B.E.

Seventh Semester Examination, May-2008

Data Ware Housing & Data Mining (IT-401-E)

Note : Attempt any five questions. All questions carry equal marks.

Q. 1. (a) What are schemas ? How these are used and useful for multidimensional database ? Explain with examples.

Ans. The structure for organizing and storing a database is known as schema. The database schema stores the data in various forms, such as fact table and dimension table. Various data warehouse schemas are star schema, snowflake schema and star flake schema. The database schema enables to store data in multi-dimensional models.

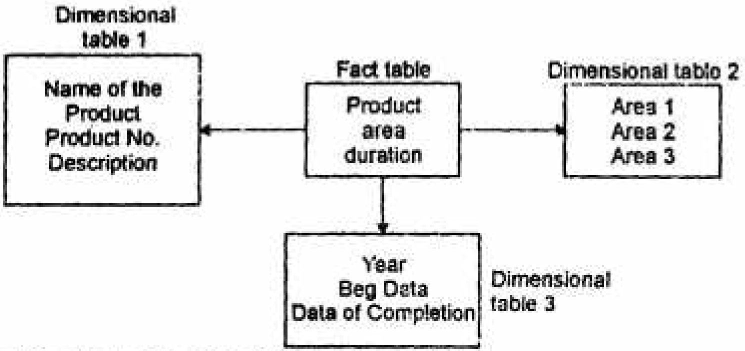
Multi-Dimensional Models for Data Storage :

	Area 1	Area 2	Area 3
Product 1			
Product 2			
Product 3			
Product 4			

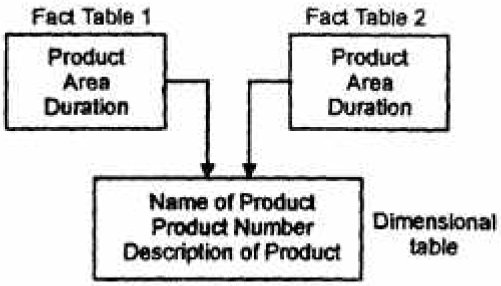
Two domensional model for data storage.

The multi-dimensional models are the structures that stores data using various dimensions. The three dimensional model is known as data table.

Figure shows the 2 dimensional model to store data about the sales of products in different areas.



The schema with a fact table, table 2 dimensional tables.



The fact construction

Determining Fact Table : The process involves :

(i) Identifying the transactions of interest.

(ii) Identifying the dimensions of each fact.

(iii) Verifying that a fact is not a dimension table.

(iv) Verifying that a dimension table is not a fact.

Q. 1. (b) Differentiate between data warehouse, data cubes and meta data with examples.

Ans. Data Warehouse : Data warehouse refers to a collection of information used to manage a business. Data warehousing updates business activities, such as creating customer loyality schemes, promoting products and identifying potential market segments.

Objectives of Data Warehousing :

Operational Data : Refers to the data that constitutes the daily processing of a company.

Decision Support Data : Refers to the data that supports the managerial decisions.

External Data : Refers to the data that is compiled from another system and then the operational systems compiles the data.

Characteristics of Data Warehousing Data :

(i) Subject-oriented

(ii) Non-volatile

(iii) Time variant

(iv) Integrated.

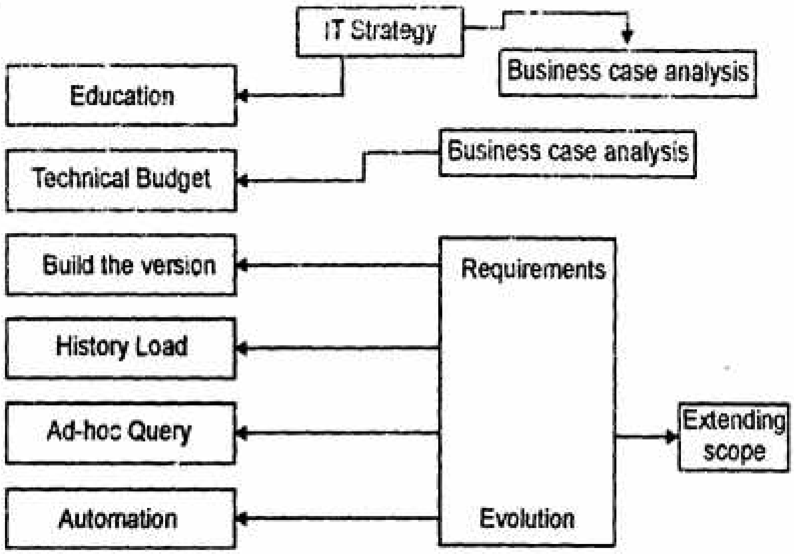
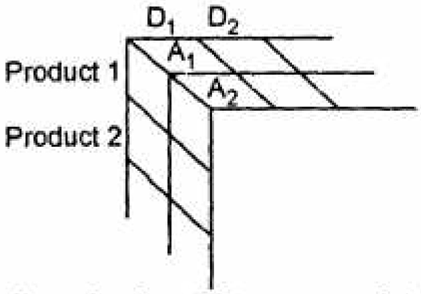


Fig. Method of data warehousing

Data Cubes : The three dimensional model is known as data cube. The data represented in cubes can be pivoted and diced, which means that the dimensions representing the data can be interchanged.

The multidimensional data models allow two types of operations, roll-up delay and rill down display.



Three dimensional model to response data pivot

Meta Data : Data that describes the structure and business meaning of data stored in CDW and DDW, as well as how they are created, accessed and used, is known as meta data or data-about-data.

Types of Meta Data :

(i) Technical meta data

(ii) Business meta data

Technical Meta Data : The static facts wound include such things as :

(i) Source file/table definitions.

(ii) Target table definitions.

(iii) Transformations and mapping rules.

Business Meta Data : Not only applies to the data in the warehouse but equally to the information about broader, class of objects such as graph, or chart view through presentation tool, a query tool, OLAP tool and web tools.

Q. 2. Explain the following briefly :

(i) Data warehouse manager

(ii) Virtual data warehouse

(iii) ROLAP vs MOLAP

Ans. (i) Data Warehouse Manager : The components of the data warehouse, such as the database, use data manager components for managing and accessing the data stored in data warehouse. The data manager is of two types, Relational Database Management (RDBMS) & Multi-dimensional Database Management System (MDBMS). The RDBMS is for designing vast enterprise data warehouse & MDBMS is for designing small size departmental data warehouse.

The database management system for building data warehouse are used according to the number of end users, complexities of the queries and the size of database.

The Management Components : The management component of the data warehousing architecture maintains the database used in the data warehouse. The various functions of management components are :

(i) Administering data acquisition operations

(ii) Managing back up copies of data

(iii) Recovering the lost data

(iv) Providing security to stored data

(v) Authorising access to stored data.

(iv) Managing the operations i.e., insert, update, delete etc.

(ii) Virtual Data Warehouse : A management information system (MIS), also known as virtual warehouse, is defined as a query & reporting system that directly performs aggregations & summarizations against the operational data, sometimes storing summarized & aggregated data privacy.

A MIS also provides query & reporting capabilities for the decision makers – managers, executives.

Characteristics of MIS :

(i) An MIS involves direct access to source data. The data access process itself does the data transformation, to the extent that it can do this.

(ii) An MIS is focussed on current data, in the first place, because the current data is what is available in the data sources.

(iii) When dealing with heterogeneous data sources, the MIS data access middleware must be able to do complex data query on a hierarchical data model or a flat file system.

(iv) An MIS usually has some difficulty dealing with external data and with unstructured data.

(v) Because of the extract facility and the local store facility built in MIS solutions, an MIS environment tends to favour proliferation of data extracts into personal stores.

(iii) ROLAP vs MOLAP : Large amount of data are maintained in the form of the relational databases. The relational database is usually used for transaction processing. For successfully executing these transactions, the database is accompanied with highly efficient data processor for many small transactions. The new trend is to form tools that make a data warehouse out of the database.

Most of the OLAP applications have three features in common. These features are :

(i) Multidimensional view of data

(ii) Calculation-intensive capabilities

(iii) Time intelligence

Relational OLAP Server (ROLAP) : The ROLAP server is placed between relational back-end server and client front-end tool. These-servers use relational database management system for storing the warehouse data. The ROLAP servers provide optimisation routines for each database management system in the back-end. It provides for aggregation and navigation tools for the database. The ROLAP has a higher scalability than multidimensional OLAP server (MOLAP).

Multidimensional OLAP Server (MOLAP) : The MOLAP servers use array-based multidimensional storage engines for providing a multidimensional view for data. These map dimensional views directly from data cube array structure. The advantage is using a data cube is that fast indexing & summarized data.

Q. 3. Differentiate between tuning and testing of data warehouse in detail. Also discuss data warehouse back end tools briefly.

Ans. Tuning the Data Warehouse : Tuning refers to measuring the performance of the data warehouse to provide the required information about various records stored in data warehouse. The factors used to measure the performance of a data warehouse are :

(a) Identifying the time required for execution of a query.

(b) Specifying the memory required by a process.

(c) Identifying the data scanning rate.

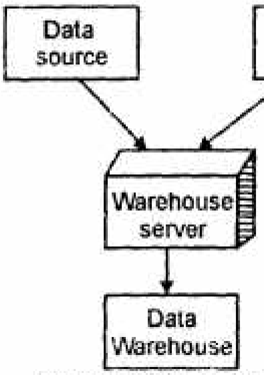
(d) Identifying the Input/Output (I/O) throughput rate.

Data Load Tuning : Data load tuning refers to recurring the delays in the data load process in a data warehouse. The factors responsible for the delay are :

(i) Transferring of data between the data sources & the data warehouse.

(ii) Maintaining indexes in data warehouse.

(iii) Performing integrity check on data.



Various stages of data flow

Testing Data Warehouse : The development of the data warehouse needs to include testing procedure to ensure that all the queries execute in time. The various types of testing in a data warehouse are :

Unit Testing : Tests the individual modules consisting of programs, procedures & the standard query language (SQL) script. A module is a self-contained component that combines with other components to develop an application.

Integration Testing : Ensures that individual modules can work together when they are integrated.

System Testing : Tests the entire database application & deals with the testing of management and query tools. The management and query tools manages the data in the data warehouse and execution of various queries.

User Acceptance Testing : Tests the system by using the inputs of the end user.

Testing of Operational Requirement :

(i) Disk capacity for storing data

(ii) Scheduling software

(iii) Testing the working of DB manager.

Sales Analysis :

(i) Determine real-time product sales to make vital pricing and distribution decisions.

(ii) Analyse historical product, sales to determine success or failures attributes.

(iii) Evaluate successful products and determine key success factors.

Financial Analysis :

(i) Compare actual to bugets on an annual, monthly basis.

(ii) Review past cash flow trends and forecast future needs.

(iii) Receive near real time, interactive financial statements.

results in the development of an associative algorithm that correlates one set of events or items with another set of events or items. Patterns desired from the algorithm are generally expressed as, for example, "eighty three percent of all records that contain A and B, C items also contain items D and E".

A common example of the use of association methods is market basket analysis. Using a linkage approach, a retailer can mine the data generated by a point-of sale system, such as the price scanner at the grosery store. By analysing the products contained in a processer's basket and then using an associative algorithm to compare hundreds of thousands of baskets, specific product affinities can be desired.

(iii) Sequence : Techniqes that use sequencing of time-series analysis relate events in time, such as the periodic per prediction of interest rate fluctuations or stock performance, based on a series of the preceding events. Though this analysis, various hidden trends, often predictive of future events, can be discovered.

Common techniques of sequence analysis method can be found in the direct mail industry. Sequences are often analysed as they relate to a specific customer/group of customers. One additional technique within this class focuses on the determination of similar sequences.

In a typical sequential analysis, the output is a pattern representing a sequence of events over time. In a similar-sequential analysis, however, the goal is to find groups of timed sequences. One of the most common examples of this approach is a large retailer using a similar-sequence approach to find unrelated departments with similar sales streams.

(iv) Cluster : In some cases, it is difficult or impossible to define the parameters of a class of data to be analysed. When parameters are elusive, clustering methods can be used to create partitions so that all members of each set are similar according to some metric or set of metricces. A cluster is simply a set of objects grouped together by virtue of their similarly or proximity to each other.

For instance, a clustering aproach might be used to mine credit card purchase data to discover that meals charged on a business-issued gold card are typically purchased on weekdays and have a mean value of greater than $ 250, whereas meals purchased using a personal platinum card occur predominately on week ends, have a mean value of $175 and include a bottle of wine more than 65 percent of the time.

Q. 5. Discuss the following briefly :

(i) Data Mining Languages

(ii) Standards of data mining

Ans. (I) Data Mining Languages :

Red Brick : Red brick offers a number of case studies that demonstrate how DM technology can be used in the real world.

The world wide customer support organisation (WCSO) within Hewlett Packard is responsible for providing support services to its hardware and software customers. For several years, the WCSO used a DM of financial, account, product and service contract information to support decision making.

Oracle : For large scale DM, orcale on SP2 offers its users robust functionality and excellent performance. Data spread across multiple SP2 processors nodes are treated as a single image oracle offers products that help customers create, administrator and use their DW. Oracles large suite of the connectivity products provides transparent access to many popular mainframe databases and enable customers to more data away from legacy mainframe applications into the DW on the SP2.

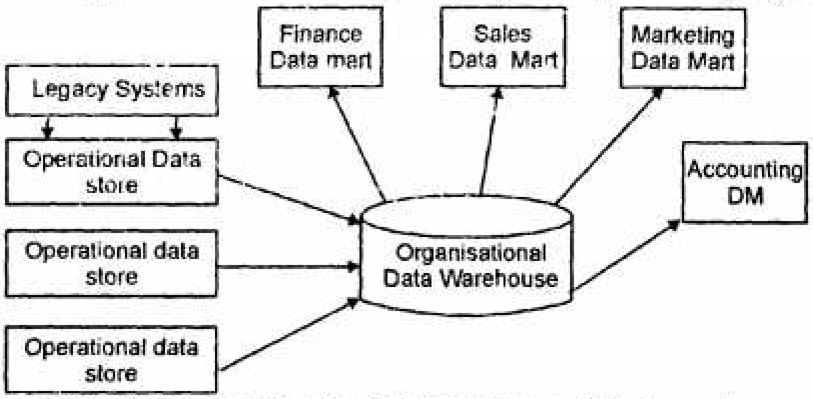
INFOMIX : As a major player in the DM field, INFOMIX can claim a number of success stories. It enables associated grocesrs to seel unique items economically, that is slow moving merchandise that is ordered monthly values daily. Rather than incrurring the high cost of warehouse these items, associated grocers created a direct link to outside especially warehouses to supply the needed items on demand. Independent stores simply order the merchandise from associated grocers. The speciality items are loaded onto associated grocers's delivery trucks.

(ii) **Standards of Data Mining :** Data mining is the process of using raw data to infer important business relationships. Once the business relationships have been discovered, they can then be used for business advantage. A data warehouse has uses other than data mining. However the fullest use of a data warehouse must include data mining. However, first we need to see the big picture.

(i) Data mining is a collection of powerfull data analysis techniques intended to assist in analysing extremely large data sets. Properly applied, data mining can reveal hidden relationships and information buried within organisational data warehouse.

(ii) There is no one data mining approach, but rather a set of techniques that often can be used in combination with each other to extract the most insight from a set of data.

Despite the seemingly recent emergence of data mining, the approach has a rich tradition of research & practice dating back over 30 years. Although most modern data mining packages still offer the classical approach, data mining has moved for beyond these first generation statistical measures to more insightful & powerful approaches that assist in explaning or predicting "what is going on in the data."



Architectural positioning of data warehouse & data marts

Infrastructure Preparation :

(i) A hardware platform

(ii) Database Management System (DBMS) platform

(iii) One or more tools for data mining.

Q. 6. What is knowledge discovery ? How is it implemented through Neural Networks and Fuzzy techniques ? Explain with examples.

Ans. Knowledge Discovery : As increasingly common synonym for DM techniques is knowledge data discovery (KDD). This more descriptive form aggregate data. Using a combination of techniques including statistical analysis, neural & fuzzy logic, Multidimensional analysis, data visualization, &

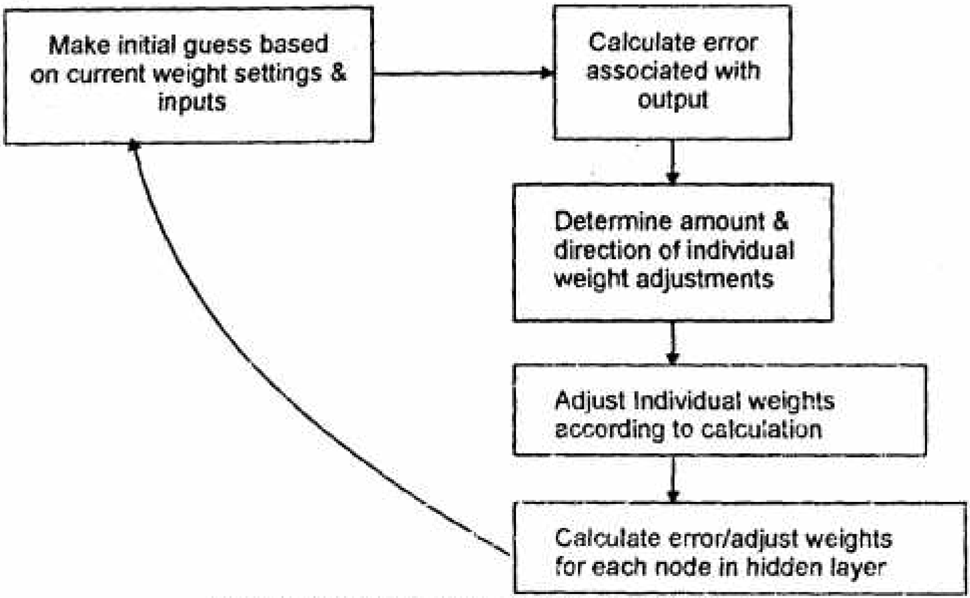
intelligent agents, KDD can discover highly useful & informative patterns within the data that can be used to develop predictive models of behaviour or consequences in a wide variety of the knowledge domains.

For example, form using KDD techniques. we know that left handed women tend to buy right handed golf-gloves. This pattern relates a customer's attributes to product sales. It is also known that AJ & T's stock price predictably rises at least 2 percent after every 3-day slump in the Dew Jones Industrial Average.

Implementation Using Fuzzy Logic & Neural Networks : A Neural network attempts to mirror the way the human brain works in recognizing patterns by developing mathematical structures with the ability to learn.

By studying combinations of variables & how different combinations affect datasets, it is possible to develop non-linear predictive models that "learn". Machine learning techniques, such as genetic algorithms and fuzzy logic, can derive meaning from complicated & imprecise data & can extract patterns from & detect trends with the data that are far too complex to be noticed by either humans or more conventional automated analysis techniques because of this ability, neural computing & machine learning technologies demonstrate broad applicability to the world of DM & thus a wide variety of complex business problems.

In contrast to the seemingly complex approaches just presented, decision trees offer a conceptually simple mathematical method of following the effect of each event, or decision or successive event. For example, in a simple decision tree involving the performance of an activity indoors or outdoors, if "indoors" is selected from the initial choice set, then the next decision will more likely be "upstairs/down stairs" rather than "sun/shade". By continually breaking datasets into separate smaller groups, a predictive model can be built. Decision trees used in DM applications assist in the classification of items or events, contained within the DW.


Make initial guess based on current weight settings & inputs
Calculate error associated with output
Determine amount & direction of individual weight adjustments
Adjust Individual weights according to calculation
Calculate error/adjust weights for each node in hidden layer

Typical training sequence for Neural networks

Advantages of Fuzzy Systems :

(i) Modelling with contraction

(ii) Increased system autonomy

Limitations of Fuzzy systems

(i) Obstacles to system.

(ii) Fuzzy Systems Lack Memory.

Neural Computing : Training the ANN : The process of training net to associate certain input patterns with correct output responses involves the use of repetitve examples and feedback, much like the training of a human being.

The operation begins by setting all of the connection weights in the net to small random values, which allows the net to begin with no 'specific memory' or imprint.

Next, the net is presented with a single data example drawn from a training set with known outputs.

The net processes this example & then provides a "guess" at the answer based on the examples provided.

Benefits Derived from Neural Computing :

(i) Avoids explicit programming & detailed IF-THEN rule base.

(ii) Reduces need for expensive or limited availability experts.

(iii) ANNs are inherently adaptable & donot require update when inputs change.

(iv) Eliminates need for redefined knowledge base.

(v) Able to process erroneous, in consistent or even incomplete data.

(vi) Allows for generalisation from specific information context.

(viii) Allows for inclusion of "common sense" with problem-solving domain.

Q. 7. Differentiate between Spatial database, Multimedia databases and World wide web with their relative merits, demerits and applications in detail.

Ans. Spatial Database : Spatial data are elements that can be stored in map form. These elements correspond to a uniquely defined location on the earth's surface. Spatial data contain three basic components : points, lines & polygons.

Points are single locations in two or three dimensional space, example a dot representing a city on a map of United States.

Lines can be isolated within a tree structure, or elements of a network structure.

Example, a river/road system.

Polygons can be isolated, adjacent or nested.

Example, state boundaries or counter lines on a map.

A GIS must be able to handle the attribute data. Simply but attribute data are the description of the spatial data seen on a map.

Examples of decision-making the scenarios in which a GIS may be appropriate include the following :

(i) Does it sensible to put a megawall in a particular locations ?

(ii) Where should legislative district boundaries be located ?

(iii) Do we expand the existing airport or build a new one in a different location ?

(iv) Will the current school facilities be sufficient for the number of students expected 10 years from now ?

(v) Which pockets of endangered environment should be protected ?

(vi) What is the impact of waste facilities can local heatlh patterns ?

Multimedia Databases : An executive information system requires a database component for retrieving, analysing, manipulating & updating files. A multimedia database management system (MMDBMS) can increase the future EIS user's resources to manipulate text, voice & images effectively within an integrated database structure. MMDBMSs provide the traditional benefits of a database management system, as well as voice concatenation, transformation of informatiion, rotation of images, scaling of objects, & merging of various data types.

The problem with these systems, especially for the executive user, is the complex interface. As the functionality of these systems, especially for the executive user, is the complex interface. As the functionality of these systems continues to increase, more applications for their use will be developed. By combining more applications with an easy-to-use system, an opportunity for a competitive advantage develops.

Literally thousands of private sector & government maintained databases currently exist, & the present volume is expected to double by the turn of the centuary. The information in these databases is available but the scanning, filtering & extraction tools that allow for efficient & effective use of those data are still being refined.

World-Wide-Web : As with everything else, it comes into contact with, the world wide web undoubtedly have a profound impact on the data warehousing. The ability to access and transfer large numbers of data relatively easily and economically will make the internet and world web ideal vehicles to integrate external data, into the DW environment.

If this is to become a reality, however a myriad of issues–relating to data integrity, accuracy & quality will have to be addressed and resolved.

It is conceivable that third party informediaries will evolve whose sole purpose is to evaluate & rate the quality & integrity of external data sources. Such quality ratings could be used to determine the degree of value to the be placed on the integration of a particular source of external data into the DW. Equally conceivable is the use of a quality rating to determine the price to be paid for access to such data, the higher the quality rating, the higher the price.

The other part of this dilemma that is painfully clear is that these same market is business forces that have led us to this place are not about to go away easily or without a hefty price tag. In contrast, these forces will be increasing, & no IT led argument about the elegance or potential dollar savings of a long-term investment in the idealist vision of a single hub-and-spoke enterprise architecture is going to slow them down.

Finally, the federated approach increases business information flexibility by creating an architecture that can accomodate the shifting needs and priorities of the business, members of the business information team are viewed as the key players in responding to quick changes in competitive or regulatory contexts. Rather than being the obstructionist team hiding behind the shield of a one-six-fits-all, hub-and-spoke architecture, the DM architects can quickly and successfully accomodate custom, turnkey and proprietary system.

Q. 8. Write short notes on the following :

(i) Rough sets

(ii) Clustering techniques

(iii) 3-Tier data warehouse architecture

Ans. (i) Rough Sets : Rough sets can be used for classification to discover structured relationships within imprecise/noisy data. It applies to discrete-valued attributes continuous-valued attributes must therefore be descriptive before its use.

Rough set theory is based on the establishment of the equivalence classes within the given training data. All of the data tuples forming an equivalence class are indiscernable, that is, the samples are identical with respect to the attributes describing the data.

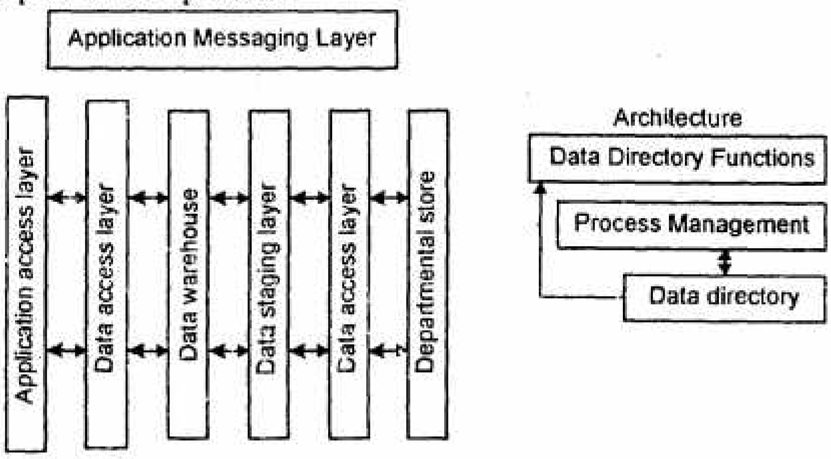
Rough sets can also be used for attributes subset collection where attributes that do not contribute towards the classification of the given training and relevance analysis or significance of each attribute is assessed with respect to the classification task.

However, algorithms to reduce the computation intensity have been proposed. Rather than searching on the entire irg. set, the matrix is instead searched to detect redundant attributes.

(ii) Clustering Techniques : In some cases, it is difficult or impossible to define the parameters of a class of data to be analyzed. When parameters are elusive, clustering methods can be used to create partitions so that all members of each set are similar according to some metric or set of metrics. A cluster is simply a set of objects grouped together by virtue of their similarity or proximity to each other. For instance, a clustering approach might be used to mine credit card purchase data to discover that meals charged on a business-issued gold card are typically purchased on weekdays and have a mean value of greater than $ 250, whereas meals purchased using a personal platinum card occur predominately on week-ends have a mean value of $ 175 & include a bottle of wine more than 65 percent of the time.

Clustering processes can be based on a particular event, such as the cancellation, of a credit card by a customer. By analysing the characteristics of members of this class, clustering might derive certain rules, that could assist the credit card issuer in reducing the number of card cancellations in future.

(iii) 3-Tier Data Warehouse Architecture : A data warehouse architecture (DWA) is a method by which the overall structure of data, communication, processing and presentation for end-user computing within the enterprise can be represented.



The operational and external database layer represents the source data for the DW. This layer comprises, primarily, operational transactions processing systems and external secondary databases. The goal of the data warehouse is to free the information locked up in the operational databases and to mix it with information from other, often external sources. An additional objective of the DW is to have a minimal impact on the performance & operation of the systems found in this layer.

Information Access Layer : The end user deals directly with the information access layer of the DWA. In particular, it represents the tools that the end user normally uses day to day to extract & analyse the data contained within the DW.

Data Access Layer : Serves as a sort of interface or intermediary between the operational and information access layers and DW itself.

Meta Data Layer : In order to provide for universal data access, it is absolutely necessary to maintain some form of data directory or repository of metadata information.

Process Management Layer : The process management layer focuses on scheduling the various tasks that must be accomplished to build and maintain the data warehouse and data directory information.

Application Messaging Layer : The application messaging layer transports information around the enterprise computing networks. This layer is also referred to as "middleware", but it can typically involve more that just networking protocols and request routing.