MEIC/MECD 2023/2024

Data Analysis and Integration

1st semester

Sample exam (with answers)

1. In this course you have learned how to create SQL views over a database.

   a) What is the relationship between views and a mediated schema?

   A mediated schema defines a common set of attributes that serve as an abstraction over multiple data sources. SQL views are used to define the schema mapping between the data sources and the mediated schema.

   b) A view can be defined based on other views. What is the purpose of doing this when defining a mediated schema?

   It is possible to define local views as wrappers over data sources. The mediated schema can then be defined based on a set of global views over local views. Views over views simplify the definition of global views, and also provide an abstraction layer over data source schemas.

   c) Explain the concept of query unfolding and describe when it should be applied.

   Query unfolding is the process of replacing and expanding a query with the definitions of the views used in that query. This process can be repeated until the original query is fully expanded into a query that no longer refers to views.

2. In this course you have learned and used ETL tools, such as Pentaho Data integration (PDI).

   a) ETL tools are used to build materialized or non-materialized views? Justify your answer.

   An ETL process consists in extracting data from data sources, transforming the data, and loading (i.e. storing) the output somewhere, for example in a file, in database table, or in a data warehouse. The output is materialized because it is stored in a persistent way and becomes available without re-computation.

   b) Suppose you use PDI to migrate data from an input table to an output table. The input table is 20 GB in size, but you only have 8 GB of RAM. Do you think it will work? Justify your answer.

   PDI works in streaming mode, meaning that some records are being processed and written to the output at the same time as more records are being read from the input. So, in principle, it should not be necessary to have all records in memory at any point in time. It should work.

   c) In PDI, there is a dialog to define a database connection. In which other tools have you seen a similar dialog? What was the purpose of defining a database connection in those other tools?

   The same dialog was present in Schema Workbench (to connect to the data warehouse tables that are needed to define the OLAP cube), in Pentaho Server (to connect to the data warehouse in order to run MDX queries), and Report Designer (to connect to a database or to a data warehouse in order to run SQL or MDX queries to generate a report).

3. Suppose you are building a transformation to detect approximate duplicate records in a database table with customers (first name, last name, e-mail, phone).

   a) How can you reduce the number of comparisons that need to be done between those records?

   In general, given two records A and B, it is not necessary to compare A with A, or B with B, or both A with B and B with A. Usually, in a transformation, we will use a join step with a condition that compares each record with the following ones, but not with the previous ones that have already been compared. If records are identified by an id, then the join condition could be A.id < B.id.

b) If you had to choose a different string matching technique to compare each field, which techniques would you choose for each field and why?

To compare first names, we could use Jaro or Jaro-Winkler, which are used to compare short strings. For the last name, we could use Soundex, where the comparison is based on phonetics rather than exact spelling. For the e-mail, we could use edit distance to align characters such as @ or dots. For phone we could use Jaccard, which considers transpositions and could be useful to compare phone numbers with the same digits but in a different order.

c) After you have computed the string matching result for each field, how do you decide if two records are duplicates or not? Please explain.

*We need to use a formula with weights (e.g. sim_total = 0.3*sim1 + 0.4*sim2 + 0.3*sim3) to combine the string matching results for each field into one overall measure. Then we need to decide on a threshold for that measure. Pairs of records with an overall measure above or below the threshold (above if similarity; below if distance) will be considered potential duplicates.*

4. In the same customers table as before, suppose that some customers may not have an e-mail or phone (or both), and other customers may have multiple e-mails and multiple phones.

a) How would you use a data profiling tool to discover these anomalies? Please explain.

With a profiling tool such as DataCleaner it is possible to check the number of NULL values in each column (completeness analyzer). This could be used to discover the customers without e-mail or phone. For those with multiple e-mails or phones, we could check for example the string length or word count (string analyzer). Records with long strings or multiple words could then be checked for the presence of multiple values.

Answer the following questions in a separate sheet of paper

---

5. Consider a data warehouse that stores 3-D facts such as “customer *C* bought product *P* on date *D*”.

a) Suppose you have the option of defining a customer hierarchy with two levels (customer, country) or three levels (customer, city, country). What is the impact of this decision on the OLAP operations that you will be able to perform on the data warehouse? Justify your answer.

With city as an additional level, it will be possible to perform OLAP operations such as roll-up from customer to city, drill-down from country to city, and slice or dice on city. Without the city level, roll-up, drill-down, slice and dice will be limited to the customer and country levels.

b) If the customer dimension has four levels (customer, city, state, country) and the state level can be skipped, what do you call this kind of hierarchy? Also, how would you implement (in the data warehouse schema) this possibility of skipping the state?

This is a ragged hierarchy. In the data warehouse schema, skipping the state can be implemented by allowing that attribute to be NULL (star structure) or by having a foreign key between the city table and the country table (snow structure) that skips the state table.

c) If the customer dimension has three levels (customer, city, country), what could make you decide between having a star structure or having a snowflake structure for this dimension?

The decision is based on the presence of additional attributes for those levels. For example, if for a city we want to store additional attributes such as area and population, it would make sense to use a snow structure with a separate table for city, where those attributes are stored only once, rather than having to replicate their values whenever the same city appears.

6. In the topic of data warehousing, you were introduced to the concept of surrogate keys.
   a) Why use a surrogate key instead of the same primary key as in the original database?

      Because the primary keys of data sources may be inconsistent or change over time. In the data warehouse, we adopt our own surrogate keys to be independent from the keys in the original data sources. It is also good practice to have all keys in the data warehouse as integers for faster processing, rather than other data types that may be used in the original database.

   b) Why do slowly-changing dimensions always require a surrogate key?

      In a slowly-changing dimension there may be multiple versions of the same record. All of these versions will have the same natural key, so the natural key cannot be used to identify them. An additional key (the surrogate key) is necessary.

   c) When you use a surrogate key instead of the natural key, what do you have to change in the transformation that populates the fact table?

      In the transformation that populates the fact table, it is necessary to include a database lookup. Given the natural key of the record, the database lookup will fetch the corresponding surrogate key for that record. That surrogate key can be found in a dimension table that has been populated before the fact table.

7. Once the data warehouse has been created, we need to analyze the data contained therein.
   a) What is the purpose of a tool called Pentaho Schema Workbench (PSW)? What is the relationship between PSW and an OLAP tool such as Saiku Analytics?

      The purpose of PSW is to allow the user to define an OLAP cube with dimensions, hierarchies, levels and measures. The cube definition is exported as an XML file that can be imported into an OLAP tool such as Saiku. This OLAP tool will allow querying the cube with MDX.

   b) If you use only ROWS and COLUMNS in an MDX query, how can you analyze data in three or more dimensions?

      *In this case, those additional dimensions must be nested either in ROWS or COLUMNS. This can be done using a CROSSJOIN operation (abbreviated as *) to join multiple dimensions in the same axis. For example: SELECT Customer.Country.Members * Time.Year.Members ON ROWS...*

   c) What is the purpose of the WHERE clause in an MDX query? Give two different examples.

      The WHERE clause can be used for slicing or dicing by selecting particular dimension levels. For example: WHERE Customer.Country.Italy. It can also be used to specify the measures that should be included in the result. For example: WHERE Measures.Sales.

   d) If a reporting tool can get data either from SQL or from MDX, why not querying the database directly with SQL, instead of querying the data warehouse with MDX?

      A reporting tool is typically used to gather analytical data. For that purpose, it is better to query the data warehouse using MDX, because it facilitates the analysis of multi-dimensional data, and also because the data warehouse contains pre-computed measures that can be used for such analysis. When querying the database with SQL, those measures would have to be re-computed every time.