CSE 40647/60647: Data Science **Fall 2017**

Homework 2

Handed Out: September 12, 2017 *Due: September 26, 2017 11:59 pm*

1 General Instructions

- This assignment is due at 11:59 PM on the due date. We will be using Sakai (https://sakailogin.nd.edu/portal/site/FA17-CSE-40647-CX-01) for collecting this assignment. Contact TA if you face technical difficulties in submitting the assignment. We shall NOT accept any late submission!
- The homework MUST be submitted in pdf format. Handwritten answers are not acceptable. Name your pdf file as YourNetid-HW2.pdf
- You need to explain the logic of your answer/result for every question. A result/answer without any explanation will not receive any point.
- It is OK to discuss the problems with the TA and your classmates, however, it is NOT OK to work together or share code. Plagiarism is an academic violation to copy, to include text from other sources, including online sources, without proper citation. To get a better idea of what constitutes plagiarism, consult the Honor code on academic integrity violations, including examples, and recommended penalties. There is a zero tolerance policy on academic integrity violations. Any student found to be violating this code will be subject to disciplinary action.
- Please use Piazza if you have questions about the homework. Also feel free to send TA emails and come to office hours.

2 Question 1 (25 points)

Suppose the base cuboid of a data cube contains two cells

$(a_1, a_2, a_3, a_4, \ldots, a_{10}) : 1$, $(a_1, b_2, a_3, b_4, \ldots, b_{10}) : 1$

where $a_i \neq b_i$ for any i. Obviously here we have 10 dimensions and each has only one level (no concept hierarchy).

1. How many **nonempty cuboids** are there in this data cube?
2. How many **nonempty aggregate closed cells** are there in this data cube?
3. How many **nonempty aggregate cells** are there in this data cube?
4. If we set minimum support = 2, how many **nonempty aggregate cells** are there in the corresponding iceberg cube?

3 Question 2 (25 points)

Which of the following algorithms: (i) Multiway array aggregation, (ii) BUC, cannot support the following operations efficiently? and explain why. This could be multiple-choice questions.

1. Computing an iceberg cube;
2. Supporting efficient OLAP query processing on a large dataset with 50 dimensions.

4 Question 3 (25 points)

Assume a base cuboid of 10 dimensions contains only three base cells[1]:

- $(a_1, d_2, d_3, d_4, \ldots, d_{10}) : 1$,
- $(d_1, b_2, d_3, d_4, \ldots, d_{10}) : 1$,
- $(d_1, d_2, c_3, d_4, \ldots, d_{10}) : 1$,

where $a_1 \neq d_1$, $b_2 \neq d_2$ and $c_3 \neq d_3$. The measure of the cube is *count*. Here we have 10 dimensions and each has only one level (no concept hierarchy).

1. How many **nonempty cuboids** will a full data cube contain?
2. How many **nonempty aggregate cells** will a full cube contain?
3. How many **nonempty aggregate cells** will an iceberg cube contain with the condition $count \geq 2$?

5 Question 4 (25 points)

We have a data array containing 3 dimensions A, B and C shown in Figure 2. The 3-D array is divided into 27 small chunks. Each dimension is divided into 3 equally sized partitions. The cardinality (size) of the dimensions A, B, and C is 900, 300, and 600. Since we divide each dimension into 3 parts with equal size, the sizes of the chunks on dimensions A, B, and C are 300, 100, and 200 respectively. Suppose we want to use **Multiway Array Aggregation Computation** to materialize the 2-D cuboids AB, AC and BC.

1. What is the ordering of chunk scanning that achieves the maximum computation efficiency, i.e. requires the least memory units?
2. Following the ordering you give in the above subquestion, what is the minimum memory requirement for holding all the 2-D planes?

[1] Here when we say "only three base cells", we assume that we have ONE transaction in each of the base cells AND we only have these THREE transactions in the cuboid.

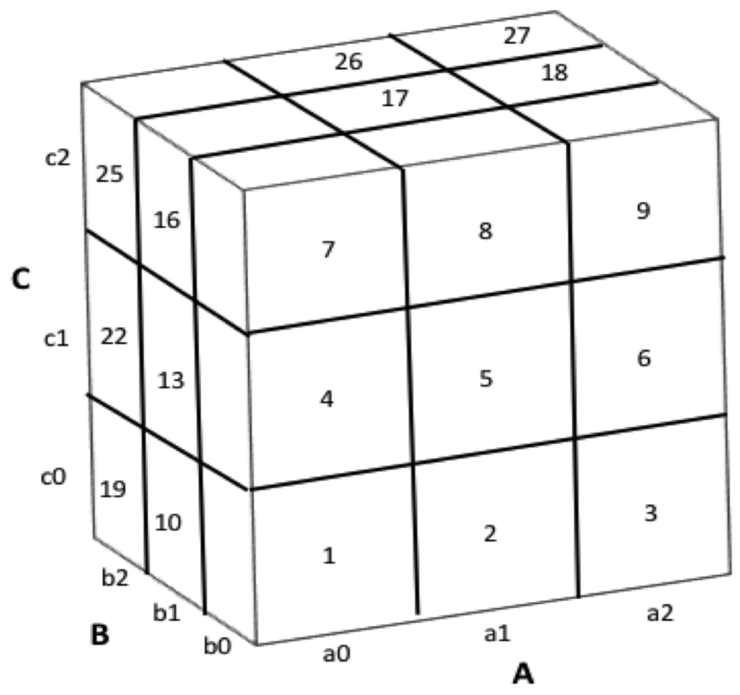

Figure 1: 3-D array of a data cube containing dimensions A, B, C and is divided into 27 small chunks.