Answer any three out of the following six questions

1. Answer the following three parts.

   a. Discuss the unique identification of entities in a database by means of primary key values versus their unique identification by means of object identity.

   [11 marks]

   b. Answer the following two parts.

      i. Recall that the manipulative part for a *nested relational database* includes the two operators $NEST$ and $UNNEST$. The operator $NEST$ transforms a nested relation into a "more deeply" nested relation while the operator $UNNEST$ transform a nested relation into a "flatter" nested relation. Explain with the aid of an example how these operators are used, given that you have available a flat relational database.

      [5 marks]

      ii. Let R be a relation schema R with schema(R) = $\{$A, B$\}$. Give an example of a nested relation r over R such that

      $$NEST_A(UNNEST_{(A)*}(r)) \neq r.$$

      [6 marks]

      [Total 11 marks]

   c. Write two to three sentences on each of the following:

      1. physical level of the database
      2. physical data independence
      3. conceptual level of the database
      4. conceptual data independence
      5. view level of the database

      [11 marks]

      [Total 33 marks]

2. Answer the following three parts.

a. Let r_1, r_2 and r_3 be three relations over relation schemas R_1, R_2 and R_3, respectively, with schema(R_1) = schema(R_2) = $\{$EN, DN$\}$ and schema(R_3) = $\{$DN, PN$\}$. (EN stands for Employee Name, DN stands for Department Name and PN stands for Project Name.)

Express the following queries both in the relational algebra and in SQL:

1. Output the tuples of employees that are either in r_1 or in r_2.
2. Output the tuples of employees that are in r_1 but not in r_2.
3. Output the names of employees in r_1 and projects they work on in r_3.
4. Output the names of employees in r_1 working on the "database" project in r_3.

[11 marks]

b. Give an incremental algorithm for testing whether a relation r over a relation schema R satisfies entity integrity, where X $\subseteq$ schema(R) is the primary key of R. Such an incremental algorithm assumes that prior to the insertion of a new tuple t into r, the relation r satisfies entity integrity.

[11 marks]

c. Give an efficient (i.e. polynomial time) algorithm for finding one key for a relation schema R with respect to a set of functional dependencies F over R.

[11 marks]

[Total 33 marks]

CONTINUED

3. Answer the following three parts.

   a. Suppose that we have a database schema consisting of two relation schemas EMP and DEPT such that schema(EMP) = {EN, DN} and schema(DEPT) = {DN, MN}. (EN stands for Employee Name, DN stands for Department Name and MN stands for Manager Name.) The primary key of EMP is EN and the primary key of DEPT is MN.

   Let r be a relation over EMP and s be a relation over DEPT. Discuss different policies of maintaining referential integrity given that the tuple <Jack, History> is inserted into r and the tuple <Computing ,Jill> is deleted from s.

   [11 marks]

   b. A view is **materialised** if it is physically stored in the database system. The **view maintenance problem** is the problem of appropriately updating a materialised view when an update is performed on the conceptual database.

   A materialised view is *self-maintainable* with respect to an update if it can be maintained without accessing the conceptual database relations.

   Suppose that a materialised view is defined as a projection of a single relation onto a subset of the attributes of its relation schema, say R.

   1. Show that such a views is self-maintainable with respect to insertions.
   2. Show that such a view is self-maintainable with respect to deletions from a relation r over R, if the primary key K for R is included in the view definition,

   [11 marks]

   c. Prove the assertion that a set of attributes X $\subseteq$ schema(R) is a superkey for a relation schema R if and only if for all relations r over R the number of tuples in π_X(r) is the same as the number of tuples in r.

   [11 marks]

[Total 33 marks]

4. Answer the following four parts.

a. Explain what kind of inefficiency in the handling of data the B-tree data structure is intended to reduce. Determine an upper bound on the height of a B-tree of order M containing N records, and use this result in support of your explanation.

[9 marks]

b. Suppose that a node in a B-tree is is denoted by an expression beginning with (and ending with), and that x indicates the position of a pointer field. Draw the B-tree of order 3 that has the nodes (x L x), (x M x N x), (x R x V x), (x S x U x), (x Y x), (x A x B x), (x E x) and (x G x H x), where the ordering on the upper-case keys is alphabetic.

[7 marks]

c. Explain what happens when a record with the key J is inserted into the B-tree in (4.b), and show the resulting B-tree after the insertion.

[9 marks]

d. How does a B^+-tree differ from a B-tree?

Suppose that a collection of records of data in an application that will involve significant amounts of insertion and deletion as well as accesses is so large that the records themselves must be held in disk storage. You are told that you can use B^+-trees, dynamic hashing or any other method of your choice to maintain an index to these records, provided that the overall system will then work efficiently. What method would you choose, and why? (If you need more information about the application in order to make a good choice, say what information you would need, and how it would influence your choice).

[8 marks]

[Total 33 marks]

CONTINUED

5. Answer the following four parts.

a. Certain problems may occur in the use of "unnormalised" data in a database. State what they are, describe how to put data into first normal form (1NF), second normal form (2NF) and third normal form (3NF), and identify the problems that are removed by each normalisation.

[9 marks]

b. The notation x $\rightarrow$y indicates that the value(s) for the item(s) y are determined once one knows the value of x, or of all the items within the brackets of x if x is an expression containing brackets.

Each of the examples of relations r below has a highest normal form: 1NF, 2NF, 3NF or something higher. Identify which is the highest form for each one, and justify your answers.

Relation	Primary Key	Dependencies
r1(A,B,C,D)	A	A $\rightarrow$ B,C,D B $\rightarrow$ C
r2(A,B,C)	A	A $\rightarrow$ B,C
r3(A,B,C,D)	(A,B,C)	B $\rightarrow$ C (A,B,C) $\rightarrow$ D
r4(A,B,C,D)	(A,B)	(A,B) $\rightarrow$ C,D C $\rightarrow$ D
r5(A,B,C,D)	(A,B)	(A,B) $\rightarrow$ C B $\rightarrow$ D

[10 marks]

c. If successively higher normal forms are successively better because they have fewer problems, why should a database designer bother to learn the details of any normal form except the highest one?

[5 marks]

d. "Normalisation and normal forms may be important for databases, but the designer of a knowledge base does not need to pay any particular attention to them". Discuss the validity of this comment.

[9 marks]

[Total 33 marks]

TURN OVER

6. Answer the following four parts.

a. Describe informally the differences between a database and a knowledge base.

[6 marks]

b. Give a Prolog-like or other notation suitable for expressing knowledge in a form that is immediately usable to make deductions from existing data. Explain the notation, and justify the claim that it is suitable for deductions.

[8 marks]

c. Suppose that you are helping to design the plot of a 19th century detective serial for television, in which the hero needs to travel from Calais to Istanbul by train, under the following conditions:

- to reduce the risk of identification by villains, the trip must not pass through any town that is a major junction for railway routes;
- as the serial will be financed by national agencies for tourism if some filming takes place against impressive scenery, at least one station where the hero can stop and/or change trains, in each country on the route, should be in a place that the relevant national agency has declared to be scenic or touristic;
- it is possible that somebody in any place where he may stay overnight is in the pay of the villains and may report his presence by telegraph to the national villain headquarters in the capital city of the relevant country - so that he needs to ensure that the earliest train connection from that capital to any town T where his first train of the next day will stop will not arrive in T until after his train has left it.

Assume that you have a database of European railway timetables available, and any databases maintained by national agencies for tourism. State what database relations you will need in order to derive a route for the hero, write a description of the knowledge that you will use in order to find the route, and then write that knowledge in the notation of (6.b).

[14 marks]

CONTINUED

d. Either explain why your answer in (6.c) specifies the most efficient way of computing the route or indicate what changes you would make in order to ensure an efficient computation.

[5 marks]

[Total 33 marks]

END OF PAPER