MIDTERM EXAMINATION
Spring 2010
CS614- Data Warehousing (Session - 6)

Ref No: 1368137
Time: 60 min
Marks: 40

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 (but not sure)**
- ► 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 (but not sure)**
- ► 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.

Ans:

Snowflake Schema: snowflaking is a method of normalizing the dimension tables in star schema. When we completely normalize all the dimension tables, then the resultent structure resemble a snowflakewith the fact table in the middle.

****

Snowflake Schema: Sometimes a pure star schema might suffer performance problems. This can occur when a de-normalized dimension table becomes very large and penalizes the star join operation. Conversely, sometimes a small outer-level dimension table does not incur a significant join cost because it can be permanently stored in a memory buffer. Furthermore, because a star structure exists at the center of a snowflake, an efficient star join can be used to satisfy part of a query. Finally, some queries will not access data from outer-level dimension tables. These queries effectively execute against a star schema that contains smaller dimension tables. Therefore, under some circumstances, a snowflake schema is more efficient than a star schema.

Question No: 22 (Marks: 2)

Why both aggregation and summarization are required?

Although summarization and aggregation are
sometimes used interchangeably

Summarization and aggregationare typically used for the following reasons:
They are required when the lowest level of detail stored in the data warehouse is at a higher level than the detail arriving from the source. This situation occurs when data warehouse queries do not require the lowest level of detail or sometimes when sufficient disk space is not available to store all the data for the time frame required by the data warehouse.
• They can be used to populate data marts from the data warehouse where the data mart does not require the same level of detail as is stored in the warehouse.
• They can be used to roll up detail values when the detail is removed from the warehouse because it is

Question No: 23 (Marks: 3)

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

Ans:
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. However, for this to work, the metrics must be additive (e.g., no ratios, averages, etc.). 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 scrapping? Give some of its uses.

Web scrapping is a process of applying screen scrapping techniques to the web. There are several web scrapping products in the market and target business users who want to creatively use the data, not write complex scripts. Some of the uses of scrapping are:
- **Building contact lists**
- **Extracting product catalogs**
- **Aggregating real-estate info**
- **Automating search Ad listings**
- **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.

Significance of Data Loading Strategies

- **Need to look at:**
  - Data freshness
  - System performance
  - Data volatility
- **Data Freshness**
  - Very fresh low update efficiency
  - Historical data, high update efficiency
  - Always trade-offs in the light of goals
- **System performance**
  - Availability of staging table space
  - Impact on query workload
- **Data Volatility**
  - Ratio of new to historical data
  - High percentages of data change (batch update)

Question No: 26 (Marks: 5)

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

Drawbacks of MOLAP:

☐ Long load time (pre-calculating the cubemay take days!).

☐ Very sparse cube (wastage of space) for highcardinality (sometimes in small hundreds).e.g. number of heaters sold in Jacobabad or Sibi.

AMOLAP is in no way a win-win situation, it has its won intrinsic pitfalls,which does notmake it an overall winner. The biggest drawback is the extremely long timetaken to
pre-calculatethe cubes, remember that in a MOLAP all possible aggregates are computed.The number of aggregates suffers from something called as the "curse of dimensionality"i.e. as the number of dimensions increases, the number of possible aggregatesincreases exponentially, this will befurther clarified in an upcoming slide. Becauseof long calculation times, cube creation is a batch process and usually has the
lowestpriority and scheduled to be done late in the night. This actually turns out tobe a
bigmistake (as we will discuss subsequently) as it may so happen that the cube generationmay not actually take place, and thedecision makers are presented with the
oldand stale data, and they tend to lose faith in the OLAP paradigm. Although the numberof possible aggregates is very large, but NOT all the aggregates may havevalues,
therecan be and will be quite a few aggregates which will have null values. Forexample,
manyof the items sold in winter are not sold in summer and not even kept in thestore

(andvice-a-versa). Consequently, there are no corresponding sales, and if the cubeis generatedthat includes all the items, there will be many null aggregates, resulting in a
verysparse cube. This will result in requirement of large amount of memory, most of whichwould be wasted. For these very reasons cube compression is a hot area ofresearch
anddevelopment. We will not discuss it any further.

673dec papers

DWH is a

- Unstructured and Heterogeneous
- **Structured and Heterogeneous**
- Unstructured and Homogenous
- Structured and Homogenous

ER is a logical design technique that seeks to remove the redundancy in data. P#98

Is also the understanding gained through experience or study

Knowledge

Companies collect and record their own operational data, but at the same time they also
use reference data obtained from ..... sources such as codes, prices etc.

1) external corect p(21)

2) operational

3) internal

As the system moves from stage-1 to stage-5 it becomes what is called as an .... data warehouse

1) active corect hai p#18

Primary key select ....

Vertical
Horizontally (shayd yai shai hai)

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 correct

Storage issue: As dimensions get less detailed (e.g., year vs. day) cubes get much
smaller, but storage consequences for building hundreds of cubes can be
significant. Lot of space.

After Modifictaion of data in the form of
1) 1N
2) 2N
3) 3N
4) 4N

Syntactically Dirty Data is
Lexical Errors
Irregularities
Integrity Constraint Violation
Business rule contradiction

Which is correct
1 and 2 **correct** P#160

2 and 3
2 3 4
4 1

Semantically Dirty Data options thin k kon sa correct hai
Integrity Constraint Violation
Business rule contradiction
Duplication
Lexical Errors
Which is correct
1 and 2 and 3 correct P#160
2 and 3
2 3 4
4 1

How data is extracted From legacy system in data? **3(marks)**

Solution:

Taken jointly, the extract programs or naturally evolving systems formed a spider web, also called "legacy systems" architecture.

HOLAP Featuers?(3marks thy shayd)page#78
Solution
HOLAP: OLAP implemented as a hybrid of MOLAP and ROLAP.
HOLAP provides a combination of relational database access and "cube" data structures
within a single framework. The goal is to get the best of both MOLAP and ROLAP:
scalability (via relational structures) and high performance (via pre-built cubes).

Some use of web scraping?(3maks) p#146

solution
Building contact lists
Extracting product catalogs
Aggregating real-estate info
Automating search Ad listings
Clipping news articles etc.

Simple many-to-many element transformations. 3marks yai mujy tu nahi mila kahin sy yai hai jo paste kar rahi houn
solution
The most complex is many-to-many element transformations. Good examples are house
holding and individualization. This is achieved by using candidate keys and fuzzy
Matching to determine which individuals are the same individuals, and which individuals
go in the same household and so on. This is a very complex transformation and will be
discussed in BSN lecture.

Diff between ER and DM? 5marks p#102
solution
ER vs. DM

ER
Constituted to optimize OLTP performance
Models the micro relationships among data elements
A wild variability of the structure of ER models.
Very vulnerable to changes in the user's querying habits, because such schemas are asymmetrical.

DM
Constituted to optimize DSS query performance.
Models the macro relationships among data elements with an overall deterministic strategy
All dimensions serve as equal entry points to the fact

table
Changes in user querying habits can be catered by automatic QL generators.

Explain Orr's Laws of Data Quality ?5marks(p#181)
solution
Law 1: "Data that is not used cannot be correct!"
Law 2: "Data quality is a function of its use, not its collection!"

Law 3: "Data will be no better than its most stringent use!"

Law 4: "Data quality problems increase with the age of the system!"

Law 5: "The less likely something is to occur, the more traumatic it will be
when it happens!"