THE UNIVERSITY OF NEW SOUTH WALES
Final Exam

COMP9311
Database Systems

TERM 2, 2024

- Time allowed: **3 hours 30 minutes (1:55PM to 5:25PM, 15 August, Sydney Time.)**
- Total number of questions: **6**
- Total number of marks: **100**
- You can answer the questions in **any order**, allocate your time wisely.
- Start each question on a **new page**.
- We accept any format: directly answering using Word or handwriting and converting to PDF. We only require the file to be clear and in **.pdf** format.
- Please place all your answers in a single file.
- Submit your answer file via **Moodle**. You may submit multiple times by adding files in your submission, and we will mark your last submission before 5:25PM.
- **Check your submission before you submit.**
- You are not allowed to share the exam paper online or with anyone else.

Question 1 (16 marks)

(a) (8 marks) Draw an ER diagram to represent the following set of application requirements for a museum database. You must use the drawing conventions in the lecture notes. Clearly state any additional reasonable assumptions if you make any.

- Each painting is uniquely identified by its ID. For each painting, we also record its title and rarity. All paintings are on display in a gallery. We also want to keep the names of all the painting's artists.
- Each painting has at least one supporting document. A supporting document contains information including rating, condition, and the year of the report. Each document has a document number that is only unique between documents belonging to the same painting.
- A gallery is a room within a museum, it is uniquely identified by a gallery ID. For each gallery, we record the floor it is located on, its capacity, and its area code. Paintings are displayed in galleries: a gallery can have zero or more paintings on display. A gallery must be looked after by at least one museum staff member. A gallery can have multiple equipment.
- We also want to store information about the museum staff, we want to keep their working hours. Each museum staff is uniquely identified by their staff ID, and we also record their name. Each museum staff must look after one or more galleries.
- We also want to keep information for museum equipment. Each equipment is uniquely identified by its staff ID, and we also record equipment name, dimensions (height, width, length), and status (i.e., whether in use or not). Each equipment must be stored in a gallery and must be managed by one or more museum staff.

(b) (8 marks) Translate the following ER diagram into a relational model. You must use the drawing conventions in the lecture notes.

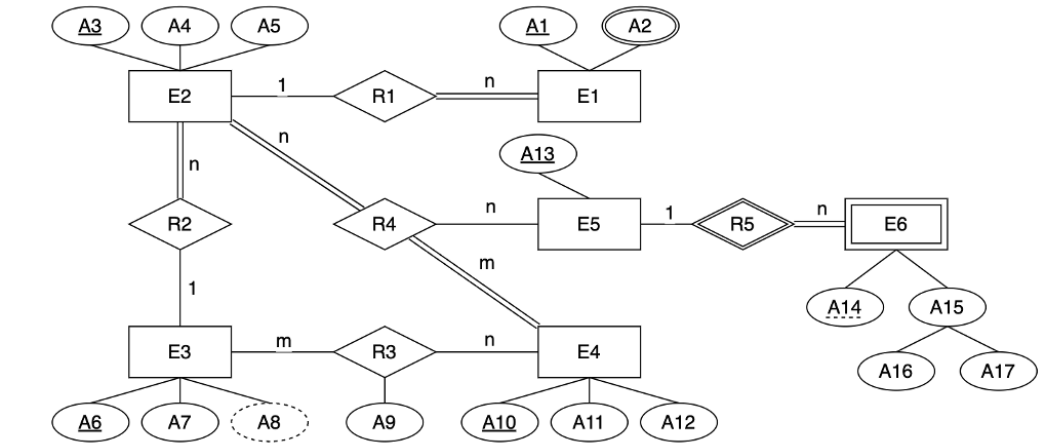

Figure 1: ER Diagram of Q1(b).

Question 2 (20 marks)

Consider the following relational database schema for a car sale database. The database schema consists of 6 relations, with the their names and attributes shown below. The underlined attribute names in each relation indicate that the combination of their values uniquely identifies each tuple (i.e., primary key). Note that attributes other than the primary key may not be unique among all tuples.

- Customer (cusID, cusName, phone)
- Manufacturer (manuID, makName, foundedYear, country)
- Car (carID, manuID, model, year, bodyType, status (available/sold))
- Sale (carID, cusID, salpID, saleYear, salePrice)
- Salesperson (salpID, salpName, rating)
- Service (serID, carID, sYear, sCost)

Answer the following queries (if not possible, provide a brief explanation). State any reasonable assumptions you made.

(a) (5 marks) Express the following query using **SQL**: Find all the names of the salesperson who have ever sold a car manufactured by Toyota.

(b) (5 marks) Express the following query using **SQL**: Find all the names of the salespersons who have ever sold cars from every manufacturer (i.e., the salespersons who have sold cars from all brands).

(c) (5 marks) Express the following query using **relational algebra**: Find the names of customers who have purchased cars that are both 'SUV' (bodyType) and were manufactured by German manufacturers. Besides, the cars have undergone servicing last year (suppose the current year is 2024).

(d) (5 marks) Express the following query using **relational algebra**: Find the names of salespersons for whom the average sales price of the cars they have sold is greater than the average sales price of cars sold by salespersons with a rating over 4.9.

Question 3 (28 marks)

(a) (18 marks) Consider the relational schema $R(A, B, C, D, E, G, H, I)$ and a set of functional dependencies $F = \{AH \rightarrow EG, AC \rightarrow DH, BDH \rightarrow A, DE \rightarrow I, G \rightarrow BC, E \rightarrow D\}$. Note that A, B, C, D, E, G, H, I and J are attributes. Justify your answer to each following question.

i. Check if $ABE \rightarrow DI$. Justify your answer. *(2 marks)*

ii. Find all the candidate keys for R. *(4 marks)*

iii. Determine the highest normal form of R. Justify your answer. *(2 marks)*

iv. Find a minimal cover F_m for F. *(3 marks)*

v. Provide a lossless-join and dependency-preserving decomposition of R into 3NF. *(3 marks)*

vi. Provide a lossless-join and dependency-preserving decomposition of R into BCNF, if possible. Justify your answer. *(4 marks)*

(b) (10 marks) **A functional dependency $X \rightarrow Y$ means that a value of X can determine the value of Y.** Based on this concept, we introduced several normal forms, e.g., 2NF, 3NF, BCNF, to reduce the redundancy in our course. However, these normal forms are not always sufficient in real applications. Consider a relation $Teaching(course, lecturer, student, grade)$ that represents the relationship between the course, staff, student, and grade in UNSW. Each course is delivered by one or more lecturers, and there are at least one student enrolled in each course. Grade records the grade of each student in each course.

Example. *Below is an example of this relation. There are two courses C_1 and C_2, C_1 has two lecturers L_1 and L_2 and two students S_1 and S_2, C_2 has one lecturer L_1 and two students S_1 and S_3.*

course	*lecturer*	*student*	*grade*
C_1	L_1	S_1	80
C_1	L_1	S_2	70
C_1	L_2	S_1	80
C_1	L_2	S_2	70
C_2	L_1	S_1	85
C_2	L_1	S_3	75

Give your answer to the following questions and justify your answer.

i. Below lists some possible functional dependencies in the relation, which is a valid functional dependency and which is not? Please explain why. *(2 marks)*
   - $course \rightarrow lecturer$
   - $course \rightarrow student$
   - $lecturer, student \rightarrow grade$
   - $course, student \rightarrow grade$

ii. Based on your answer to (i), calculate the candidate keys for the relation. *(2 mark)*

iii. Based on your answer to (i) and (ii), provide a lossless-join decomposition of R into BCNF. *(2 marks)*

iv. BCNF removes redundancies caused by functional dependencies, but other forms of redundancy may still exist even after decomposing a relation into BCNF. Please identify and explain the remaining redundancy in the BCNF decomposition of the given relation and provide a lossless-join decomposition of the relation that eliminates all remaining redundancy. *(4 marks)*

Question 4 (11 marks)

(a) (11 marks) In a database system, assume that the buffer pool is initially empty and there are **3** buffer frames in the buffer pool. Consider the following page requests from a transaction:

$$5, 2, 6, 5, 2, 4, 7, 3, 6, 1, 7, 4, 3, 7, 1$$

(Page 5 is first read from disk, then page 2, page 6, ...)

i. Sketch the process of how blocks are replaced in the *Least Recently Used (LRU)* policy. For each page request, indicate whether it's a 'hit' or a 'miss'. *(3 marks)*

ii. Sketch the process of how blocks are replaced in the *Most Recently Used (MRU)* policy. For each page request, indicate whether it's a 'hit' or a 'miss'. *(3 marks)*

iii. Consider a new page replacement policy where the *Least Frequently Used (LFU)* pages are replaced. In this policy, each page in the buffer has an associated counter. Each time a reference is made to a page, its counter is increased by one. When the cache reaches capacity and a new block needs to be inserted, the system will search for the buffer slot with the lowest counter and remove that page from the cache. If two pages have the same counter value (i.e., frequency), the page that was least recently used among them will be removed. Sketch the process of how blocks are replaced under this policy. For each page request, indicate whether it's a 'hit' or a 'miss'. *(3 marks)*

iv. Calculate the page hits rates of the above three policies and justify which policy performs better. *(2 marks)*

Question 5 (17 marks)

(a) (4 marks) Consider the following two schedules of the set of transactions T_1, T_2, T_3, T_4. ($R_i(X)$ means transaction T_i reads the value of X from the database, and $W_i(X)$ means transaction T_i writes the value of X to the database.)

$$S1 : R_1(B), R_2(A), W_3(D), W_3(B), W_1(C), R_4(D), R_3(A), R_2(C), W_3(A), W_2(C)$$

$$S2 : W_3(D), W_3(B), W_1(C), R_3(A), R_2(C), R_4(D), R_2(A), W_2(C), R_1(B), W_3(A)$$

i Is $S1$ conflict serializable? If yes, please provide the equivalent serial schedule. If no, please justify your answer using the precedence graph. *(2 marks)*

ii Is $S2$ conflict serializable? If yes, please provide the equivalent serial schedule. If no, please justify your answer using the precedence graph. *(2 marks)*

(b) (7 marks) Consider the following two transactions, suppose the initial values of X, Y, Z are 10, 20 and 30 respectively.

T_1	T_2
Read(Z)	Read(Z)
Read(Y)	Read(X)
Read(X)	Read(Y)
Y=Y-X+Z	X=X-Y+Z
Write(Y)	Write(X)

i If there is no control over the execution of T_1 and T_2, what might be the values of X, Y and Z after the execution of the two transactions, respectively? List all the possible values. *(2 marks)*

ii Add appropriate locks to the schedule to make it satisfy the two-phase locking protocol and list possible values of X, Y and Z after the execution of the two transactions, respectively. *(3 marks)*

iii Give an example schedule of the two transactions T_1 and T_2 which causes deadlock by using the two-phase locking protocol in (ii). *(2 marks)*

(c) (6 marks) Consider the following log file of a database system. The log file contains the log records of the transactions T_1, T_2, T_3, T_4. The log records are ordered by their timestamps. (e.g., [read_item, T, X, 3] means transaction T reads the value 3 of the item X in the database. [write_item, T, X, 3, 4] means transaction T change the value 3 of the item X in the database to 4.])

Time	Log record
t_1	[start_transaction, T_1]
t_2	[read_item, T_1, X, 2]
t_3	[start_transaction, T_2]
t_4	[write_item, T_1, X, 2, 4]
t_5	[read_item, T_2, Y, 8]
t_6	[commit, T_1]
t_7	[read_item, T_2, X, 2]
t_8	[checkpoint]
t_9	[start_transaction, T_3]
t_{10}	[write_item, T_2, Y, 8, 7]
t_{11}	[commit, T_2]
t_{12}	[write_item, T_3, Z, 15, 20]
t_{13}	[start_transaction, T_4]
t_{14}	[write_item, T_4, X, 4, 8]
t_{15}	[read_item, T_3, X, 8]

i If the system crashes immediately after t_9, what should be done to the transactions T_1, T_2, T_3, T_4 when the system is restarted? Please briefly explain the reason. *(3 marks)*

ii The database system that generated the log file above is a buggy database implemented by a student in the COMP9315 Database Implementation course, and thus has several issues. Please help the student identify these issues and explain why the database violates the ACID properties. (The four ACID properties are: atomicity, consistency, isolation, and durability.) *(3 marks)*

Question 6 (8 marks)

Which type of database model (*relational, key-value, document, column-family, graph, or "it depends"*) is the most suitable choice for each of the following scenarios? Justify your answer.

- Build a database for storing pictures in an online forum. *(2 marks)*
- Build a movie recommendation system that needs to analyse a large amount of movie rating data. *(2 marks)*
- Build a system for friend recommendation for a social app. *(2 marks)*
- Build an app for a startup company with rapidly evolving business needs. *(2 marks)*

END OF EXAM PAPER