**Birla Institute of Technology & Science, Pilani**

**Work-Integrated Learning Programmes Division**

**Second Semester 2013-2014**

**Mid-Semester Test**

**(EC-2 Regular)**

Course No. : SS ZG515

Course Title : DATA WAREHOUSING

Nature of Exam : Closed Book

No. of Pages = 1

# No. of Questions = 6

Weightage : 35%

Duration : 2 Hours

Note:

1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.
4. Briefly explain why do you need a separate data staging component in a data warehouse environment? [3]

Answer: **[3 marks for listing the pointes noted below]**

Data Staging component is critical in DW environment because:

* Facilitate error recycling and prevents bad data entering into DW tables.
* Helps data transformation.
* Data staging area is isolated from DW and Data Mart so that inbound data gets accumulated without any impact on DW reporting

1. Explain at least six characteristics or features of a data warehouse? [6]

Answer: **[1 marks for each of the points listed below with brief explanation ]**

* Subject oriented data
* Integrated data
* Time-Variant data
* Nonvolatile data
* Granular data so that all possible reporting is possible.
* Facilitates multi-dimensional and ad hoc queries/reporting.

1. Briefly explain any three advantages of STAR schema. Can you think of any disadvantages of the STAR schema with respect to reporting? [6]

Answer: **[1.5 marks for each of the advantages listed below. ]**

**Advantages:**

* Easy for users to Understand
  + Simplified business reporting logic
  + Simpler queries
* Optimizes Navigation
  + Query performance gains
  + Fast Aggregations
* Most suitable for Query processing

Disadvantages: **[Award 1.5 marks if any one of the below points are explained.]**

* Data integrity is not enforced as like in Normalized Database.
* One off inserts/updates can result in data anamolies which is better controlled in Normalized DB.

1. Explain the significance of slowly changing dimensions. Explain each type with an example. [7]

Answer:

Significance: **[1 marks]**

Slowly changing Dimensional data changes slowly rather than changing on time-based.

Due to this fact, SCD allows to track historical changes in business and track results against those historical changes.

Explain Type 1, Type 2 and Type 3 with an example. **[2 marks for each SCD type explonation]**

1. Briefly explain initial load, incremental load and full refresh with an example. [5]

Answer: **[1.5 marks for each of the refresh types, 0.5 marks for satisfactory answer]**

* Initial Load: This is the first opportunity to test interface validation and historical data load from source systems. During this load time, we need to control user access so that this work can be completed soon. On an average size DW, this activity might take any ware from 2 – 4 days.
* Incremental Load: This is the data load that pulls delta update from previous data loads. 90% of daily volume will fall into incremental load. Well established ETL process will benefit Incremental load. Delta extract mechanism could be established on Source system or from DW system.
* Full Refresh: Full data refresh is typical for Master/dimensional data. Entire data will be extracted and loaded. This works great on dimensional data.

1. Explain the differences between ETL vs. ELT, which is best suitable for a global data warehouse with users accessing the system from different time zones round the clock. [8]

Answer: **[4 marks each for ETL AND ELT explanation]**

* ETL is extract, transform, load in which transformation takes place on a transformation server using either an “engine” or by generated code.
  + First, load “raw” data into empty tables using RDBMS block slamming utilities.
  + Next, use SQL to transform the “raw” data into a form appropriate to the target tables.
  + Ideally, the SQL is generated using a meta data driven tool rather than hand coding.
  + Finally, use insert-select into the target table for incremental loads or view switching if a full refresh strategy is used.
* ELT is extract, load, transform in which data transformations take place in the relational database on the data warehouse server.
  + ELT Processing obviates the need for a separate transformation server.
  + Assumes that spare capacity exists on DW server to support transformation operations.
  + ELT leverages the build-in scalability and manageability of the parallel RDBMS and HW platform.
  + Must allocate sufficient staging area space to support load of raw data and execution of the transformation SQL.
  + Works well only for batch oriented transforms because SQL is optimized for set processing.

\*\*\*\*\*\*\*\*\*