

**Data Analysis and Integration**Sample exam

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?
 - b) A view can be defined based on other views. What is the purpose of doing this when defining a mediated schema?
 - c) Explain the concept of query unfolding and describe when it should be applied.
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.
 - 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.
 - 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?
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?
 - 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?
 - 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.
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.

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.
 - 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?
 - 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?
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?
 - b) Why do slowly-changing dimensions always require a surrogate key?
 - 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?
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?
 - b) If you use only ROWS and COLUMNS in an MDX query, how can you analyze data in three or more dimensions?
 - c) What is the purpose of the WHERE clause in an MDX query? Give two different examples.
 - 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?