

Assignment 2

CS 412: Introduction to Data Mining (Spring 2023)

Instructor: Hanghang Tong

Release date: Feb. 2nd, 2023

Due date: Feb. 14th, 2023

- This assignment will cover the content of Chapter #3 (Data Warehousing and Online Analytical Processing).
- Feel free to discuss with other members of the class when doing the homework. You should, however, write down your own solution **independently**. ***Very Important Notes***: (1) **there is a fine line between collaboration and completing the assignment by yourself** and (2) **aiding others to cheat would have the same consequence as the cheating itself**. Please try to keep the solution brief and clear.
- Please use Piazza first if you have questions about the assignment. Also feel free to send us e-mails and come to office hours.
- The assignment is due at 11:59 PM on the due date. We will be using Piazza for collecting homework assignments. **Please do not hand in a scan of your hand-written solution, only the typed solution (e.g., Microsoft Word, Latex, etc) will be graded**. Contact the TAs if you are having technical difficulties in submitting the assignment. We do **NOT** accept late assignment!
- The assignment should be submitted as a **single** PDF file using the name convention **yourNetID_HW2.pdf**. If you use additional source code for solving problems (Note that simply using Excel for computation is not allowed, as the Excel table might not show the intermediate steps well. Please use programming languages, e.g., python, Java, C++, etc.), you are required to submit them and use the file names to identify the corresponding questions. For instance, **yourNetID_HW2_P1.py** refers to the python source code for Problem 1, replace netid with your netid. Compress all the files (PDF and source code files) into one zip file. Submit the compressed file **ONLY**. (If you did not use any source code, submitting the PDF file without compression will be fine)
- For each question, **you will NOT get full credits if you only give out a final result**. Necessary calculation steps are required. If the result is not an integer, round your result to 4 decimal places.

Problem 1. OLAP (12 points)

We would like to build a data cube of the fruit data, and want to include the following dimensions and measures:

- 4 dimensions: (Variety, Location, Size, Maturity)
- 3 measures: (Price.mean, Price.min, Fruit.num_eq_min)

It indicates that the fruit price is a function of variety, location, size and maturity. Suppose we are aggregating from the Maturity 1-D cuboid to the apex cuboid. **The meanings of these measures are explained below.**

- (1) (6 points) Show that Price.mean is an algebraic measure. What distributive measures do you need to use in calculating Price.mean? Here Price.mean denotes the mean value of all prices.
- (2) (6 points) Explain whether Fruit.num_eq_min is an algebraic measure or holistic measure. Please justify your answer. Here Fruit.num_eq_min means the number of fruits that have the lowest price. For example, given 5 fruits with prices: 10, 11, 23, 10, 23, then the lowest price is Price.min = 10 and Fruit.num_eq_min = 2.

Problem 2. Data Cube Concepts (26 points)

Suppose the base cuboid of this data cube contains two cells:

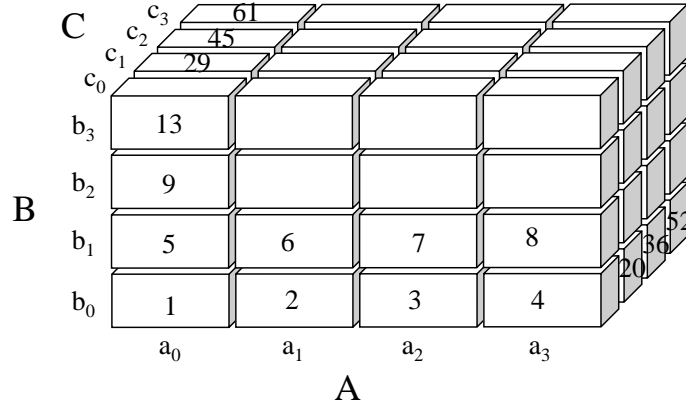
$$(a_1, a_2, a_3, a_4, a_5, a_6, a_7) : 1, (a_1, b_2, a_3, b_4, a_5, b_6, a_7) : 1$$

where $a_i \neq b_i$ for $i = 2, 4, 6$. Assume each dimension contains no concept hierarchy.

- (1) (6 points) Please list all the (nonempty) closed cells in this data cube.
- (2) (4 points) How many (nonempty) aggregate cells are there in this data cube?
- (3) (4 points) How many (nonempty) aggregate closed cells are there in this data cube? Please list them.
- (4) (4 points) If we set minimum support = 2, how many (nonempty) aggregate cells are there in the corresponding iceberg cube?
- (5) (8 points) What are the differences among star schema, snowflake schema, and fact constellations for modeling the data warehouses? Which schema do you suggest to model this cube? Please justify your answer.

Problem 3. Data Cube Computation (20 points)

Assume our data is stored in a data cube with 3 dimensions A, B and C. We would like to do full cube computation using multi-way array aggregation. The lengths of dimensions A and B are 4000 and 400, respectively. The length of dimension C is an unknown value x ($x > 0$). We cut each dimension into quarters and get 64 chunks as follows.



- (1) (4 points) If we follow the scan order $1 - 2 - 3 - 4 - 5 - 6 - \dots$, what is the memory requirement to compute the whole cube?
- (2) (4 points) If we follow the scan order $1 - 5 - 9 - 13 - 2 - 6 - \dots$, what is the memory requirement to compute the whole cube?
- (3) (4 points) If we follow the scan order $1 - 17 - 33 - 49 - 5 - 21 - \dots$, what is the memory requirement to compute the whole cube?
- (4) (4 points) What is the maximum **integer value** of x if (3) is the most memory-saving order **among (1), (2) and (3)**?
- (5) (4 points) Besides the scan orders in (1)-(3), are there any other scan orders? If yes, please list all the other scan orders by specifying the first 10 chunks in their orders (you can assume the order always starts from chunk 1). If not, please justify your answer.

Problem 4. Snowflake Schema (18 points)

Suppose that a data warehouse for Big Ten universities consists of the four dimensions: *student*, *course*, *semester*, and *instructor*, and two measures: *avg grade* and *count*. Each dimension has a concept hierarchy, e.g., for *student*, the concept hierarchy is “*student* < *major* < *college* < *university* < *all*”; for *course*, the concept hierarchy is “*course* < *department* < *college* < *university* < *all*”, etc.

At the lowest conceptual level, i.e., for a given *student*, *course*, *semester*, *instructor* combination, the *avg grade* measure stores the actual course grade of the student. At higher levels of the concept hierarchy for one or more dimensions, *avg grade* stores the average grade for the given combination.

- (1) (8 points) Assume that *student* has attributes $S_1, S_2, S_{3,key}, S_4$, where $S_{3,key}$ has attributes $S_{3,key}, S_{3,1}, S_{3,2}, S_{3,3}$; *course* has attributes $C_1, C_{2,key}, C_3$, where $C_{2,key}$ has attributes $C_{2,key}, C_{2,1}$; *semester* has attributes T_1, T_2 , and *instructor* has attributes I_1, I_2 . Draw a *snowflake schema* diagram for the data warehouse, where the dimension tables will be based on the attributes of the dimensions.
- (2) (5 points) Starting with the base cuboid [*student*, *course*, *semester*, *instructor*], what specific OLAP operations (e.g., roll-up, drill down, etc.) should you perform in order to list the average grade of Computer Science courses (i.e., department = “Computer Science” for dimension *course*) for each student.
- (3) (5 points) If each of the four dimensions in the data warehouse has five levels (including all), e.g., “*student* < *major* < *status* < *university* < *all*” for student, how many cuboids will the data cube contain (including the base and apex cuboids)? Clearly justify your answer. and we have the option of not having a dimension.

Problem 5. OLAP Operations (24 points)

In this problem, we explore some OLAP operations¹ with `pandas`², a famous data analysis tools for Python.

We analyze a dataset of student performance on Mathematics (file: `student-mat.csv`), which is a part of the Student Performance Data Set³. In the dataset, each record (row) is a student, each attribute (column) has different meaning with explanation in file `student.txt`.

- (1) (4 points) Load the data. How many attributes does the dataset have? How many students (records) does the dataset have?

Hint: You need to set an appropriate delimiter `sep` when using `pandas.read_csv()`.

- (2) (4 points) What is the mean of all students' final grades (G3)? What is the median?

Hint: Here is a tutorial for calculating summary statistics.

- (3) (8 points) Slicing and dicing.

(a) How many students are 16 years old? What is the mean of their final grade?

(b) How many students have age from 16 to 18 (both inclusive) and study time more than 5 hours a week simultaneously? What is the mean of their final grade?

Hint: Check Boolean indexing to see how to select data and write conditions and logical operations. Also remember to check the definition of `studytime` attributes.

- (4) (8 points) Pivot table. The pivot table API in `pandas` can help us pivot and visualize two or more dimensions of the data cube. Draw a pivot table with two dimensions: study time and travel time, and show the mean of final grades. It is not necessary to convert the index (e.g., `traveltime=1`) to actual meaning (e.g., < 15 min). You can type the table or take a screenshot.

Hint: Check the pivot table API and specify `values`, `index`, `columns`, and `aggfunc`.

¹`pandas` is not an OLAP system. We only use it to understand some operations on multidimensional data.

²<https://pandas.pydata.org/docs/index.html>

³The full dataset contains both Mathematics and Portuguese performance. We only use the Mathematics performance.