PRIFYSGOL ABERYSTWYTH UNIVERSITY

SEMESTER ONE 2026

COMPUTER SCIENCE

CS27020 : MODELLING PERSISTENT DATA

Time allowed: **2 HOURS**

Answer all **THREE** questions

All questions carry equal marks.

YOU MAY BRING THE FOLLOWING INTO THE EXAMINATION:
UP TO 3 SHEETS OF NOTES ON A4 PAPER (WHICH MAY BE DOUBLE-SIDED). THERE ARE NO ADDITIONAL RESTRICTIONS ON CONTENT OR FORMATTING OF THESE NOTES.
NO ELECTRONIC DEVICES.

CALCULATORS ARE NOT PERMITTED.

STUDENTS MAY SUBMIT ANSWERS TO THIS PAPER IN EITHER WELSH OR ENGLISH

Once seated, students may complete the front cover of the answer book(s) and the attendance slip ONLY.

DO NOT TURN THE EXAMINATION QUESTION PAPER OVER UNTIL INSTRUCTED TO DO SO.

1. **This question is about data integrity and SQL.**

a) A laptop lending service stores information about laptop loaning . The Laptops table holds device details. The Loans table records each loan. Below is the given skeleton SQL definition for relations of Laptops and Loans.

```
CREATE TABLE laptops (
  laptop_id INT,
  model VARCHAR(100)
);

CREATE TABLE loans (
  loan_id INT,
  laptop_id INT,
  user_id VARCHAR(10),
  loan_date DATE,
  due_date DATE,
  return_date DATE
);
```

(i) Add basic constraints using SQL DDL only (no functions or triggers) to enforce the integrity rules below. Implement the following (slight deviations from SQL syntax will not be penalised):

- Primary keys on both tables.
- Loans.laptop_id must reference Laptops.laptop_id.
- In Loans table, laptop_id, user_id, loan_date, due_date are NOT NULL.
- In Loans table, due_date must be on/after loan_date, and return_date must be NULL (for active loans) or on/after loan_date.

[9 marks]

(ii) Recommend one way to ensure at most one active loan per laptop (`return_date IS NULL`). Briefly describe the approach and justify your choice. No implementation code is required. [3 marks]

b) Transaction management and concurrency control are key RDBMS functions for ensuring data integrity.

(i) Consider the following interleaved schedule S for transactions T_1 and T_2.

$$S = [W_1(B),\ R_2(B),\ R_1(A),\ W_2(B)]$$

Here, $R_i(X)$ and $W_i(X)$ denote a read and a write of item X by transaction T_i.

Is this schedule serialisable? Justify your answer. [6 marks]

(ii) Suppose you need to excecute a single-statement query in PostgreSQL to list each laptop's current status (see Question 1a). Choose an appropriate isolation level from: READ COMMITTED, REPEATABLE READ, or SERIALIZABLE, and justify your choice. [3 marks]

c) Consider the following `Laptops` and `Loans` tables, which adhere to the specified integrity constraints in Question 1a.

Laptops Table

laptop_id	model
1	Dell X1
2	Dell X1
3	Apple MacBook
4	Lenovo ThinkPad

Loans Table

loan_id	laptop_id	user_id	loan_date	due_date	return_date
1001	1	abc1	2025-01-10	2025-04-10	2025-04-10
1002	1	abc1	2025-04-10	2025-07-10	2025-07-10
1003	1	abc1	2025-07-10	2025-10-10	NULL
1004	2	cat2	2025-03-05	2025-09-05	NULL
1006	2	mad2	2024-09-01	2025-03-01	2025-02-15
1005	4	sun1	2025-08-15	2025-11-15	2025-10-31

(i) List the results for the following SQL query, which is intended to find all `user_ids` who have been loaned a `Dell X1` model. [4 marks]

```
SELECT DISTINCT T2.user_id
FROM Laptops T1
INNER JOIN Loans T2 ON T1.laptop_id = T2.laptop_id
WHERE T1.model = 'Dell X1';
```

(ii) What is the output of the query from Question 1c(i) if `INNER JOIN` is replaced with `LEFT OUTER JOIN`? [4 marks]

(iii) Consider the two SQL Queries A and B listed below, both intended to find users associated with multiple loans. Explain the difference between the results that Query A and Query B would produce, and state what each query is actually counting. [4 marks]

Query A

```
SELECT user_id
FROM Loans
GROUP BY user_id
HAVING COUNT(loan_id) > 1;
```

Query B

```
SELECT user_id
FROM Loans
GROUP BY user_id
HAVING COUNT(DISTINCT laptop_id) > 1;
```

2. **This question is about theory and normalisation.**

Table 1 shows sample data on patient visits to various hospital departments. Each row represents a diagnosis or treatment given during a single patient visit.

Table 1: Hospital_Visits Table (Sample Data)

Visit ID	Patient ID	Patient Name	VisitDate	Doctor ID	Doctor Name	Department Name	Diagnosis	Treatment Code	Treatment name
V101	PAT1	Smith	2025-10-01	DOC1	Evans	Cardio	Hypertension	C5	ECG
V101	PAT1	Smith	2025-10-01	DOC1	Evans	Cardio	Hypertension	C2	BP
V102	PAT2	Doe	2025-10-01	DOC5	Patel	Ortho	Ankle Injury	C3	X-Ray
V103	PAT1	Smith	2025-10-01	DOC1	Evans	Cardio	Follow-up	C2	BP
V104	PAT3	Jones	2025-10-02	DOC5	Patel	Ortho	Knee Pain	C3	X-Ray
V104	PAT3	Jones	2025-10-02	DOC5	Patel	Ortho	Knee Pain	C1	MRI

Assumptions:

- PatientID uniquely identifies each patient (PK).
- DoctorID uniquely identifies each doctor (PK).
- VisitID uniquely identifies each patient visit (PK).
- TreatmentCode uniquely identifies each treatment procedure (PK).
- Each doctor belongs to exactly one department.

Key functional dependencies:

- VisitID $\rightarrow$ {PatientID, PatientName, VisitDate, DoctorID, DoctorName, DepartmentName}
- PatientID $\rightarrow$ PatientName
- DoctorID $\rightarrow$ {DoctorName, DepartmentName}
- TreatmentCode $\rightarrow$ TreatmentName
- {VisitID, TreatmentCode} $\rightarrow$ Diagnosis

Note: For all parts of this question, please provide only the database schema or table headers, clearly identifying all Primary Keys (PK) and Foreign Keys (FK). Populating the tables with data is not required.

a) Explain the relationship between a superkey and a candidate key. [3 marks]

b) Identify the candidate key(s) for Table 1. [4 marks]

c) Using Table 1, identify and explain one type of data anomaly that could occur due to its structure. Provide a specific example related to the hospital data. [4 marks]

d) What normal form is Table 1 currently in? Justify your answer. [4 marks]

e) Is Table 1 in Second Normal Form (2NF)? If not, perform the step-by-step normalization of Table 1 into 2NF. Clearly show the schema indicating PKs and FKs, and justify your changes. [7 marks]

f) Are the tables generated in Question 2e in Third Normal Form (3NF)? If not, convert the 2NF schema into 3NF. Justify any changes made. [7 marks]

g) Are all the tables in your final 3NF schema (from Question 2f) also in Boyce-Codd Normal Form (BCNF)? Explain why or why not for each table. [4 marks]

3. **This question is about semi-structured data and NoSQL.**

a) Examine the following XML file that lists specifications for the laptops shown in Question 1c).

```
<laptops>
  <laptop id="1" model="Dell X1" specs="">
  </laptop>
  <laptop id="2" model="Dell X1" serial="DX1-002">
    <specs>
      <ram unit="GB">32</ram>
    </specs>
  </laptop>

  <laptop id="3" brand="Apple" model="MacBook">
    <specs chipset="M2" ramGB="8"/>
  </laptop>

  <laptop id="4" model="ThinkPad" brand="Lenovo">
    <specs>
      <cpu>Ryzen 7 7840U</cpu>
      <ram unit="GB">16</ram>
      <storage type="SSD" capacityGB="512"/>
      <os>Windows 11</os>
    </specs>
  </laptop>
</laptops>
```

(i) List two characteristics of the XML above that make it semi-structured, and give a brief example for each. [4 marks]

(ii) XML uses elements, tags, and attributes. Match each item to its category (element, tag, or attribute).

(1) `model="Dell X1"`

(2) `<ram unit="GB">16</ram>`

(3) `<laptop>`

[3 marks]

(iii) JSON is an alternative to XML for semi-structured data. Write down a json representation for the laptop with `id = 3`. [5 marks]

(iv) Using the JSON object for the laptop `id = 3` from Question 3a(iii), fill in the JSONPath expressions to extract:

```
(1) the chipset
$._______

(2) the full specs as a JSON object
$._______
```

[4 marks]

b) A webinar platform runs its services (live video stream and chat) across two regions, with asynchronous replication. During a network partition, the live stream continues in one region, but audience in the other lose the stream updates.

(i) State the CAP theorem and, and specify, under a partition, which two of the three properties the platform can still provide. [4 marks]

(ii) For the live chat feature, should the platform prioritise CP or AP during the partition? Briefly justify. [4 marks]

c) Evaluate the following statements regarding NoSQL systems and their data models. Indicate whether each statement is True or False. Briefly explain your answers. [9 marks]

S1. A key-value store does not require a predefined schema for values.

S2. MongoDB natively enforces foreign key constraints between collections.

S3. A webinar platform handles a 10x traffic spike by adding 8 more identical app servers aiming to increase its vertical scalability.

END OF PAPER