MEIC

2018/2019

Data Analysis and Integration

1st semester

Exam – January 16, 2019

Part I

1. Suppose you have two query expressions: Q1(X) :- Q2(X, Y), Q3(Y, Z), X>6 and Q'1(R) :- Q2(R, S), Q3(S, T), R=7. What is the relationship between the two expressions? Justify.

2. In Global-As-View (GAV) schema mapping, we write expressions in the form $Gi(\bar{X}) \supseteq Q(\bar{S})$. In Local-As-View (LAV) we write them in the form $Si(\bar{X}) \subseteq Qi(G)$. What do these formulas mean? Explain both cases.

3. Consider the task of *schema mapping*. Explain how SQL views are useful for *schema mapping.*

4. Explain what is the purpose of the query reformulation component of a virtual data integration system.

Part II

Suppose you are integrating two different web sites that publish songs and artists. Let us call one of those sites A and the other B.

5. You find that *song* in A matches *track* in B (i.e. *A.song ≈ B.track*), even though the set of songs in A is not exactly the same as the set of tracks in B. How could you use the Jaccard coefficient (or Jaccard measure) to find this match? Explain.

6. The same artist can be written in different ways in both systems (e.g., “The Pretenders” and “Pretenders”). To detect this kind of matches, which of the following string matching algorithms would you chose and why: edit distance, Jaro measure, or Soundex measure (only one of them)?

7. If you need to detect approximate duplicate records based on song title, artist name, and record label (e.g. Columbia Records), how can you combine multiple string matching measures in a single similarity score? Give an example.

8. When comparing records from A and B, why is it inefficient to compare all records from one system with all records from the other system? Explain one possible optimization to make such comparison more efficient.

9. Suppose you have already identified clusters of approximate duplicates based on song title, artist name, and record label. How can you obtain a single record (with those three attributes) from a cluster of duplicates with slightly different song titles, artist names, and record labels?

10. Suppose that, before integrating A and B, you perform data profiling on the data from each system. What kind of useful insights can you obtain from data profiling? Give some examples.

➔ Note: answer the questions in the next page in a separate sheet of paper

Part III

Suppose that you are asked to design a data warehouse to manage data concerning the occurrences of medical emergencies, their location, the time and the means (emergency car, motocycle, etc) sent to the location. The data warehouse schema must have three dimensions, d_time, d_location, d_mean, and one fact table named f_occurrences that stores the priority of the occurrence (on a scale 1 to 5, being 5 the top priority)
For each dimension, we want to store the following attributes:

dim_location: latitude, longitude, city, district (with hierarchy: city -> district)
dim_time: min, hour, day, month, year (with hierarchy: min -> hour -> day -> month -> year)
dim_mean: plate_number, type

11. Present the relational schema (star schema) for the data warehouse above using the following notation, where FK means foreign key:
table1(primary_key, attribute1, attribute2) *attribute2: FK(table2)*
Use surrogate keys in dimensions dim_location and dim_time.

12. Write a single query in SQL/OLAP, using GROUPING SETS, ROLLUP or CUBE, that returns the union of all the following results:
    - The number of occurrences with top priority.
    - The number of occurrences with top priority per district.
    - The number of occurrences with top priority per year
    - The number of occurrences with top priority per district and per year.

13. Give a reason for using a surrogate key in dimension d_location instead of the natural key city.

14. Which dimensions of the data warehouse are good candidates for Slowly Changing Dimensions? Justify.

15. When loading the tables of the data warehouse during the ETL process, is there a specific order that should be satisfied? Why?

Part IV. Miscellaneous

16. “Data warehouses often sacrifice data normalization”. Is this statement true or false? Justify

17. Usually, a data warehouse is designed as a star schema. What would be a possible reason for using a snowflake schema or a starflake schema instead of a star schema? Explain in each case.

18. Dice and Slice are two OLAP operations that can be performed over an OLAP cube. Describe in words what these operations do.

19. Data profiling is typically applied before a data cleaning and transformation process. What kind of output can it produce that is useful for data cleaning?

20. Among the data cleaning tasks that we have mentioned in the lectures, indicate one for which it would be useful to use the map-reduce paradigm to execute it and justify.