CS412: An Introduction to Data Warehousing and Data Mining **Fall 2014**

Assignment 2

Due Date: 10/08/2014

General Instruction

- Feel free to talk to other members of the class in doing the homework. We are more concerned that you learn how to solve the problem than that you demonstrate that you solved it entirely on your own. You should, however, write down your solution yourself. Please try to keep the solution brief and clear.
- Please use Piazza first if you have questions about the homework. Also feel free to send us e-mails and come to office hours.

Assignment Submission

- The homework is due at **11:59 PM on the due date**. **We do NOT accept late homework!**
- We will be using Compass for collecting the homework assignments. Please submit your answers via Compass (http://compass2g.illinois.edu). Please do NOT hand in a hard copy of your write-up. Contact the TAs if you are having technical difficulties in submitting the assignment.
- The homework **MUST** be submitted in **pdf** format. Answers to the written part and mini-MP should be included in one .pdf file.
- If scripts are used to solve problems, you are required to submit the source code, and use the file names to identify the corresponding questions. For instance 'Question2.netid.py' refer to the python source code for Question 2, replace 'netid' with your netid. Please **DO NOT** zip the files.
- For each question, you will **NOT** get full credit if you only give out a final result. Necessary calculation steps and reasoning are required.

Dataset

- The data for Q2 can be download from compass2g or the link.

Question 1

[30'] Assume a base cuboid of 10 dimensions contains only two base cells:

$$(1)\ (a_1, a_2, a_3, b_4, ..., b_9, b_{10}),\ \text{and}\ (2)\ (b_1, b_2, b_3, b_4, ..., b_9, b_{10})$$

where a_i is not equal to b_i (for any i). The measure of the cube is count.

Purpose Have a better understanding of cubes, multidimension view of data, and cuboid structures.

Requirement Include final results and state reasons how you calculate the cells. Keep it brief and clear.

1. [5'] **How many nonempty aggregated (i.e., non-base) cells a complete cube will contain?**

2. [5'] **How many nonempty aggregated cells an iceberg cube will contain, if the condition of the iceberg cube is** $count \geq 2$**?**

3. [10'] **How many (non-star) dimensions does the closed cell with count 2 have?**

4. [10'] **How many closed cells in the full cube?**

Question 2

[50'] You can find an artifically generated dataset here. It contains 50 lines, with each line representing a business. For each business, it has (business_id, City, State, Category, Rating, Price). We now want to construct a cube over four dimensions (Location, Category, Rating, Price) with "Count" as the measure. Note that in the Location dimension, there is a concept hierarchy, i.e. City and State. Please process the data, and answer following questions.

Purpose Have a better understanding of measures, and cuboid structures.

Requirement

- You should write scripts to manipulate the data. No restrictions on programming language.
- You can use any built-in functions.
- Include your source code in your final submission.
- You also need to include **necessary** steps in your .pdf file, to show how you get the results. Keep it brief and clear.

1. [10'] **How many cuboids are there in this cube**

2. [8'] **How many cells are there in the cuboid (Location(city), Category, Rating, Price)**

3. [8'] **Now let's drill up by climbing up in the Location dimension, from City to State. How many cells are there in the cuboid (Location(State), Category, Rating, Price)?**

4. [8'] **Further, how many cells in the cuboid (* , Category , Rating , Price)?**

5. [8'] **What is the count for the cell (Location(state) = 'Illinois' , * , rating = 3 , Price = 'Moderate')?**

6. [8'] **What is the count for the cell (Location(city) = 'Chicago' , Category='food' , * , *)?**

Mini-MP

[20'] CubesViewer is a visual, web-based tool application for exploring and analyzing OLAP databases served by the Cubes OLAP Framework[1]. The CubesViewer Explorer demo can be found at the link.

Purpose
- Have a better understanding of OLAP operations, multi-dimension view of data.
- Get some hands-on experience on OLAP.

Requirement
- Write down your OLAP operation, include which operation(s) (Roll-up, drill-down, slice, dice) on which dimension(s) (product, geo, browser, etc.).
- Include the snapshots of your results.
- For subquestion 3, you need to write out the cube you are interested in, you also need to write down OLAP operations to get your results, and snapshot of your results.
- Necessary analysis is required for subquestion 3.

1. [5'] **With regard to the cubes of "Webshop / Sales" on the demo page, find the product under category Sports with most revenue in North American from September 2012 to September 2013, and what is the least?**

2. [5'] **With regard to the cubes of "Website / Visits" on the demo page, what's the popular way (source + browser) of visiting the web store from Europe? Do the visiting counts change along time in Europe? what's best granularity of visualization?**

[1] https://github.com/jjmontesl/cubesviewer

3. [10'] **For each data set ("Webshop / Sales" and "Website / Visits"), come up with a interesting cube that might help decision making. Also, you need to reason about what kind of decision making would benefit from your cube(s).**