**DWH\_Assingment-2\_Solution:**

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:

*Q1• Which courses are most popular? Which instructors are most popular?*

Soln: Course- CS 276a is most popular.

Manning and Raghavan are most popular.

*Q2• Which courses are most popular among graduate students? Undergraduates? • Are there courses for which the assigned classrooms is too large or too small?*

Soln: Course- CS 276a is most popular among graduate students and undergraduates students.

No there are not any courses for which the assigned classrooms is too large or too small.

*Problem Statement*:

*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*

**Soln:** I have choosen option A

*OptionA*

*Modify the Instructor dimension by adding special rows representing instructorteams. Forexample, CS276a is 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.*

**Dimension tables**: Course, Department, Student, Term, and Classroom

**Fact table**: Instructor

**Snowflake schema of student enrollment in Stanford courses :**

Instructor(Fact)

Dept\_id

Dept\_Name

No\_of\_instructors

HOD\_name

Instructor\_name

Allocated\_classrooms

Department(Dim)

Course(Dim)

Course\_id

Name

No\_of\_subjects

Duration

Enrollment\_count

First\_name

Last\_name

Course\_id

Dept\_id

TenuredFlag

Instructor\_teams

Enrollment\_count

Student\_id

Student\_id

Enrolled\_course\_name

No\_of\_Term\_student\_passed

First\_name

Last\_name

Term(Dim)

Allocated\_classrooms

Vacant\_classrooms

Classroom\_number

Classroom(Dim)

Student(Dim)

Syllabus

Schedule

Term\_no

No\_of\_Term\_student\_passed

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

Ans:- **Option A**

*Strengths of option A are:-*

* It makes the point clear that which all are dimension tables and which all are fact tables
* It has taken instructor dimension as a instructor fact by including instructorteams as one of the column of fact instructor

*Weaknesses of option A are:*-

* It cannot allocated enrollments equally among the multiple instructors.
* It cannot given equally chance to each and every instructor.

**Option B**

*Strenghts of option B are:-*

* It has allocated enrollments equally among the multiple instructors.
* It has given equally chance to each and every instructor.

*Weaknesses of option B are :-*

* It is not clear about what it is actually want to convey
* It has not mention about what should we do about instructor dimension that we should instructor dimension in fact table or not.

**Option C**

*Strenghts of option C are :-*

* It has divided the situations into two parts.
* It has giving two fact tables.

*Weaknesses of option C are:-*

* It has given the complete solution opposite to the scenario requirement.
* It cannot allocated enrollments equally among the multiple instructors.

***Question 2.*** *Which option would you choose and why?*

Ans:-

* I choose option-A because this option is perfect solution and also completing the schema of student enrollment in Stanford courses.

**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.)

Ans:-

No My answer to Question 2 is not different if the majority of classes had multiple instructors because I have made a instructorteams per course .If only one or two classes had multiple instructors than also not a problem because every single instructor is capable of taking each and every classes.

**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?

**Scenario II** In this scenario, we are building a data warehouse for an online brokerage company. The company makes money by charging commissions when customers buy and sell stocks. We are planning to have a Trades fact table with the grain of one row per stock trade. We will use the following dimensions: Date, Customer, Account, Security (i.e. which stock was traded), and TradeType.

Ans:- Dimension tables:-Date, Customer, Account,

Security (i.e. which stock was traded), and TradeType.

Fact table:- Trades

**Snowflake schema of online brokerage company**

Trades(Fact)

Security\_key

Stock\_QRcode

Stock\_id

Security(Dim)

Created\_at

Updated\_at

Account\_subscription\_type

Account\_id

Account(Dim)

Date\_of\_stock\_traded

Date\_of\_customer\_join

Date\_of\_account\_open

Date\_of\_securitykey\_generated

Date(Dim)

Trade\_import

Trade\_export

Date\_of\_stock\_traded

Account\_id

Stock\_id

Trade\_id

#### Online \_trading\_id

Customer\_id

Updated\_scores

Customer(Dim)

Customer\_id

Account\_number

Customer\_Activity\_Segments

Personal\_details

Stock(Dim)

Transactions(Dim)

Stock\_id

Stock\_name

Stock\_category

No\_of\_stock\_traded

Customer\_Activity\_Segments

Average\_transaction\_size

Recency\_of\_transactions

TradeType(Dim)

#### Short-term Trading

#### Medium Trading

#### Long-term Trading

#### Online \_trading\_id

**CustomerScores(Dim)**

Activity scores

Profitability scores.

Updated\_scores

* data analysts have told us that they have developed two customer scoring techniques that are used extensively in their analysis.
* Each customer is placed into one of nine Customer Activity Segments based on their frequency of transactions, average transaction size, and recency of transactions.
* EachcustomerisassignedaCustomerProfitabilityScorebasedontheprofitsearnedasaresultofthatcustomer’s trades. The score can be either 1,2,3,4, or 5, with 5 being the most profitable.

These two scores are frequently used as filters or grouping attributes in queries. For example:

Q.How many trades were placed in July by customers in each customer activity segment?

**Ans:-** Number of trades were placed in july by customers in each customer activity segment is based on the scores earned by the customers.

Q. What was the total commission earned in each quarter of 2003 on trades of IBM stock by customers with a profitability score of 4 or 5?

**Ans:-** The total commission earned in each quarter of 2003 on trades of IBM stock by customers with a profitability score of 4 or 5 is not mentioned in this scenario.

There are a total of 100,000 customers, and scores are recalculated every three months. The activity level or profitability level of some customers changes over time, and users are very interested in understanding how and why this occurs.

We are considering several options for dealing with the customer scores:

**OptionA Option B Option C**

**Option D**

**Solution:-** I have choosen option C

**Option C**

The scores are stored in a separate Customer Scores dimension which contains 45 rows, one for each combi- nation of activity and profitability scores. The Trades fact table includes a foreign key to the Customer Scores dimension.

*Please answer the following questions.*

Question 5. What are the strengths and weaknesses of each option?

Ans:- **Option C**

The strengths of option C are:-

* Customer scores dimension table helps to calculate the exact and correct scores of the customers.
* The foreign key of the fact table referring to the customer scores dimension table which give the exact value of the trade placed.

The Weaknesses of option C are:-

* In this it is not mention that how many trades were placed by the each and every customers.

**Option A**

The strengths of option A are:-

* The old scores will overwritten by the new updated scores
* The scores row will be added in the customer dimension itself.

The weakness of option A are:-

* We can’t able to know about the old scores once it will overwritten by the new scores.
* In this it is not mention that how many trades were placed by the each and every customers.

**Option B**

The strengths of option B are :-

* There are two separate rows scores and updated scores are included in the customer dimension table.

The weaknesses of option B are:-

* In this it is not mention that how many trades were placed by the each and every customers.

**Option D**

The strengths of option D are:-

* It includes outrigger table .
* The scores are stored in a Customer Scores outrigger table which contains 45 rows.

The weaknesses of option D are:-

* It has completely changed the nature of both dimension table and fact table
* It is unrecognizable that what it is trying to say.
* In this it is not mention that how many trades were placed by the each and every customers.

*Question 6. Which option would you choose and why?*

Ans:- I would choose option C because it helps in calculating the perfect and correct scores of all customers in a more clear and perfect way.

Question 7. Would your answer to Question 6 be different if the number of customers and/or the time interval between score recalculations was much larger or much smaller? (Explain your answer.)

Ans:- No my answer to Question 6 is not different if the number of customers and /or the time interval between score recalculations was much larger or much smaller because customer scores dimension table helps to calculate the exact and correct scores of the customers.