MEIC

2019/2020

Data Analysis and Integration

1st semester

Exam Resolution – January 15, 2020

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.

Answer*: Q’1(X) is contained in Q1(R) because, for any database D, Q’1(D) ⊆ Q1(D). The two query expressions in the body of Q1(X) and Q1’(X) are the same except the last predicate. However, R=7 is contained in R>6.*

2. 4. 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.

Answer*: The first expression means that the global schema (mediated schema) is composed by the union of ‘i’ sets of queries/views over data sources, i.e., is defined as views over the local schemas (data sources). The second expression means that the schema of each data source (local schema) is defined as a view over the global schema (mediated schema).*

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

Answer*: Given a schema matching (i.e., correspondences), the schema mapping specifies the operations/transformations/queries, which correspond to SQL views, to be performed over the source data so that they can be transformed into the target data.*

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

Answer*: Given a query over a schema, we translate it into a query that refers to the data sources in order to find the combination of local queries that are required to answer a query posed against the mediated schema. Also, it can be used to check if there are equivalent queries and determine if a query can be answered as combination of existing views.*

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.

Answer*: We can use the Jaccard coefficient to compare each attribute in A with each attribute in B (e.g., all songs in A against all tracks in B). We build a similarity matrix between the attributes of both sites where Jaccard(A, B) is the ratio between the number of common values of both attributes (A intersection with B) over the total number of distinct values for both attributes (A union with B), and this value ranges from 0 (no similar) to 1 (similar). We can say that two attributes match if their Jaccard values is higher than or equal to a given threshold.*

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)?

***Answer**: Jaro measure, since it is used to compare short strings and because it accounts more for characters matches independently of their position, it does not penalize much the "The" in the example, Jaro=0.77. The Edit-distance is not a good measure for this case because it would heavily penalize the "The", i.e., the distance in the example would be to insert or delete "The " giving 4 of distance. The Soundex converts each string to a code that represents phonetics, in the example we have two words in one string against one word in the other, as such, the codes for both strings would be distinct, e.g., T163 != P635.*

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

***Answer**: For song title and artist name, which are short strings, I would choose Jaro/Jaro-Winkler. For record label, I would use soundex, which take into consideration the phonetics of words. All these measures could be combined using a weighted average to obtain a single value of total similarity, i.e.: sim_total = 0.2*sim_1 + 0.4*sim_2 + 0.4*sim_3.*
The total similarity could be compared with a given threshold (set up by user) in order to classify two records as similar or not.

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.

***Answer**: If we have many records, comparing all with all may be very expensive. Methods such as Sorted Neighborhood Method can be used to speed up the task.*

9. Suppose you have already identified clusters of 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?

***Answer**: Based on clusters, one can define for each attribute the criteria to use when selecting a value and consider it as the representative of that cluster. For instance, by choosing the most frequent value or the longest value for each attribute in each cluster and combine these values as a single record which represents that cluster.*

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.

***Answer**: One can gather statistics and discover anomalies in data by doing data profiling task. For instance, to detect a country which has more artists, more songs etc. Find that countries in A are represented in a different format than in B, e.g., full name v.s. abbreviation which we would have to transform the data into a common format before integrating A with B.*

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.

***Answer**:*
dim_location(location_key, latitude, longitude, district)
dim_time(time_key, timestamp, day, month, year)
dim_mean(mean_key, plate_number, type)
f_occurrences(location_key, time_key, mean_key, priority) location_key: FK(dim_location), time_key: FK(dim_time), mean_key: FK(dim_mean)

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.

***Answer**:*

```
Select l.district, t.year, count(f.priority) as 'top priority'
From f_occurrences f
natural join dim_location l
natural join dim_time t
Where f.priority = 5
Group by cube (l.district, t.year)
```

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

***Answer**:*
A surrogate key exists to turn the dimension table data independent of the data source and to uniquely identify each record of the table. A possible reason is that one may have the same location belonging to different districts. Another possible reason is that district is a string and if it was considered as the key of d_location, it would be used when joining this table with the fact table. Join attributes that are strings lead to a less efficient join operation than join attributes of type integer.

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

***Answer**:*
Dim_mean, if we consider that a plate number is assigned by a state, thus the same plate number can be assigned for a different type of vehicle in different states.

Dim_location, if we consider that a district can be splitted or merged, or even has its name changed along the time.

Note: any of the two dimensions considered above was considered as a correct answer to the question.

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

***Answer**:*
Yes. First one must load the dimension tables, and after the fact table – to guarantee integrity constrains.

Part IV. Miscellaneous

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

***Answer**:*
TRUE. Data in DW are typically non-normalized. For example, dimensions may contain non-normalized data in order to avoid join operations and thus make OLAP queries running faster.

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.

***Answer**:*
A possible reason to use a snowflake or starflake schema could be to reduce the storage for the DW. And if we don´t mind reducing the performance of the DW.

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

***Answer**:*
Slice selects a particular value of a dimension level (ex: sales by product in 2010), and dice applies conditions on two or more dimensions of the OLAP cube
Or
Slice removes a dimension in an OLAP cube so a cube of n-1 dimensions is obtained. Dice keeps the cells of a cube that satisfy a Boolean condition.

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?

***Answer**:*
Data profiling can be used as a step to detect inaccurate records to be corrected (data clean). For instance, the tuples containing NULL values in an attribute and/or statistic about data that can be used to correct (clean) them.

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.

***Answer**:*
Map-reduce can be used, for instance, to execute the approximate duplicate detection task, i.e. to find records that correspond to the same real world entity.