London Business School

Database Management

AM05
Masters in Analytics and Management (MAM)

Dr. David Tilson

Date: 17 October 2019
Start Time: 9.00 AM
Duration: 90 minutes

INSTRUCTIONS FOR STUDENTS:

- **Write your LBS number and stream (if applicable) in the spaces below:**

LBS no.
(printed on your seat label and the Academic Honour Code)

Stream

- This is a closed exam.
- You may use a calculator. However, devices that store data or connect to the internet are not allowed.
- You are responsible for ensuring that you hand your completed answers to the invigilator with all relevant answer sheets stapled together.
- If the question is not clear, state your assumptions and if they are reasonable you will be given credit.
- A total of 50 points are available for this exam. Allocate your time optimally.

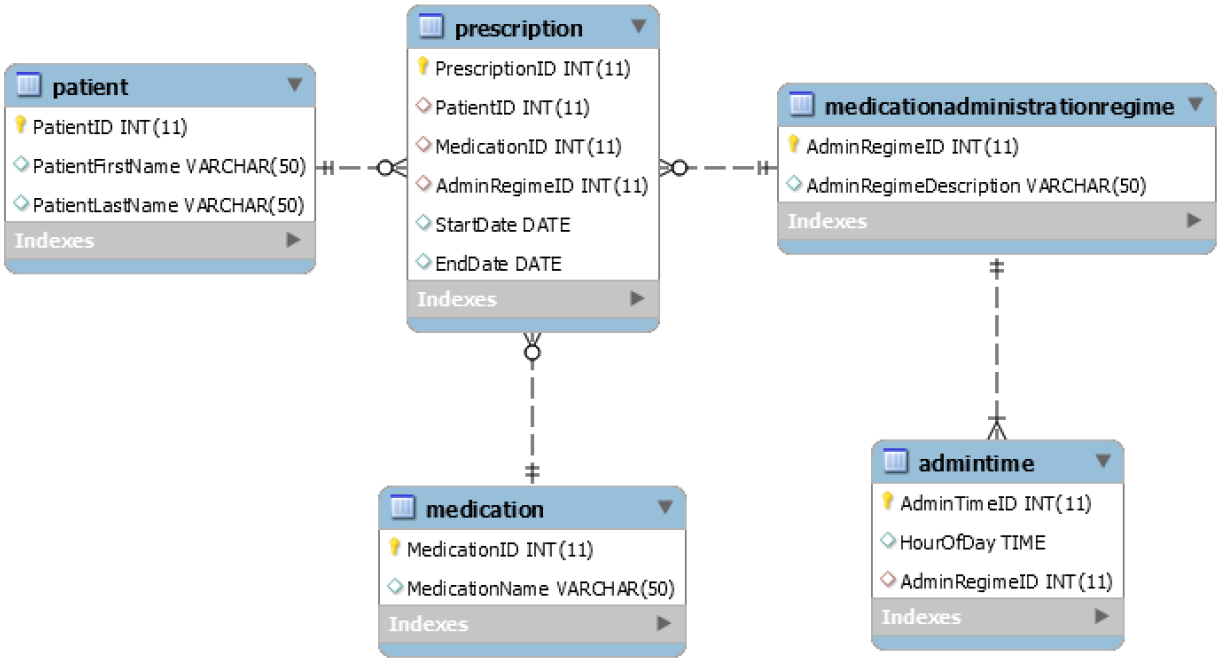
FOR OFFICE USE ONLY – Please complete with total score for the exam AND score achieved per question.

Question	Points	Score	Question	Points	Score
1	20		**6**		
2	10		**7**		
3	10		**8**		
4	10		**9**		
5			**10**		
			Total	50	

Grader initials:	IM sample: Yes ☐	Central Services initials:

QUESTION 1

This question is set in the context of a small database that stores information about patients, medications, and prescriptions. A partial schema is shown in the figure below:



The tables are populated with the following data:

SELECT * FROM Prescription;

PrescriptionID	PatientID	MedicationID	AdminRegimeID	StartDate	EndDate
2	1	1	3	2019-10-12	2019-10-15
3	1	3	2	2019-10-13	2019-10-16
4	2	3	1	2019-10-14	2019-10-16
5	4	2	2	2019-10-10	2019-10-11

SELECT * FROM Patient;

PatientID	PatientFirstName	PatientLastName
1	Alice	Blue
2	Bob	Costa
3	Charlie	Darwin
4	Derek	Eagle

SELECT * FROM Medication;

MedicationID	MedicationName
1	Tylenol
2	Nyquil
3	Ibuprofen
4	Paracetamol

SELECT * FROM MedicationAdministrationRegime;

AdminRegimeID	AdminRegimeDescription
1	Morning and Night
2	Every 4 hours
3	Every 6 hours

SELECT * FROM AdminTimes;

AdminTimeID	HourOfDay	AdminRegimeID
1	09:00:00	1
2	21:00:00	1
3	12:00:00	2
4	16:00:00	2
5	20:00:00	2
6	00:00:00	2
7	04:00:00	2
8	08:00:00	2
9	08:00:00	3
10	14:00:00	3
11	20:00:00	3
12	02:00:00	3

1-1 [2 points] Enter the results of this query in the table provided

```
SELECT PatientLastName, COUNT(*) AS Num
FROM Patient pat INNER JOIN Prescription pre ON
pat.PatientID = pre.PatientID
GROUP BY pat.PatientLastName
ORDER BY pat.PatientLastName ASC;
```

PatientLastName	Num
Blue	2
Costa	1
Eagle	1

1-2 [2 points] Enter the results of this query in the table provided

```
SELECT PatientLastName, COUNT(*) AS Num
FROM Patient pat LEFT OUTER JOIN Prescription pre ON
pat.PatientID = pre.PatientID
GROUP BY pat.PatientLastName
ORDER BY pat.PatientLastName ASC;
```

PatientLastName	Num
Blue	2
Costa	1
Darwin	1
Eagle	1

1-3 [2 points] Enter the results of this query in the table provided

```
SELECT PatientLastName, COUNT(pre.MedicationID) AS Num
FROM Patient pat LEFT OUTER JOIN Prescription pre ON
pat.PatientID = pre.PatientID
GROUP BY pat.PatientLastName
ORDER BY pat.PatientLastName ASC;
```

PatientLastName	Num
Blue	2
Costa	1
Darwin	0
Eagle	1

1-4 [2 points] Are the results of the last two queries the same or different? Explain why that is the case.

Answer: The results of the queries are not the same. Query in 1-2 counts all the rows in the Prescriptions table that are associated with rows in the Patient table. Query in 1-3 **counts all the non-Null entr**ies in the MedicatonID field of the Prescription table that are associated with the corresponding rows of the Patient table. So, the NULL in that field for Darwin is not included in the count.

1-5 [2 points] Enter the results of this query in the table provided

```
SELECT HourOfDay, COUNT(*) AS NumRegimes
FROM AdminTime
GROUP BY HourOfDay
HAVING COUNT(*) > 1;
```

HourOfDay	NumRegimes
20:00:00	2
08:00:00	2

1-6 [2 points] Enter the results of this query in the table provided

```
SELECT AVG(x.NumRegimes) AS AvgOfNumRegimes
FROM (SELECT HourOfDay, COUNT(*) AS NumRegimes
FROM AdminTime
GROUP BY HourOfDay
HAVING COUNT(*) <2 )  AS x;
```

AvgOfNumRegimes	
1	

1-7 [2 points] Enter the results of this query in the table provided

```
SELECT COUNT(PatientLastName) AS NumPatientNames
FROM Prescription pre INNER JOIN Patient pat
ON pre.PatientID = pat.PatientID;
```

NumPatientNames	
4	

1-8 [2 points] Enter the results of this query in the table provided

```
SELECT PatientLastName
FROM Patient pat WHERE NOT EXISTS
(SELECT * FROM Prescription pre
WHERE pre.PatientID = pat.PatientID);
```

PatientLastName	
Darwin	

1-9 [2 points] Questions 1-6 and 1-8 include subqueries. Which one is a correlated subquery? How do you know?

Answer: Question 1.8 contains a correlated subquery. The internal query (subquery) is correlated because it uses a value from the outside query (pat.PatientID)

1-10 [2 points] Enter the results of this query in the table provided

```
SELECT PatientLastName, PrescriptionID
FROM Patient pat LEFT OUTER JOIN Prescription pre
ON pat.PatientID = pre.PatientID
WHERE PrescriptionID IS NULL;
```

PatientLastName	PrescriptionID
Darwin	NULL

QUESTION 2 [10 points]

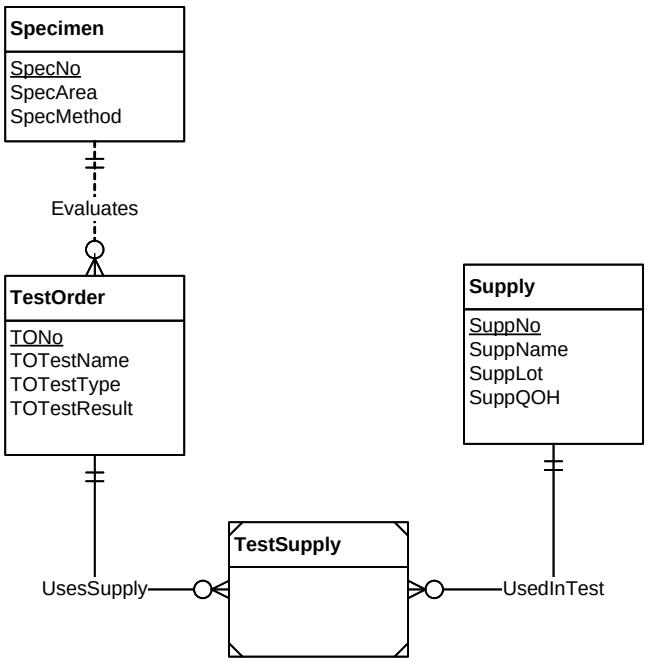
Draw an ERD for the following situation

- A laboratory collects specimens that may later be analysed. For each specimen collected, the database should record a unique SpecNo. It should also specify SpecArea, and SpecCollMethod
- A specimen is analysed when a test order is issued. A specimen may not have a test order until after a considerable delay
- A test order contains a unique test order number (TONo), TOTestName, TOTestType and TOTestResult
- A test order is created for exactly one specimen
- The database should keep track of supplies needed for test orders
- A test order can use a collection of supplies (0 or more) and a supply can be used on a collection of test orders (0 or more). The Supply entity type contains a unique SuppNo, SuppName, SuppLotNo, and SuppNoInStock

Notes

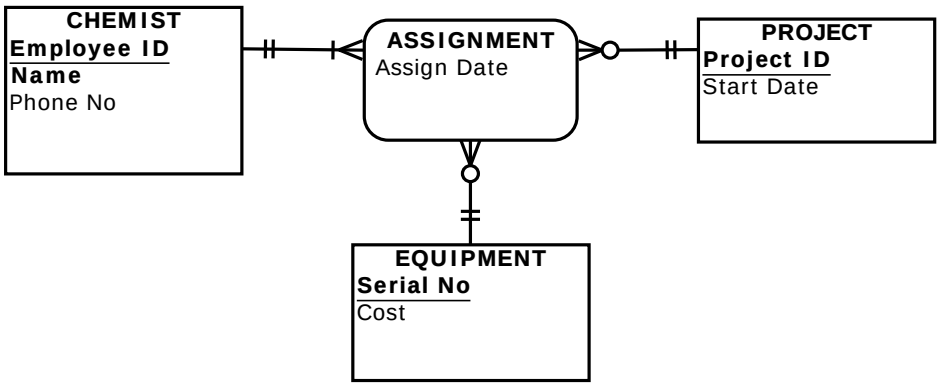
- M:N relationships should be modelled with associative entities
- Choose appropriate names for all relationships and entity types based on your common knowledge of test orders and supplies
- Use doubled line relationships and rectangles to represent weak entities. Underline identifiers that are likely to become primary keys

Answer: would include these entities, relations, cardinalities, and attributes.



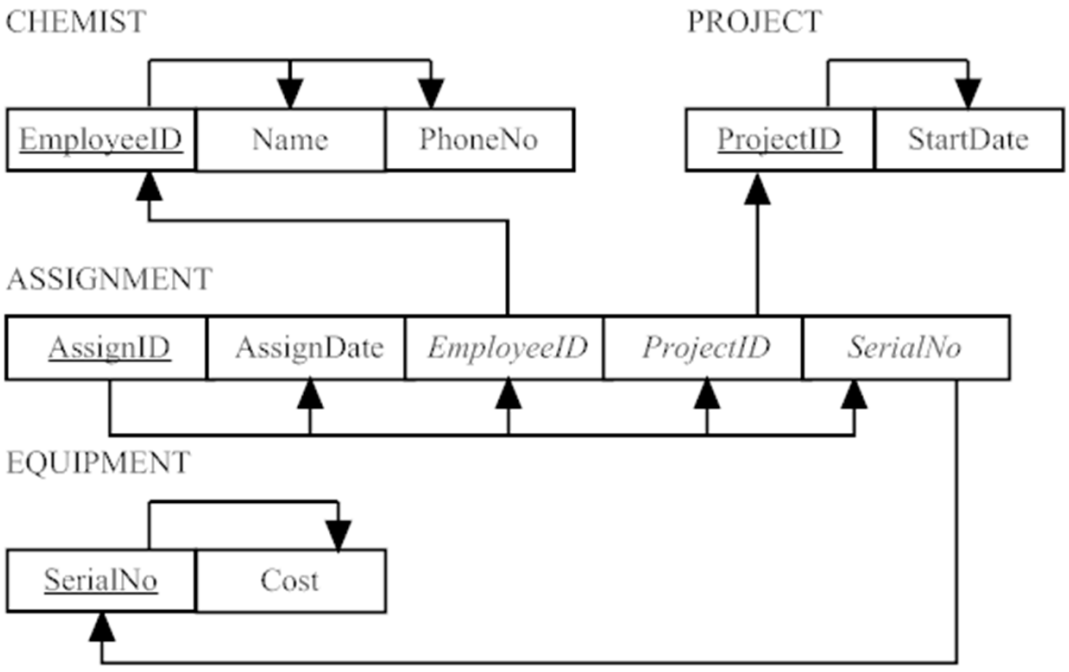
QUESTION 3

The following ERD represents a data model for tracking the allocation of laboratory equipment to chemists working on projects.



3-1 [8 points] Convert the ERD into a set of relational schemas. Indicate the functional dependencies, and the PK-FK relationships with arrows. Convert all relations into third normal form (3NF). Use this sort of format to represent relations.

Answer



3-2 [2 points] How do you know that the schema that you created for question 3-1 is in third normal form?

Answer: Because there no remaining functional dependencies on anything other than primary keys in all of the relations.

QUESTION 4

4-1 [2 points] Give two disadvantages of the independent data mart architecture relative to the enterprise data warehouse architecture.

Answer: Any two of the following for full credit

1. Five claimed limitations of independent data marts:
2. A separate ETL process has to be developed for each data mart. This
3. A clear, enterprise-wide view of data may not be provided because data marts may not be consistent with one another.
4. Analysis is limited because there is no capability to drill down into greater detail or into related facts in other data marts.
5. Scaling costs are excessive as each new application creates a separate data mart, which repeats all the extract and load steps.
6. Attempting to make separate data marts consistent is expensive.

4-2 [2 points] What does the term data independence mean, and why is it an important goal?

Answer: Data independence refers to the separation of data descriptions from the application programs that use the data. It is an important goal because it allows an organization's data to change and evolve without changing the application programs that use the data. Additionally, data independence allows changes to application programs without requiring changes in data storage structure.

4-3 [2 points] What is the main thing that HDFS does that traditional file systems do not?

Answer: The main thing is that it provides a view of the data across all the nodes in the cluster (i.e. it is a distributed file system). It also provides other features like scalability and fault tolerance. Full credit for any of these.

4-4 [2 points] Briefly describe the main steps in ETL

Answer (1 point for just stating what ET and L stand for)

- Extract – capture snapshot of chosen source data
- Scrub\Cleanse – fix errors, misspellings, wrong dates, incorrect field usage, mismatched address, missing or duplicate sata,
- Transform – Convert data from format of operational system to dataware house as the right level of granuality
- Load - place transformed data into the warehouse (and create indexes

4-5 [2 points] Fill in the blanks

The __ **Repository** _ provides centralized storage for all data definitions, data relationships, and other system components in a RDBMS.

The **_Star Schema (or possibly the dimensional model)_** is a simple database design in which dimensional data are separated from fact or event data.