Model Test Paper

Question no. 1 is compulsory. Attempt all parts.

Q1. Each question carries equal marks. (5*5 marks)

A) What are Knowledge discovery phases?

Solution: *Step 1: Define Business Objectives.* This step is similar to any information system project. First of all, determine whether you really need a data mining solution. State your objectives. Are you looking to improve your direct marketing campaigns? Do you want to detect fraud in credit card usage? Are you looking for associations between products that sell together? In this step, define expectations. Express how the final results will be presented and used in the operational systems.

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 has to be selected. Include appropriate metadata about the selected data. By now, you also know what type of mining algorithm you will be using. The mining algorithm has a bearing on data selection. The variables selected for data mining are also known as active variables. Preprocessing is meant to improve the quality of selected data. When you select from the data warehouse, it is assumed that the data is already cleansed. Preprocessing could also involve enriching the selected data with external data. In the preprocessing substep, remove noisy data, that is, data blatantly out of range. Also ensure that there are no missing values. Clearly, if the data for mining is selected from the data warehouse, it is again assumed that all the necessary data transformations have already been completed. Make sure that this really is the case.

Step 3: Perform Data Mining. Obviously, this is the crucial step. The knowledge discovery

engine applies the selected algorithm to the prepared data. The output from this step is a set of relationships or patterns. However, this step and the next step of evaluation may be performed in an iterative manner. After an initial evaluation, you may adjust the data and redo this step. The duration and intensity of this step depend on the type of data mining application. If you are segmenting the database, not too many iterations are needed. If you are creating a predictive model, the models are repeatedly set up and tested with sample data before testing with the real database.

Step 4: Evaluate Results. You are actually seeking interesting patterns or relationships.

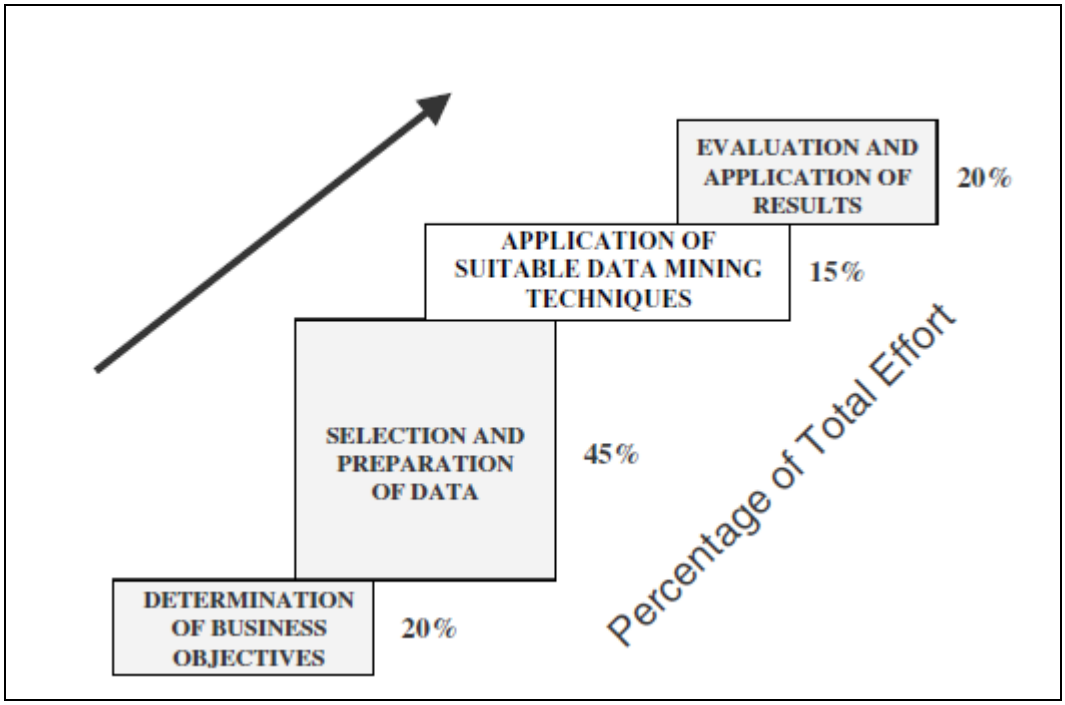
These help you in the understanding of your customers,products, profits, and markets. In the selected data, there are potentially many patterns or relationships. In this step, you examine all the resulting patterns. You will apply a filtering mechanism and select only the promising

Neha Sharma
Assistant Professor
IT

patterns to be presented and applied. Again, this step also depends on the specific kind of data mining algorithm 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 the knowledge base for repeated use.

Step 6: Incorporate Usage of Discoveries. The goal of any data mining operation is to understand the business, discern new patterns and possibilities, and also turn this understanding into actions. This step is for using the results to create actionable items in the business. You assemble the results of the discovery in the best way so that they can be exploited to improve the business.



B) Why machine learning is done?

Solution: Following are the reasons:

1. To understand and improve the efficiency of human learning.
2. To discover new things or structure that is unknown to human beings.
3. To fill in skeletal or computer specifications about a domain.

Neha Sharma
Assistant Professor
IT

C) What are the steps in the data mining process?

Solution: Steps of data mining process are:

a. Data cleaning
b. Data integration
c. Data selection
d. Data transformation
e. Data mining
f. Pattern evaluation
g. Knowledge representation

D) Differentiate between different data mining techniques.

Solution:

Data Mining Technique	Underlying Structure	Basic Process	Validation Method
Cluster Detection	Distance calculations in n-vector space	Grouping of values in the same neighborhood	Cross validation to verify accuracy
Decision Trees	Binary Tree	Splits at decision points based on entropy	Cross validation
Memory-based Reasoning	Predictive structure based on distance and combination functions	Association of unknown instances with known instances	Cross validation
Link Analysis	Based on linking of variables	Discover links among variables by their values	Not applicable
Neural Networks	Forward propagation network	Weighted inputs of predictors at each node	Not applicable
Genetic Algorithms	Not applicable	Survival of the fittest on mutation of	Mostly cross validation

E) What are the requirements of clustering?

Solution:

Neha Sharma
Assistant Professor
IT

- Scalability
- Ability to deal with different types of attributes
- Ability to deal with noisy data
- Minimal requirements for domain knowledge to determine input parameters
- Constraint based clustering
- Interpretability and usability

Each unit is of 12.5 marks.

Unit- I

Ques 2. What are the advantages of Dimensional modeling?

Solution:

- Ease of use.
- High performance
- Predictable, standard framework
- Understandable
- Extensible to accommodate unexpected new data elements and new design decisions

Ques 3. What are the characteristics of the strategic information?

Solution: Following are the characteristics of strategic information:

INTEGRATED: Must have a single, enterprise-wide view.
DATA INTEGRITY: Information must be accurate and must conform to business rules.
ACCESSIBLE: Every business factor must have one and only one value.
CREDIBLE: Easily accessible with intuitive access paths, and responsive for analysis.
TIMELY: Information must be available within the stipulated time frame.

Ques 4. Explain the history of Decision-Support Systems.

Solution:
Depending on the size and nature of the business, most companies have gone through the

following stages of attempts to provide strategic information for decision making.

Ad Hoc Reports. This was the earliest stage. Users, especially from Marketing and Finance, would send requests to IT for special reports. IT would write special programs, typically one for each request, and produce the ad hoc reports.

Special Extract Programs. This stage 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. IT would create and keep the extract files to fulfill any requests for special reports. For any reports that could not be run off the extracted files, IT would write individual special programs.

Small Applications. In this stage, 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. The report printing programs would print the information based on user-specific parameters. Some advanced applications would also allow users to view information through online screens.

Information Centers. In the early 1970s, some major corporations created what were called information centers. The information center typically was a place where users could go to request ad hoc reports or view special information on screens. These were predetermined reports or screens. IT personnel were present at these information centers to help the users to obtain the desired information.

Decision-Support Systems. In this stage, companies began to build more sophisticated systems intended to provide strategic information. Again, similar to the earlier attempts, these systems were supported by extracted files. The systems were menu-driven and provided online information and also the ability to print special reports. Many of such decision-support systems were for marketing.

Executive Information Systems. 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 every day and provide ability to request simple, straightforward reports. However, only preprogrammed screens and reports were available. After seeing the total countrywide sales, if the executive wanted to see the analysis by region, by product, or by another dimension, it was not possible unless such breakdowns were already preprogrammed. This limitation caused frustration and executive information systems did not last long in many companies.

Unit- II

Ques 5. Differentiate between operational and informational system.

Solution:

Neha Sharma
Assistant Professor
IT

	OPERATIONAL	INFORMATIONAL
Data Content	Current values	Archived, derived, summarized
Data Structure	Optimized for transactions	Optimized for complex queries
Access Frequency	High	Medium to low
Access Type	Read, update, delete	Read
Usage	Predictable, repetitive	Ad hoc, random, heuristic
Response Time	Sub-seconds	Several seconds to minutes
Users	Large number	Relatively small number

Ques 6. Explain data warehouse architecture.

Solution:

Source Data Component

Source data coming into the data warehouse may be grouped into four broad categories:

Production Data. This category of data comes from the various operational systems of the enterprise. Based on the information requirements in the data warehouse, you choose segments of data from the different operational systems.

Internal Data. In every organization, users keep their "private" spreadsheets, documents, customer profiles, and sometimes even departmental databases. This is the internal data, parts of which could be useful in a data warehouse.

Archived Data. Operational systems are primarily intended to run the current business. In every operational system, you periodically take the old data and store it in archived files.

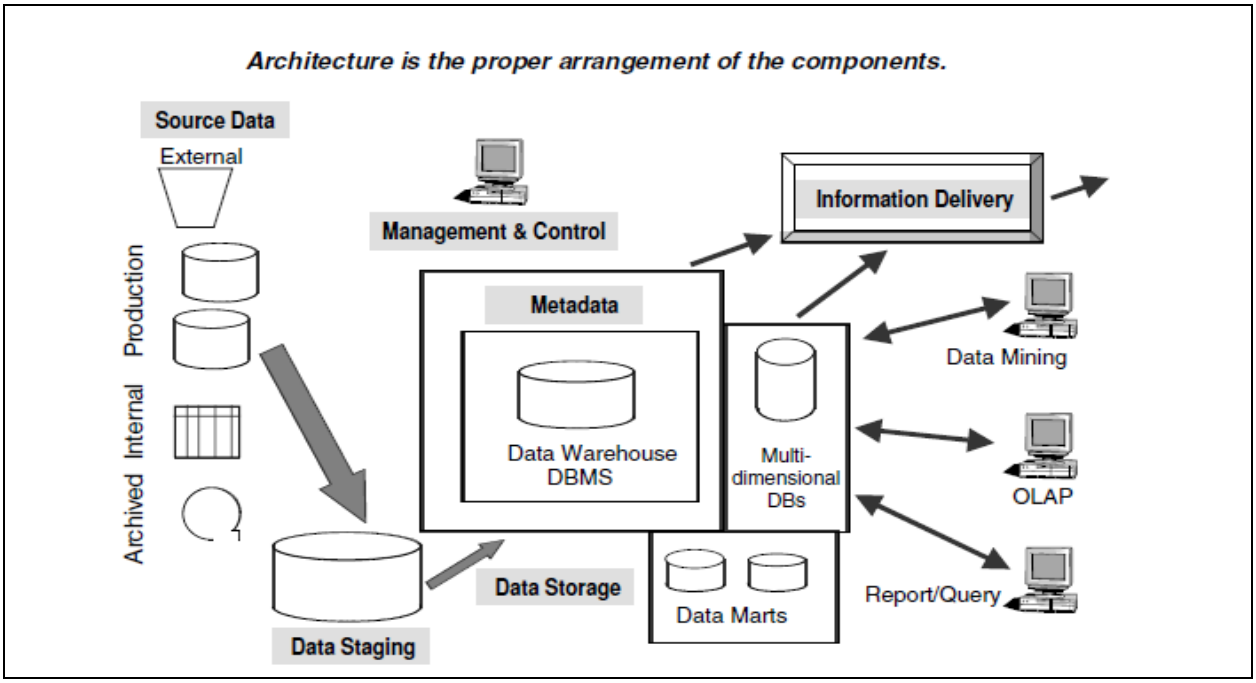
External Data. Most executives depend on data from external sources for a high percentage of the information they use. They use statistics relating to their industry produced by external agencies.

Data Staging Component:

Three major functions need to be performed for getting the data ready. You have to extract the data, transform the data, and then load the data into the data warehouse storage. These three major functions of extraction, transformation, and preparation for loading take place in a staging area. The data staging component consists of a workbench for these functions. Data staging provides a place and an area with a set of functions to clean, change, combine, convert, deduplicate, and prepare source data for storage and use in the data warehouse.

- Data extraction
- Data transformation
- Data loading

Neha Sharma
Assistant Professor
IT



Data Storage Component:

The data storage for the data warehouse is a separate repository. The operational systems of your enterprise support the day-to-day operations. These are online transaction processing applications. The data repositories for the operational systems typically contain only the current data. Also, these data repositories contain the data structured in highly normalized formats for fast and efficient processing.

Information Delivery Component:

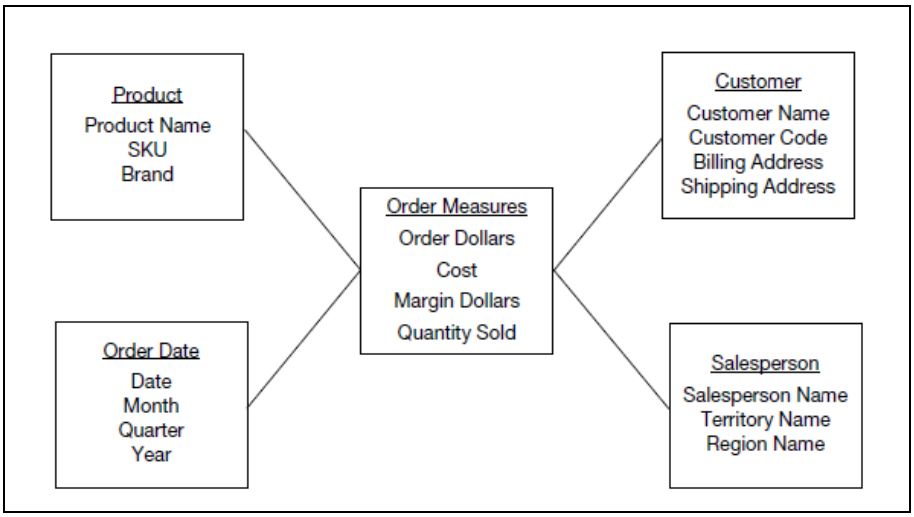
In order to provide information to the wide community of data warehouse users, the information delivery component includes different methods of information delivery like as internet, intranet, online, email etc.

Unit- III

Ques 7. Create star schema corresponding to order analysis.

Solution:

Neha Sharma
Assistant Professor
IT



Ques 8. Explain type-1, type-2 and type-3 changes.

Solution:

Type 1 Changes: Correction of Errors

Nature of Type 1 Changes. These changes usually relate to the corrections of errors in the source systems. For example, suppose a spelling error in the customer name is corrected to read as Michael Romano from the erroneous entry of Michel Romano. Also, suppose the customer name for another customer is changed from Kristin Daniels to Kristin Samuelson, and the marital status changed from single to married.

Applying Type 1 Changes to the Data Warehouse.

- Overwrite the attribute value in the dimension table row with the new value.
- The old value of the attribute is not preserved
- No other changes are made in the dimension table row
- The key of this dimension table or any other key values are not affected
- This type is easiest to implement

Type 2 Changes: Preservation of History

Nature of Type 2 Changes. Go back to the change in the marital status for Kristin Samuelson. Assume that in your data warehouse one of the essential requirements is to track orders by marital status in addition to tracking by other attributes. If the change to marital status happened on October 1, 2000, all orders from Kristin Samuelson before that date must be included under marital status: single, and all orders on or after October 1, 2000 should be included under marital status: married.

Applying Type 2 Changes to the Data Warehouse.

Neha Sharma
Assistant Professor
IT

- Add a new dimension table row with the new value of the changed attribute
- An effective date field may be included in the dimension table
- There are no changes to the original row in the dimension table
- The key of the original row is not affected
- The new row is inserted with a new surrogate key

Type 3 Changes: Tentative Soft Revisions

Nature of Type 3 Changes. Almost all the usual changes to dimension values are either Type 1 or Type 2 changes. Of these two, Type 1 changes are more common. Type 2 changes preserve the history. When you apply a Type 2 change on a certain date, that date is a cut-off point. In the above case of change to marital status on October 1, 2000, that date is the cut-off date. Any orders from the customer prior to that date fall into the older orders group; orders on or after that date fall into the newer orders group. An order for this customer has to fall in one or the other group; it cannot be counted in both groups for any period of time.

Applying Type 3 Changes to the Data Warehouse

- Add an "old" field in the dimension table for the affected attribute
- Push down the existing value of the attribute from the "current" field to the "old"field
- Keep the new value of the attribute in the "current" field
- Also, you may add a "current" effective date field for the attribute
- The key of the row is not affected
- No new dimension row is needed
- The existing queries will seamlessly switch to the "current" value
- Any queries that need to use the "old" value must be revised accordingly
- The technique works best for one "soft" change at a time
- If there is a succession of changes, more sophisticated techniques must be devised

Ques 9. What is Descriptive and predictive data mining?

Solution:

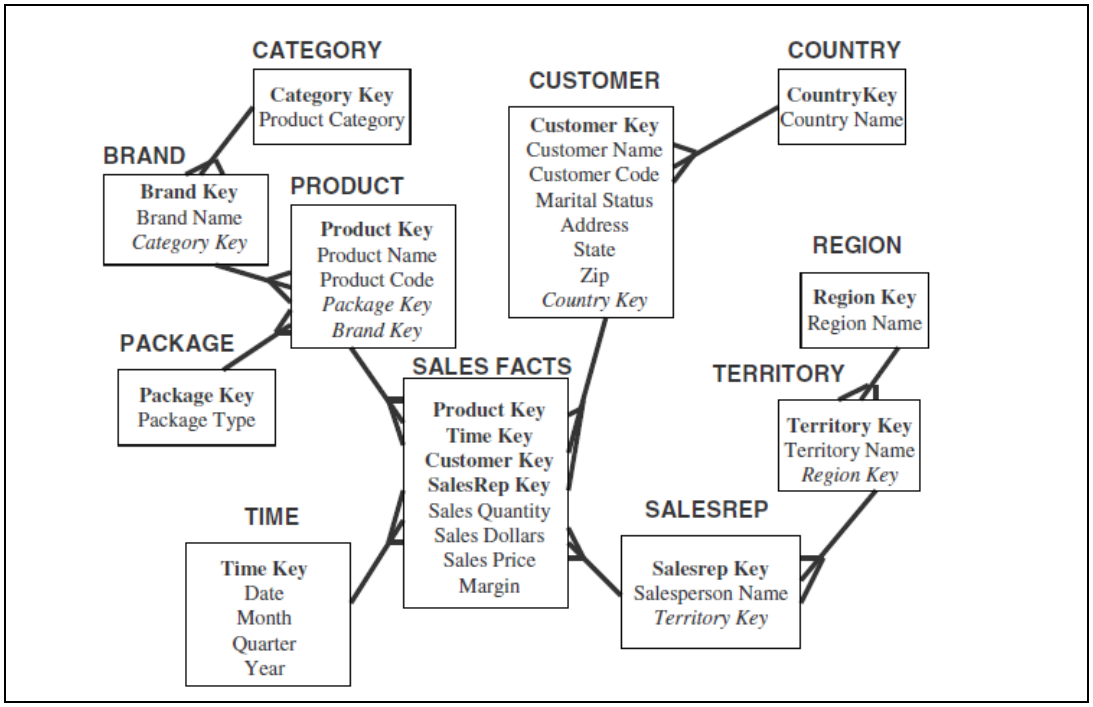
Descriptive data mining describes the data set in a concise and summertime manner and Presents interesting general properties of the data. Predictive data mining analyzes the data in order to construct one or set of models and attempts to predict the behavior of new data sets

Unit- IV

Ques 10. Create snowflake schema corresponding to sales.

Neha Sharma
Assistant Professor
IT

Solution:



Ques 11. Explain Codd's guidelines.

Solution:

Twelve guidelines for an OLAP system, given by Dr. E.F.Codd are:

Multidimensional Conceptual View. Provide a multidimensional data model that is intuitively analytical and easy to use. Business users' view of an enterprise is multidimensional in nature. Therefore, a multidimensional data model conforms to how the users perceive business problems.

Transparency. Make the technology, underlying data repository, computing architecture, and the diverse nature of source data totally transparent to users. Such transparency, supporting a true open system approach, helps to enhance the efficiency and productivity of the users through front-end tools that are familiar to them.

Accessibility. Provide access only to the data that is actually needed to perform the specific analysis, presenting a single, coherent, and consistent view to the users. The OLAP system must map its own logical schema to the heterogeneous physical data stores and perform any necessary transformations.

Consistent Reporting Performance. Ensure that the users do not experience any significant degradation in reporting performance as the number of dimensions or the size of the database

increases. Users must perceive consistent run time, response time, or machine utilization every time a given query is run.

Client/Server Architecture. Conform the system to the principles of client/server architecture for optimum performance, flexibility, adaptability, and interoperability. Make the server component sufficiently intelligent to enable various clients to be attached with a minimum of effort and integration programming.

Generic Dimensionality. Ensure that every data dimension is equivalent in both structure and operational capabilities. Have one logical structure for all dimensions. The basic data structure or the access techniques must not be biased toward any single data dimension.

Dynamic Sparse Matrix Handling. Adapt the physical schema to the specific analytical model being created and loaded that optimizes sparse matrix handling. When encountering a sparse matrix, the system must be able to dynamically deduce the distribution of the data and adjust the storage and access to achieve and maintain consistent level of performance.

Multiuser Support. Provide support for end users to work concurrently with either the same analytical model or to create different models from the same data. In short, provide concurrent data access, data integrity, and access security.

Unrestricted Cross-dimensional Operations. Provide ability for the system to recognize dimensional hierarchies and automatically perform roll-up and drill-down operations within a dimension or across dimensions. Have the interface language allow calculations and data manipulations across any number of data dimensions, without restricting any relations between data cells, regardless of the number of common data attributes each cell contains.

Intuitive Data Manipulation. Enable consolidation path reorientation (pivoting), drill-down and roll-up, and other manipulations to be accomplished intuitively and directly via point-and-click and drag-and-drop actions on the cells of the analytical model. Avoid the use of a menu or multiple trips to a user interface.

Flexible Reporting. Provide capabilities to the business user to arrange columns, rows, and cells in a manner that facilitates easy manipulation, analysis, and synthesis of information. Every dimension, including any subsets, must be able to be displayed.

with equal ease.

Unlimited Dimensions and Aggregation Levels. Accommodate at least fifteen, preferably twenty, data dimensions within a common analytical model. Each of these generic dimensions must allow a practically unlimited number of user-defined aggregation levels within any given consolidation path.

Ques 12. What is Temporal Database?

Solution: Temporal database store time related data .It usually stores relational data that include time related attributes. These attributes may involve several time stamps, each having different semantics.

Neha Sharma
Assistant Professor
IT