Birla Institute of Technology & Science, Pilani

Work Integrated Learning Programmes Division

M.S. Software Engineering at Wipro Technologies (WASE)

Second Semester 2014 - 2015 (April - September 2015)

Mid-Semester Exam (Regular) – Answer Key

Course Number   :   **SEWP ZG514**

Course Title   :   Data Warehousing

Type of Exam   :   Closed Book

Weightage   :   30 %

Duration   :   90 Minutes

No. of Pages : 3
No. of Questions : 5

Date of Exam   :   14th June 2015   Session :   FN

Note:
1. Please read and follow all the instructions given on the cover page of the answer script.
2. Start each answer from a fresh page. All parts of a question should be answered consecutively.

1. Short Questions:-   5 * 2 = 10
    a. How do you characterize Data Warehouse?
    Inmon's definition or equivalent explanation is acceptable.
    b. Compare and contrast Type 2 and Type 3 SCD (slowly changing dimensions)
    In case of type 2 SCD, a new row is created with changed attribute with a new surrogate key and effective date is used. The original row stays unchanged.
    In case of type 3 SCD, a new column is added to the row to store existing value of the attribute and new value is stored in the original column to represent current value.
    c. What is meant by grain of the business process? Explain with examples.
    Grain indicates level of data detail in a dimensional model. The granularity of fact table constrains the level of analysis. A retail store recording daily sale of soaps can analyze product performance, but not which customer prefers which soap.
    d. What is a conformed dimension?
    Conformed dimensions are the dimension tables that have consistency w.r.t. dimension keys, attribute column names, attribute definitions, and attribute values across data marts in an enterprise.
    e. Define initial load, incremental load, and full refresh
    Initial Load : Populating all the data warehouse tables for the very first time
    Incremental Load : Applying ongoing changes periodically
    Full refresh : Completely erase existing data & reload with fresh data

2. Define Star and Snowflake schemas. List relative advantages of each. (2+3 marks)

Star schema organizes the DW tables into facts and corresponding dimensions such that each dimension is a completely denormalized table and each dimension has foreign key relation with the fact table. In case of snowflake, the dimension entity is normalized. (2 marks)

Advantages of Star Schema are :
  Easy for Users to Understand
  Optimizes Navigation
  Most Suitable for Query Processing
  Can implement STARjoin and STARindex for performance enhancement

Advantages of snowflake schema are :
  Small saving in storage space
  Normalized structure are easier to update and maintain.

3. Name five types of major transformation tasks. Give an example for each. (2+3 marks)
List and explain any five of the following:
Format revisions, Decoding of fields, Calculated and derived values, Splitting of single fields, Merging of information, Character conversion, Unit conversion, Date conversion, Summarization, Key restructuring, Deduplication

4. Differentiate top-down and bottom-up approaches for DW development. List relative advantages. (2+3 marks)

Top-down approach looks at big picture of the organization, and build a mammoth enterprise DW
- Advantages are
    - A truly corporate effort, an enterprise view of data
    - Inherently architected-not a union of disparate data marts
    - Single, central storage of data about the content
    - Centralized rules and control
    - May see quick results if implemented with iterations
- Disadvantages are
    - Takes longer to build even with an iterative method
    - High exposure/risk to failure
    - Needs high level of cross-functional skills
    - High outlay without proof of concept

Bottom-up approach looks at individual local and departmental requirements, and builds departmental data marts to evolve into enterprise solution.
- Advantages are
    - Faster and easier implementation of manageable pieces
    - Favorable ROI and POC
    - Less risk of failure
    - Inherently incremental, can schedule important data marts first
    - Allows project team to learn and grow
- Disadvantages are
    - Each data mart has its own narrow view of data
    - Permeates redundant data in every data mart
    - Perpetual inconsistent and irreconcilable data
    - Proliferates unmanageable interfaces

5. A call center offers support to customers who had purchased household items of a large manufacturer of the items. The center is interested in understanding
    a. Which products are requiring maximum support from the customers
    b. Which support staff are responding to most calls
  Design a Data Warehouse that helps the call center management.

The grain for DW is individual calls attended by support staff. The measure can be call duration or simply call count. (1 mark)

The dimensions are product, customer, time, support staff. (2 marks)

A schema diagram with appropriate attributes (2 marks)