

Student name: Hsu, Cheng-Chun Student ID: s5390209

*Family name, Given name*

# School of Information and Communication Technology (ICT)

## Database Design

Trimester 2, 2024

**Final Exam**

### Writing time

120 minutes

### Reading time

10 minutes

There are multiple sets of questions. You will be randomly given a set. Please download your set of questions to your computer and work on this document.

Submission consists of one file:

• A Word or PDF document

• You can use draw.io to illustrate your ERD and normalisation step; then attach the picture in this Word document. Alternatively, you can draw on a paper, take picture of that, and then insert into this Word document.

**Question 1 (10 points):**

Given the ERD below, answer the following questions.

A diagram of a software company

Description automatically generated

a) The following query returns 14. **[3 points]**

SELECT COUNT(Apt\_Lvl\_ID)

FROM STORESTAFF

WHERE Apt\_Lvl\_ID <> (

SELECT Apt\_Lvl\_ID

FROM APPOINTMENT

WHERE Apt\_Lvl\_Desc = 'Supervisor'

);

Student A concludes that “Our database contains information on 14 people who are not supervisors”.

***Question***: Do you think the conclusion from A is correct? Provide a clear justification for your reasoning.

***Answer***: No, because the subquery can return many values of “Apt\_Lvl\_ID”. The correct query is like: SELECT COUNT(Apt\_Lvl\_ID)

FROM STORESTAFF

WHERE Apt\_Lvl\_ID NOT IN (

SELECT Apt\_Lvl\_ID

FROM APPOINTMENT

WHERE Apt\_Lvl\_Desc = 'Supervisor'

);

b) The following query attempts to return the max hourly rate of each store. **[3 points]**

*Query 1:*

SELECT Str\_Num,

(SELECT MAX(Hourly\_Rate)

FROM SALARYRATE sr

WHERE sr.Str\_Num = s.Str\_Num) AS Max\_Hourly\_Rate

FROM STORE s;

*Query 2:*

SELECT Str\_Num, Hourly\_Rate as "Max\_Hourly\_Rate"

FROM SALARYRATE

GROUP BY Str\_Num

***Question***: Given any data, will these two queries return the same results? If yes, provide a clear justification for your reasoning. If no, give an example and explanation. Compare the performance of these two queries.

***Answer***: These 2 queries will not return the same results. As the query 2 don’t use the MAX() function to find the max hourly rate. The query 1 has better performance as it uses subquery to filter first.

c) Given these two queries: **[4 points]**

*Query 1:*

SELECT Str\_Num, MAX(Hourly\_Rate) AS Max\_Hourly\_Rate

FROM SALARYRATE

GROUP BY Str\_Num

HAVING MAX(Hourly\_Rate) >= (SELECT MAX(Hourly\_Rate) FROM SALARYRATE WHERE Str\_Num = SALARYRATE.Str\_Num);

*Query 2:*

SELECT DISTINCT Str\_Num,

(SELECT MAX(Hourly\_Rate)

FROM SALARYRATE sr

WHERE sr.Str\_Num = s.Str\_Num) AS Max\_Hourly\_Rate

FROM SALARYRATE s;

***Question***: Given any data, will these two queries return the same results? If yes, provide a clear justification for your reasoning. If no, give an example and explanation. Compare the performance of these two queries.

***Answer***:

Yes, it will return the same results. Query 2 will have better performance as the having in query 1 seems unnecessary.

**Question 2 (10 points):**

Your company has been engaged by Phones'R'Us, a retail chain specializing in mobile phones, to design a database system for managing customer orders. Your colleague, Colleague A, has proposed the following ERD.

\**Due to time constraints for the exam, additional factors such as color options and storage capacities will not be included in the question.*

A diagram of a product

Description automatically generated

A snapshot of the database from the design of Colleage A:

|  |  |  |
| --- | --- | --- |
| **Product** | | |
| **PhoneTypeID** | **PhoneTypeName** | **WarrantyPeriod** |
| apple16 | Apple iPhone 16 | 2 years |
| apple16pro | Apple iPhone 16 Pro | 2.5 years |
| samsungnote7 | Samsung Galaxy Note 7 | 5 years |
| … | … | … |

|  |  |  |  |
| --- | --- | --- | --- |
| **OrderItem** | | | |
| **OrderItemID** | **OrderID** | **PhoneTypeID** | **Quantity** |
| 1001 | 132 | apple16pro | 2 |
| 1046 | 555 | samsungnote7 | 2 |
| 1064 | 555 | Apple16 | 3 |
| … | … | … |  |

In the ERD above, the "Product" entity in the ERD represents specific phone models, such as iPhone 16, iPhone 16 Pro, Samsung Galaxy Note 7,….

After reviewing the proposed ERD, your Colleague B raised concerns about its ability to track the store that sold each phone and the corresponding customer order numbers, which would be essential if a customer wants to return a product. Colleague B suggests that we should be able to trace this information using the phone's serial number. The below is the version of ERD made by Colleague B.

A diagram of a product

Description automatically generated

A snapshot of the database from the design of Colleage B:

|  |  |  |
| --- | --- | --- |
| **Product** | | |
| **PhoneSerialNumber** | **PhoneTypeName** | **WarrantyPeriod** |
| 350750724344136 | Apple iPhone 16 | 2 years |
| 35-332907 | Google Pixel 9 | 4 years |
| … | … | … |

|  |  |  |
| --- | --- | --- |
| **OrderItem** | | |
| **OrderItemID** | **OrderID** | **PhoneSerialNumber** |
| 1001 | 131 | 350750724344136 |
| 1045 | 190 | 35-332907 |
| 1493 | 131 | 350734231412343 |
| … | … | … |

However, your Colleague C, argues that recording every phone’s serial number during the stock-in process would be impractical, as it would require scanning thousands of phones.

Do you agree with Colleague C’s idea? If yes, draw the ERD representing that idea. The ERD must comply with the Third Normal Form (3NF), ensuring there are no signs of data redundancy. Draw the revised ERD and attach it below.

***Answer***: Yes, I agree with colleague c.

A diagram of a phone number

Description automatically generated

**Question 3 (8 points):**

The following is the relation schema of part of the database that XYZ Solution has recently implemented for their client EzyMovie:

**MOVIE**(Movie ID, MovieName, *StudioID*, MovieBudget)

**STUDIO**(StudioID, StudioName, City)

**ACTOR**(ActorID, ActorName, ActorDateOfBirth)

**MOVIEACTOR**(*MovieID*, *ActorID*, OrderOfAppearance)

Using the above schema, write SQL queries that will answer/process the following:

1. List studio names and their locations for those studios that have the letter r in the third position of their names.
2. List the actor names and the movie names they have starred in.
3. List actor names in order of their appearance in the movie Fast and Not Furious.
4. List all actor names who have worked for the studio Paramount Pictures.
5. List the total number of movies produced by each studio.
6. List the names of actors who have not appeared in any movie
7. Find the average budget of movies produced by each studio.
8. List all studios located in Burbank and the total number of actors who have worked for these studios.

***Answer***:

A . SELECT StudioName, City FROM STUDIO WHERE StudioName LIKE ‘\_\_r%’

B . SELECT act.ActorName, mv.MovieName FROM ACTOR AS act JOIN MOVIEACTOR AS movact ON act.ActorID = movact.ActorID JOIN MOVIE AS mv ON movact.MovieID = mv.MovieID

C . SELECT act.ActorName FROM ACTOR AS act JOIN MOVIEACTOR AS movact ON act.ActorID = movact.ActorID JOIN MOVIE AS mv ON movact.MovieID = mv.MovieID WHERE mv.MovieName = ‘Fast and Not Furious’ ORDER BY movact.OrderOfAppearance

D . SELECT act.ActorName FROM ACTOR AS act JOIN MOVIEACTOR AS movact ON act.ActorID = movact.ActorID JOIN MOVIE AS mv ON movact.MovieID = mv.MovieID JOIN STUDIO AS sto ON mv.StudioID = sto.StudioID WHERE sto.StudioName = ‘Paramount Pictures’  
  
E . SELECT sto.StudioName, COUNT(\*) FROM STUDIO AS sto JOIN MOVIE AS mv on sto.StudioID = mv.StudioID GROUP BY sto.StudioName

F . SELECT ActorName FROM ACTOR AS act JOIN MOVIEACTOR AS movact ON act.ActorID = movact.ActorID WHERE movact is NULL

G . SELECT sto.StudioName, AVG(mv.MovieBudget) FROM STUDIO AS sto JOIN MOVIE AS mv ON sto.StudioID = mv.StudioID GROUP BY sto.StudioName

H . SELECT sto.StudioName, COUNT(\*) FROM STUDIO AS sto JOIN MOVIE AS mv ON sto.StudioID =mv.StudioID JOIN MOVIEACTOR AS mvact ON mv.MovieID = mvact.MovieID JOIN ACTOR ON AS act mvact.ActorID = act.ActorID WHERE sto.City = ‘Burbank’ GROUP BY sto.StudioName

**Question 4 (2 points):**

MediCore Health Services has gathered detailed data on patient visits from the past five years, containing millions of records with attributes like medical history, test results, prescriptions, and treatment outcomes. Due to limited computational resources and the need for faster processing, the analytics team needs to prepare this data for further analysis to identify trends and insights.

Considering the following requirements:

* The team needs to minimize computational costs while ensuring that the most important features of the data are retained.
* The data must be processed in such a way that the analysis can effectively focus on relevant patterns without being overwhelmed by noise or irrelevant information.

Which approach—data sampling or data reduction—would be most appropriate for preparing this dataset for analysis?

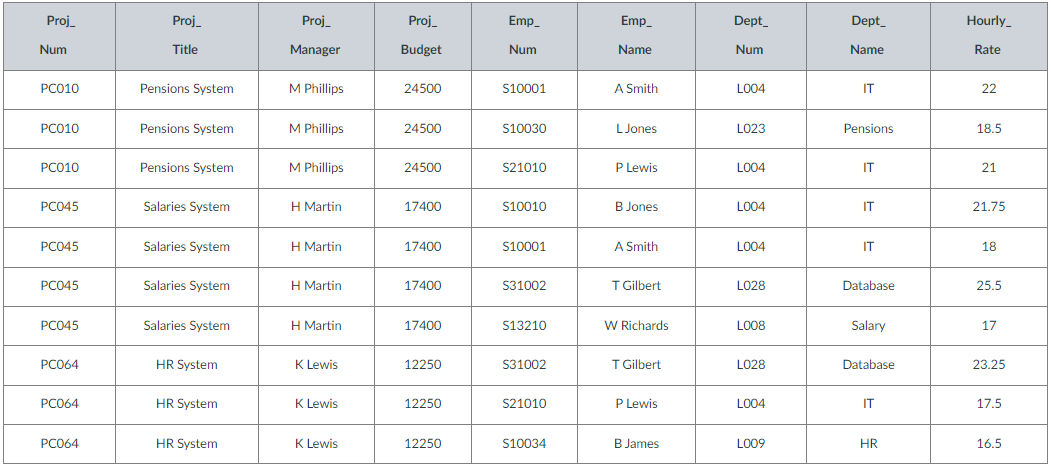
Explain your choice with a clear justification.

***Answer***:

I would choose data reduction. Though both the approaches can reduce computational cost, but data sampling could cause important features of the data be missing. In other hand, data reduction can help reduce noise or irrelevant information.

**Question 5 (5 points):**

Given the following table and functional dependencies:



**a.** Proj\_Num → Proj\_title, Proj\_Manager, Proj\_Budget

**b.** Emp\_Num → Emp\_Name, Dep\_Num, Dept\_Name

**c.** Dept\_Num → Dept\_Name

**d.** {Proj\_Num, Emp\_Num} → Hourly\_Rate

Convert the table into 3NF tables showing all primary keys and foreign keys.

Note that you can either use dependency diagram or functional dependencies for each relation while applying normalisation steps.

**Answer:**

Project(Proj\_Num , Proj\_title, Proj\_Manager, Proj\_Budget)

Employee(Emp\_Num, Emp\_Name)

Department(Dept\_Num, Dept\_Name)

ProjectEmployee(Proj\_Num, Emp\_Num, Hourly\_Rate)

**Question 6 (10 points):**

The PharmaCity is a chain of pharmacies, below is the information:

* Pharmacies appoint a supervisor for each contract. There must always be a supervisor for each contract, but the contract supervisor can change over the lifetime of the contract.
* Each pharmacy has a name, address, and phone number.
* Every patient has a primary physician. Every doctor has at least one patient.
* Each pharmacy sells several drugs and has a price for each. A drug could be sold at several pharmacies, and the price could vary from one pharmacy to another.
* Patients are identified by their Medicare numbers, and their names, addresses, and ages must be recorded.
* Doctors are identified by a Practitioner number. For each doctor, the name, specialty, and years of experience must be recorded.
* Each pharmaceutical company is identified by name and has a phone number.
* For each drug, the trade name and formula must be recorded. Each drug is sold by a given pharmaceutical company, and the drug name identifies a drug uniquely. If a pharmaceutical company is deleted, you need not keep track of its products any longer.
* Doctors prescribe drugs for patients. A doctor could prescribe one or more drugs for several patients, and a patient could obtain prescriptions from several doctors.
* Each prescription has a date and a quantity associated with it. You can assume that, if a doctor prescribes the same drug for the same patient more than once, only the last such prescription needs to be stored.
* Pharmaceutical companies have long-term contracts with pharmacies. A pharmaceutical company can contract with several pharmacies, and a pharmacy can contract with several pharmaceutical companies. For each contract, you have to store a start date, an end date, and the text of the contract.

According to the information above, draw a corresponding logical ERD using the Crow’s Foot notations. Please note:

* You can decide the names for entities, attributes and relationships, but you should use UPPERCASE to name your entities and use Capitalisation to name attributes.
* Identity the primary key (PK) and foreign key (FK) attributes.
* Clearly indicate connectivity and participation between entities.
* Document all assumptions that you have made. Be realistic and not contradict anything that is in the description.

In particular, in the ERD you need to show:

1. All entities with their attributes, including PK and FK attributes,
2. Relationship (with name) between entities, and
3. Connectivity and participation between entities

**Answers**:

A diagram of a patient

Description automatically generated

*End of questions*