MIDTERM EXAMINATION
Spring 2010
CS614- Data Warehousing

Question No: 1 (Marks: 1) - Please choose one

The need to synchronize data upon update is called

- ► Data Manipulation
- ► Data Replication
- ► **Data Coherency**
- ► Data Imitation

Question No: 2 (Marks: 1) - Please choose one

Taken jointly, the extract programs or naturally evolving systems formed a spider web, also known as

- ► Distributed Systems Architecture
- ► **Legacy Systems Architecture**
- ► Online Systems Architecture
- ► Intranet Systems Architecture

Question No: 3 (Marks: 1) - Please choose one

For good decision making, data should be integrated across the organization to cross the LoB (Line of Business). This is to give the total view of organization from:

- ► Owner's Perspective
- ► Customer's Perspective
- ► **Decision Maker's Perspective**
- ► Employee's Perspective

Question No: 4 (Marks: 1) - Please choose one

Node of a B-Tree is stored in memory block and traversing a B-Tree involves _____ page faults.

- ► O (n)
- ► O (n^2)
- ► O (n lg n)
- ► **O (lg n)**

Question No: 5 (Marks: 1) - Please choose one

Which statement is true for De-Normalization?

- ► Redundant data is a performance liability at query time, but is a performance benefit at update time.
- ► Redundant data is a performance benefit at both query time and update time.
- ► Redundant data is a performance liability at both query time and update time.
- ► **Redundant data is a performance benefit at query time, but is a performance liability at update time.**

Question No: 6 (Marks: 1) - Please choose one

Pre-join technique is used to avoid

- ► **Run time join**
- ► Compile time join
- ► Load time join

Question No: 7 (Marks: 1) - Please choose one

Cube is a _____ entity containing values of a certain fact at a certain aggregation level at an intersection of a combination of dimensions.

- ► **Logical**
- ► Physical
- ► Analytical
- ► **None of these**

Question No: 8 (Marks: 1) - Please choose one

The goal of star schema design is to simplify _____

- ► Logical data model
- ► **Physical data model**
- ► Conceptual data model
- ► None of these

Question No: 9 (Marks: 1) - Please choose one

Grain is the _____ level of data stored in the warehouse.

- ► **Atomic**
- ► Summarized
- ► Aggregated
- ► Cube

Question No: 10 (Marks: 1) - Please choose one

Transactional fact tables do not have records for events that do not occur. These are called

- ► **Not Recording Facts**
- ► Fact-less Facts
- ► Null Facts
- ► None of these

Question No: 11 (Marks: 1) - Please choose one

A _____ dimension is a collection of random transactional codes, flags and/text attributes that are unrelated to any particular dimension. The _____ dimension is simply a structure that provides a convenient place to store the _____ attributes.

- ► **Junk**
- ► Time
- ► Parallel
- ► None of these

Question No: 12 (Marks: 1) - Please choose one

During ETL process of an organization, suppose you have data which can be transformed using any of the transformation method. Which of the following strategy will be your choice for least complexity?

- ► **One-to-One Scalar Transformation**
- ► One-to-Many Element Transformation
- ► Many-to-Many Element Transformation
- ► Many-to-One Element Transformation

Question No: 13 (Marks: 1) - Please choose one

Change Data Capture is one of the challenging technical issues in _____

- ► **Data Extraction**
- ► Data Loading
- ► Data Transformation
- ► Data Cleansing

Question No: 14 (Marks: 1) - Please choose one

Rearranging the grouping of source data, delivering it to the destination database, and ensuring the quality of data are crucial to the process of loading the data warehouse. Data _____ is vitally important to the overall health of a warehouse project.

1. **Cleansing**
2. Cleaning
3. Scrubbing

Which of the following options is true?

- ► **Option 1 only**
- ► Option 2 only
- ► Option 1 & 2 only
- ► Option 1, 2 & 3

Question No: 15 (Marks: 1) - Please choose one

When performing objective assessments, companies follow a set of principles to develop metrics specific to their needs, there is hard to have "one size fits all" approach. Which of the following statement represents the pervasive functional forms?

- ► **Simple Ratio, Min or Max Operation, Weighted Average**
- ► Only Complex Ratio, Min Operation, Max Operation
- ► Only Simple Ratio, Min or Max Operation
- ► Only Min or Max Operation, Weighted Average

Question No: 16 (Marks: 1) - Please choose one

The input to the data warehouse can come from OLTP or transactional system but not from other third party database.

- ► True
- ► **False**

Question No: 17 (Marks: 1) - Please choose one

Normalization effects performance

- ► True
- ► **False**

Question No: 18 (Marks: 1) - Please choose one

Collapsing tables can be done on the _____ relationships

- ► One-to-One
- ► Many-to-Many
- ► **Both One-to-One and Many-to-Many**
- ► None of these

Question No: 19 (Marks: 1) - Please choose one

_____ breaks a table into multiple tables based upon common column values.

- ► **Horizontal splitting**
- ► Vertical splitting

Question No: 20 (Marks: 1) - Please choose one

If w is the window size and n is the size of data set, then the complexity of merging phase in BSN method is_____

- ► O (n)
- ► O (w)
- ► **O (w n)**
- ► O (w log n)

Question No: 21 (Marks: 2)

Briefly describe snowflake schema.

1. A **Snowflake Schema** is a logical arrangement of tables in a multidimensional database such that the entity relationship diagram resembles a snowflake in shape. The snowflake schema is represented by centralized fact tables which are connected to multiple dimensions. **OR**
2. The **Snowflake Schema** is an extension of the star schema, where each point of the star explodes into more points. In a star schema, each dimension is represented by a single dimensional table, whereas in a snowflake schema, that dimensional table is normalized into multiple lookup tables, each representing a level in the dimensional hierarchy.

Question No: 22 (Marks: 2)

Why both aggregation and summarization are required?

1. **Summarization** is the addition of like values along one or more business dimensions. An example of summarization is adding up detail revenue values by day to arrive at weekly totals.
2. **Aggregation** refers to a summarization coupled with a calculation across different business elements. An example of aggregation is the addition of bimonthly salary to monthly commission and bonus to arrive at monthly employee compensation values.

3. Summarization with calculation across business dimension is aggregation. Example Monthly compensation = monthly sale + bonus

Question No: 23 (Marks: 3)

Under what condition smart tools work properly to construct a less detailed aggregate from more detailed aggregate?

1. Smart tools will allow less detailed aggregates to be constructed from more detailed aggregates (full aggregate awareness) at run-time so that we do not go all the way down to the detail for every aggregation.
2. However, for this to work, the metrics must be additive (e.g., no ratios, averages, etc.).
3. More detailed pre-aggregates are larger, but can also be used to build less detailed aggregates on-the-go.

Question No: 24 (Marks: 3)

What is web scraping? Give some of its uses.

1. **Web Scrapping** is a process of applying screen scrapping techniques to the clean the junk information from a web page. Some of its **uses are**:-
   a. Building contact lists
   b. Extracting product catalogs
   c. Aggregating real-estate info
   d. Automating search Ad listings
   e. Clipping news articles etc.

Question No: 25 (Marks: 5)

After completing the transformation task, data loading activity is started. How many types of data loading strategies are and when each type of strategy is adopted? Explain.

1. Once we have transformed data then there are three primary loading strategies, which are as under:-
   a. **Full Data Refresh** with BLOCK INSERT or 'block slamming' into empty table.
   b. **Incremental Data Refresh** with BLOCK INSERT or 'block slamming' into existing (populated) tables.
   c. **Trickle/Continuous Feed** with constant data collection and loading using row level insert and update operations.

Question No: 26 (Marks: 5)

What are the drawbacks of MOLAP? Also explain the curse of Dimensionality?

1. **Drawbacks of MOLAP**
   a. Long load time (pre-calculating the cube may take days!).
   b. Very sparse cube (wastage of space) for high cardinality (sometimes in small hundreds). e.g. number of heaters sold in Jacobabad or Sibi.
2. **Curse of Dimensionality**. When the number of dimensions increases, the number of possible aggregates increases exponentially this is called "curse of dimensionality"