CS245 – Databases. Programming Project

Dec 2, 2015

The goal of the project is to compute the following SQL query:

```
SELECT TOP 10 ID (colX + colY + … + colZ) as mySum
FROM Data D
ORDER BY mySum DESC
```

You can build your own code or use *any* existing system, including a commercial DBMS. You are free to copy/paste source code from the web. You can preprocess the data in any way and create any number of additional structures. **The only limitation is that the program must run on the Ubuntu-15 machine in the CS lab.**

This project counts for 25% of your final grade for CS245. If your program generates wrong results, you will get 0 (zero) marks. If your program does not finish within 5 minutes, you will get 0 (zero) marks. If your program runs correctly (i.e., generates exactly the same result as the SQL query in less than 5 minutes), then we will evaluate the performance (see below for details). The most efficient program will get 25 marks and the least efficient 15 marks.

This is not a team project. Everybody is expected to provide his/her own code. Cases of plagiarism will be taken very seriously and will result to 0 (zero) marks plus disciplinary actions from the university. If plagiarism is suspected, I reserve the right to ask the student to prove his/her abilities by implementing in my office within limited time a similar operator.

Deliverables:

(i) Set up the entire environment you need (including the installation of any software) on the lab machine

(ii) Prepare a project report. The report must describe the basic idea of your solution. Your report must be maximum 1 page in the IEEE format: http://www.ieee.org/publications_standards/publications/authors/author_templates.html

You will submit the PDF file of the report to panos.kalnis@kaust.edu.sa .

Deadline: Tue, 15-Dec-2015, 23:59. This is a strict deadline! If you miss the deadline you will get 0 (zero) marks. There will be no extension for any reason.

Dataset

There is one relation with the following schema:

Data(ID:Int, col1:Int, col2:Int,… ,col20:Int), where ID is the primary key

You will receive a python program that generates the data file `data-c20.txt`. Execute as follows:

- `python test.py`

The program will generate 50M tuples (around 10GB). It takes quite some time to finish (at least 30min); be patient.

Input parameters

At runtime you will be given the set C of columns that participate in mySum. For example:

C = {col1, col3, col7}

Any combination of 1 up to 20 columns may be given.

Evaluation

Because of the diverse environments/systems that each one may employ, we will use either the linux command 'time' or a chronometer to measure the real (wall) time.