AM05 Data Mgmt – Assignment 4

There are two parts to this homework. Please submit your answers to both parts in one Word document.

Part 1. Star schema and data warehouse

Problem 1. Millenium College wants you to help design a star schema to record grades for courses completed by students. There are four dimension tables, with attributes as follows:

Dimension	Dimension Attributes	Dimension Size comment
CourseSection	1. CourseID 2. SectionNumber 3. CourseName 4. Units 5. RoomID 6. RoomCapacity	During a given semester, the college offers an average of 600 course sections
Professor	1. ProfID 2. ProfName 3. Title 4. DepartmentID 5. DepartmentName	There are typically 300 professors at Millenium at any given time
Student	1. StudentID 2. StudentName 3. Major	Each course section has an average of 30 students, and students typically take five courses per period.
Period	1. SemesterID 2. Year	The database will contain data for 30 periods (a total of 10 years).

The only fact that is to be recorded in the fact table is CourseGrade.

a. Design a *star schema* for this problem. See example in lab 4 for the format you should follow.

b. Estimate the **number of rows** in the fact table, using the assumptions stated previously and estimate the **total size of the fact table (in bytes),** assuming that each field has an average of 5 bytes.

c. If you didn't want to or didn't have to stick with a strict star schema for this data mart, how would you change the design? Why?

d. Various characteristics of sections, professors, and students change over time. For example, people sometime change their names, new major emerge while others

disapper, professors may move department or get promoted, etc. How do you propose designing the star schema to allow for these changes? Why?

Problem 2. Millenium College now wants to include new data about course sections: *the department* offering the course, the *academic unit* to which the department reports, and the *budget unit* to which the department is assigned. Change your answer to Problem 1 to accommodate these new data requirements. Submit your solution with normalized dimension tables (3NF). In other words use the normalization principles we covered earlier in the course to develop a *snowflake schema* (see https://en.wikipedia.org/wiki/Snowflake_schema for a reminder of what that means). Provide a brief rationale for the choices you made to implement the required changes.

Part 2. Text Parsing with Regular Expressions

Please download the `AM05.Assignment04.Data.sql` file from Canvas. Open it in MySQL Workbench and run the SQL statements that creates the `calls_hw4` table and populates it with 516 rows of data in the `original_text` column.

1. Use **regular expressions** and other functions to parse the text in the `original_text` column into the following fields:
    - Date
    - Time
    - Event
    - Address
    - Responding Agency
    - Event ID

    You should use UPDATE queries to store the extracted values in the corresponding columns of the table. You will have to *cast* some of extracted text (e.g. Date) into other data types before you can store the values in the table. The solution for lab 4 give examples of doing this.

    Submit the SQL code you used to perform these tasks along with a screen shot that shows the first 15 rows of the updated `calls_hw4` table using the following query:

    ```
    SELECT call_id, call_date, call_time , event_descrip,
    event_address, resp_agency, event_id FROM calls_hw4 LIMIT 15;
    ```

2. Create an SQL query to report the 5 events types that have the largest number of occurrences in the database. Your results, which should be provided in descending order, should look something like the table below. Submit your SQL code and a screenshot of your results.

Event	Count
MVA with unknown injuries	100
dangerous conditions	99
hit and run	87
parking violation	70
dumpster fire	36

3. Create an SQL query to report the 5 responding agencies that responded to the largest number of calls over the two dates in the database. Summarize the number of calls by date, and calculate the total. Provide results in descending order (the question is answered below using an Excel Pivot Table). It is acceptable to submit the results as three tables, one for the total number of calls on each date and one for the total across the two days. Submit your SQL code and screenshot(s) of your results.

Responding Agency	# Calls responded to on 10/6/2019	# Calls responded to on 10/7/2019	Grand Total
Rochester City Police	66	85	151
Monroe County Police	38	74	112
Rochester City Fire	21	31	52
Regional Traffic Operations Center	19	24	43
Greece Police	9	9	18