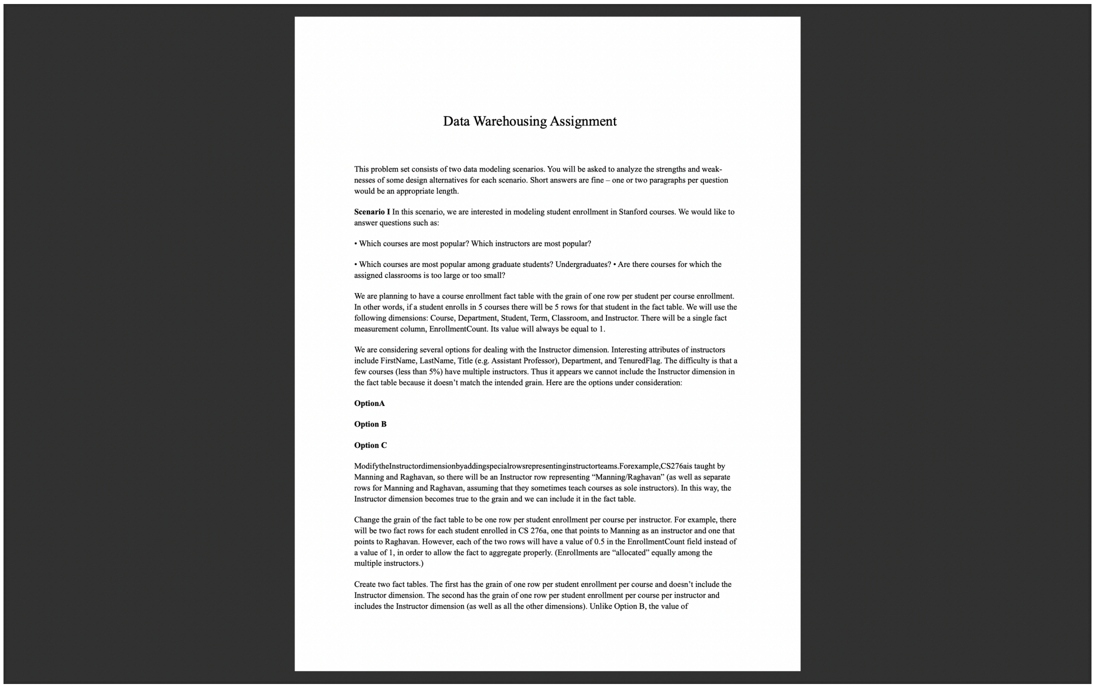
Data Warehousing Assignment 2



Data Warehousing Assignment

This problem set consists of two data modeling scenarios. You will be asked to analyze the strengths and weaknesses of some design alternatives for each scenario. Short answers are fine – one or two paragraphs per question would be an appropriate length.

Scenario I In this scenario, we are interested in modeling student enrollment in Stanford courses. We would like to answer questions such as:

• Which courses are most popular? Which instructors are most popular?

• Which courses are most popular among graduate students? Undergraduates? • Are there courses for which the assigned classrooms is too large or too small?

We are planning to have a course enrollment fact table with the grain of one row per student per course enrollment. In other words, if a student enrolls in 5 courses there will be 5 rows for that student in the fact table. We will use the following dimensions: Course, Department, Student, Term, Classroom, and Instructor. There will be a single fact measurement column, EnrollmentCount. Its value will always be equal to 1.

We are considering several options for dealing with the Instructor dimension. Interesting attributes of instructors include FirstName, LastName, Title (e.g. Assistant Professor), Department, and TenuredFlag. The difficulty is that a few courses (less than 5%) have multiple instructors. Thus it appears we cannot include the Instructor dimension in the fact table because it doesn't match the intended grain. Here are the options under consideration:

OptionA

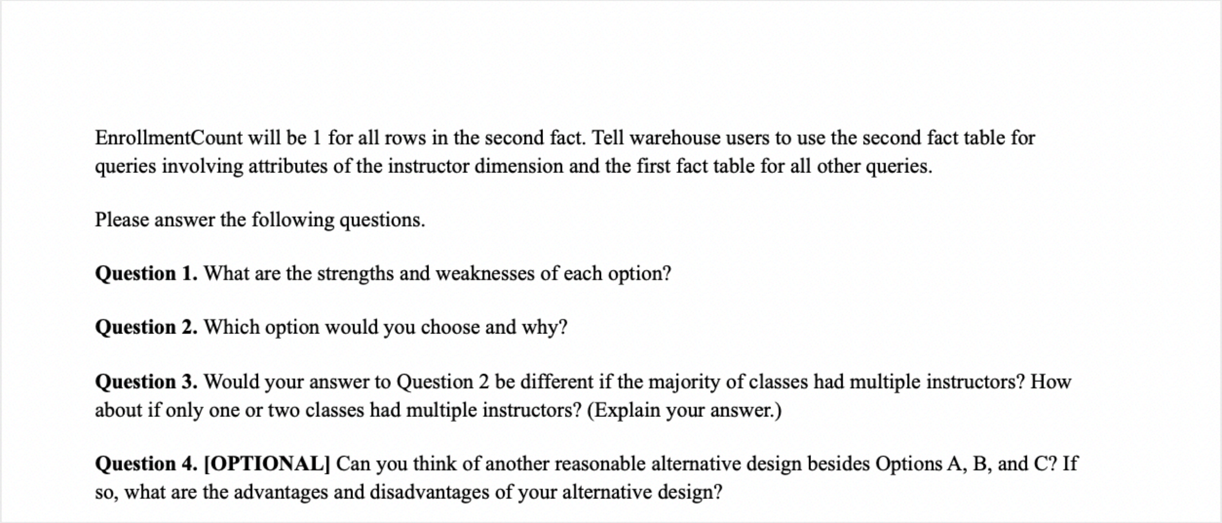
Option B

Option C

ModifytheInstructordimensionbyaddingspecialrowsrepresentinginstructorteams.Forexample,CS276ais taught by Manning and Raghavan, so there will be an Instructor row representing "Manning/Raghavan" (as well as separate rows for Manning and Raghavan, assuming that they sometimes teach courses as sole instructors). In this way, the Instructor dimension becomes true to the grain and we can include it in the fact table.

Change the grain of the fact table to be one row per student enrollment per course per instructor. For example, there will be two fact rows for each student enrolled in CS 276a, one that points to Manning as an instructor and one that points to Raghavan. However, each of the two rows will have a value of 0.5 in the EnrollmentCount field instead of a value of 1, in order to allow the fact to aggregate properly. (Enrollments are "allocated" equally among the multiple instructors.)

Create two fact tables. The first has the grain of one row per student enrollment per course and doesn't include the Instructor dimension. The second has the grain of one row per student enrollment per course per instructor and includes the Instructor dimension (as well as all the other dimensions). Unlike Option B, the value of



EnrollmentCount will be 1 for all rows in the second fact. Tell warehouse users to use the second fact table for queries involving attributes of the instructor dimension and the first fact table for all other queries.

Please answer the following questions.

Question 1. What are the strengths and weaknesses of each option?

Question 2. Which option would you choose and why?

Question 3. Would your answer to Question 2 be different if the majority of classes had multiple instructors? How about if only one or two classes had multiple instructors? (Explain your answer.)

Question 4. [OPTIONAL] Can you think of another reasonable alternative design besides Options A, B, and C? If so, what are the advantages and disadvantages of your alternative design?

Answer1:

Option A:

Strength: Accurately displays the record without even storing more data, optimized way to solve the Instructor issue.
Weakness: It may become difficult to understand the names of two Instructors

together, but I think this is the only way to represent. Also, if there are multiple Instructors, that might create issue in creating data schema eg: we need to define space "varchar(30)" which may take more space or may throw an error if the names are bigger than expected.

Option B:
Strength: Separately defines the names of each Instructor in each line without congesting the names in just 1 line.
Weakness: A lot of unnecessay entries in the table for a student per course who is being taught by multiple instructors.

Option C:
Strength: We solved the issue of putting Instructor names to the Fact table as in the second table, we can find the information related to Instructor and in the first, related to all other not related to Instructor.
Weakness: storing more data in this way, some duplicate columns in both tables therefore space requirement issue, requires more computational power as 2 fact tables are used instead of one. Also, it will display wrong counts in the output for multiple instructors per student per course enrollment.

Answer 2:
I would choose option A if the Instructors are not greater than 2 because it will save a lot of memory and computation power of CPU. Also, it would be easir to understand in single line of vizualisation in the output.

Answer 3:
If there are multiple Instructors, I would choose Option D as it will display more Instructors data precisely than any other Option available.

Answer 4:

Option D:
We can duplicate Option C but the only difference is that we will not assign Enrollment Count as 1 in front of each Instructor in the second fact table. Instead of that, we would do 0.5, 0.33 or some other number dependent upon the number of Instructors taking that course.
Advantages: We solved the issue of putting Instructor names to the Fact table as in the second table, we can find the information related to Instructor and in the first, related to all other not related to Instructor.
DisAdvantages: storing more data in this way, some duplicate columns in both tables therefore space requirement issue, requires more computational power as 2 fact tables are used instead of one.