HERIOT WATT UNIVERSITY
UK | DUBAI | MALAYSIA

F28DM Coursework 2

Deadline: **Mon 7th April March – 3:30pm (Week 13)**
Submission: F28DM Canvas – CW2 – online quiz (found under ASSIGNMENTS in CANVAS)

Overview of the coursework

This is an **individual** assignment whereby you will submit your answers to a CANVAS online quiz. You also need to submit a single text file of your SQL workings – please use the template provided.

Collaboration and Plagiarism

This coursework should be carried out individually. Students must never give hard or soft copies of their coursework/code to other students. Students must always refuse any request from another student for a copy of their code.

Sharing a coursework or code with another student is collusion, and if detected, this will be reported to the School's Discipline Committee. If found guilty of collusion, the penalty could involve voiding the course.

Required Tasks

Answer the following questions using SQL and the database provided.

- Submit your answers via the Canvas Quiz in the F28DM course, CW2.
- Not all questions are worth the same marks, this reflects question difficulty
- Questions will be marked correct or incorrect, no partial marks
- In additional submit a script of your SQL statements to Canvas.
    - You must ensure that your submitted code follows the XML template provided and is valid using the XSD Schema provided.
    - You can check your code file is valid on this webpage: https://www.freeformatter.com/xml-validator-xsd.html#google_vignette
- **Each query must be answered with a single SQL statement.** In other words you must not calculate the answer from using several separate SQL statements, putting the answer from one in to the next by hand.

XML LINKS (same as links above but shown in full):

XML template: https://github.com/spatial-intelligence/F28DM/blob/main/F28DM_CW2_answer_template.xml

XSD Schema: https://github.com/spatial-intelligence/F28DM/blob/main/F28DM_CW2_answer_schema.xsd

XML Template:

- Download the XML template provided (link on previous page).
- Ensure you substitute your username, and add your answers with the corresponding SQL statement, and example has been provided as question 0.
- The first few lines of the template (downloaded from link above) should look like this:

```
<data>

<!--UPDATE WITH YOUR USERNAME -->
    <USERNAME>abc123</USERNAME>

<!--START OF EXAMPLE -->
    <question id="0">
        <answer>1094</answer>
        <sql>
            SELECT COUNT(*) FROM trajectory.tracking;
        </sql>
    </question>
<!--END OF EXAMPLE -->

<!--START OF YOUR ANSWERS and SQL STATEMENTS-->

    <question id="1">
        <answer></answer>
        <sql>

        </sql>
    </question> …
```

- Submit your answers via Canvas CW2 in the assessment section. You must include a single text file upload of your SQL workings (last part of the quiz)
- If you do not include your code then you will receive **0 marks for this coursework.**
- You can save the quiz, but only submit once.
- Each answer must be 1 SQL statement, without using VIEWS or creating additional tables. However you can make use of CTE, Windowing, CASE statements, etc
- Your SQL code should give the exact answer (e.g. if asking for a movie title it must return just the movie title as text and not the movie ID or other columns).
- Answers are to be calculated based on the data provided and no additional sources of information
- Explore the data before answering the questions to understand the structure.

Database to use

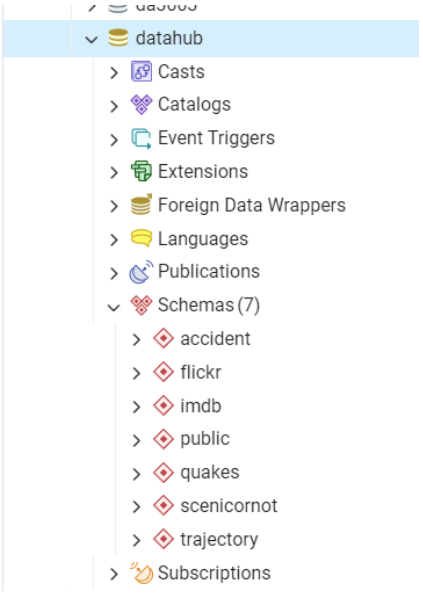
Use the **datahub** database for these questions and the appropriate schema as identified in the questions.

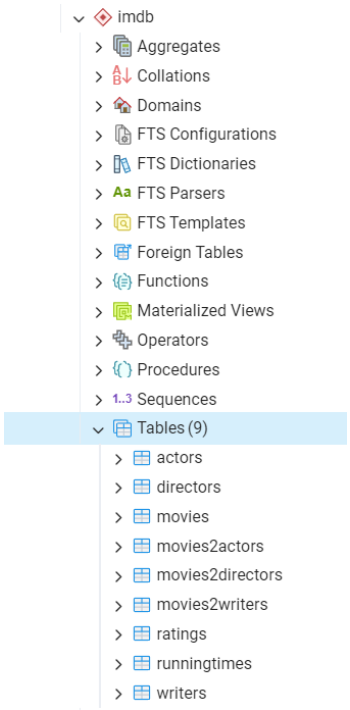
Connect using pgAdmin4 as before but instead of opening your own database select '**datahub'** instead. You should have SELECT access but not be able to create any new tables, views etc in this database. You will notice that there are a few different schemas (e.g. accident, flickr, imdb) – these each contain different sets of tables.

They can be access from SQL using: `SELECT * FROM schema.table;`
For example: `SELECT COUNT(*) FROM imdb.actors;`

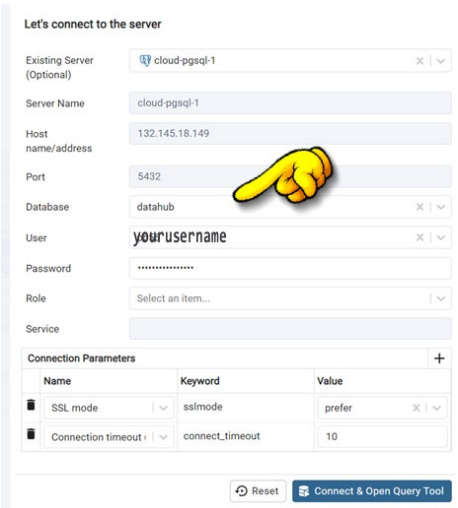
pgAdmin4

Select the Database as: **datahub**



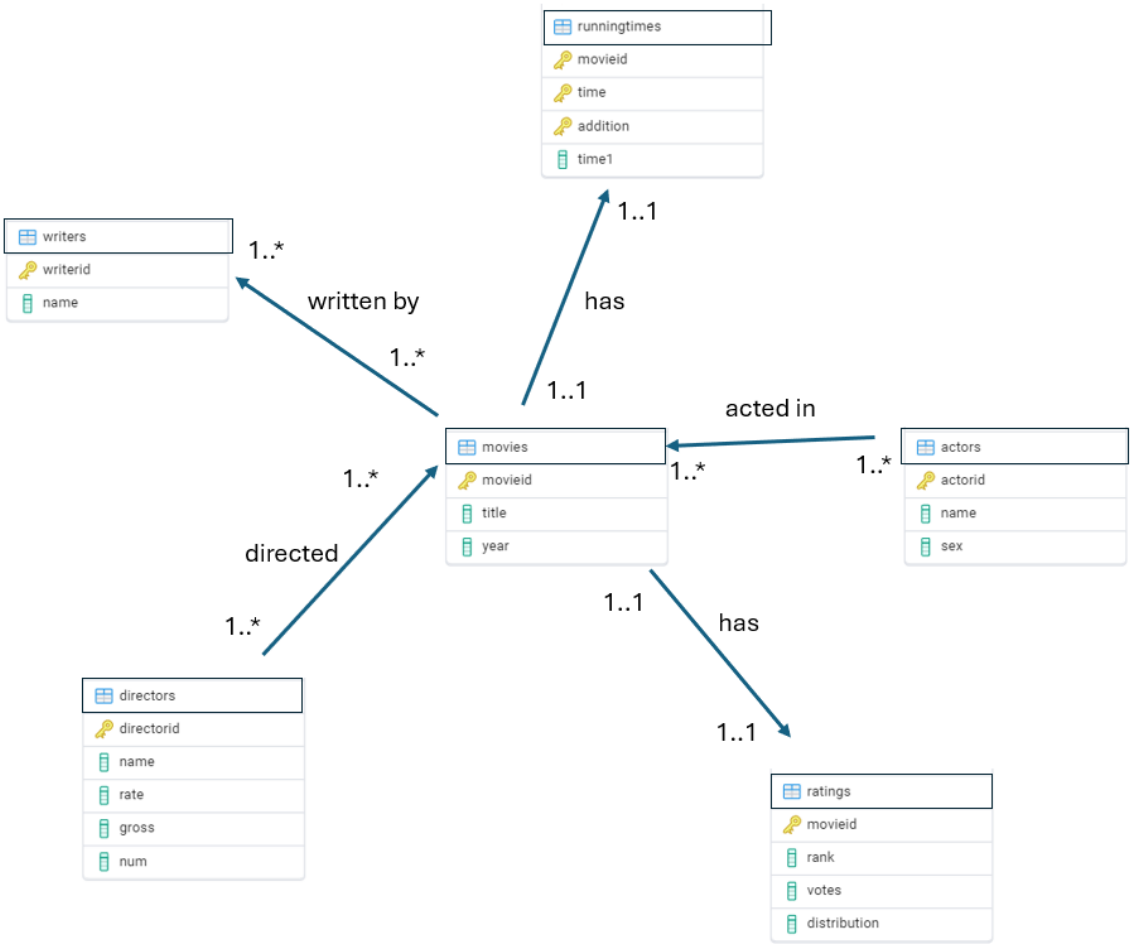


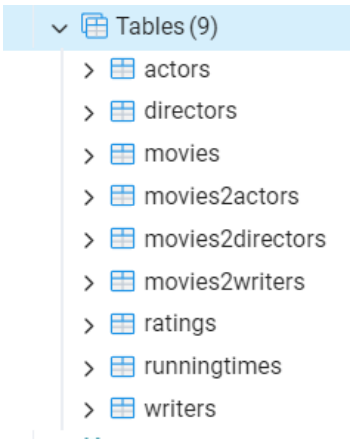
You can open a schema to see the tables



<u>Alternative way to connect:</u>
In pgAdmin v9.1 you can open a query window directly connecting to the **datahub** database if you wish, rather than scrolling down the list of databases. Ensure you change the Database to *datahub* and do not leave it on the default *postgres*.

IMDB schema


runningtimes
movieid
time
addition
time1
1..1
writers
writerid
name
1..*
written by
1..*
has
1..1
acted in
movies
movieid
title
year
1..*
1..*
actors
actorid
name
sex
1..*
directed
1..*
1..1
has
1..1
directors
directorid
name
rate
gross
num
ratings
movieid
rank
votes
distribution


Tables (9)
actors
directors
movies
movies2actors
movies2directors
movies2writers
ratings
runningtimes
writers

Questions

#	Question	Points
Using the **'imdb'** schema to answer these questions		
1	What is the TITLE of the earliest movie in the dataset?	1
2	How many movies have more than 7 directors?	1
3	Give the YEAR for the movie which has the most directors?	1
4	Which actor (give the actor ID) has appeared most often with Tom Hanks?	1
5	How many actors have worked on more than 10 films together?	2
6	Which decade has the highest average ranked movies? (put the first year from the decade, so for 1900-1909 you would put 1900)	2
7	How many different movies [based on movie id] are missing a genre in the dataset?	2
8	How many movies have an actor/actress written and directed but not starred in? (i.e. the person that wrote and directed the movie is an actor/actress but they didn't star in their own movie).	2
9	Identify the Director (by director ID) with the most movies over 160 minutes long.	2
10	Find the actor id for the person who starred in over 43 movies before 2015, including at least 1 directed by Steven Spielberg	2
Use the '**scenicornot'** schema to answer these questions		
11	How many photo locations are within 5km of a point in London (530000, 180000) using British National Grid [EPSG: 27700] ?	1
12	Based on using the hexgrid (10km cells) layer provided, which region (cell_id) with at least 3 sample locations has the highest average scenic scores?	2
13	What is the largest total number of votes cast within a single hex cell?	2
14	Find the cell_ids of the 2 neighbouring hexagons which have the biggest difference in average scores. Your answer should consists of just the 2 cell_ids.	3
Use the '**quakes'** schema to answer these questions in conjunction with **'scenicornot'**		
15	Which 48 hour period* since 1st Jan 2010 00:00:00 has had the highest number of quake events? [give your answer as yyyy-mm-dd hh:mm:ss] * this means divide the time into 48 hour periods from the datum of 1 Jan 2010 00:00:00, not a sliding 48 hour window	1
16	How many different scenicornot location points are within 100 metres of an earthquake location?	2
17	How many spatial clusters are there in the quake dataset when using a threshold of 15km, and a minimum of 5 quakes per cluster?	2
Use the '**trajectory'** schema to answer these questions		
18	What was the average walking speed for userid = 'p3' ?	1
19	Which trajectory has the most sample points, give the userid?	1
20	Which user has the highest sinuosity ratio for their route ?	1
21	How many meters did person 1 (userid='p1') walk? *(Answers will be marked to the nearest metre, but you can leave decimal places)*	2

(continued on next page)

Use the **'flickr'** schema to answer these questions		
22	Which month is the most popular for photos?	1
23	Which user_id visited the most locations based on the 100m grid provided?	2
Use the **'accident'** schema to answer these questions		
24	How many 'avenue' streets (based on osm_name) of the same name have more than one osm_id?	1
25	Which of the routes [a,b,c] is the most dangerous based on a measure of crashes per kilometre, within a distance of 15 metres around (each side) the route?	2

Reminder:

1.Complete the Canvas quiz with your answers

2. Submit your XML answer script with answers/SQL code as the final question

--- END OF COURSEWORK 2 ---