(Please write your Roll No. Immediately) Roll. No----------

Sample Paper for End Term Examination

B.Tech- VI Semester **Feb., 2014**
Paper Code: ETCS 310 **Subject: DWDM**
Time: 3 Hours **Max. Marks: 75**

Note: Question No. 1 is compulsory. Attempt any four Questions from rest.

Q1. Answer the following questions:

a. What are Hypercube? (2)
b. Differentiate between OLAP and Data Warehouse (4)
c. Data Warehouse is an environment and not a product. Explain. (3)
d. What are junk dimensions? (3)
e. What do you mean by Slicing and Dicing? Give examples. (4)
f. Every data Structure in Data warehouse contains time element. Explain why? (3)
g. What is Pivoting? Give an example? (4)
h. What is significance of metadata in data warehouse? (2)

Q2.

a. What are various ways by which a corporate can provide strategic information for Decision making? Explain each in detail. (7)
b. Explain snapshot and transaction fact tables. How are they related? Give example. (5.5)

Q3. Describe slowly changing dimensions? Explain type each in detail. (12.5)

Q4. What are the basic principles of neural networks? Give an example and use this to show How technique works. (12.5)

Q5. A. What is metadata? Explain. (3)

B. What are data marts? How are they different from traditional data warehouse? (2.5)

C. What is multidimensional data model? Explain. (7.5)

Q6. A.What are the features of snowflake schema? What is fact constellation? (7.5)

B. What are necessary requirements of clustering in data mining? (5)

Q7. A. Explain five characteristic features of Data warehouse in detail. (5.5)

B. Explain Knowledge discovery process in detail. (7)

SOLUTIONS

Q1

a. In geometry, a hypercube is an n-dimensional analogue of a square (n = 2) and a cube (n = 3).
Hypercube - data cube of dimension > 3, helps in analyzing data in multiple dimensions. Operations like slice dice pivoting can be performed.

b. Data Warehouse

Data from different data sources is stored in a relational database for end use analysis.
Data organization is in the form of summarized, aggregated, non volatile and subject oriented patterns.
Supports the analysis of data but does not support data of online analysis.

Online Analytical Processing

is characterized by relatively low volume of transactions. Queries are often very complex and involve aggregations. For OLAP systems a response time is an effectiveness measure. OLAP applications are widely used by Data Mining techniques. In OLAP database there is aggregated, historical data, stored in multi-dimensional schemas (usually star schema).

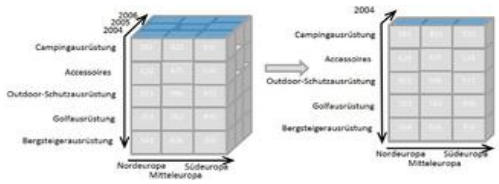
c. A data warehouse is not a single software or hardware product you purchase to provide strategic information. It is, rather, a computing environment where users can find strategic information, an environment where users are put directly in touch with the data they need to make better decisions. It is a user-centric environment.

Let us summarize the characteristics of this new computing environment called the data warehouse:

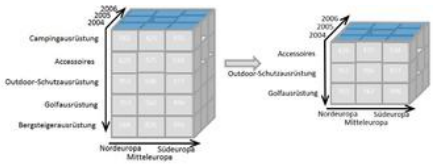
- An ideal environment for data analysis and decision support
- Fluid, flexible, and interactive
- 100 percent user-driven
- Very responsive and conducive to the ask–answer–ask–again pattern
- Provides the ability to discover answers to complex, unpredictable questions

d. A junk dimension is a convenient grouping of typically low-cardinality flags and indicators. By creating an abstract dimension, these flags and indicators are removed from the fact table while placing them into a useful dimensional framework. A Junk Dimension is a dimension table consisting of attributes that do not belong in the fact table or in any of the existing dimension tables. The nature of these attributes is usually text or various flags, e.g. non-generic comments or just simple yes/no or true/false indicators. These kinds of attributes are typically remaining when all the obvious dimensions in the business process have been identified and thus the designer is faced with the challenge of where to put these attributes that do not belong in the other dimensions.

e. *Slice* is the act of picking a rectangular subset of a cube by choosing a single value for one of its dimensions, creating a new cube with one fewer dimension. picture shows a slicing operation: The sales figures of all sales regions and all product categories of the company in the year 2004 are "sliced" out of the data cube

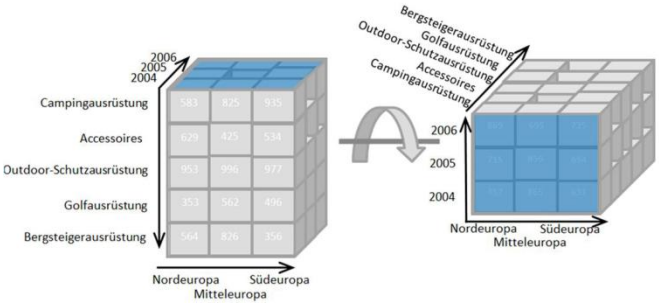


The dice operation produces a subcube by allowing the analyst to pick specific values of multiple dimensions. The picture shows a dicing operation: The new cube shows the sales figures of a limited number of product categories, the time and region dimensions cover the same range as before.



f. A data warehouse, because of the very nature of its purpose, has to c s. Data is stored as snapshots over past and current periods. Thus every data st nent.

g. Pivot allows an analyst to rotate the cube in space to see its various faces. For example, cities could be arranged vertically and products horizontally while viewing data for a particular quarter. Pivoting could replace products with time periods to see data across time for a single product
The picture shows a pivoting operation: The whole cube is rotated, giving another perspective on the data.



h. Metadata in a data warehouse contains answers to the questions about the data in the data warehouse. We keep the answers in a place called metadata repository. Before users can run their queries they need to know about the data in data warehouse. So they need metadata. Without metadata support users of large data warehouses are totally handicapped

Q2

a. Following are the attempts to provide strategic information for decision making
- Ad Hoc Reports: this was the earliest stage. Users would send requests to IT for some special reports. It would write special programs to produce ad hoc reports.
- Special Extract programs: This was an attempt by IT to anticipate somewhat the types of reports that would be requested from time to time. It would write a suite of programs and run the programs periodically to extract data from the various applications.
- Small applications: In this IT formalized the extract process. IT would create simple applications based on the extracted files. The users could stipulate the parameters for each special report.
- Information Centers: these were places where users could go and request ad hoc reports or view special information on screens. IT personals were present to help users obtain desired information.
- Decision – support Systems: These systems were supported by extracted files. The systems were menu driven and provided online information and also the ability to print special reports
- Executive Information system: This was an attempt to bring strategic information to the executive desktop. The main criteria were simplicity and ease of use. The system would display key information everyday and provide ability to request simple, straightforward reports.

b. 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.

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.

Q3. With **Slowly Changing Dimensions (SCDs)** data changes slowly, rather than changing on a time-based, regular schedule. For example consider customer demographics dimension table. What happens when a customer status changes from rental home to own home? The corresponding row in dimension table changes.

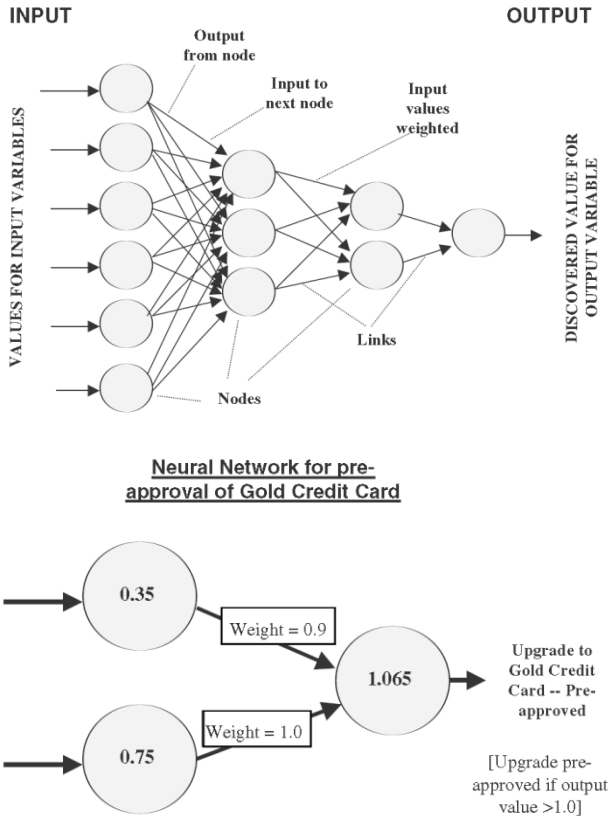
- Type 1 changes: Correction of errors
  These changes usually relate to correction of errors in the source systems. For example, suppose a spelling error in customer name. here we need not preserve old values as it is erroneous and must be discarded
  Applying type 1 changes
  - Overwrite the attribute value in the dimension table row with new value.
  - Old value is not preserved
  - No other changes are made in dimension table row
  - No keys are affected

- Type 2 Changes: Preservation of history
  In this case the historical value must be preserved. For example if address of a person is changed to other state. If it is requirement in your data warehouse that you must be ale to track orders by state, then this change along with old value must be preserved.
  Applying Type 2 changes
  - Add a new dimension table row with new value of changed attribute
  - An effective date field may be included I dimension table
  - There are no changes to original row.
  - No keys are affected
  - The new row is inserted with a new surrogate key.
- Type 3 Changes: Soft Revisions
  These are tentative or soft changes. For example if marketing department is contemplating a realignment of territorial assignments for salespersons. Before making a permanent realignment they want to count the orders in two ways: according to current territorial alignment and also to proposed realignment. This type of change is type 3 change
  Applying Type 3 change
  - Add an old value field in dimension table of affected attribute
  - Push down the existing value from current field to old field
  - Keep new value of attribute in new field
  - Also add current effective date field
  - The Key of row is not affected
  - No new dimension row is needed
  - Existing queries will seamlessly switch to new values.
  - Any query that needs to use old value must be revised accordingly

Q4.

In computer science and related fields, artificial neural networks (ANNs) are computational models inspired by animals' central nervous systems (in particular the brain) that are capable of machine learning and pattern recognition. They are usually presented as systems of interconnected "neurons" that can compute values from inputs by feeding information through the network.

Neural networks mimic the human brain by learning from a training dataset and applying the learning to generalize patterns for classification and prediction. These algorithms are effective when the data is shapeless and lacks any apparent pattern. The basic unit of an artificial neural network is modeled after the neurons in the brain. This unit is known as a node and is one of the two main structures of the neural network model. The other structure is the link that corresponds to the connection between neurons in the brain.



How Neural network works

The neural network receives values of the variables or predictors at the input nodes. There may be several inner layers operating on the predictors and they move from node to node until the discovered result is presented at the output node. The inner layers are also called hidden layers because as input dataset is running through many iterations, the inner layer rehash the predictors over and over again.

Q5.

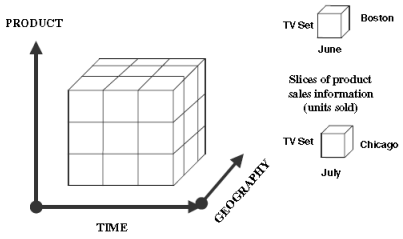
a. Metadata in a data warehouse is similar to data dictionary or data catalog in a database management system. In the data dictionary you keep the information about the logical structure, the information about the files and addresses, the information about indexes and so on. Data dictionary consists of data about the data in the data base

   Similarly metadata component is data about the data in data warehouse.

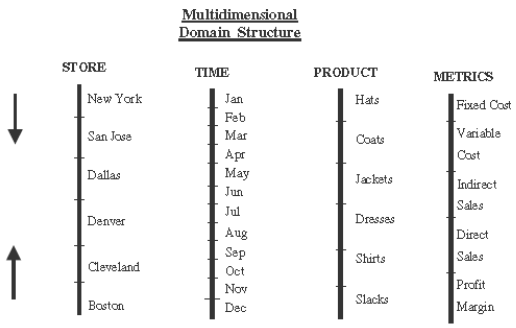
b. A data mart is the access layer of the data warehouse environment that is used to get data out to the users. The data mart is a subset of the data warehouse that is usually oriented to a specific business line or team. Data marts are small slices of the data warehouse. Whereas data warehouses have an enterprise-wide depth, the information in data marts pertains to a single

department. In some deployments, each department or business unit is considered the owner of its data mart including all the hardware, software and data.
This enables each department to use, manipulate and develop their data any way they see fit; without altering information inside other data marts or the data warehouse.

c. The multidimensional data model is an integral part of On-Line Analytical Processing, or OLAP. Because OLAP is on-line, it must provide answers quickly; analysts pose iterative queries during interactive sessions, not in batch jobs that run overnight. And because OLAP is also analytic, the queries are complex. The multidimensional data model is designed to solve complex queries in real time.
Logical cubes provide a means of organizing measures that have the same shape, that is, they have the exact same dimensions. Measures in the same cube have the same relationships to other logical objects and can easily be analyzed and displayed together. **Dimensions** contain a set of unique values that identify and categorize data. They form the edges of a logical cube, and thus of the measures within the cube. Because measures are typically multidimensional, a single value in a measure must be qualified by a member of each dimension to be meaningful. For example, the Sales measure has four dimensions: Time, Customer, Product, and Channel. A particular Sales value (43,613.50) only has meaning when it is qualified by a specific time period (July), a Geography(USA), a product (TV set). For eg.

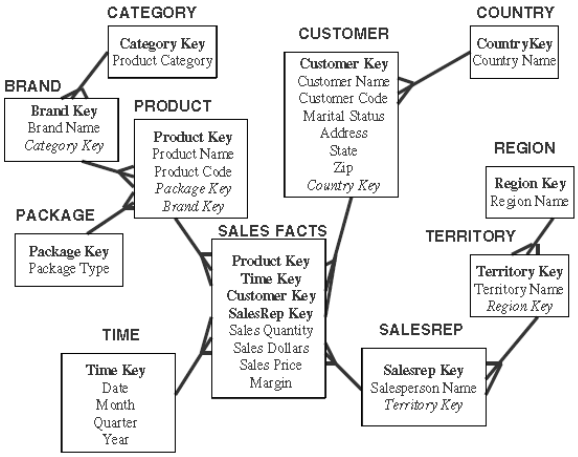


More than 3 dimensions can be displayed using Multidimension ; two or more dimensions to one so that all dimension can be viewed in 2 dim



Q6.

a. In computing, a snowflake schema is a logical arrangement of tables in a multidimensional database such that the entity relationship diagram resembles a snowflake shape. The snowflake schema is represented by centralized fact tables which are connected to multiple dimensions. "Snowflaking" is a method of normalising the dimension tables in a star schema. When it is completely normalised along all the dimension tables, the resultant structure resembles a snowflake with the fact table in the middle. The principle behind snowflaking is normalisation of the dimension tables by removing low cardinality attributes and forming separate tables.
Example of sales snowflake schema



The snowflake schema provides some advantages over the star schema in certain situations, including:

- Some OLAP multidimensional database modeling tools are optimized for snowflake schemas.
- Normalizing attributes results in storage savings, the tradeoff being additional complexity in source query joins.

Disadvantages:

- Schema less intuitive and end-users are put off by complexity .
- Ability to browse through contents difficult
- Degraded querry performance.

b. Here is the typical requirements of clustering in data mining:

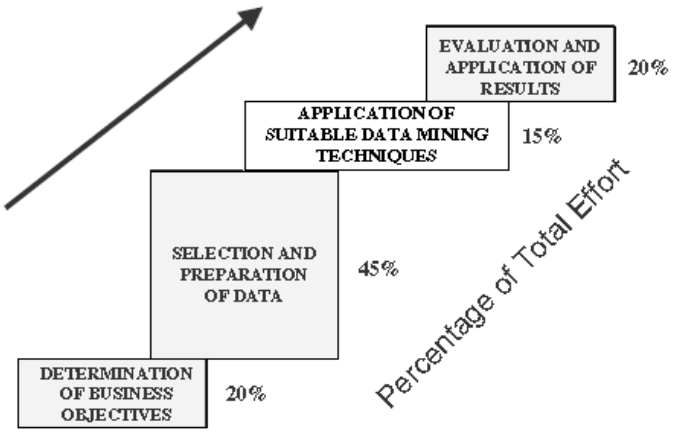
- Scalability - We need highly scalable clustering algorithms to deal with large databases.
- Ability to deal with different kind of attributes - Algorithms should be capable to be applied on any kind of data such as interval based (numerical) data, categorical, binary data.
- Discovery of clusters with attribute shape - The clustering algorithm should be capable of detect cluster of arbitrary shape. The should not be bounded to only distance measures that tend to find spherical cluster of small size.
- High dimensionality - The clustering algorithm should not only be able to handle low- dimensional data but also the high dimensional space.
- Ability to deal with noisy data - Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
- Interpretability - The clustering results should be interpretable, comprehensible and usable.

Q7.

a. Characteristic features of data warehouse are

- Subject Oriented - The Data warehouse is subject oriented because it provides us the information around a subject rather the organization's ongoing operations. These subjects can be product, customers, suppliers, sales, revenue etc. The data warehouse does not focus on the ongoing operations rather it focuses on modeling and analysis of data for decision making.
- Integrated - Data Warehouse is constructed by integration of data from heterogeneous sources such as relational databases, flat files etc. This integration enhances the effective analysis of data.
- Time-Variant - The Data in Data Warehouse is identified with a particular time period. The data in data warehouse provide information from historical point of view.
- Non Volatile - Non volatile means that the previous data is not removed when new data is added to it. The data warehouse is kept separate from the operational database therefore frequent changes in operational database are not reflected in data warehouse.
- Data Granularity – In data warehouse it is efficient to keep data summarized at different levels. Depending on the query you can then go at particular level of detail and satisfy the query. Data granularity refers to the level of detail. The lower the level of detail the finer the granularity

b. Knowledge discovery in databases (KDD) is the process of discovering useful knowledge from a collection of data. This widely used data mining technique is a process that includes data preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results.



Knowledge discovery phases

STEP 1 Define Objectives This is similar to any information system project. First of all determine whether you really need a data miming solution. State your objectives. Are you looking to improve your direct marketing campaigns?
Do you want to detect fraud in credit card usage?
In this step define the expectations/

STEP 2 Prepare data: This step consists of data selection, preprocessing of data, and data transformation. Select the data to be extracted from the data warehouse. Use the business objectives to determine what data is to be selected. Include appropriate metadata about selected data.

STEP 3 : Perform data mining: The knowledge discover engine applies the selected algorithmto the prepared data. The output from this step is a set of relationships or patterns.

STEP 4: Evaluate the results: In the selected data, there are potentially many patterns or relationships. In this step, examine all the resulting patterns. Select only the promising patterns to be presented and applied.

STEP 5: Present Discoveries: Presentation of knowledge discoveries may be in the form of visual navigation, charts, graphs, or free form texts.
Presentation also includes storing of interesting discoveries in knowledge base for repeated use.

STEP 6: Incorporate the usage of discoveries. The goal of any data mining operation is to understand the business, discover new patterns and possibilities, and also turn this understanding into actions. This step is for using tesults to create actionable items in the business.