# BUS5DWR: Assignment 1 (40%)

Answers to this assignment may be provided in a copy of this Word document under the question or in the space provided.

**Part 1 [7 marks]**

A fictitious train company has the following data extract (imagine there are thousands of rows similar to these).



Using the ERDPlus modelling tool (<https://erdplus.com/#/standalone> **—** no need to register), create an ER diagram and relational schema diagram of a structured database design in second normal form consisting of two or more named tables capturing the above data. [1+1=2 marks]

* In the ER diagram indicate the cardinality of each relationship (one-one, one-many etc). [1 mark]
* In the relational schema diagram show the attribute names and primary key(s) of each table. [1 mark]

Include screen captures/image exports of the diagrams in your submission.

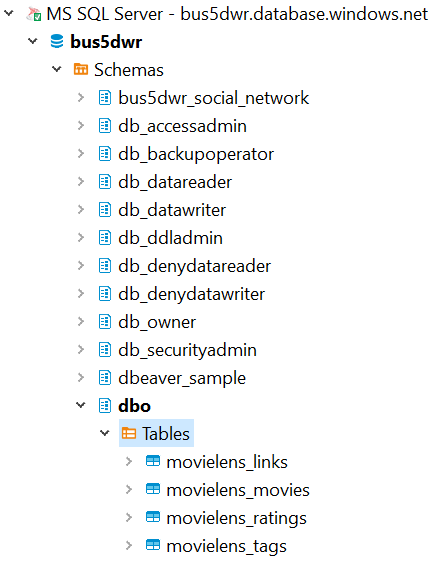
Note the design needs to enable the following types of queries easily:

1. How many trains are simultaneously running at 2pm on a weekday?
2. How many trains depart from a particular station?

Justify your design choice and explain why it is in second normal form. [2+1=3 marks]

**Part 2 [33 marks in 12 parts]**

We will make use of the MovieLens dataset contained in bus5dwr.dbo within the bus5dwr.database.windows.net server. Please refer to Lecture 1 slides for instructions on how to access the server via DBeaver if you have not accessed it already. The data involves movie ratings and tags and is described in more detail here: <http://files.grouplens.org/datasets/movielens/ml-latest-small-README.html>.



You will see four tables:

    i) bus5dwr.dbo.movielens\_links

    ii) bus5dwr.dbo.movielens\_movies

    iii) bus5dwr.dbo.movielens\_ratings

    iv) bus5dwr.dbo.movielens\_tags

Start by exploring the contents of the tables understanding the meaning of their attributes.

Write SQL queries (using MS SQL Server) that use one or more of the tables to answer the following questions. Please include your code as text rather than an image to enable it to be copy-pasted for assessment.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Question | Answer | Code | Marks |
| 1 | How many users contributed tags? |  |  | 2 |
| 2 | How many movies had exactly 3 genres associated with them? (Comedy|Drama|Romance would be one example.) |  |  | 2 |
| 3 | How many users gave ratings that are not whole numbers (e.g. 3.5, 4.5)? |  |  | 2 |
| 4 | List the movies whose title (and year) appear more than once in the movielens\_movies table. |  |  | 2 |
| 5 | How many movies were released between 1985 and 1995 inclusive? |  |  | 2 |
| 6 | Which movies in the movielens\_movies table did not receive a rating? Write a query to list five of these movies. |  |  | 2 |
| 7 | How many users tagged three or more different movies? |  |  | 3 |
| 8 | What are the names of the five movies that had the most ratings? Show the title and number of ratings in two columns. | Title Num\_ratings  … … |  | 3 |
| 9 | What are the five tags that had the most distinct users contributing them? Show the tag and number of users in two columns. | Tag Num\_users  … … |  | 3 |
| 10 | What are the three highest rated movies by average rating where one of the genres is ‘Mystery’ and  the number of ratings for the movie is at least five? Show the movie name and average rating in two columns. | Title Average\_rating  … … |  | 4 |
| 11 | Find the five genres with the most distinct tags. Show the genre and tag count in two columns. For example, a movie having two genres Action|Thriller will have its tags contribute to each of the Action and Thriller genres.  Hint: see example B in <https://docs.microsoft.com/en-us/sql/t-sql/functions/string-split-transact-sql?view=sql-server-2017> to see how to split the genres field. | Genre Tag\_count  … … |  | 4 |
| 12 | Let us attempt to find the highest rated movies, but not simply by computing the average rating, since that would favour movies with a small number of ratings.  We will make use of the following formula.  Weighted rating of a movie = (V\*R+10\*C)/(V+10), where   * V = number of votes for the movie * R = average (mean) rating for the movie * C = the average rating across all movies   This formula represents the equivalent of adding 10 additional ratings to each movie, each equal to the average rating across all movies, then finding the resulting average. Hence a movie with a single rating will not have too high a weighted rating.  Find the names and ratings of the five highest rated movies according to the formula above having the tag ‘funny’. | Title Weighted\_rating  … … |  | 4 |

Total: 40 marks

## Marking rubrics

The following marking guide will be used by the marker in assessing your work. Please have a look to understand what you need to cover for each question in this assignment.

**Part 1:**

For each part where marks are indicated, full marks will be given for a correct answer, 0.5 marks deducted for each minor error up to the value of that part.

**Part 2:**

Full marks for each faultless SQL statement (i.e. correct output). For each error picked up by the marker, 0.5 marks will be deducted for minor errors and 1 mark for critical errors up to the value of that part.