Database Management Systems

February 23th, 2016

1. (7 Points) The following relations are given (primary keys are underlined):

   USER (UId, Name, Surname, Email, City, Region, Bith_Date)
   PURCHASE (Timestamp, UId, Shop, Amount, Cashback, #Items)
   DONATION (Timestamp, UId, Amount, IID)
   INITIATIVE(IId, Name, Description, City, Region, Category)

   Assume the following cardinalities:

   - card(USER)= 10^5 tuples,
     number of cities: 500,
     number of regions: 21,
     MIN(Bith_Date) = 1/1/1945, MAX(Bith_Date) = 31/12/1994,
   - card(PURCHASE)= 10^8 tuples,
     MIN(TO_DATE(Timestamp)) = 1/1/2014, MAX(TO_DATE(Timestamp)) = 31/12/2015,
     number of shops: 10^3,
     MIN(Amount) = 1 Euro, MAX(Amount) = 1000 Euro
     MIN(Cashback) = 0 Euro, MAX(Cashback) = 10 Euro
   - card(DONATION)= 10^4 tuples,
     MIN(Amount) = 1 Euro, MAX(Amount) = 100 Euro
     MIN(TO_DATE(Timestamp)) = 1/1/2014, MAX(TO_DATE(Timestamp)) = 31/12/2015,
   - card(INITIATIVE)= $5 \cdot 10^2$ tuples,
     number of cities: 50,
     number of regions: 21,
     number of categories: 10

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

   - having count(*)$\leq$100 $\simeq \frac{1}{10}$.

   Consider the following SQL query:

```
select Region, SUM(Amount)
from DONATION D, INITIATIVE I
where D.IId = I.IId and Category = 'Animals'
and ((TO_DATE(D.Timestamp) ≥ 1-12-2014 and TO_DATE(D.Timestamp) ≤ 31-12-2014)
or (TO_DATE(D.Timestamp) ≥ 1-12-2015 and TO_DATE(D.Timestamp) ≤ 31-12-2015))
and D.UId NOT IN (select U.UId
    from PURCHASE P, USER U
    where P.UId = U.UId and P.Cashback ≤ 1 and U.Birth_Date ≥ 1-1-1985
    group by U.UId
    having count(*)≤100)
group by Region;
```

   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. (8 Points) The following relations are given (primary keys are underlined, optional attributes are denoted as *).

```
BOOK(ISBN, Editor, PubblicationDate)
AUTHOR(AuthorCode, Name, Surname, BirthDate)
TOTAL_SALES_IN_BOOKSHOP(BookshopCode, ISBN, TotalNumberSoldCopies, TotalRevenue)
AUTHOR_REVENUE(ISBN, AuthorCode, AuthorContribution, AuthorRevenue)
BOOK_SALE(BookshopCode, ISBN, Date, NumberSoldCopies, Revenue)
BEST_SELLING_BOOK(BookshopCode, Timestamp, ISBN, TotalRevenue, NumberOfAuthors)
```

You want to automatically manage the information about the book sales. The `BOOK` table lists the books on sale at the various bookshops. For each bookshop, the `TOTAL_SALES_IN_BOOKSHOP` table shows the total number of sold copies for each book and the corresponding total revenue. Each book can have one or more authors. The `AUTHOR_REVENUE` table lists the authors of each book. Each author of the book receives a revenue proportional to its contribution to the book (attribute `AuthorContribution` in the `AUTHOR_REVENUE` table, numeric attribute with values in the range between 0 and 1). The `AuthorRevenue` attribute indicates the revenue obtained by each author on the book. It is computed as the product of the revenue achieved on the book multiplied by the author's contribution to the book.

Write the triggers to manage the following activities.

(1) *Update of the book sales in a bookshop and of the revenues for the authors of the book.* Every day the information about the daily sales of a book in a bookshop are stored in the `BOOK_SALE` table (insertion of a new record). The following operations should be carried out.

(a) You must update in the `TOTAL_SALES_IN_BOOKSHOP` table the total number of copies of the book sold at the bookshop and the total amount grossed by the bookshop for that book. Consider that, in the `TOTAL_SALES_IN_BOOKSHOP` table, for each bookshop is already present the record for a book only for books for which it has been previously sold at least one copy in the bookshop. You must also record (insert into the `BEST_SELLING_BOOK` table) the book that has currently received the highest revenue in the bookshop. Assume that there is at most one book holding this revenue value at the bookshop. For this book you must record the total revenue at the bookshop and the number of authors of the book. The value of the `timestamp` attribute is given by the system variable `sysdate`.

(b) You must update the total revenue for each of the authors of the book (update of the `AuthorRevenue` attribute) taking into account the new sold copies.

(2) *Integrity constraint on the contribution of the authors of a book.* The integrity constraint requires that the sum of the contributions (attribute `AuthorContribution`) of the different authors to the same book is always lower than or equal to 1. Any modification operation on the `AUTHOR_REVENUE` table that causes the violation of constraint must not be executed.

1. Data Warehouse design

   The University and Research Ministry would like to analyze how much time the researchers in its country spend on funded-project activities. Each hour spent in research activities on a project by a researcher has a nominal cost.

   Each research project consists of disjoint groups of activities, named Work Packages. The research project activities belong to one of the following three types: Management, Research & Development, Other. Each Work Package includes only activities of the same type.

   The research projects are divided by topic (e.g., Big Data, Smart City); each topic belongs to a specific discipline (e.g., computer science, urban design). Each research project is identified by a name, has a length in time (one or more years), and starts in response to a call for funding. Each call is characterized by a specific funding scheme (e.g., grant, co-financing, venture capital), it is managed by a funding body (e.g., European Union, Ministry, foundations), and it belongs to a specific sector (e.g., ICT, Nanotechnologies, Energy).

   The individual researchers fill in a monthly report, named timesheet, where they state how many hours they worked on the activities of each Work Package on the different projects they are involved in. Each individual researcher belongs to a Research Unit (e.g., the university department). Each Research Unit belongs to a specific Research Body (e.g., Politecnico di Torino, Istituto Superiore Mario Boella).

   The Ministry would like to analyze the total number of hours, the total nominal cost, and the average hourly cost, according to the following dimensions:

   - project and Work Packages;
   - type of project activity, project topic, and discipline;
   - project lenght in time (number of years, whose value can be 1, 2, 3, 4, 5, or more than 5);
   - call, and sector of the call;
   - funding scheme, and funding body;
   - research unit, and research body;
   - month, month of the year, 2-month period, 3-month period, 6-month period, year, academic year (starting in September and ending in August).

   Design

   (a) (6 Points) Design the data warehouse, including both the conceptual model and the fact and dimension tables, to address the given specifications. The data warehouse must also allow efficient execution of the following queries.

   (b) (8 Points) Write the following queries using extended SQL language.

(a) For each 2-month period and for each project, select the average hourly cost, and the average number of hours per Work Package. Select also the 2-monthly cumulative total costs for each academic year and project.

(b) Select the average hourly cost of each Work Package, the percentage of hours of each Work Package with respect to the total of each project, and the percentage of hours that each project devoted to each activity type with respect to the total of the project.