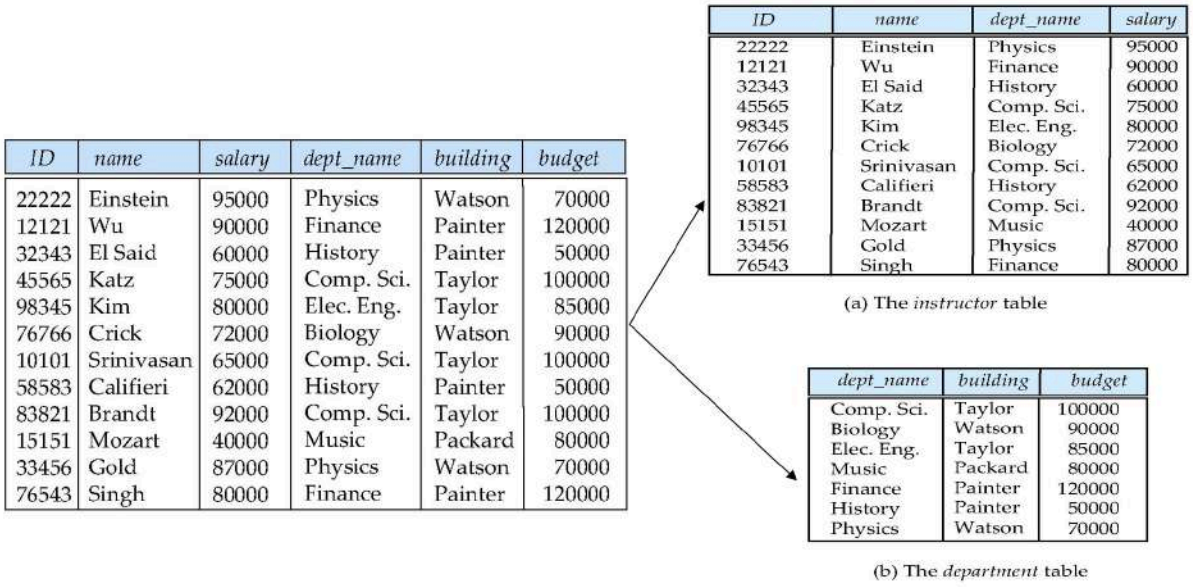
- Stringency index is a composite measure based on nine response indicators including school closures, workplace closures, and travel bans. 管控嚴格指數
- 7-day moving average of daily confirmed new cases means the level of the pandemic in average 個案數的七日移動平均。需要計算移動平均的原因是每天新增的個案數波動太大（有假日、監測沒有每天回報數字等等的影響），所以會算七日移動平均，將曲線滑化。以google查詢案例數為例（下圖），可以看到每隔幾天都有個低點，通常都是週一，因為週日診所/醫院沒開，想被確診也無法（？），而七日平均值則是相對較平滑的曲線，有助益判斷疫情走向。



Q: When calculating the over Stringency index, how to deal with the 0 moving average issue? (Stringency index/0)?
A:
The idea of over Stringency index is showing the unreasonable regulation of the government. If the case number is 0, the government should not apply any regulation to society. To express this idea, you can replace the 0 moving average with 0.1, or the original suggeston 0.5, to show the “overreaction” of a specific county.

Q: Can we remove redundant data when we perform decomposition?
A:

Yes, this is one of the reasons for normalization and decomposition (remove redundancy). See the figure below. You can remove redundant records using any method.



ID	name	salary	dept_name	building	budget
22222	Einstein	95000	Physics	Watson	70000
12121	Wu	90000	Finance	Painter	120000
32343	El Said	60000	History	Painter	50000
45565	Katz	75000	Comp. Sci.	Taylor	100000
98345	Kim	80000	Elec. Eng.	Taylor	85000
76766	Crick	72000	Biology	Watson	90000
10101	Srinivasan	65000	Comp. Sci.	Taylor	100000
58583	Califieri	62000	History	Painter	50000
83821	Brandt	92000	Comp. Sci.	Taylor	100000
15151	Mozart	40000	Music	Packard	80000
33456	Gold	87000	Physics	Watson	70000
76543	Singh	80000	Finance	Painter	120000

ID	name	dept_name	salary
22222	Einstein	Physics	95000
12121	Wu	Finance	90000
32343	El Said	History	60000
45565	Katz	Comp. Sci.	75000
98345	Kim	Elec. Eng.	80000
76766	Crick	Biology	72000
10101	Srinivasan	Comp. Sci.	65000
58583	Califieri	History	62000
83821	Brandt	Comp. Sci.	92000
15151	Mozart	Music	40000
33456	Gold	Physics	87000
76543	Singh	Finance	80000

(a) The *instructor* table

dept_name	building	budget
Comp. Sci.	Taylor	100000
Biology	Watson	90000
Elec. Eng.	Taylor	85000
Music	Packard	80000
Finance	Painter	120000
History	Painter	50000
Physics	Watson	70000

(b) The *department* table

Q: Some columns have only null values and are hard to define functional dependency. Can we ignore these columns that have only null values?
A:
Yes, and please mentioned the reasons for ignoring/removing these columns,

Q: Do we need to include all the data (from the 2 .csv files) in the ER diagram?
A:
Yes. However, the 2 tables are in UNF with redundant data. Please redesign it with an ER diagram and then convert it to a table schema. **Some relations might have too many attributes and is hard to include all of them in the diagram. If so, you can use "a range of attributes", such as C1 to C10, or the other ways, to list attributes.**

Q: Do we need to include all the columns in the 2 .csv files?
A:
Yes. The aims of this HW are to let you know the real-world data is dirty and we need some work before dumping the data into a database. Even though some of the columns are not used in the query steps, we still need to include the data for the other users (in the real-world setting). If you think there is a derived attribute that can be removed from the data schema, please indicate these derived attributes and then you can remove them from the data schema. **Some relations might have too many attributes and is hard to include all of them in the diagram. If so, you can use "a range of attributes", such as C1 to C10, or the other ways, to list attributes.**

Q: When I executed COPY in SQL to import the data into AWS RDS, I got an error and it failed to import the data. How can I address this issue?
A:
Mostly, you may encounter the errors about role permissions. COPY requires the role to be accessible to the database server, not the client. Please refer to PostgreSQL Official

Document (EN) or PostgreSQL中文手冊(中) (5th paragraph, Section “Notes”) for more details. Alternatively, you can import your csv data by right-click on the table and choose “Import/Export Data…”, and import with GUI, or you can use `\copy` command in psql commandline.

Q: When we draw the ER diagram, can we draw the diagram directly from the 2 .csv files?
A:
No, you need to draw a “good” ER diagram, without redundancy, based on the 2 .csv files.

Q: Is it possible that we cannot find any functional dependencies on attributes?
A:
You always can find at least one functional dependency which links PK and the attribute. If that is the only functional dependency you have, that is reasonable.

Q: Some countries have different names in “country-and-continent-codes-list-csv.csv”, for example, United States or United States of America. Do we need to keep all the names?
A:
You should keep the name consistent throughout the databases, so please modify the country name in **“country-and-continent-codes-list-csv.csv”**

Q: Some countries are not included in “country-and-continent-codes-list-csv.csv”, for example, Kosovo. Can I add it manually?
A:
Yes, you can modify the **“country-and-continent-codes-list-csv.csv”**, or you can just ignore the country.

Q: The ER diagram is huge and hard to fit the A4 size, may I submit a .png file separately? Or can I draw multiple ER diagrams?
A:
Yes, you can submit the ER diagram separately and you can also use different ER diagrams to show the whole picture. If you draw multiple diagrams, please describe the links between different diagrams in the document.

Q: I got an error message when creating a lambda function. The error message said I cannot create IAM role. How to solve this problem?
A:
In the Learner lab document, it mentioned that you need to “Attach the existing LabRole to any function that you create if that function will need permissions to interact with other AWS services.”

Q: May I have some hints for the bonus part?
A:
You should understand what is lambda function, then try to connect your lambda function with the database. If you can retrieve data from the database through the lambda function, then try to insert a record. Finally, try to use the lambda function to download the updated .csv, then insert them into your database.
If you can only finish part of the request, please also include the part you have done in your submission.

Q: What is 7-day moving average?
A:
https://www.georgiaruralhealth.org/blog/what-is-a-moving-average-and-why-is-it-useful/