Answer any three out of the following five questions

1. Answer the following three parts.

   a. Discuss the differences between the concepts of unique and referential identification of entities in the hierarchical and relational data models.

   [11 marks]

   b. A set of relational algebra operators is independent if *no* operator in this set is redundant, i.e. *no* operator in this set can be expressed by the others. For example, the intersection operator is not independent of the difference operator.

   Give a set of relational algebra operators which are both relationally complete and independent, and justify your choice of operators.

   [11 marks]

   c. Write a sentence or two on each of the following:

      1) physical data independence
      2) growth independence
      3) ISA entity types
      4) ID entity types
      5) the Universal Relation Schema Assumption

   [11 marks]

[Total 33 marks]

2. Answer the following three parts.

a. When we join two relations on the common attributes of their underlying relation schemas, tuples which do not match are discarded from the output. For example, suppose that we have in our database two relations, one containing the set of enrolled students and the other containing the set of students having registered for their exams. Then, when joining these two relations, tuples representing students who have not registered for their exams will be discarded. The *outer join* is an additional relational algebra operator which returns, in a single output relation, both the matched and discarded tuples; the discarded tuples are appropriately padded with null values.

   Show, with an example, how the outer join operator can be implemented by a relational algebra expression.

   [11 marks]

b. Assume that we represent a null value of type value "exists but is unknown" by a default value. For example, we could represent the fact that we do not know the salary of an employee by a default salary of zero pounds.

   Give two examples how this approach can go wrong.

   [11 marks]

c. A data warehouse is a subject-oriented, integrated and time-varying database which is primarily used for decision making purposes. As an example a sales data warehouse may contain the product sold, the time of sale, the place of sale and the sales person. Typically such a data warehouse will be of orders of magnitude larger than a database which does *not* contain specific sales data, but rather contains the company details, including details about product range, outlet locations and personnel.

   Argue whether or not such data warehouses should be normalised beyond first normal form.

   [11 marks]

[Total 33 marks]

CONTINUED

3. Answer the following three parts.

   a. Briefly compare the relational algebra to SQL and suggest why SQL has become the industry standard relational query language as opposed to the relational algebra.

   [11 marks]

   b. Give an efficient (polynomial time) algorithm for finding one key for a relation schema R with respect to a set of Functional Dependencies (FDs) F over R.

   [11 marks]

   c. What is the difference between a minimum and optimum cover of a set of FDs?

   Give examples of

   1) a set of FDs which is not minimum,
   2) a set of FDs is minimum but not optimum, and
   3) a set of FDs which is optimum.

   [11 marks]

[Total 33 marks]

4. Answer the following three parts.

a. Consider the World-Wide-Web (the Web) as a database, and a search engine over the Web as the query processor specifying the retrieval of information from the Web database.

Highlight the differences between querying the Web database using a search engine and querying a relational database using SQL.

[11 marks]

b. Why is Boyce-Codd Normal Form (BCNF) preferable, in principle, to Third Normal Form (3NF)?

In practice why would a database designer be content with a 3NF database schema that is not in BCNF ? (Justify your answer with an example.)

[11 marks]

c. You are given the following attributes concerning an employee relation schema, R: E# (Employee number), EN (Employee Name), ESD (Employee Start Date), EED (Employee End Date), EJT (Employee Job Title), P# (Project number), PN (Project Name), PSD (Project Start Date), PED (Project End Date), D# (Department number), DN (Department name) and M# (Manager number).

The semantics of R are given in terms of the following functional dependencies (FDs):

1) $\{\text{E\#, ESD}\} \rightarrow \{\text{EJT, EED}\}$
2) $\text{E\#} \rightarrow \{\text{EN, P\#, PN, D\#}\}$
3) $\text{P\#} \rightarrow \{\text{PN, PSD, PED}\}$
4) $\text{D\#} \rightarrow \{\text{DN, M\#}\}$
5) $\text{M\#} \rightarrow \text{D\#}$

Decompose R into a BCNF database schema which is a lossless join and dependency preserving decomposition with respect to F.

[11 marks]

[Total 33 marks]

CONTINUED

5. Answer the following three parts.

   a. Discuss the reasons why the relational data model is currently more successful than the hierarchical data model.

      Do you think that the relational data model will give way in the future to a variation of the hierarchical data model? (Justify your answer.)

      [11 marks]

   b. Suggest how ERDs can help in incorporating foreign key information into database design.

      Discuss the significance of foreign keys in database design using a simple example.

      [11 marks]

   c. The Functional Dependency (FD) inference problem is the data mining problem of discovering a cover of the set of non-trivial FDs that are satisfied in a relation.

      Give a simple algorithm whose input is a relation r over a relation schema R and whose output is a cover of the set of non-trivial FDs that are satisfied in r.

      [11 marks]

[Total 33 marks]

END OF PAPER