Beijing-Dublin International College
COMP2013J - Databases
Individual Database Assignment

1 Assignment Details

Submission Type: Individual assignment
Language: Assignment must be implemented in MySQL DBMS
Assignment Weight: 70% of your final grade.

This assignment is broken into ~~multiple parts~~ five tasks ~~each to be submitted separately~~ to be submitted in two parts. This document will introduce the concept that we will be modelling and the requirements for submission.

2 Problem Description

An online movie and TV streaming platform requires an information to keep track of its business. Within this system, it needs a database to record information about its customers, purchases and usage of the service. In addition, in order to try and understand what content the customers want to see more, they also record information about the movies and TV shows in their database.

Description

For each customer, we record the name (given name and family name), address, phone number and email address. In addition, there are two types of customers. The first is a purchasing customer, they pay individually for each item that they rent or buy. The second type is subscription customers, they pay a monthly fee and do not have any requirement to pay for the items that they watch. We must record two pieces of extra information about subscription customers, their next renewal date and the price that they pay.

Customers without subscriptions when they want to view a show or movie have two choices, they can rent it or they can buy it. When an item is rented it us usually available for a short period, such as two weeks, and the customer may watch it as many times as they want in that period of time. When an item is purchased, a customer gets to access the item as often as they want to forever. Every time a an item (episode of a TV show or movie) is watched on the service, this is recorded in the database. This allows the service to determine which items are the most or least popular.

For each movie, the company records the movie title, year of release, rental price, purchase price, length in minutes and names and jobs of the cast and crew. Cast and crew refers to the actors/actresses as well as writers and directors. There may be multiple of any of these associated with a single movie or indeed multiple movies.

For each TV show, the company records the name of the TV show, the list of episodes and the year it started. For each episode in each TV show, the company records the season number, episode number, title, length of the episode in minutes and names and jobs of the cast and crew. Cast and crew can change between episodes of a TV show, so this information is recorded for each episode.

3 Tasks

This assignment is split into 5 parts. The overall goal is to complete a database design for supporting the operation of this business based on the description above. As a part of this goal, the following tasks must be completed:

1. An entity relationship diagram to model the database for this company.
2. Map the entity relationship diagram to the relational model.
3. SQL code to create the appropriate tables.

4. A database dump with fake data for the tables.
5. Java code to interface with the database and using SQL to find the result of a number of queries.

3.1 Parts

~~I will be assessing your work based on the previous parts that you have submitted. So your mapping should make sense for the ER diagram that you submit in the first part, your create table statements should make sense for the logical mapping you submit in part 2, your database structure in part 4 should match the create table statements that you submitted in part 3 and your SQL in part 5 will be executed on the data you submit in part 4.~~ You work will be assessed as a logical sequence. So all parts should be consistent. E.g. The mapping in part 2 should be based on the ER diagram that you created in part 1, and the create table statements should be based on you mapping from part 2. This applies to all parts of the assignment.

3.2 Frequently Asked Questions

3.2.1 What if I make a mistake?

If you make a mistake in an earlier part, then you should explain this in the later part. Say what the mistake was, what you changed and why. If you do not include the explanation I will just have to assume that you have made a mistake...

3.2.2 What If I miss a deadline?

1. Don't
2. But if you do, submit the missing part with your next submission. You will not get graded for it, but it means that I can grade your next part correctly and you will not lose any marks in that part too.

I am giving really long deadlines for this assignment and you know a really long time in advance what they are so do not leave the assignment to the last minute. If you are worried about problems with your internet or other problems that might prevent you from submitting close to the deadline, then submit your partially completed assignment early. You can then replace this with later versions as you get more of the assignment done. Then even if you do have a problem that prevents you from submitting your final version, you will still be graded on the last version that was submitted.

3.2.3 Are there opportunities to get extra credit?

Yes, the final part of the assignment has some optional parts that you can complete to get extra credit in the assignment. Look at the marking scheme in that section for more details.

3.2.4 How will my grade be calculated?

Your grade will be calculated using the Computer Science grading scheme available here: `https://csintranet.ucd.ie/CSGrading/` This is also known as the Alternative Linear Conversion Grade Scale and is available here: `https://www.ucd.ie/students/assessment/gradesexplained.html` along with more comprehensive details about how your final grade will be calculated from your assignment and worksheet grades.

3.2.5 Can I get my friend to help?

This is an individual assignment, and I will be checking very closely for plagiarism. When discussing the assignment, asking questions or helping friends, make sure that you do not share information about your solution. Try to keep your discussion on the general concept. If you are having a specific technical issue, then you should discuss it with a teaching assistant (first option) or me.

4 Assignment Part 1: ER Diagram

Due date: ~~1st of May 2020 @ 17:00~~ 29th of May 2020 @ 17:00 Beijing Time (No Exceptions)
Format: This part must be submitted as a single PDF document named using your UCD student number and the part number E.g. `06373313.part1.pdf`
Assignment Weight: 20% of your assignment grade.

4.1 Instructions

Complete an Entity Relationship diagram to represent the concepts in the problem description. This should be submitted as a single pdf document containing your name and student number as well as the following:

1. A glossary of terms

2. A list of assumptions that you made about the data and relationships

3. An Entity relationship diagram

When creating the entity relationship diagram, you should show clearly the process of developing the model.

You should use some appropriate software to develop the ER diagram, hand drawn diagrams are not acceptable. This can be done using either image creation software or software specifically designed for representing ER diagrams. For example, I use the software OmniGraffle on Mac, but any good image creation or diagramming software should work.

You should make sure when you are designing your model that the queries in section 8 can be answered using this model.

4.1.1 Assumptions

The choices that we make during the process of creating the ER diagram are often based on information in the problem description, but some times there is not enough information to decide something and we need to make a choice ourselves. This choice is an assumption and typically it will be based on our best guess about the data. For example, consider the problem of representing the cast of a movie. Here are some possible assumptions that I might make;

1. I have assumed that the combination of given name and family name are unique across all cast and can be used as the key

2. I have assumed that the there is no candidate key that can be used to uniquely identify cast and crew and have added a surrogate key

Here both of these assumptions are reasonable, associations like the screen actors guild do not allow professional actors to use the same name. However, because there is the possibility that writers and directors might not have the same constraint, we could equally assume the opposite.

The requirement to include these assumptions is only for parts of the problem where there is no definite description of the way something should be represented. In these situations you should make your choice and then explain your reasoning as an assumption. This way I will be able to understand the choices that you have made in how you represent things.

4.2 Marking Scheme

Item	Fail (E/F)	Pass (D/C)	Excellent (A/B)
Glossary of Terms	No glossary of terms is included	Glossary of terms is included, but it is not complete	Glossary of terms is completed satisfactorily
Assumptions	No explanation of assumptions is included	Some assumptions are explained, but the list is not complete	All assumptions made are explained satisfactorily
Quality of ER Diagram	Diagram is hand drawn and scanned or a picture was taken	Diagram was produced using software, but the image quality is poor	Diagram was produced in excellent quality using software
Entities	Not all entities are represented correctly	All entities are represented correctly, but there are some errors such as they are weak when they should be strong or the reverse	All entities are represented correctly and they are of the correct type
Attributes	No Attributes are represented correctly	Not all attributes are represented correctly, or some are not the correct type (derived, multiple values, composite)	All attributes are represented and they are of the correct type
Relationships	Not all of the required relationships are represented correctly	The relationships are represented correctly, but attributes on the relationships are not represented	Relationships and the attributes are represented correctly
Cardinality	Cardinalities of relationships are not included or they do not make any sense	Some cardinalities are included and not others or some cardinalities do not make any sense	All cardinalities are included and they make logical sense
Inheritance	No Inheritance used	Inheritance used in incorrect place	Inheritance used correctly

Table 1: Indicative Marking Scheme for Assignment Part 1: ER Diagram

5 Assignment Part 2: Relational Model

Due date: ~~15th of May 2020 @ 17:00~~ 29th of May 2020 @ 17:00 Beijing Time (No Exceptions)
Format: This part must be submitted as a single PDF document named using your UCD student number and the part number E.g. `06373313.part2.pdf`
Assignment Weight: 30% of your assignment grade.

5.1 Instructions

Complete the process of logical design by mapping the ER diagram produced in Assignment 1 to a logical schema. This logical schema should be normalised to at least third normal from or Boyce-Codd Normal Form (BCNF). When mapping the ER diagram to the relational model, show clearly the process used and state the reasons for your choice of primary keys and foreign keys in your model. Again, document any assumptions you have made.

Your submission should include the following:

1. The logical schema.
2. An explanation of your choices of primary and foreign keys for each table.
3. Your explanation of the normal form that your schema is in.

5.2 Marking Scheme

Item	Fail (E/F)	Pass (D/C)	Excellent (A/B)
Logical Schema	Logical schema not included	Logical schema included, but not complete	Complete logical schema included
Normal Form	Schema is in 1NF	Schema is in 2NF	Schema is in 3NF or BCNF
Assessment of Normal Form	Student explained the incorrect form (E.g. said it was BCNF, but it was 1/2/3NF)	Student said it was the correct from, but did not explain why	Student explained the correct form (Whichever form it was)
Primary Keys	Choice of primary keys was not adequately explained		Choice of primary keys was well explained
Foreign Keys	Choice of foreign keys was not adequately explained		Choice of foreign keys was well explained

Table 2: Indicative Marking Scheme for Assignment Part 2: Relational Model

6 Assignment Part 3: CREATE TABLE Statements

Due date: 29th of May 2020 @ 17:00 Beijing Time (No Exceptions)
Format: This part must be submitted as a text file containing the SQL statements and a PDF document containing the explanations. These should both be named using your UCD student number and the part number E.g. `06373313.part3.pdf` and `06373313.part3.txt`
Assignment Weight: 10% of your assignment grade.

6.1 Instructions

Write appropriate CREATE TABLE statements for your relational model. For each table you should include the following:

1. The SQL statement to create the table
2. An explanation of your choice of data type for each attribute
3. An explanation of all constraints on this table (including both intra and inter relational constraints)

6.2 Marking Scheme

Item	Fail (E/F)	Pass (D/C)	Excellent (A/B)
CREATE TABLE Statements	Many statements were not complete or were not correct SQL	All statements were submitted, but some were not correct SQL	All statements were submitted and worked correctly
Data Types	Choices of data types were not explained	Choices of data types were explained, but not well	Choices of data types were well explained E.G. why CHAR or VARCHAR etc.
Explanation of Intra-Relational Constraints	Student did not explain the constraints	Student explained the constraints, but some were not a good choice or did not make sense	Student explained the constraints and all were sensible
Explanation of Inter-Relational Constraints	Student did not explain the constraints	Student explained the constraints, but some were not a good choice or did not make sense	Student explained the constraints and all were sensible

Table 3: Indicative Marking Scheme for Assignment Part 3: CREATE TABLE Statements

7 Assignment Part 4: SQL Dump

Due date: ~~5^{th} of June 2020 @ 17:00~~ 19^{th} of June 2020 @ 17:00 Beijing Time (No Exceptions)
Format: Assignment must be submitted as a single SQL dump file exported from MySQL This file should be named using your UCD student number and the part number. E.g. `06373313.part4.sql`.
Assignment Weight: 10% of your assignment grade.

7.1 Instructions

You should generate some data and insert it into the tables. There should be at least enough data for you to test the queries in section 8 and the amount and quality of the data will determine your grade in this component.

A recommended strategy for this would be to use write a program to insert data into your database (you do not need to submit the program if you use one). You can use a list of made up information that you compile yourself or data that you have downloaded from the internet for the basic information in the database. Usage, rental and purchase information can be generated using random functions of any programming language (but you may need to query the database to get the right keys to use).

7.1.1 Data Sources:

- Fake customer data: `https://www.fakenamegenerator.com/`
- On-line data sets: `https://www.kaggle.com/`

You should insert enough data to make sure that you can test your queries well and satisfy the criteria in the marking scheme, but be aware that the submission limit for your database is 50 Mb. Note that all data should be inserted for the year 2019 and make sure that there is data in your database so that every query will return a result.

7.2 Marking Scheme

Item	Fail (E/F)	Pass (D/C)	Excellent (A/B)
Customers	Less than 20 customers inserted	Between 20 and 50 customers	50 or more customers inserted with a mix of subscription and purchase customers
Purchases and Rentals (movies)	Less than 50 purchases and rentals (combined)	Between 50 and 250 purchases and rentals	More than 250 purchases and rentals
Purchases and Rentals (TV Episodes)	Less than 50 purchases and rentals (combined)	Between 50 and 250 purchases and rentals	More than 250 purchases and rentals
Movies	Less than 15 Movies	Between 15 and 30 movies	More than 30 movies
TV Episodes	Less than 5 TV Shows with episodes (average 10 episodes)	Between 5 and 10 TV Shows with episodes (average 10 episodes)	More than 10 TV Shows with episodes (average 10 episodes)
Cast and Crew	No Cast and Crew included	Some cast and crew for some movies/TV episodes (but not all)	At least 3 cast and crew for each movie/TV episode
Views	Less than 500 views (in total) of movies and TV episodes	More than 500 views (in total) of either movies or TV episodes (but not mixed)	More than 500 views of movies and TV episodes (randomly mixed across the various movies and TV episodes)

Table 4: Indicative Marking Scheme for Assignment 4: SQL Dump

8 Assignment Part 5: Integration

Due date: 19th of June 2020 @ 17:00 Beijing Time (No Exceptions)
Format: Assignment must be submitted as a single zip file exported from Eclipse and should be named using your UCD student number and the part number. E.g. `06373313.part5.zip`
Assignment Weight: 30% of your assignment grade.

8.1 Instructions

Based on the tables that you have designed, write SQL queries to find the answer to each of the following questions. Each question should be answered using 1 query.

1. How much money did the company make in 2019, broken down by Movies and TV in 2019?
2. How much money did the company make in 2019, broken down by purchases and rentals?
3. Which movie was watched the highest number of times?
4. Which user spent the most amount of time watching TV and movies in 2019?
5. Which TV show had the lowest average price for each episode?
6. List all of the movies that can be bought for less than 50 RMB.
7. What episodes of any TV show have never been watched?
8. List the top 5 actors/actresses that appeared in the most movies
9. List the top 10 actors/actresses that appeared in the most episodes of TV shows
10. What is number of subscription customer and their average payment for 2019 as well as the number of non-subscription customers and the average spend for 2019 (rentals and purchases)?

8.2 Important Notes

- These queries should be implemented within a Java program that connects to the database, performs the queries and then outputs the results
- The results should be well formatted and output in a clear and consistent way
- If you are unable to get a query to work, you should comment out that query and write a comment explaining how you were attempting to get the result
- Your code should be written to query a MySQL database that is named UCD followed by your UCD student number e.g. `UCD06373313`. This is where I will load your database dump from part 4, so your code should query this database.
- The username your code should use is `restrictedstudent`
- The password your code should use is `sqlpassword`

8.3 Marking Scheme

Item	Fail (E/F)	Pass (D/C)	Excellent (A/B)
Connection	Database connection was not implemented or only SQL queries were submitted	Database connection was implemented, but errors resulted from queries or it was not set up correctly	Database connection worked and queries were returned
Queries (for each one)	The query did not work	The query did not work correctly but was close or there was a commented explanation how it was attempted	The query worked correctly
Formatting	The data was just printed	The data was somewhat formatted	The query and column headers were printed, the columns in the results were well aligned and the queries were separated

Table 5: Indicative Marking Scheme for Assignment 5: Integration

9 Marking Schemes

This marking schemes given in the previous sections give the details of the expectations I have for your assignments and the grades you will receive. These marking schemes are subject to change. This means that it may be changed at any time without notice if I feel some parts were too easy or too hard and additional criteria may be added. The final grade from this section will be based on a weighted sum of the individual parts. The weights for each part will be based on the difficulty and importance.

The text explaining each can be used as a guide to the amount of work expected for the different parts of the assignment.

10 Extra Credit

I am providing some opportunities for you to increase your grade by doming some extra work for more credit. These are completely optional and it is possible to get full marks without completing any extra credit. However, this could prove difficult. The extra credit can add up to a maximum of 10% to your assignment grade. You will only be graded on a single piece of extra credit work, so complete and submit at most one of the following options. There are three options to get extra credit in the assignment:

1. Complete a graphical user interface (GUI) for the final part of the application (as well as the text based output)
2. In addition to the MySQL dump for part 4, create a similar list of command for creating and storing the same data as a NoSQL database in MongoDB
3. Write a report detailing the indexes that you would recommend adding to your tables to speed up the queries

10.1 GUI

Create a graphical user interface for your application that interacts with the database. This application should have the following features

- Allow the user to enter their username and password to access the database
- Allow the user to choose (through menus or buttons) which queries to execute
- Display the results of the queries in the GUI, these should be nicely formatted

This is probably the most difficult of the three options. It requires you to learn how to create the GUI entirely on your own. However, there is a potential benefit to this. Creating a database and connecting it to the database will be a part of the tasks that you are required to complete in COMP3013J - Object-Oriented Design in the first semester of next year. So having some experience with this will give you some experience before you need to complete these tasks.

10.2 NoSQL

For this option, you are required to Implement a similar database in MongoDB and write the commands to insert the same information into you design. More specifically, the following is required:

- A PDF document describing the collections and document structures you are using giving examples of each document type
- Included in this document, an explanation of how you implemented the relationships and why you chose that implementation
- A text file containing all of the commands that would be required to create you collections and insert all of your documents

10.3 Indexes

Write a report detailing the secondary indexes that you recommend adding to your database. This (PDF document) report should be based on the database design you have specified and include the following:

- A description of each of the indexes that you would recommend
- For each index recommended, you should describe the type of index created (unique/non-unique)
- For each query in part 5, you should describe which indexes **might** improve the performance and where possible support this by including the output of the `EXPLAIN` of the query