Uppsala universitet
Institutionen för informationsteknologi
Kjell Orsborn, Tore Risch

Tentamen 2003-12-18
DATABASTEKNIK - 1DL116, 1MB025

```
Datum ..............................Torsdagen den 18 December, 2003
Tid ..................................................8:00-13:00
Jourhavande lärare ...Kjell Orsborn, tel. 471 11 54 eller 070 425 06 91

Hjälpmedel ..............................................miniräknare
```

Anvisningar:

- Läs igenom hela skrivningen och notera eventuella oklarheter innan du börjar lösa uppgifterna. Förutom anvisningarna på skrivningsomslaget så gäller följande:
  - Skriv tydligt och klart. Lösningar som inte går att läsa kan naturligtvis inte ge några poäng och oklara formuleringar kan dessutom misstolkas.
  - Antaganden utöver de som står i uppgiften måste anges. Gjorda antaganden får förstås inte förändra den givna uppgiften.
  - Skriv endast på en sida av papperet och använd ett nytt papper för varje uppgift för att underlätta rättning och minska risken för missförstånd.
- För godkänt krävs det cirka 50% av maxpoäng.

1. **Database terminology:** 4pts

Concisely explain the following concepts (in a database context):

(a) meta data

Answer: Meta data, or the database schema, include data about data, i.e. a description of the database stored in the system catalog. Meta-data consist of information about structure of files, type and storage format of each data item, various constraints on the data and other types of information about data such as authorization privileges and access statistics. For the relational model this include descriptions of the relation names, attribute names, data types, primary keys, secondary keys, foreign keys, other constraints, views, storage structures and indexes, and security and authorization information.

(b) full functional dependency

Answer: A functional dependency between two sets of attributes X and Y that are subsets of the relation schema R, denoted $X \rightarrow Y$, is said to be a full functional dependency if the determinant X is minimal, that is no attribute can be removed from X without losing the dependency.

(c) transaction

Answer: A transaction is a logical unit of database processing that is performed in its entirety or not at all.

(d) candidate key

Answer: A candidate key is a minimal set of attributes in a relational schema that is a potential primary key, i.e. uniquely can identify all non-key attributes as well as candidate keys.

2. **Conceptual data modeling:** 4pts

Enhanced Entitity-Relationship modeling, to various degree, supports features to group entities. Explain what information the following two features are representing and how they can be represented in EER:

(a) specialization

(b) aggregation

Answer:

(a) Specialization is a process to conceptually refine a general entity type called a superclass by specifying a set of subclasses. The subclasses are created by identifying some distinguishing characteristic among subsets of entities of the superclass that is the basis to form the subclasses.

(b) Aggregation is an abstraction concept to group entities into composite objects from their components. In three cases can aggregation be related to the EER model. The 1st case is an aggregation of attribute values of an object to form the whole object. The 2nd case is the representation of an aggregation relationship using an

ordinary relationship. The 3rd case is not explicitly supported in EER but involve the possibility to combine related objects using a particular relationship instance into a higher-level aggregate object.

3. **Relational algebra and SQL:** 4pts

Assume that we have two relations (tables) with the following relational schemas, where *ID determines keys:

```
CIRCLE(CID,CNAME,RADIUS,POINTID)
POINT(PID,PNAME,X-COOR,Y-COOR)
```

(a) Express in relational algebra the following query: Which keys, names, radii and x-coordinates of their centre do those circles have that have a radius under 50.0 (cm), and that have its centre point in the positive half plane $x > 0$. (1pt)

Assume that we in a database have one relation (table) with the following relation schema:

```
RECTANGLE(RID,RNAME,X1COOR,Y1COOR,X2COOR,Y2COOR)
```

where RID determine the key, RNAME is a name, and **COOR:s are x- and y-coordinates for the rectangle in the xy-plane. All edges to single rectangles are parallel to an x- or y-axis, i.e. no rectangle is rotated with respect to an xy-coordinate system. Formulate the following queries in SQL:

(b) How big is the area of each rectangle? (1pt)

(c) Which pair of rectangles (e.g. keys and names) overlap each other (duplicates are allowed in the result)? (2pts)

Answer:

(a) $\Pi_{<cid,cname,radius,x-coor>}(\sigma_{radius<50.0 AND x-coor>0.0}(CIRCLE *_{pointid=pid} POINT))$
where * represents the (natural) join operator.

(b)
```
SELECT RID, RNAME, (X2COOR-X1COOR)*(Y2COOR-Y1COOR) AS AREA
FROM RECTANGLE;
```

(c)
```
SELECT DISTINCT r1.RID,  r1.RNAME, r2.RID,  r2.RNAME
FROM
RECTANGLE r1, RECTANGLE r2
WHERE
(r2.X1COOR < r1.X2COOR) AND
(r2.X2COOR > r1.X1COOR) AND
(r2.Y1COOR < r1.Y2COOR) AND
(r2.Y2COOR > r1.Y1COOR) AND
r1 < r2;
```

, where `DISTINCT` and the $r1 < r2$ condition removes duplicates. Using $\leqslant$ and $\geqslant$ would add "overlapping oneself" which is correct as well.

4. **Recovery:** 4pts

Describe the basic steps in the recovery procedure according to the deferred update model in a multi-user version.

Answer:

Recovery according to the deferred update model (no-undo/redo):

1. Start from the last record in the log file and traverse backwards.

Create two lists:

C transactions that have reached their commit points

NC transactions that have not reached their commit points.

2. Start from the beginning of the log file and redo all (Write,T,...) for all transactions T in the list C.

3. Restart all transactions in the list NC.

If the log file is long, step 2 will take long time. An improvement of this method is accomplished by introducing what's called check points.

5. **Database application interfaces:** 4pts

(a) What is ODBC? (1 pt)

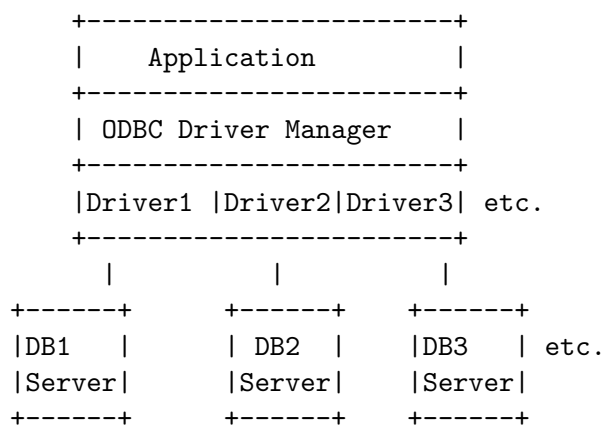
(b) Describe the architecture of ODBC. Draw a picture. (2 pts)

(c) What is the difference between JDBC and ODBC? (1 pt)

Answer:

(a) ODBC = Open DataBase Connectivity is a DBMS independent client server API for communication with a relational DBMS by passing SQL strings between application and DBMS server.

(b)



DB1, DB2, DB3 different kinds of DBMSs, e.g. DB2, Oracle, MySQL, Mimer

(c) JDBC is ODBC for the programming language Java, i.e. JDBC provides DBMS independent relational database interface from Java.

6. **Query optimization:** 4pts

A large company maintains a table of the effectiveness of their sales force:

```
SALES(SSN, SALES, ...)
```

containing the SALES (in $) of each sales person identified with SSN (social security or 'person' number), primary clustered B-tree index on SSN and secondary unclustered B-tree index on SALES. There are 10000 sales persons and 10 rows fit in a disk block while 100 keys fit in each index node block. The management needs to regularly know the top ten sales performers and you are asked to design an application program to quickly get those stars. SQL nowadays provides an extension to express such top-10 queries and with it the query in this case would look like:

```
select ssn, sales
stop after 10 rows
from sales
order by sales descending
```

This means that the application will get at most the 10 first rows back from the query.

(a) What is the advantage from a performance point of view to have such a 'stop after n rows' clause in select? (1 pts)

(b) What execution plan is optimal? Show why it is optimal. (3 pts)

Answer:

(a) If the query optimizer knows that the user needs only a few first rows from the sorted list other strategies can be chosen than if all sorted rows are needed. E.g., the optimizer can choose between traversing the sorted but unclustered index on sales or scanning the entire table in cluster order and then sort the result.

(b) The optimal plan would traverse the index on SALES in reverse order (assuming it is sorted in ascending order) and read the table rows to which the index entries point. When 10 rows have been read the traversal is stopped.

NOTICE that the stream oriented communication between applications and databases solves the problem of avoiding to compute all result tuples when only the first 10 are required by the application!

An alternative plan would be to scan the table and sort the result.

The cost to execute the optimal can be estimated as:

Access 2 index blocks + 10 table blocks. Only 2 index blocks need to be accessed since 100 keys fit in each index block, and since the

index will have a depth of 2 because there are 10000 sales persons. One can assume that 10 table rows need to be accessed since the index is unclustered.

The alternative to scan the table would cost cost 1000 blocks just to scna the table + the time to sort the result.

7. **Database integrity:** 4pts

   (a) Give examples of 3 kinds of actions that can be taken when referential integrity constraints are violated in SQL. (3 pts)

   (b) What are 'domain constraints' in SQL? (1 pt)

   Answer:

   (a) When integrity constraints are violated SQL allows the specification of the following actions:

   ```
   SET NULL value in other relation
   DELETE row in other relation
   SET DEFAULT value in other relation
   CASCADE referential integgrity check to other relation.
   RESTRICT update so that updating transaction will fail
   NO ACTION
   ```

   (b) Domain constraints are restrictions on simple values of user defined types (domains).

8. **Multi-media Databases:** 4pts

   (a) Why is object-relational technology good for storing multi-media objects in databases? (2 pts)

   (b) What are BLOBs and what are they used for? (1 pt)

   (c) Why is RAID good for storing databases? (1 pt)

   Answer:

   (a) Object-relational databases allow extensions (plug ins) of the DBMS engine with user defined query functions, data representations, indexing methods, and query optimization. Multi-media databases require representation, indexing, and querying over new kinds of very voluminous data. Object-relational technology provides extensibility of the DBMS required for this.

   (b) A BLOB means Binary Large OBject and is a datatype in modern relational databases holding very long uninterpreted bit strings. Uninterpreted means that it is up to the application to fully manage the contents of the BLOB and the DBMS does not allow queries over the contents of the BLOBs (unless the DBMS is extended with new query functions using object-relational technology).

(c) RAID (Redundant Array of Inexpensive Disks) or 'disk arrays' constist of a set of disks connected to a computer. 'Striped' files are files that are split into pieces that are stored on different disks in a disk array. This allows to speed up parallel reading from a striped file proportional to the number of stripes, in order to beat the slow access by mechanic disk arms. 'Mirroring' means that the same file is stored in more than one place of the disk array, which increases availability and access speed. Often files in RAID systems are both striped and mirrored.

Good luck and Merry Christmas!

/ Kjell och Tore