Exam Advanced Database Systems (211090)

Wednesday 7 November 2007, 9:00 – 12.30 hour (WA-4)

The exam consists of 5 questions divided into subquestions
It is NOT permitted to use book or notes

1 XML (7 points)

Below you find an XML document taken from `http://www.atak.nl/pub/atak.xml`
Answer the following questions.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<rss version="2.0">
  <channel>
    <title>Atak Poppodium</title>
    <link>http://www.atak.nl/</link>
    <description>Informatie van de aktiviteiten in Poppodium Atak, Enschede
    <!-- Atak is the pop temple of Enschede with lot's of life music! -->
    </description>
    <item>
      <title>GOG-BOT Media Art Fest</title>
      <link>http://www.atak.nl/pub/link.php?id=184</link>
      <description>GOG-BOT Media Art Fest Enschede 2007 is een jaarlijks
        terugkerend platform voor elektronische geluidskunst en muziek,
        beeldende kunst, film-, en computerkunst.</description>
    </item>
    <item>
      <title>Fiddlehead en Trinity Knot</title>
      <link>http://www.atak.nl/pub/link.php?id=185</link>
      <description>Het is weer tijd voor een Keltische special. Een X-It
        Phoenix die geheel in het teken zal staan van de Keltische voorloper
        van Halloween, te weten Samhain.</description>
    </item>
  </channel>
</rss>
```

(a) Give the/a DTD for the document above. Is there more than one DTD possible? Explain your answer.

(b) Give at least three differences between HTML documents and XML documents.

(c) RSS is a standard XML format for news and the content of news-like sites like Wired, news-oriented community sites like Slashdot, and personal weblogs. Why would Atak publish its news in this standard XML format?

2 Using SQL in applications (8 points)

Below you find a fragment of example code that produces the RSS feed of part 1.

```
(...)
10   Connection db = DriverManager.getConnection(url, login, passw);
11   Query = "SELECT * FROM news WHERE source = ?";
12   PreparedStatement ps = db.prepareStatement(Query);
13   ps.setString(1, source);
14   ResultSet rs = ps.executeQuery();
15   System.out.println("<rss version=\"2.0\"><channel>");
16   System.out.println("<title>Atak Poppodium</title>");
17   System.out.println("<link>http://www.atak.nl/</link>");
18   System.out.println("<description>Informatie van... </description>");
19   while(rs.next()) {
20     System.out.println("<item>");
21     System.out.println("  <title>" + rs.getString("title") + "</title>");
22     System.out.println("  <link>http://www.atak.nl/pub/link.php?id=" +
23                        rs.getString("id") + "</link>");
24     System.out.println("  <description>" + rs.getString("description") +
25                        "</description>");
26     System.out.println("</item>");
27   }
28   System.out.println("</channel></rss>");
(...)
```

Answer the following questions:

(a) Is the programme code using a so-called "Statement-Level Interface" or a "Call-Level Interface" to include the SQL constructs? Explain your answer.

(b) Is this form of embedding SQL statements referred to as (answer with "yes" or "no"):

i) dynamic SQL
ii) static SQL
iii) embedded SQL
iv) SQL/PSM
v) a stored procedure

(c) Which line numbers contain the so-called "cursor"? In this case, the cursor is insensitive. What is an "insensitive cursor"?

(d) Suppose we remove line 13 and replace line 11 with

```
Query = "SELECT * FROM news WHERE source = '" + source + "'";
```

i) Would this work as well? Explain why/why not.
ii) Would the line introduce a security risk? If so, explain why.

3 Query Optimization (12 points)

Consider the following relational schema:

```
Employee(Id, Name)
WorksIn(EmpId, DeptName)
Department(DeptName, Building)
```

Assume that keys are underlined, and:

- The relation Employee is sorted on the Id attribute.
- In Department: Unclustered index on DeptName; clustered on Building. No other indices.
- Dozens of departments can reside in the same building.

Answer the following questions:

(a) Write an SQL query that generates a table of employee names who work for department "Sales" in building E213.

(b) Give the "naive" (unoptimized) translation of your SQL query into the relational algebra (as given by the general translation of SQL to relational algebra.)

(c) Describe carefully and completely how you would actually most efficiently evaluate the above query. That is, list the precise sequence of relational operations to be performed, the relations on which these operations are to be performed, and the methods to be used to compute the result of each operation. (As part of the solution you will need to decide which index to use and why.)

(d) Give the relational algebra expression for this query that most closely corresponds to the way you chose (in the previous subproblem) to efficiently evaluate this query.

4 Distributed Databases (9 points)

Suppose that we have two relations:

```
Order(CustomerId, ItemId, Quantity)
Inventory(ItemId, ItemName, Price)
```

The `Order` relation does not have non-trivial keys. In `Inventory`, `ItemId` is a key, but `ItemName` is not. The database is distributed such that `Order` is at site *A*, `Inventory` is at site *B*, and some queries are issued at site *C*. Assume the following statistics:

- Every attribute is 20 bytes long.
- `Order` has 10,000 tuples; 500 different item numbers `ItemID`.
- `Inventory` has 1,000 tuples; 200 different item names.

The answer to the following query is issued at site *C*.

```
SELECT CustomerId, Quantity
FROM Order O, Inventory I
WHERE O.ItemId = I.ItemID
AND I.ItemName = 'screwdriver'
```

Answer the following questions:

(a) Following a *multidatabase solution with local schemas*, find the best evaluation plan for this distributed query and estimate its cost in terms of bytes that need to be sent around.

(b) Write the relational algebra expression that most closely corresponds to your answer to the previous question.

(c) Now, following a *integrated distributed database solution supporting a global schema*, find again the best evaluation plan for this distributed query and estimate its cost in terms of bytes that need to be sent around.

5 On-Line Analytical Processing (9 points)

Consider the following schema for an on-line analytical processing (OLAP) application. The database contains for each timestamp the number of sold products of a super market.

```
Sales(MarketId, ProductId, TimeId, SalesAmount)
Market(MarketId, City, Province, Country)
Product(ProductId, Name, Brand, Manufacturer)
Time(TimeId, Day, Month, Year)
```

Suppose the dimension tables each have 100 tuples. Answer the following questions.

(a) What is the maximum number of rows in the fact table of this application?

(b) Suppose that we use the CUBE operator on this fact table to perform aggregations on all dimensions as follows.

```
SELECT MarketId, ProductId, TimeId, SUM(SalesAmount)
FROM Sales
GROUP BY CUBE(MarketId, ProductId, TimeId)
```

What is the maximum number of rows that result from this query?

(c) What is the maximum number of rows that result from the query if we use the ROLLUP operator instead of the CUBE operator?

(d) Give the SQL query that present per product, per brand, and per manufacturer the total number of sold items.