CS 441: Midsemester Exam, Fall 1999

Write your Roll No. on the top left of every additional sheet, including this one.

I. Complete the following sentences in the space below. *Return this page with the rest of your answer script.* ...18

1. When using MRU, a block is pinned so that

2. The disk performance measure that is affected when related data is clustered on a disk is

3. When hardware swizzling is done for page P, swizzling overheads can be reduced if for every page P' referred to by a persistent pointer in P,

4. Secondary indexes must be dense because

5. Closed hashing (with oveerflow chaining) is preferable to open hashing when

6. An example of a situation where it is not good to adopt the "perform projections early" heuristic is

II. In data warehousing environments, data is stored in a relational database using a *star schema*. A star schema usually consists of a single *fact* table and set of *dimension* tables. Consider the star schema presented in Figure 1. The schema models the activities of a world-wide wholesale supplier over a period of 7 years. The fact table is the `SALES` table, and the dimension tables are the `PART`, `SUPPLIER`, `CUSTOMER`, and `TIME` tables. The fact table contains foreign keys to each of the dimension tables.

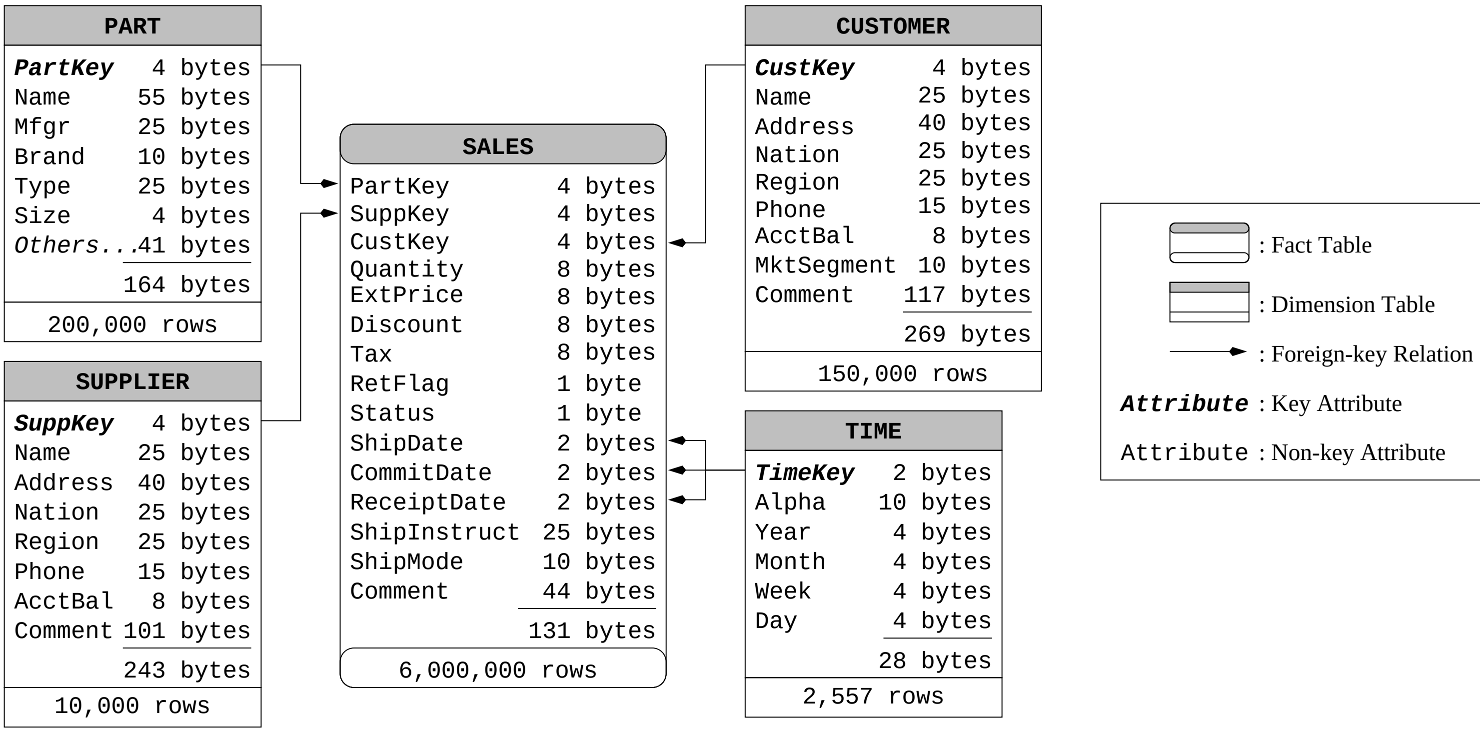

Figure 1: A Sample Warehouse Star Schema

A common type of query in data warehousing systems is the *star-join* query. In a star-join, one or more dimension tables are joined with the fact table. Consider the following query:

Query 1

```
SELECT      U.Name, SUM(S.ExtPrice)
FROM        SALES S, TIME T, CUSTOMER C, SUPPLIER U
WHERE       T.Year BETWEEN 1996 AND 1998
AND         U.Nation='India' AND C.Nation='India'
AND         S.ShipDate = T.TimeKey AND S.CustKey = C.CustKey
AND         S.SuppKey = U.SuppKey
GROUP BY    U.Name
```

1. In unambiguous English (!) state what the above query does. ...6

2. Give the (extended) relational Algebra expression for Query 1. ...6

3. Give the query DAG corresponding to this expression. ...6

4. Suppose primary indexes are maintained for each of the dimension tables involved in the query and secondary indexes are available for the foreign key columns of the SALES relation. ...25

   (a) Give an evaluation plan for the query.

   (b) In the plan graph, annotate each JOIN operation node with the specific algorithm you recommend for that operation.

   (c) Mark the pipelinable edges, if any.

   (d) Explain the reason behind your choice *for the overall plan* and *for each JOIN algorithm.* Clearly state any assumptions made.

   (e) Choose one of the Join operation nodes and state which buffer management strategy you would recommend while performing that operation.

5. What additional information – beyond what is given in Figure 1 – would you need to estimate the size of the result for Query 1? ...6

6. Suppose the above query is posed repeatedly, with different parameters. Would you recommend an ordered index or a hash structure for (a) the ReceiptDate attribute and (b) the CustKey attribute of the SALES relation? Why? ...6

7. Several data warehouse products use an index structure called *bitmapped index.* Recall that in a B^+-tree, given a search key value, a successful search leads you to the set of matching records (rows) from the indexed relation. This set is usually called the *rowset.* A bitmapped index structure is identical to a conventional B^+-tree except that each rowset may be represented as a bit vector instead of a set of pointers to the matching records *when doing so leads to reduced storage requirements.* The bit vector size is the same as the cardinality of the relation; a particular bit is *set* if the corresponding record matches the search key.

   - Under what conditions will using bitmaps for a particular key value lead to reduced storage requirements (compared to the usual approach of storing pointers to matching records)? ...7
   - How would one process range queries with respect to an attribute for which all leaves store bitmaps? ...7

8. The data warehouse should be continuously available, that is, without any downtime. You are asked to choose between (a) Mirrored disks and (b) Block interleaved Distributed Parity RAID. ...8

   (a) Which would you choose and why?

   (b) Would your answer change if you are told, in addition, that the warehouse is updated very rarely.

9. Suppose a new product is offered and a sale of that product is transacted. How should the warehouse be updated so that no integrity violation occurs? ...5

Total Marks = 100