CIS 671, Winter 2002 **Thanks to Jeff Blackburn**

Homework 7: OLAP[*] Solutions

(*) Adapted from Garcia-Molina, Ullman & Widom, *Database System Implementation*, Prentice-Hall, 2000, pp. 612-632.

1. An on-line seller of computers wishes to maintain data about orders. Customers can order their PC with any of several processors, a selected amount of main memory, any of several disk units, and several CD or DVD readers. The fact table for such a database might be:

```
        Orders(cust, date, proc, memory, disk, cd, quant, price)
```

where cust is the customer ID and a foreign key for a dimension table about customers, and proc (processor), disk and cd are similar. For example, a disk ID might be elaborated in a dimension table giving the manufacturer and several disk characteristics. The memory attribute is simply an integer: the number of megabytes of memory ordered. The quant attribute is the number of machines of this type ordered by this customer, and price is the total cost of each machine ordered.

a. Which are the dimension attributes, and which are the dependent attributes?

```
dimension: cust, date, proc, mem, disk, cd
dependent: quant, price
```

b. For some of the dimensions attributes, a dimension table is likely to be needed. Suggest appropriate schemas for these dimension tables.

```
Custs (custid, name, address1, city, state, zip, phone, company)
Dates (day, week, month, year)
Procs (procid, pman, pmodno, speed, l2sz, sckttyp, pprice)
Mems (memid, mman, size, spd, mtype, pin, mprice)
Disks (diskid, dman, dmodno, size, seek, lat, interface, dprice)
```

2. Suppose that we want to examine the data above to find trends and thus to predict which components the company should order more of. Describe a series of drill-down and roll-up queries that would lead to the conclusion that customers are beginning to prefer a DVD drive to a CD drive. *Do NOT assume that the user starts with the question of comparing DVD vs. CD drives. Instead assume the user is just looking for some trends, eventually discovering the change from CD to DVD drives.*

```
select    month, sum(price)
from      Orders natural join Dates
where     year = ’2001’
group by  month

select    disk, week, sum(price)
from      Orders natural join Disks join Dates on date = day
where     year = ’2001’
group by  disk, week

select    cd, week, sum(price)
from      Orders natural join Cds join Dates on date = day
where     year = ’2001’
group by  cd, week

select    cd, month, sum(price)
from      Orders natural join Cds join Dates on date = day
where     ((year = ’2001’ or year = ’2000’)
          and cd = ’DVD’ )
group by  cd, month

select    cd, month, sum(price)
from      Orders natural join Cds join Dates on date = day
where     ((year = ’2001’ or year = ’2000’)
            and cd = ’CD’)
group by  cd, month
```

3. To apply the CUBE operator to the example above, we might find it convenient to break several dimensions more finely. For example, instead of one processor dimension, we might have one dimension for the type (e.g., AMD K-6 or Pentium-III), and another dimension for the speed. Suggest a set of dimensions and dependent attributes that will allow us to obtain answers to a variety of useful aggregation queries. In particular, what role does the customer play? Also the price above referred to the price of one machine, while several identical machines could be ordered in a single tuple. What should the dependent attribute(s) be?

```
a) From page 1:
       Procs (procid, pman, modno, speed, l2sz, sckttyp, pprice)
   New:
       Procs (procid, type, speed, pprice)
       Types (typeid, pman, modno, skttyp, isa, #bits)
       Speeds (speedid, mhz, frntbs, L1sz, L2sz)
```

b) Grouping. Can create target market segments based on processor purchasing trends of customers found by viewing aggregations.

c) price is generated for each configuration and needs to be multiplied by quant to get the total price (totprice) for the order. totprice would be dependent on quant and price in addition to cust (could have special discount) and date (could be having a sale).

We could partition customers geographically or by industry type.

4. What tuples of the cube above would you use to answer the following queries?

a. Find, for each processor speed, the total number of computers ordered in each month of the year 2000.

```
(proc, speedid, date,       mhz, month, quant totnum)(group by month)
  *       *    >’12/31/1999  m     *      q    m*q
               <’1/1/2001’
```

b. List for each type of hard disk (e.g., SCSI or IDE) and each processor type the number of computers ordered.

```
(disk, proc, typeid, interface, quant, totnum) (group by interface, typeid)
  *     *      *         *        q    sum(q)
```

c. Find the average price of computers with 400 megahertz processors for each month from January 1999.

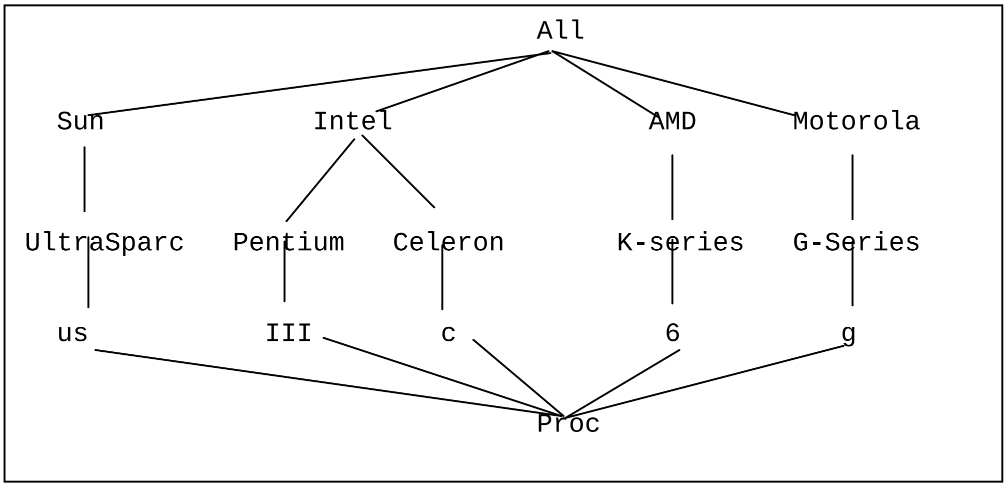
```
(proc, speedid,  date,     mhz, month, quant, price, avgpr)
                                                  (group by month)
  *       *   >’1/31/1999’ 400    *      q     p    (sum(p)/q)
```

5. The computers described above do not include monitors. What dimensions would you suggest to represent monitors? You may assume the price of the monitor is included in the price of the computer.

```
Orders(cust, date, proc, mem, disk, cd, mon, quant, price)
Mons(monid, montype, sz)
Montypes(montypeid, monman, modno, refresh)
```

6. The design above is suitable for the CUBE operator. However, some of the dimensions could also be given a nontrivial lattice structure. In particular, the processor type could also be organized by manufacturer (e.g., Sun Intel, AMD, Motorola), series (e.g., Sun UltraSparc, Intel Pentium or Celeron, AMD K-series, or Motorola G-series), and model (e.g., Pentium-III or AMD K-6).

a. Design the lattice of processor types following the examples described above.



b. Define a view that groups processors by series, hard disks by type, and CD's by speed, aggregating everything else.

```
insert    into V1
select    pseries, dtype, cdspeed
from      (((Orders join Procs on proc = procid) join Disks on
            disk = diskid) join Cds on cd = cdid)
group by  pseries, dtype, cdspeed
```

c. Define a view that groups processors by manufacturer, hard disks by speed, and aggregates everything else except memory size.

```
insert     into V2
select     pman, dspeed, mem
from       ((Orders join Procs on proc = procid)
           join Disks on disk = diskid)
group by   pman, dspeed, mem
```

d. Give examples of queries that can be answered from the view of (b) only, the view of (c) only, both, and neither.

(b) only. Average cd speed of computers with Pentium processors and SCSI disk drives.

(c) only. Average memory size of computers with Intel processors and 8ms drives.

Both. Total number of computers ordered.

Neither. Most popular disk manufacturer.