Politecnico di Torino
Database Management Systems

September 21^{st} 2011

1. (6 Points) The following relations are given (primary keys are underlined):

```
PLAY-ACTOR(AId, Name, Nationality, BirthDate)
COMEDY(ComId, Title, Director, SceneNumber, Year)
PLAY-ACTOR-IN-COMEDY(ComId, AId, Role)
PLANNING(ComId, Theater, Date, StartTime, LengthOfTime)
```

Assume the following cardinalities:

- card(PLAY-ACTOR)= 10^4 tuples,
  MIN(BirthDate) = 1-1-1960, MAX(BirthDate) = 31-12-1999,
- card(COMEDY)= 10^3 tuples,
  distinct values of SceneNumber $\simeq 15$,
- card(PLAY-ACTOR-IN-COMEDY)= 10^6 tuples,
  distinct values of Role $\simeq 30$,
- card(PLANNING)= 10^8 tuples,
  MIN(Date) = 1-1-2010, MAX(Date) = 31-12-2010,
  MIN(LengthOfTime) = 81, MAX(LengthOfTime) = 180,

Furthermore, assume the following reduction factor for the group by condition:

- having count(DISTINCT Theater)≥50 $\simeq \frac{1}{10}$.

Consider the following SQL query:

```
select Title, Director
from COMEDY C, PLANNING P, PLAY-ACTOR-IN-COMEDY AC
where P.ComId=C.ComId and AC.ComId=C.ComId
      and LengthOfTime=180 and SceneNumber>12
      and AC.Aid in (select Aid from PLAY-ACTOR
                     where BirthDate ≥ 1996)
group by ComId, Title, Director
having count(DISTINCT Theater) ≥ 50
```

For the SQL query:

(a) Report the corresponding algebraic expression and specify the cardinality of each node (representing an intermediate result or a leaf). If necessary, assume a data distribution. Also analyze the group by anticipation.

(b) Select one or more secondary physical structures to increase query performance. Justify your choice and report the corresponding execution plan (join orders, access methods, etc.).

2. (7 Points) The following relations are given (primary keys are underlined, optional attributes are denoted with *):

```
EMPLOYEE(ECode, EName, Qualification)
HOURLY_PAY_EMPLOYEE(TypeOfActivity, Qualification, HourlyPay)
TypeOfActivity={Ordinary, Overtime}
DAILY_SUMMARY(ECode, Date, HoursOfWork)
PAY_PACKET(ECode, Month, TotalAmount)
PAY_PACKET_REQUEST(RCode, ECode, Month)
```

A company wants to manage the calculation of the monthly pay-packets for the employees.

For each employee, the monthly pay-packet is computed based on her hourly pay and the number of hours she has worked in the month. The hourly pay depends on the qualification of the employee and the type of activity (ordinary or overtime) she carried out. Each month, the hours of work of the employee are paid as 'ordinary' activity up to 160 hours. The hours of work of the employee exceeding 160 hours are paid as 'overtime' activity.

The `EMPLOYEE` table contains the qualification for each employee. The `HOURLY_PAY_EMPLOYEE` table contains the hourly pay based on the type of activity (ordinary or overtime, attribute `TypeOfActivity)` and the qualification of the employees. The `DAILY_SUMMARY` table contains, for each employee, the hours of work in each date.

Write the triggers managing the following activities .

(1) *Calculation of a new pay-packet.* The calculation of a new pay-packet is requested (a new record is inserted in the `PAY_PACKET_REQUEST` table), for the hours of work of a given employee in a specific month. To calculate the pay-packet, all the hours worked by the employee in the month have to be considered, including both ordinary and overtime activity. A new record including the information on the new pay-packet must be inserted in the `PAY_PACKET` table. The month can be extracted from the `Date` attribute in the `DAILY_SUMMARY` table by using the `TOMONTH` function.

(2) *Integrity constraint on the working day.* The integrity constraint requires that any new record inserted in the `DAILY_SUMMARY` table contains the number of hours worked by the employee in the *current date*. Insertion of a new record with the hours of work related to dates preceding or following the current date is not possible. If a new record concerning different dates than the current date is inserted, the insert operation must not be executed. The current date can be extracted from the system date (function `sysdate`) by using the `TODAY` function. Write the trigger enforcing this integrity constraint.

3. Data Warehouse design

Problem specifications

A video surveillance company manages telecontrol systems installed at several customer's sites all over the world. The company wants to design a data warehouse to efficiently analyze data collected from different sources. In every monitored site some surveillance cameras are installed. The monitored site, owned by a customer, can be public (e.g., parks) or private (e.g., offices).

Surveillance cameras record a video and process it to recognize human silhouettes. They can associate one of the following categories with each silhouette: adult males, adult females, children. Every association is characterized by a confidence level (low, medium, high) measuring the probability of a correct identification for the silhouette by the classifier. In the current database of the company, this information is stored to count the transit of people (males, females, children) in the different locations where the surveillance camera are installed. The location of each surveillance camera can be: indoor, outdoor, entrance, exit. Each surveillance camera is also characterized by the type of adopted technology (e.g., infrared, high-resolution, etc., one kind for each camera). Moreover, for each recognition the kind of lighting (solar or artificial) is known.

The surveillance company is interested in analyzing, for each surveillance camera, the number of transits of people, males, females and children according to:

- location and technology of surveillance camera
- site where the surveillance camera is installed, customer and type of monitored site (public or private)
- city, region, state of monitored site
- confidence level of silhouette recognition (high, medium, low) and kind of lighting (solar, artificial)
- day hour and time slot (i.e., 0-8, 8-12, 12-18 e 18-24)
- date, day of the month, day of the week, month, year
- the working days or the holidays.

The data warehouse will store information about years 2006-2010. Moreover, the following statistics are known (the candidate may estimate missing information that she deems relevant):

- there are on average 10 surveillance cameras for each monitored site
- there are about 10 monitored sites in each city
- the company has installations in around 1000 cities for each of the 10 states in which it operates
- the company has 10 thousand customers
- there are 4 different locations for the installation of surveillance cameras and the surveillance cameras are built with 4 different technologies

The following are some of the frequent analyses the company is interested in:

(a) Select the percentage of males and the percentage of females with respect to the total number of recognized people (children included), separately for each day of the week and customer. Consider only the data related to 2010 with a “high” confidence level.

(b) Select the percentage of people on holidays with respect to the total, separately for each monitored site and considering only confidence level “high”.

(c) For each month, select the cumulative number of people from the beginning of the year, the monthly total number of people and the percentage of recognized people during the month with respect to the yearly total, separately for each monitored site.

Design

(a) (7 Points) Design the data warehouse to address the described issues. In particular, the designed data warehouse must allow efficient execution of all the queries described in the specifications.

(b) (4 Points) Write frequent query (b) of the “problem specifications” using the extended SQL language.

(c) (*Optional:* 5 Points) Write frequent query (c) of the “problem specifications” using the extended SQL language.