Birla Institute of Technology & Science, Pilani

Work Integrated Learning Programmes Division
M. S (Software Engineering) at Wipro Technologies (WASE)
II Semester 2014 - 2015

Comprehensive Examination (Regular_ ANSWER KEY)

Course Number:	SEWP ZG514
Course Title:	DATA WAREHOUSING
Type of Exam:	Open Book
Weightage:	60%
Duration:	3 Hours
Date of Exam:	23 Aug 2015

Session – FN

No. of Pages: 2
No. of Questions: 12

1. Differentiate Data Warehousing and Data Mining. (2 marks)

   DW enables OLAP by doing various summarization and forecasts. Data mining discovers hidden patterns in data. Data mining operates at a detail level instead of a summary level

2. Differentiate dependent and independent data marts. (2 marks)

   Dependent data marts draw data from a central data warehouse that has already been created.

   Independent data marts, in contrast, are standalone systems built by drawing data directly from operational or external sources of data, or both.

3. Define "Data Staging". At what stage of the Data Warehousing cycle, will staging happen? (2 marks)

   Staging is the "temporary" storage of data extracted from various sources, into an intermediate work area, for carrying out the transformation; and also before loading the transformed data into the Data Storage area of the Data Warehouse.

   At what stage in the DWH cycle will Data Staging happen:

   - Extraction (after)
   - Transformation (during)
   - Load or Transfer (before).

4. Why it is recommended to have a separate time dimension rather than having the date as one of the attributes in the fact table? (3 marks)

   Performance analysis generally involves comparisons of year-to-year, month-to-month etc. DWs always have require time in fact table. There are many date attributes not supported by the SQL date function, including fiscal periods, seasons, holidays, and weekends. Instead of attempting non-standard calendar calculations in a query, it is effective to have date dimension table.

5. Differentiate aggregations done using 'Group By' and Cube statements. (3 marks)

Group By clause causes computation of aggregations across specified dimensions at a particular level of conceptual hierarchy. The CUBE computes aggregations at all combinations of levels of hierarchy.

6. Data warehouse is subject oriented. For following companies what will be critical business subjects? (3 marks)
   a. Manufacturing company
   b. Insurance company
   c. Bank
   d. Retail chain
   e. University
   f. Telecom company

   Ans: 0.5 marks for each subject area identified
   Manufacturing company - Sales, inventory, shipment
   Insurance company – Claims, sales and marketing
   Bank – Loans, fixed deposit.
   Retail chain – Sales, procurement
   University – Registration, Attendance
   Telecom – Billing, customer support calls

7. For following statements, indicate True or False with proper justification: (5 marks)
   A. It is a good practice to drop the indexes before the initial load.
      True, Index can slow down loading
   B. The choice of index type depends on cardinality.
      True, Bit-map index can work well when cardinality is low
   C. The importance of metadata is the same for data warehouse and an operational system.
      False, in case of DW users access information in ad-hoc way. Meta data becomes more important
   D. Backing up the data warehouse is not necessary because you can recover data from the source systems.
      False, Data is loaded into DW after significant processing.
   E. The essential difference between ROLAP and MOLAP is in the way data is stored.
      True, ROLAP stores using relational structure, while MOLAP stores using MDDB.

8. Compare outriggers & mini-dimensions. Give situations where you would use each one with justification. (5 marks)

   Outrigger is characterized by Large number of attributes, Different grain, and Different update frequency. Outriggers are connected to fact table via dimension, i.e. there is no foreign key for outrigger in fact table. Outriggers save space, but introduce more joins and complicate interface for business users.

   Mini-dimension separates out frequently analyzed or frequently changing attributes into a separate dimension. We include foreign keys to both main dimension & mini-dimension in fact table. Mini-dimension typically consists of frequently changing attribute subset within the large multimillion-row dimension.

9. State why, for the integration of multiple heterogeneous information sources, many companies in industry prefer the update-driven approach (which constructs and uses data warehouses), rather than the query-driven approach (which applies wrappers and integrators). Describe situations where the query-driven approach is preferable over the update-driven approach. (5 marks)

For decision-making queries and frequently-asked queries, the update-driven approach is more preferable. This is because expensive data integration and aggregate computation are done before query processing time. In order for the data collected in multiple heterogeneous databases to be used in decision-making processes, data must be integrated and summarized with the semantic heterogeneity problems among multiple databases analyzed and solved. If the query-driven approach is employed, these queries will be translated into multiple (often complex) queries for each individual database. The translated queries will compete for resources with the activities at the local sites, thus degrading their performance. In addition, these queries will generate a complex answer set, which will require further filtering and integration. Thus, the query-driven approach is, in general, inefficient and expensive. The update-driven approach employed in data warehousing is faster and more efficient since most of the queries needed could be done off-line. For queries that are used rarely, reference the most current data, and/or do not require aggregations, the query-driven approach would be preferable over the update-driven approach. In this case, it may not be justifiable for an organization to pay heavy expenses for building and maintaining a data warehouse, if only a small number and/or relatively small size databases are used; or if the queries rely on the current data, since the data warehouses do not contain the most current information.

10. Metadata is the most important information to manage DW. Most DW tasks are aided by metadata. Please identify ten DW tasks for which metadata is crucial. State the relevant metadata for each task. (10 marks)

Ten tasks (& corresponding metadata) can be

How to cleanse data while populating the warehouse? (A cleansing condition e.g. if Pin Code is non-numeric, reject)
How to control runaway queries? (If takes more than 6 hours, stop).
Which summary table to be used for responding to a query? (e.g. Summary by region, period)
How to transform input data? (e.g. Convert state code into state name)
Which summary table to create? (e.g. Summary by product based on query statistics)
Which index to create? (e.g. Index on brand, based on query statistics)
Which partition to be used for a query (Based on query analysis)
When to schedule backup?
Which archive to be used for recovery?
What is Query execution plan (Based on table statistics)

11. Suppose that a data warehouse for Big University consists of the following four dimensions: student, course, semester, and instructor, and two measures count and

avg_grade. When at the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the avg_grade measure stores the actual course grade of the student. At higher conceptual levels, avg_grade stores the average grade for the given combination.

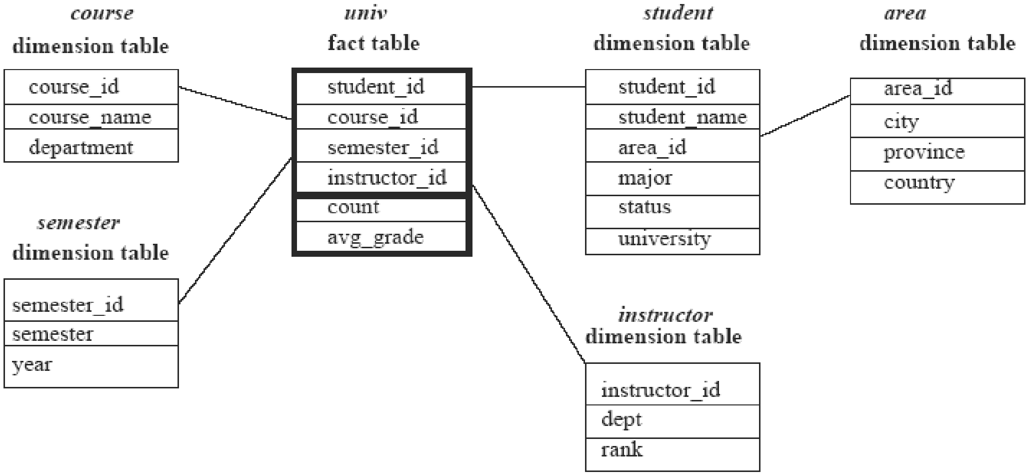
(a) Draw a snowflake schema diagram for the data warehouse.
(b) Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations (e.g., roll-up from semester to year) should one perform in order to list the average grade of CS courses for each Big University student.
(c) If each dimension has five levels (including all), such as "student < major < status < university < all", how many cuboids will this cube contain (including the base and apex cuboids)?

(4+3+3 = 10 Marks)

Solution:

(a). snowflake schema diagram for the data warehouse.

Students should specify all suitable attributes and required keys



(b)

Starting with the base cuboid [student, course, semester, instructor]

1. roll-up on course from (course_key) to department
2. roll-up on student from (student_key) to university
3. Dice on course, student with department ="CS" and university="Big University"
4. Drill-down on student from university to student name

(c) The cube will contain 5^4=625 cuboids.

12. An electronics stores chain is keen on obtaining the following information

I. Sale of products by make, model, location, price range

II. Profitability of various products. Here consider design options for a stable cost price, and volatile cost price scenarios.

a. Design a dimensional model that can help management obtain the answers.

b. Management have noticed that sale of products drop when there are promotional offers from competitors. Incorporate necessary elements in the design so that impact of competitor promotion can be identified. (6+4 marks)

a) The dimensional model can have Sales fact table with location, product, and time dimensions. Analysis by price range can be achieved by having unit price as a fact table measure, or makng it a product attribute. Profitability can be obtained by including cost price in fact table. If cost price is stable, it can be made an attribute in product dimension. A volatile cost price can be handled by including it in fact table, or by creating a starflake dimension that intersects with product and time dimensions.

b) The promotional offers from competitors can be incorporated in time dimension to measure their impact on store sales. Or competitors' promotion can also be incorporated as separate dimension.