END TERM Examination(Model Test Paper)

Sixth Semester[B.Tech]

Paper Code: ETCS - 310	**Subject: DWDM**
Time: 3 hrs	**Maximum Marks: 75**
Note: Q.No.1 is compulsory. Attempt any four questions of remaining	

Q1. Answer the following questions in brief . :- [2.5 * 10]

a.) What do you mean by enterprise data warehouse?

Ans. An **enterprise data warehouse** is a unified database that holds all the business information an organization and makes it accessible all across the company.(**Explain below mentioned features in brief**)

- Unified approach for organizing and representing data
- Ability to classify data according to subject
- Give access according to divisions (sales, finance, inventory and so on)
- Normalized design
- Robust infrastructure with contingency plans
- High level of security Scalability

b.) What are the requirements of cluster analysis ?

Ans. Requirements of cluster analysis in data mining are :

- Scalability
- Ability to deal with different kind of attributes
- High dimensionality
- Ability to deal with noisy data
- Interpretability

c.) What are junk dimensions. Explain?

Ans. When developing a dimensional model, we often encounter miscellaneous flags and indicators. These flags do not logically belong to the core dimension tables. A **junk dimension** is grouping of low cardinality flags and indicators. This junk dimension helps in avoiding cluttered design of data warehouse. Provides an easy way to access the dimensions from a single point of entry and improves the performance of sql queries.

(**Explain with suitable example**)

d.) How are users of data warehouse classified?

Ans. The following are the approaches that can be used to classify the users.

- The users can be classified as per the hierarchy of users in an organisation i.e. users can be classified by department, section, group, and so on.
- The user can also be classified according to their role, with people grouped across departments based on their role.

e.) Explain materialized view of data warehouse ?

Ans.

1. Materialized view is normal database object like "table,index"
2. It is basically use for Data warehouse or Replication purpose.
3. Snapshot is synonym for materialized view.
4. A materialized view can be stored in the same database as its base tables or in a different database
5. A materialized view provides access to table data by storing the results of a query in a separate schema object. Unlike an ordinary view, which does not take up any storage space or contain any data, a materialized view contains the rows resulting from a query against one or more base tables or views.
6. A materialized view improve response time through query rewrite or reduce execution time.

f.) Differentiate between OLAP and OLTP ?

Ans. Online transaction processing, or OLTP, refers to a class of systems that facilitate and manage transaction-oriented applications, typically for data entry and retrieval transaction processing.

- Current data
- Changeability: Frequent data changes
- Priority: High availability, High data volume
- Database Operation: Online update/insert/delete and read
- Normalization is very high
- Data Structure: Relational (flat tables)
- Integration: Minimal
- Data Set: 6-18 months

OLAP stands for On Line Analytical Processing, a series of protocols used mainly for business reporting. Using OLAP, businesses can analyze data in all manner of different ways, including budgeting, planning, simulation, data warehouse reporting, and trend analysis

- Current and historical data
- Changeability: Data frozen
- Priority: Simple to use, flexible access to data
- Database Operation: Read
- Less Normalization due to data staging and less performance
- Data Structure: Multi Dimensional format
- Integration: Comprehensive
- Data Set: 2-7 years

g.) What type of processing takes place in a data warehouse ?

Ans. There are at least four levels of analytical processing requirements:
1. Running of simple queries and reports against current and historical data.
2. Ability to perform “what if” analysis in many different ways.
3. Ability to query, step back, analyze, and then continue the process to any desired
length.

4. Ability to spot historical trends and apply them in future interactive processes.

h.) Differentiate between DSS and OLAP ?

Ans. DSS, Decision Support System, as the name suggests, helps in taking decisions for top executive professionals. Data accessing, time-series data manipulation of an enterprise's internal / sometimes external data is emphasized by DSS. The manipulation is done by tailor made tools that are task specific and operators and general tools for providing additional functionality.

OLAP, Online Analysis Processing, is capable of providing highest level of functionality and support for decision which is linked for analyzing large collections of historical data. The functionality of an OLAP tool is purely based on the existing / current data.

i.) What is ETL process ?

Ans. Extraction, transform, and loading (ETL) refers to a process in database usage and especially in data warehousing that:

- Extracts data from outside sources.
- Transforms it to fit operational needs, which can include quality levels
- Loads it into the end target (database, more specifically, operational data store, data store, or data warehouse).

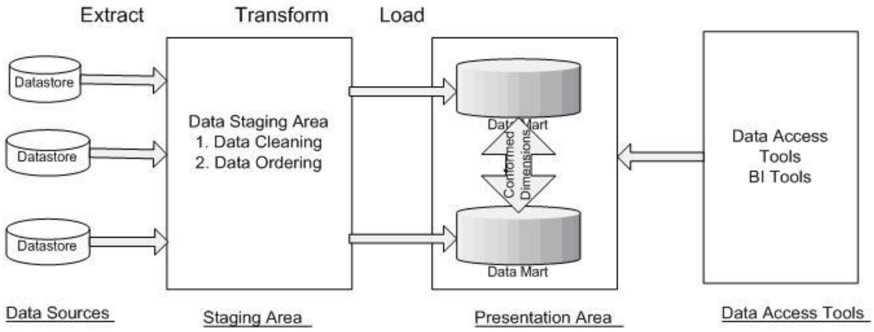
ETL systems are commonly used to integrate data from multiple applications, typically developed and supported by different vendors or hosted on separate computer hardware. The disparate systems containing the original data are frequently managed and operated by different employees.

j.) What are star index and star join ?

Ans. STARjoin is a high-speed, single-pass, parallelizable, multitable join. It can join more than two tables in a single operation. This special scheme boosts query performance.
STARindex is a specialized index to accelerate join performance. These are indexes created on one or more foreign keys of the fact table. These indexes speed up joins between the dimension tables and the fact table.

Q2 a.) What are different components of data warehouse. Explain with a neat diagram? [5]

Ans. Following diagram depicts different components of Data Warehouse architecture:-



(Explain each component in detail)

- Operational Source System
- Data Staging Area
- Data Presentation Area
- Data Access Tools

Q2 b.) What is online transaction processing(OLTP) ? Describe the evaluation of OLTP. What are the critical features of OLTP systems ? [7.5]

Ans. Online transaction processing, or **OLTP**, is a class of information systems that facilitate and manage transaction-oriented applications, typically for data entry and retrieval transaction processing.

The following issues are important for evaluating an OLTP system:

- rollback segments
- indexes, clusters, and hashing
- discrete transactions
- data block size
- dynamic allocation of space to tables and rollback segments
- transaction processing monitors and the multithreaded server
- the shared pool
- well-tuned SQL statements

- integrity constraints
- client/server architecture
- dynamically changeable initialization parameters
- procedures, packages, and functions

Critical features of OLTP systems are:-

- An OLTP system is customer-oriented and is used for transaction and query processing by clerks, clients, and information technology professionals.
- An OLTP system manages current data that, typically, are too detailed to be easily used for decision making.
- An OLTP system usually adopts an entity-relationship (ER) data model and an application-oriented database design.
- The access patterns of an OLTP system consist mainly of short, atomic transactions.

Q3 a.) List out the reasons why traditional method of analysis provided in the data warehouse are not sufficient .? [5]

Ans. Explain following given points in detail.

- No provision for aggregate navigation
- Restrictions on presentation and alternation of resulted report sets.
- No support for multidimensional data
- SQL is ill-suited for analyzing data and exploring relationships
- OLTP systems do not provide much analysis data

Q3 b.) What are the various types of metadata. Explain in detail ? [3]

Ans. The metadata can be broadly categorized into three categories:

- **Business Metadata** - This metadata has the data ownership information, business definition and changing policies.
- **Technical Metadata** - Technical metadata includes database system names, table and column names and sizes, data types and allowed values. Technical metadata also includes structural information such as primary and foreign key attributes and indices.

- **Operational Metadata** - This metadata includes currency of data and data lineage. Currency of data means whether data is active, archived or purged. Lineage of data means history of data migrated and transformation applied on it.

Q3 c.) What are the possible aggregates in which fact tables may be formed.? [4.5]

Ans. Explain each with suitable example

- **One-Way Aggregates** When you rise to higher levels in the hierarchy of one dimension and keep the level at the lowest in the other dimensions, you create one-way aggregate tables.
- **Two-Way Aggregates** When you rise to higher levels in the hierarchies of two dimensions and keep the level at the lowest in the other dimension, you create two-way aggregate
- **Three-Way Aggregates** When you rise to higher levels in the hierarchies of all the three dimensions, you create three-way aggregate tables.

Q4 a.) Explain the following :-

(i) MOLAP (ii)ROLAP (iii) HOLAP [7.5]

Ans. In the OLAP world, there are mainly two different types: Multidimensional OLAP (MOLAP) and Relational OLAP (ROLAP). Hybrid OLAP (HOLAP) refers to technologies that combine MOLAP and ROLAP.

MOLAP

This is the more traditional way of OLAP analysis. In MOLAP, data is stored in a multidimensional cube. The storage is not in the relational database, but in proprietary formats.

Advantages:

- **Excellent performance:** MOLAP cubes are built for fast data retrieval, and are optimal for slicing and dicing operations.
- **Can perform complex calculations:** All calculations have been pre-generated when the cube is created. Hence, complex calculations are not only doable, but they return quickly.

Disadvantages:

- **Limited in the amount of data it can handle:** Because all calculations are performed when the cube is built, it is not possible to include a large amount of data in the cube itself. This is not to say that the data in the cube cannot be derived from a large amount of data. Indeed, this is possible. But in this case, only summary-level information will be included in the cube itself.
- **Requires additional investment:** Cube technology are often proprietary and do not already exist in the organization. Therefore, to adopt MOLAP technology, chances are additional investments in human and capital resources are needed.

ROLAP

This methodology relies on manipulating the data stored in the relational database to give the appearance of traditional OLAP's slicing and dicing functionality. In essence, each action of slicing and dicing is equivalent to adding a "WHERE" clause in the SQL statement.

Advantages:

- **Can handle large amounts of data:** The data size limitation of ROLAP technology is the limitation on data size of the underlying relational database. In other words, ROLAP itself places no limitation on data amount.
- **Can leverage functionalities inherent in the relational database:** Often, relational database already comes with a host of functionalities. ROLAP technologies, since they sit on top of the relational database, can therefore leverage these functionalities.

Disadvantages:

- **Performance can be slow:** Because each ROLAP report is essentially a SQL query (or multiple SQL queries) in the relational database, the query time can be long if the underlying data size is large.
- **Limited by SQL functionalities:** Because ROLAP technology mainly relies on generating SQL statements to query the relational database, and SQL statements do not fit all needs (for example, it is difficult to perform complex calculations using SQL), ROLAP technologies are therefore traditionally limited by what SQL can do.

ROLAP vendors have mitigated this risk by building into the tool out-of-the-box complex functions as well as the ability to allow users to define their own functions.

HOLAP

HOLAP technologies attempt to combine the advantages of MOLAP and ROLAP. For summary-type information, HOLAP leverages cube technology for faster performance. When detail information is needed, HOLAP can "drill through" from the cube into the underlying relational data.

Q 4 b.) What do you mean by knowledge discovery process ? [5]

Ans. List of steps involved in knowledge discovery process:

- **Data Cleaning** - In this step the noise and inconsistent data is removed.
- **Data Integration** - In this step multiple data sources are combined.
- **Data Selection** - In this step relevant to the analysis task are retrieved from the database.
- **Data Transformation** - In this step data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- **Data Mining** - In this step intelligent methods are applied in order to extract data patterns.
- **Pattern Evaluation** - In this step, data patterns are evaluated.
- **Knowledge Presentation** - In this step, knowledge is represented.

(Explain KDD process with architecture diagram)

Q5 a.) Explain top down and bottom up approaches for building a data warehouse? Describe the merits and demerits of both these approaches. ? [7.5]

Ans. Top-Down Approach

In this approach the data in the data warehouse is stored at the lowest level of granularity based on a normalized data model. The centralized data warehouse would feed the dependent data marts that may be designed based on a dimensional data model.

Advantages:

- 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:

- Takes longer to build even with an iterative method
- High exposure to risk of failure
- Needs high level of cross-functional skills
- High outlay without proof of concept

Bottom-Up Approach

In this approach data marts are created first to provide analytical and reporting capabilities for specific business subjects based on the dimensional data model. Data marts contain data at the lowest level of granularity and also as summaries depending on the needs for analysis. These data marts are joined or “unioned” together by conforming the dimensions.

Advantages:

- Faster and easier implementation of manageable pieces
- Favourable return on investment and proof of concept
- Less risk of failure
- Inherently incremental; can schedule important data marts first
- Allows project team to learn and grow

Disadvantages:

- Each data mart has its own narrow view of data
- Permeates redundant data in every data mart

- Perpetuates inconsistent and irreconcilable data
- Proliferates unmanageable interfaces

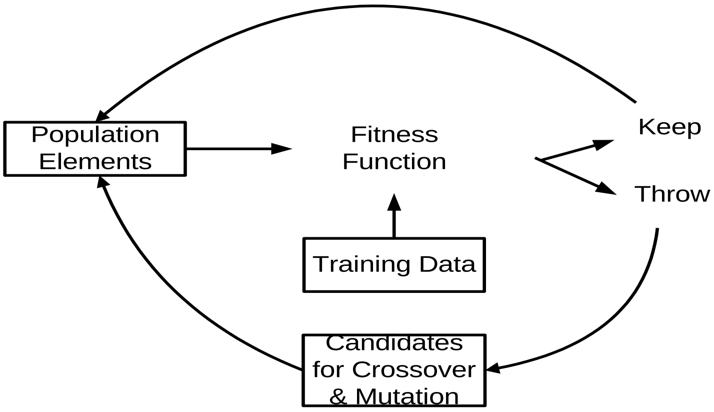
Q 5 b.) Describe slowly changing dimensions .? [5]

Ans . The Slowly Changing Dimension problem is a common one particular to data warehousing. There are in general three ways to solve this type of problem, and they are categorized as follows:

- Type 1 Changes: Correction of Errors
- Type 2 Changes: Preservation of History
- Type 3 Changes: Tentative Soft Revisions

Q6. What are the basic principles of genetic algorithms? Give an example. Use the example to describe how this technique works.? [12.5]

Ans. Genetic Algorithms (GA) apply an evolutionary approach to inductive learning. GA has been successfully applied to problems that are difficult to solve using conventional techniques such as scheduling problems, traveling salesperson problem, network routing problems and financial marketing.



- Step 1: Initialize a population P of n elements as a potential solution.
- Step 2: Until a specified termination condition is satisfied:
  - Use a fitness function to evaluate each element of the current solution. If an element passes the fitness criteria, it remains in P.

– The population now contains m elements (m <= n). Use genetic operators to create (n – m) new elements. Add the new elements to the population.

Crossover: The elements most often used for crossover are those destined to be eliminated from the population. Crossover forms new elements for the population by combining parts of two elements currently in the population

Mutation: Mutation is sparingly applied to elements chosen for elimination. Mutation can be applied by randomly flipping bits (or attribute values) within a single element.

Selection: Selection is to replace to-be-deleted elements by copies of elements that pass the fitness test with high scores. With selection, the overall fitness of the population is guaranteed to increase.

Q7 a.) Give Dr.E.F Codd's 12 guidelines for OLAP ? [7.5]

Ans. 12 rules (according to Codd) which should satisfy the OLAP software are:

- Multi-Dimensional Conceptual View
- Transparency
- Accessibility
- Consistent Reporting Performance
- Client-Server Architecture
- Generic Dimensionality
- Dynamic Sparse Matrix Handling
- Multi-User Support
- Unrestricted Cross-dimensional operations
- Intuitive Data Manipulation
- Flexible Reporting
- Unlimited Dimensions and Aggregation Levels

Q 7 b.) Explain snapshot and periodic fact tables with the help of an example. What is the difference between verification and discovery ? [5]

Ans. **Transactional fact table:** Transaction fact table is one that holds data at the grain of per transaction. e.g. If a customer buys 3 different products from a point of sale. then fact table will have three records for each transaction indicating 3 different type of product sale. Basically if three transactions are displayed on customer receipt then we have to store three records in fact table as granularity of fact table is at transaction level.

e.g. Customer Bank transaction

Customer	Transaction Type	Amount	Date
Customer1	Credit	10000	01-01-2012
Customer1	Debit	5000	02-01-2012
Customer1	Credit	1000	03-01-2012

Periodic snapshot fact table: As its name suggests periodic snapshot fact table are used to store a snapshot of data taken at particular point of time. Periodic fact tables stores one row for particular period of time.

e.g. let’s take an example of credit/debit transaction made by a customer.

Customer	Transaction Type	Amount	Date
Customer1	Credit	10000	01-01-2012
Customer1	Debit	5000	02-01-2012
Customer1	Credit	1000	03-01-2012

Above table is a transaction fact table suppose we need to create a periodic fact table whose grain is month which stores customer balance at the end of month then it should look like as below

Customer	Month	Amount
Customer1	Jan-2012	6000

Verification approach to data analysis is driven by a hypothesis or conjecture about some relationship. The analyst then forms a query to formulate the hypothesis. The resulting report will confirm or disconfirm the theory. OLAP,DSS , SQL based systems use verification approach.

Data mining uses the **discovery** approach. It sifts through the data in search of frequently occurring patterns and trends to report generalizations about the data. Verification is conventional mining whereas discovery is just like open pit gold mining.

Q8. Define the following:- [4.5 + 4 + 4]

a.) Drill down and roll up

Ans. **Roll-up:** This operation performs aggregation on a data cube in any of the following way:

- By climbing up a concept hierarchy for a dimension
- By dimension reduction. Drill-down

Drill-down operation is reverse of the roll-up. This operation is performed by either of the following way:

- By stepping down a concept hierarchy for a dimension.
- By introducing new dimension.

b.) Bayesian Classification

Ans. The classification problem may be formalized using a-posteriori probabilities:

- P(C|X) = prob. that the sample tuple
- X=<x_1,…,x_k> is of class C.
- E.g. P(class=N | outlook=sunny,windy=true,…)
- Idea: assign to sample X the class label C such that P(C|X) is maximal

Bayes theorem:

$$P(C|X) = P(X|C) \cdot P(C) / P(X)$$

- P(X) is constant for all classes
- P(C) = relative freq of class C samples
- C such that P(C|X) is maximum =
  C such that P(X|C)·P(C) is maximum

- Problem: computing P(X|C) is unfeasible!

c.) Data Staging

A **staging area**, or **landing zone**, is an intermediate storage area used for data processing during the extraction, transformation and loading(ETL) process. The data staging area sits between the data source(s) and the data target(s), which are often data warehouses data marts or other data repositories.

Data staging areas are often transient in nature, with their contents being erased prior to running an ETL process or immediately following successful completion of an ETL process. There are staging area architectures, however, which are designed to hold data for extended periods of time for archival or troubleshooting purposes.

Staging areas can be implemented in the form of tables in relational databases, text-based flat files (or XML files) stored in file systems or proprietary formatted binary files stored in file systems. Staging area architectures range in complexity from a set of simple relational tables in a target database to self-contained database instances or file systems. Though the source systems and target systems supported by ETL processes are often relational databases, the staging areas that sit between data sources and targets need not also be relational databases