This test is **a time-limited exam**. Please be brief, precise and honest. Good luck!

Please be aware that I will be subtracting points for incorrect answers given in Part 1 and 2 (no negative points in Part 3!).

**Note:** Make sure that the **whole statement** is True before you mark it as T. There is no such thing as "partially true statements", they are simply false...

**PART 1. True/False Questions** – +/- 6% of the total test’s score

**Note:** Every symbol, which does not look like T or F to me, is going to be classified as "−1 pt."

---

**F** (1 pt.) The goal of Query Processing and Optimization (QPO) module in DBMS is to always find a query execution plan with the shortest execution time.

**T** (1 pt.) One of the reasons for having a Query Processing and Optimization (QPO) module in a DBMS is the fact that there are often multiple execution strategies for the same node/building block of a query tree.

**T** (1 pt.) Synchronized Tree Traversal is one of the strategies allowing execution of Spatial Join operation. The approach is usually more time-efficient than a "brute-force" comparison.

---

**O** (1 pt.) Rule Based Approach for choosing execution strategy for a building block/node of a query tree is used more often in Spatial DBMS’s than a Cost Based Approach. However, the Cost Based evaluation is more common in regular RDBMSs.

**F** (1 pt.) During the Filter Step, in a Filter-Refine strategy for processing spatial queries, we are filtering out the objects, which are not matching our search criteria, by using exact spatial data representations.

---

**O** (1 pt.) Dijkstra’s and A* algorithms, often used to find the shortest paths in graphs, work well when the entire graph can be loaded into the primary memory. If this is not the case, their performance may degrade substantially.

**PART 2. Single Choice Questions** – +/- 10% of the total test’s score

**Note:** Every symbol, which I have hard time to identify as an a-e letter, is going to be classified as "−1 pt.". You need to **choose one answer only!!!**

1. (2 pts.) The highest level of data summarization in data cube is frequently referred to as:

   a) **✓** 0-D cuboid
   b) ___ 1-D cuboid
   c) ___ 2-D cuboid
   d) ___ 3-D cuboid
   e) ___ base-cuboid
2. (2 pts.) Filter & Refine is a common strategy to answer spatial queries. In the Filter step, SDBMS:

a) ___ may use approximate representations of data objects’ shapes, but must always stick to the actually requested spatial operators.

b) ___ may use approximate representations of data objects’ shapes and approximations of actually requested spatial operators.

c) ___ may use approximations of actually requested spatial operators, but must always operate on actual/exact data objects’ shapes.

d) ___ must always take advantage of the Synchronized Tree Traversal algorithm.

e) ___ can never take advantage of the Synchronized Tree Traversal algorithm.

3. (2 pts.) An Abstract Directed Graph:

a) ___ has all nodes representing actual spatial points.

b) ___ can never be acyclic.

c) ___ must be always acyclic.

d) ___ can be used to generate the graph’s transitive closure.

e) ___ is presented in the figure below:

![Directed Graph Image]

4. (2 pts.) Galaxy schema is a description of data warehouse model, which:

a) ___ has exactly the same meaning (and usage!) as a star schema.

b) ___ can never be modeled as a collection of snowflakes.

c) ___ has one fact table in the middle, connected to a set of normalized dimension tables.

d) ___ has one fact table in the middle, connected to a set of dimension tables.

e) ___ has multiple fact tables, connected to a set of dimension tables.
5. (2 pts.) From the list below, choose only one combination of operations (in proper order!), which characterizes in the most accurate way processing of a SQL query, before the optimization:

a) ___ parsing, identification of language tokens, validation of semantics, translation to relational algebra/building blocks, generation of query tree.

b) ___ identification of language tokens, validation of semantics, translation to relational algebra/building blocks, generation of query tree.

c) ___ identification of language tokens, parsing, validation of semantics, generation of query tree, translation of building blocks to relational algebra.

d) ___ identification of language tokens, parsing, validation of semantics, translation to relational algebra/building blocks, generation of query tree.

e) ___ identification of language tokens, parsing, validation of semantics, generation of query tree, translation to relational algebra/building blocks.

PART 3. Practical Problems – 84% of the total test’s score

1. (28%) Present a spatial query given below in the form of Optimized Query Tree.

```sql
SELECT City.Name
FROM City, Facility
WHERE City.Population > 50000 AND Facility.Type = 'NPP'
AND DISTANCE(City.Shape, Facility.Shape) < 30.
```

![Optimized Query Tree Diagram]

CS 535, Fall 2009 – 2nd Test
2. (28%) Use Snowflake Schema to design a mini-warehouse with a single fact table, which contains the following columns:

<table>
<thead>
<tr>
<th>Column</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>studentID</td>
<td>Foreign key to dimension table student (e.g. 8 digits on your ID number).</td>
</tr>
<tr>
<td>courseID</td>
<td>Foreign key to dimension table course (i.e. course code in the registrar).</td>
</tr>
<tr>
<td>instructorID</td>
<td>Foreign key to dimension table instructor (i.e. 8 digits on faculty ID number).</td>
</tr>
<tr>
<td>num_of_absences</td>
<td>Numeric column containing the number of absences.</td>
</tr>
</tbody>
</table>

Include in your design the following requirements:

a) When dealing with dimension student, we would like to be able to retrieve: student_name, department_of_major, department_of_minor, student_date_of_birth, student_address (hint: decompose the last attribute as expected in the Snowflake Schema).

b) When dealing with dimension course, we would like to be able to retrieve: course_name, number_of_credits, department_where_offered.

c) When dealing with dimension instructor, we would like to be able to retrieve: instructor_name, department_where_employed, position, faculty_date_of_birth.

d) In addition - make sure to let the future data-warehouse users retrieve the following information from your data-warehouse schema (make sure to link it to the rest of the schema properly): department_name, department_address, number_of_employees_within_department.

Note: Make sure to properly mark primary keys in all relations!

* ANSWER IN EXTRA SHEET *
1. Country
   - Zip
   - State
   - City
   - Address Line 1
   - Address Line 2
   - Address Line 3

2. Zip
   - Country

3. Student ID
   - Address
   - Student Name

4. Department
   - ID
   - Name
   - Address

5. Course
   - ID
   - Name
   - Instructor
   - Department
   - Date
   - Time
   - Year
   - Month
   - Day
   - Grade
   - Absence
   - Grade
   - Instructor
   - Course

6. Fact
   - Course ID
   - Student ID

7. Time
   - Date
   - Position
   - Dept Id
   - Instructor Name

8. Instructor
   - Course Id
   - Dept Id
   - Instructor Name
   - Course

9. Grade
   - Course Id
   - Name
   - Course

10. Absence
    - Course Id
    - Instructor
    - Course

Assuming address is only determined independently.

Assuming we are recording 1 address per student.

Note: Zipped County

Address Line 1, Address Line 2, City, State.

Assuming address is only determined independently.

Note: Assuming zip code = to city.
3. (28%) Concept of Transitive Closure of graph is crucial in many applications of spatial networks. There are different ways in which graphs may be stored in computer’s memory (e.g. Adjacency-matrix, Adjacency-List, Node and Edge Relations, Denormalized Node Relation). Which of these ways is the most convenient to represent a huge American Public Roads’ Network in spatial databases (taking under consideration the fact that a transitive closure is frequently used, when dealing with this type of network)? Present main reasons standing behind your choice.

A sample road network.

The transitive closure would be

According to me, Node and Edge relationship should be used. The reason behind my choice is given below. We should not be using Adjacency matrix, since the no. Of roads will be huge, thus maintaining an the matrix will be a problem. If there are m roads then the matrix will be an \( mxm \) matrix. Moreover it is highly difficult to calculate.
transitive closure from the adjacency matrix.

In case of adjacency list we have to perform a DFS/BFS to find out the transitive closure, which may be a huge overhead since we will be using the transitive closure frequently & since the no. of roads is huge. Moreover, the denormalized node relation will be having redundancy, which may turn out to be a huge overhead.

Now, if we look at Node and Edge relations, we can see it is important for a road network to find which roads are connected to each other, which can be easily found by Node & Edge relations. Moreover, to find the transitive closure, we need to take into account the edge relationships too.