KEELE UNIVERSITY

OPEN-BOOK EXAMINATION, 2020/21

FHEQ Level 6

May 2021

COMPUTER SCIENCE

CSC-30002

ADVANCED DATABASES AND APPLICATIONS

Candidates should attempt to answer ALL THREE questions

1.

(a) Consider the following scenario and answer the question that follows.

A magazine publishing company publishes one regional magazine in each of four countries, namely, France (FR), Italy (IT), The Netherlands (NL) and the United Kindgom (UK). The company has 300,000 customers distributed throughout the four countries. On the first of each month, an annual subscription invoice is sent to each customer whose subscription is due for renewal. The company's management maintains the following single relational database that they wish to decentralise into the four regional subsidiaries:

CUSTOMER *(C_NUMBER, C_NAME, C_ADDRESS, C_COUNTRY, C_SUBSDATE, ...)*
INVOICE *(INV_NUMBER, INV_DATE, INV_TOTAL, C_NUMBER, ...)*

The management at the company's headquarters, however, will have access to customer and invoice data to generate annual reports and to issue ad hoc queries such as listing all customers by region and reporting all invoices by customer and by region.

Given those requirements, write PAL or SQL CREATE VIEW queries that will partition the database correctly. Show the node names and all essential attributes.

[40%]

(b) Three nodal relations located at nodes N1, N2 and N3, respectively, are:

```
N1: Emp1 (Eno, Ename, Location, Bdate, Gr-sal)
N2: Emp2 (E#, En, Loc, Net-sal, Tax)
N3: Emp3 (Eno, En, City, Birthd, Net-sal, Tax)
```

Assume that:

`E#` and `Eno` are employee numbers; `En` and `Ename` are employee names; `Location`, `Loc` and `City` are locations; `Net-sal` is the net salary; `Gr-sal` represent the gross salary and `Gr-sal = Net-sal + Tax`; `Bdate` and `Birthd` represent the date of birth.

This question continues on the **NEXT** page …

In addition, the unit of currency is pound sterling in N1 and N2, and US dollar in N3. The conversion between the two currencies is given by the expression £1 = CF*$1, where CF is the conversion factor or exchange rate. Furthermore, the location "London" is written as "Londres" in `Emp2`, and "Rome" as "Roma" in `Emp3`. (English is your default working language.)

Use PAL syntax to answer questions (i) and (ii) below.

(i) Integrate the three nodal relations to create the following global relation: `EMP (Eno, Ename, Loc, Bdate, Gsal).Gsal` represents gross salary.

[30%]

(ii) The following global query retrieves the employee number, location, salary and the date of birth of all employees with a gross salary exceeding £20000:

```
Q == ?[Eno, Loc, Bdate, Gsal] %
EMP : Gsal > 20000
```

Decompose this query into three nodal queries for retrieving the required data from nodes N1, N2 and N3.

[30%]

2.

(a) In the context of distributed updates and recovery, explain why the *two-phase commit protocol* is potentially a blocking protocol and discuss briefly how the *three-phase commit protocol* is a non-blocking protocol in the absence of complete site failure.

[20%]

(b) Three relations SUP, PART and SPL at nodes N1, N2, and N3, respectively, are:

N1	SUP	S#	SN	CITY	100 tuples
		5	5	20	30 bytes
N2	PART	P#	PN	COL	200 tuples
		5	10	10	25 bytes
N3	SPL	S#	P#	QTY	2000 tuples
		5	5	5	15 bytes

Where:

S# = Supplier number, SN = Supplier name, P# = Part number,

PN = Part name, COL = Colour, and QTY = Quantity

The following query originates from N1, and intends to retrieve the quantities of blue washers supplied by Joan or Mona of Paris. The total number of those tuples is 150. The PAL query for retrieving this information is given by:

```
?[S#, SN, P#, QTY] % SUP(S#)*(S#) SPL (P#)*(P#)
PART: [(SN = “Joan” | SN = “Mona”), CITY =
“Paris”, PN = “Washer”, COL = “Blue”]
```

Assume that the final result of executing this query has 150 tuples and there are 10 supplier tuples for Joan or Mona of Paris, and there are only 50 blue washers. For simplicity, assume also that suppliers have an equal share of supplies and all joins are *natural* joins.

This question continues on the **NEXT** page …

Answer questions (i), (ii) and (iii) below.

(i) Calculate the size of nodal partitions which will be required for evaluating the **minimal** data movements for the distributed query in the three strategies given in (ii) below.

[15%]

(ii) Calculate the minimal data movement for the following three strategies. Show all steps clearly.

Strategy 1: Move data from N2 to N3 for the partial evaluation, and then from N3 to N1 for the final evaluation.

[15%]

Strategy 2: Move data from N1 to N3, process the partial result there and then move the data to N2. Process the final result at N2 and move it to N1.

[15%]

Strategy 3: Move data from N1 and N2 to N3 for the complete evaluation, and then move the result to N1.

[10%]

(iii) Identify the best strategy from the three above and justify your answer.

[5%]

(c) Web-based e-commerce and e-business systems rely heavily on distributed database architectures for transaction processing. Explain why this is the case and discuss briefly the key operational issues that need to be addressed before distributed databases can be deployed on the Web.

[20%]

3.

(a) Discuss briefly the importance of *semantic interoperability* between Relational, XML, and Object data models and explain how this can be achieved.

[30%]

(b) Explain the key issues that need to be considered during the following stages of building a data warehouse.

(i) Data extraction and loading

(ii) Data management

(iii) Query processing

[20%]

(c) In designing the data warehouse database, what is meant by the term *denormalisation*? When is this concept appropriate and when is it not appropriate in designing the data warehouse database? Relate your answer to the *star, snowflake* and *starflake* schemas.

[20%]

(d) The use of relational databases in a data warehouse generally introduces significant computational overheads in *Online Analytical Processing* (OLAP). Explain why this is the case and describe how *Relational OLAP* (ROLAP) and *Multidimensional OLAP* (MOLAP) systems overcome this challenge. Give examples for each system to support your answer.

[30%]