MAHARASHTRA STATE BOARD OF TECHNICAL EDUCATION
(Autonomous)
(ISO/IEC - 27001 - 2005 Certified)

MODEL ANSWER

SUMMER– 17 EXAMINATION

Subject Title: DATA WAREHOUSING AND DATA MINING **Subject Code:** 17520

Important Instructions to examiners:

1) The answers should be examined by key words and not as word-to-word as given in the model answer scheme.
2) The model answer and the answer written by candidate may vary but the examiner may try to assess the understanding level of the candidate.
3) The language errors such as grammatical, spelling errors should not be given more Importance (Not applicable for subject English and Communication Skills.
4) While assessing figures, examiner may give credit for principal components indicated in the figure. The figures drawn by candidate and model answer may vary. The examiner may give credit for any equivalent figure drawn.
5) Credits may be given step wise for numerical problems. In some cases, the assumed constant values may vary and there may be some difference in the candidate's answers and model answer.
6) In case of some questions credit may be given by judgement on part of examiner of relevant answer based on candidate's understanding.
7) For programming language papers, credit may be given to any other program based on equivalent concept.

Q. No	Sub Q. N.	Answer	Marking Scheme
1.	**a)**	**Attempt any three of the following :**	**(4x3= 12) Marks**
	a)	**Describe any four needs of data warehousing.**	**4M**
	Ans:	1) **Advanced query processing**: in most businesses, even the best database systems are bound to either a single server or a handful of servers in a cluster. A data warehouse is a purpose built hardware solution far more advanced than standard database servers. What this means is a data warehouse will process queries much faster and more effectively, leading to efficiency and increased productivity. 2) **Better consistency of data:** developers work with data warehousing systems after data has been received so that all the information contained in the data warehouse is standardized. Only uniform data can be used efficiently for successful comparisons. Other solutions simply cannot match a data warehouse's level of consistency. 3**) Improved user access**: a standard database can be read and manipulated by programs like SQL Query Studio or the Oracle client, but there is considerable ramp up time for end users to effectively use these apps to get what they need. Business intelligence and data warehouse end-user access tools are built specifically for the purposes data warehouses are used: analysis, benchmarking, prediction and more.	**(Any 4 Need: 1 Mark Each)**

4**) All-in-one**: a data warehouse has the ability to receive data from many different sources, meaning any system in a business can contribute its data. Let's face it: different business segments use different applications. Only a proper data warehouse solution can receive data from all of them and give a business the "big picture" view that is needed to analyze the business, make plans, track competitors and more.

5) **Future-proof**: a data warehouse doesn't care where it gets its data from. It can work with any raw information and developers can "massage" any data it may have trouble with. Considering this, you can see that a data warehouse will outlast other changes in the business' technology. For example, a business can overhaul its accounting system, choose a whole new CRM solution or change the applications it uses to gather statistics on the market and it won't matter at all to the data warehouse. Upgrading or overhauling apps anywhere in the enterprise will not require subsequent expenditures to change the data warehouse side.

6) **Retention of data history**: end-user applications typically don't have the ability, not to mention the space, to maintain much transaction history and keep track of multiple changes to data. Data warehousing solutions have the ability to track all alterations to data, providing a reliable history of all changes, additions and deletions. With a data warehouse, the integrity of data is ensured.

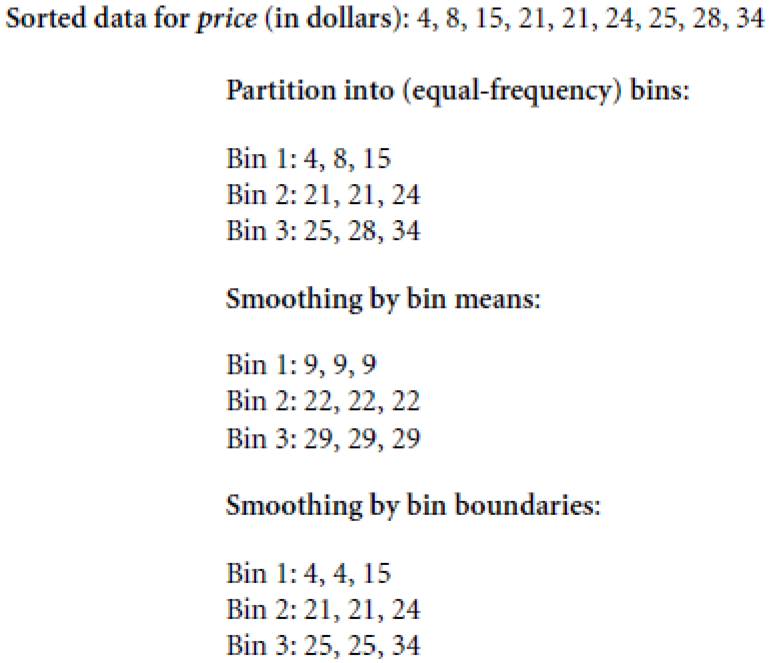
7**) Disaster recovery implications**: a data warehouse system offers a great deal of security when it comes to disaster recovery. Since data from disparate systems is all sent to a data warehouse, that data warehouse essentially acts as another information backup source. Considering the data warehouse will also be backed up, that's now four places where the same information will be stored: the original source, its backup, the data warehouse and its subsequent backup. This is unparalleled information security.

(b) **What is data cleaning technique? Explain any one technique in detail.** **4 M**

Ans: Data cleaning is performed as data preprocessing step while preparing the data for a data warehouse. Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. Noise is a random error or variance in a measured variable. Given a numerical attribute such as, say, price, we can "smooth" out the data to remove the noise by the following data smoothing techniques:

(Definition: 2 marks, Explanation of any one: 2 marks)

1. Binning: Binning methods smooth a sorted data value by consulting its "neighborhood," that is, the values around it. The sorted values are distributed into a number of "buckets," or bins. Because binning methods consult the neighborhood of values, they perform local smoothing. Figure 3.4 illustrates some binning techniques. In this example, the data for price are first sorted and then partitioned into equal-frequency bins of size 3 (i.e., each bin contains three values). In smoothing by bin means, each value in a bin is replaced by the mean value of the bin. For example, the mean of the values 4, 8, and 15 in Bin 1 is 9. Therefore, each original value in this bin is replaced by the value 9. Similarly, smoothing by bin medians can be employed, in which each bin value is replaced by the bin median. In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value. In general, the larger the width, the greater the effect of the smoothing. Alternatively, bins may be equal-width, where the interval range of values in each bin is constant.



Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Figure: Binning

2. Regression: Data can be smoothed by fitting the data to a function, such as with regression. Linear regression involves finding the "best" line to fit two attributes (or variables), so that one attribute can be used to predict the other. Multiple linear regression is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.

3. Clustering: Outliers may be detected by clustering, where similar values are organized into groups, or "clusters." Intuitively, values that fall outside of the set of clusters may be considered outliers.

	c)	**Describe Multidimensional data model.**	**4 M**
	Ans:	In the multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by concept hierarchies. This organization provides users with the flexibility to view data from different perspectives. A number of OLAP data cube operations exist to materialize these different views, allowing interactive querying and analysis of the data at hand. Hence, OLAP provides a user-friendly environment for interactive data analysis.	**(Diagram: 2 marks, Explanation: 2 marks)**

Example: OLAP operations:

Each of the operations described below is illustrated in below Figure. At the center of the figure is a data cube for *All Electronics* sales. The cube contains the dimensions *location, time*, and *item*, where *location* is aggregated with respect to city values, *time* is aggregated with respect to quarters, and *item* is aggregated with respect to item types.

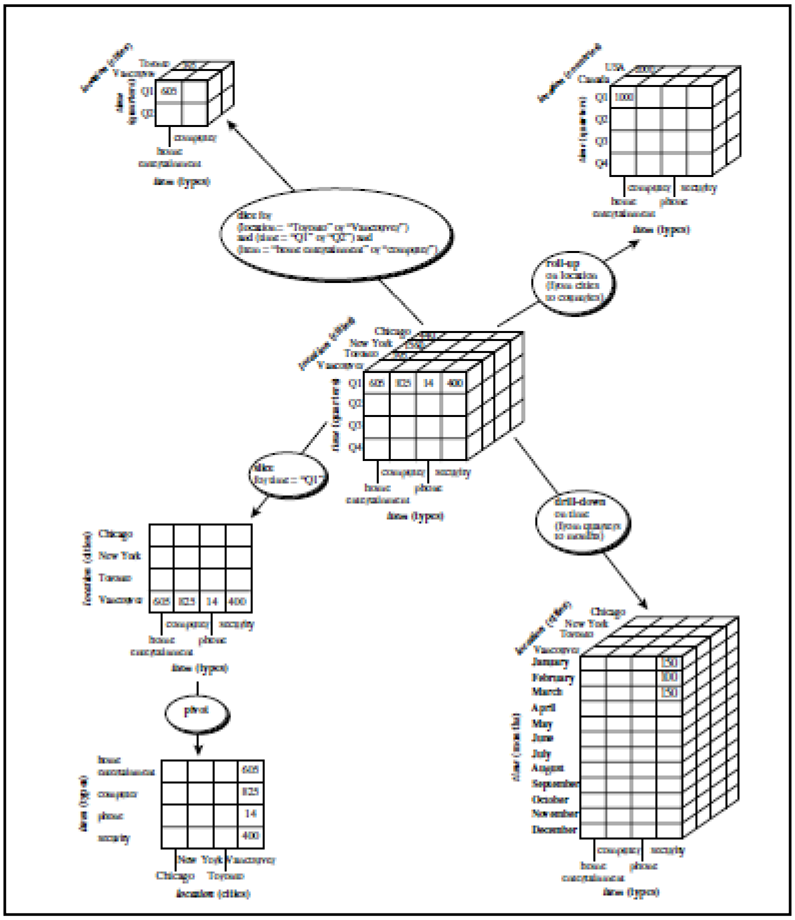


Fig: Example of typical OLAP operation on Multidimensional data

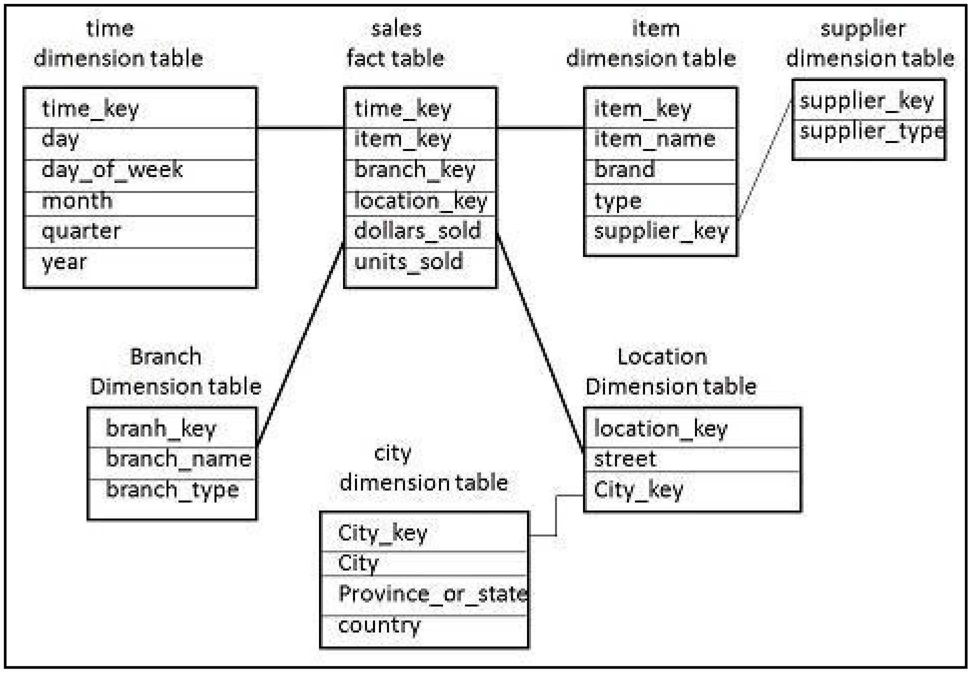
This cube is referred to as the central cube. The measure displayed is *dollars sold* (in thousands). The data examined are for the cities Chicago, New York, Toronto, and Vancouver.

	d)	**What is concept description?**	**4 M**
	Ans:	**Concept Description:** Concept description refers to the data to be associated with the classes or concepts. For example, in a company, the classes of items for sales include computer and printers, and concepts of customers include big spenders and budget spenders. Such descriptions of a class or a concept are called class/concept descriptions. These descriptions can be derived by the following two ways. 1. **Data Characterization:** this refers to summarizing data of class under study. This class under study is called as Target Class. **2. Data Discrimination:** It refers to the mapping or classification of a class with some predefined group or class. The simplest kind of descriptive data mining is concept description. A concept usually refers to a collection of data such as frequent buyers, graduate students, and so on. As a data mining task, concept description is not a simple enumeration of the data. Instead, concept description generates descriptions for characterization and comparison of the data. It is sometimes called class description, when the concept to be described refers to a lass of objects. Characterization provides a concise and succinct summarization of the given collection of the data, while concept or class comparison (also known as discrimination) provides discriminations comparing two or more collections of data. Since concept description involves both characterization and comparison, techniques for accomplishing each of these tasks will study. Concept description has close ties with the data generalization. Given the large amount of data stored in database, it is useful to be describe concepts in concise and succinct terms at generalized at multiple levels of abstraction facilities users in examining the general behavior of the data. Given the AB Company database, for example, instead of examining individual customer transactions, sales managers may prefer to view the data generalized to higher levels, such as summarized to higher levels, such as summarized by customer groups according to geographic regions, frequency of purchases per group, and customer income. Such multiple dimensional, multilevel data generalization is similar to multidimensional data analysis in data warehouses.	**(Definition: 2 marks, Concepts: 2 marks)**
	b)	**Attempt any one of the following :**	**(6X1=6) Marks**
	i)	**Describe any six characteristics of data warehouse.**	**6 M**
	Ans:	1) **Subject Oriented:** Data warehouses are designed to help you analyze data. For example, to learn more about your company's sales data, you can build a warehouse that concentrates on sales. Using this warehouse, you can answer questions like "Who was our best customer for this item last year?" This ability to define a data warehouse by subject matter, sales in this case makes the data warehouse subject oriented. 2) **Integrated:** Integration is closely related to subject orientation. Data warehouses must put data from disparate sources into a consistent format. They must resolve such problems as naming conflicts and inconsistencies among units of measure.	**(Each Characteristics: 1 mark Each, any 6)**

		When they achieve this, they are said to be integrated. 3) **Nonvolatile:** Nonvolatile means that, once entered into the warehouse, data should not change. This is logical because the purpose of a warehouse is to enable you to analyze what has occurred. 4) **Time Variant:** In order to discover trends in business, analysts need large amounts of data. This is very much in contrast to online transaction processing (OLTP) systems, where performance requirements demand that historical data be moved to an archive. A data warehouse's focus on change over time is what is meant by the term time variant. Typically, data flows from one or more online transaction processing (OLTP) databases into a data warehouse on a monthly, weekly, or daily basis. The data is normally processed in a staging file before being added to the data warehouse. Data warehouses commonly range in size from tens of gigabytes to a few terabytes. Usually, the vast majority of the data is stored in a few very large fact tables. 5) **Separate:** The DW is separate from the operational systems in the company. It gets its data out of these legacy systems. 6) **Available:** The task of a DW is to make data accessible for the user. 7) **Aggregation performance:** The data which is requested by the user has to perform well on all scales of aggregation. 8) **Consistency:** Structural and contents of the data is very important and can only be guaranteed by the use of metadata: this is independent from the source and collection date of the data	
	ii)	**What is data reduction? State its different techniques.**	**6 M**
	Ans:	Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results. Strategies for data reduction include the following: **1. Data cube aggregation**: where aggregation operations are applied to the data in the construction of a data cube. **2. Attribute subset selection**: where irrelevant, weakly relevant or redundant attributes or dimensions may be detected and removed. **3. Dimensionality reduction**: where encoding mechanisms are used to reduce the data set size.	**(Definition: 2 marks, Explanation of any 4 techniques : 1 mark each)**

		4. Numerosity reduction: where the data are replaced or estimated by alternative, smaller data representations such as parametric models (which need store only the model parameters instead of the actual data) or nonparametric methods such as clustering, sampling, and the use of histograms. **5. Discretization and concept hierarchy generation**: where raw data values for attributes are replaced by ranges or higher conceptual levels. Data discretization is a form of numerosity reduction that is very useful for the automatic generation of concept hierarchies. Discretization and concept hierarchy generation are powerful tools for data mining, in that they allow the mining of data at multiple levels of abstraction	
2.		**Answer any two of the following :**	**(8X2=16) Marks**
	(a)	**Describe discretization and concept of hierarchy generation for numeric and categorical data.**	**8M**
	Ans:	Concept hierarchies for numeric attributes can be constructed automatically based on data distribution analysis. Five methods for concept hierarchy generation are defined below- inning histogram analysis entropy-based discretization and data segmentation by natural partitioning **Binning:** Attribute values can be discretized by distributing the values into bin and replacing each bin by the mean bin value or bin median value. These technique can be applied ecursively to the resulting partitions in order to generate concept hierarchies. **Histogram Analysis**: Histograms can also be used for discretization. Partitioning rules can be applied to define range of values. The histogram analyses algorithm can be applied recursively to each partition in order to automatically generate amultilevel concept hierarchy, with the procedure terminating once a pre specified number of concept levels have been reached. A minimum interval size can be used per level to control the recursive procedure. This specifies the minimum width of the partition, or the minimum member of partitions at each level. **Cluster Analysis**: A clustering algorithm can be applied to partition data into clusters or groups. Each cluster forms a node of a concept hierarchy, where all noses are at thesame conceptual level. Each cluster may be further decomposed into sub-clusters, forming a lower level in the hierarchy. Clusters may also be grouped together to form a higher-level concept hierarchy. Segmentation by natural partitioning: Breaking up annual salaries in the range of into ranges like ($50,000-$100,000) are often more desirable than ranges like ($51, 263, 89-$60,765.3) arrived at by cluster analysis. The 3-4-5 rule can be used to segment numeric data into relatively uniform —natural intervals. In general the rule partitions agive range of data into 3,4,or 5 equinity intervals, recursively level by level based on value range at the most significant digit. The rule can be recursively applied to each interval creating a concept hierarchy for the given numeric attribute.	**(Numeric data: 4 marks and categorical data 4 marks)**

		Categorical data are discrete data. Categorical attributes have finite number of distinct values, with no ordering among the values, examples include geographic location, item type and job category. There are several methods for generation of concept hierarchies for categorical data. Specification of a partial ordering of attributes explicitly at the schema level by experts. Concept hierarchies for categorical attributes or dimensions typically involve a group of attributes. A user or an expert can easily define concept hierarchy by specifying a partial or total ordering of the attributes at a schema level. A hierarchy can be defined at the schema level such as street < city < province <state < country. Specification of a portion of a hierarchy by **explicit data grouping**: This is identically a manual definition of a portion of a concept hierarchy. In a large database, is unrealistic to define an entire concept hierarchy by explicit value enumeration. However, it is realistic to specify explicit groupings for a small portion of the intermediate-level data. Specification of a set of attributes but not their partial ordering: A user may specify a set of attributes forming a concept hierarchy, but omit to specify their partial ordering. The system can then try to automatically generate the attribute ordering so as to construct a meaningful concept hierarchy.	
	b)	**Describe the following schemes for multidimensional database.** **1)Star** **2)Snowflake**	**8M**
	Ans:	**Star Schema**: •In star schema each dimension is represented with only one dimension table. •This dimension table contains the set of attributes. •In the following diagram we have shown the sales data of a company with respect to the four dimensions namely, time, item, branch and location. •There is a fact table at the center. This fact table contains the keys to each of four dimensions. •The fact table also contain the attributes namely, dollars sold and units sold. **Snowflake Schema**: •In Snowflake schema some dimension tables are normalized. •Dimensions with hierarchies can be decomposed into a snowflake structure when you want to avoid joins to big dimension tables when you are using an aggregate of the fact table. •The normalization split up the data into additional tables. •Unlike Star schema the dimensions table in snowflake schema is normalized for example the item dimension table in star schema is normalized and split into two dimension tables namely, item and supplier table.	**(4 marks each)**



•Therefore now the item dimension table contains the attributes item key, item name, type, brand, and supplier-key.
•The supplier key is linked to supplier dimension table. The supplier dimension table contains the attributes supplier key, and supplier type.

c) State the association rules in data mining. Write applications of each rule. **8M**

Ans:

(Statement of Rule: 4 marks, Application: 4 marks)

Similar to the mining of association rules in transactional and relational databases, spatial association rules can be mined in spatial databases. A spatial association rule is of the form A)B [s%;c%] where A and B are sets of spatial or nonspatial predicates, s% is the support of the rule, and c% is the confidence of the rule. For example, the following is a spatial association rule: is a(X; “school”)^close to(X; “sports center”))close to(X; “park”) [0:5%;80%]. This rule states that 80% of schools that are close to sports centers are also close to parks, and 0.5% of the data belongs to such a case. Various kinds of spatial predicates can constitute a spatial association rule. Examples include distance information (such as close to and far away), topological relations (like intersect, overlap, and disjoint), and spatial orientations (like left of and west of). Since spatial association mining needs to evaluate multiple spatial relationships among a large number of spatial objects, the process could be quite costly. An interesting mining optimization method called progressive refinement can be adopted in spatial association analysis. The method first mines large data sets roughly using a fast algorithm and then improves the quality of mining in a pruned data set using a more expensive algorithm. To ensure that the pruned data set covers the complete set of answers when applying the high-quality data mining algorithms at a later stage, an important requirement for the rough mining algorithm applied in the early stage is the superset coverage property: that is, it preserves all of the potential answers. In other words, it should allow a false-positive test, which might include some data sets that do not belong to the answer sets, but it should not allow a false-negative test, which might exclude some potential answers. For mining spatial associations related to the spatial predicate close to, we can first collect the candidates that pass the minimum support threshold by Applying certain rough spatial evaluation algorithms, for example, using an MBR structure (which registers only two spatial points rather than a set of complex polygons), and evaluating the relaxed spatial predicate, g close to, which is a generalized close to covering a broader context that includes close to, touch, and intersect. If two spatial objects are closely located, their enclosing MBRs must be closely located, matching g close to. However, the reverse is not always true: if the enclosing MBRs are closely located, the two spatial objects may or may not be located so closely. Thus, the MBR pruning is a false-positive testing tool for closeness: only those that pass the rough test need to be further examined using more expensive spatial computation algorithms. With this preprocessing, only the patterns that are frequent at the approximation level will need to be examined by more etailed and finer, yet more expensive, spatial computation. Besides mining spatial association rules, one may like to identify groups of particular features that appear frequently close to each other in a geospatial map. Such a problem is essentially the problem of mining spatial co-locations. Finding spatial co-locations can be considered as a special case of mining spatial associations. However, based on the property of spatial autocorrelation, interesting features likely coexist in closely located regions. Thus spatial co-location can be just what one really wants to explore. Efficient methods can be developed for mining spatial co-locations by exploring the methodologies like Aprori and progressive refinement, similar to what has been done for mining spatial association rules. Mining Associations in Multimedia Data Association rules involving multimedia objects can be mined in image and video databases. At least three categories can be observed:

Associations between image content and non-image content features: A rule like “If at least 50% of the upper part of the picture is blue, then it is likely to represent sky” belongs to this category since it links the image content to the keyword sky. Associations among image contents that are not related to spatial relationships: A rule like “If a picture contains two blue squares, then it is likely to contain one red circle As well belongs to this category since the associations are all regarding image contents. Associations among image contents related to spatial relationships: A rule like “If a red triangle is between two yellow squares, then it is likely a big oval-shaped object is underneath” belongs to this category since it associates objects in the image with spatial relationships.

3.		**Answer any four of the following :**	**(4x4=16) Marks**
	(a)	**Describe Decision support system.**	**4 M**
	Ans:	Decision support systems are interactive software-based systems intended to help managers in decision making by accessing large volume of information generated from various related information systems involved in organizational business processes, like, office automation system, transaction processing system etc. DSS uses the summary information, exceptions, patterns and trends using the analytical models. Decision Support System helps in decision making but does not always give a decision itself. The decision makers compile useful information from raw data, documents, personal knowledge, and/or business models to identify and solve problems and make decisions. Programmed and Non-programmed Decisions There are two types of decisions - programmed and non-programmed decisions. Programmed decisions are basically automated processes, general routine work, where: •These decisions have been taken several times •These decisions follow some guidelines or rules Non-programmed decisions occur in unusual and non-addressed situations, so It would be a new decision There will not be any rules to follow •These decisions are made based on available information •These decisions are based on the manger's discretion, instinct, perception and judgment Decision support systems generally involve non-programmed decisions. Therefore, there will be no exact report, content or format for these systems.	**(Definition: 1 mark, Any 3 type: 1 mark each)**
	b)	**Explain benefits of data warehousing.**	**4 M**
	Ans:	Benefits of data warehousing Data warehouse usage includes: • Locating the right info • Presentation of info • Testing of hypothesis • Discovery of info Sharing the analysis The benefits can be classified into two: •Tangible benefits (quantified / measureable):It includes, o Improvement in product inventory o Decrement in production cost o Improvement in selection of target markets o Enhancement in asset and liability management •Intangible benefits (not easy to quantified): It includes, o Improvement in productivity by keeping all data in single location and	**(Explanation of any 2 benefits:2 mark each)**

		eliminating rekeying of data o Reduced redundant processing Enhanced customer relation	
	c)	**Describe need for OLAP.**	**4 M**
	Ans:	OLAP (online analytical processing) is a function of business intelligence software that enables a user to easily and selectively extract and view data from different points of view. OLAP technology is a vast improvement over traditional relational database management systems (RDBMS). Relational databases, which have a two-dimensional structure, do not allow the multidimensional data views that OLAP provides. Traditionally used as an analytical tool for marketing and financial reporting, OLAP is now viewed as a valuable tool for any management system that needs to create a flexible decision support system. Today's work environment is characterized by flatter organizations that need to be able to adapt quickly to changing conditions. Managers need the tools that will allow them to make quick, intelligent decisions on the fly. Making the wrong decision or taking too long to make it can affect the competitive position of an organization. OLAP provides the multidimensional capabilities that most organizations need today. By using a multidimensional data store, also known in the industry as a hypercube, OLAP allows the end user to analyze data along the axes of their business. The two most common forms of analysis that most businesses use are called "slice and dice" and "drill down".	**(Any 4 need, Each need carries: 1 mark)**
	d)	**Describe market basket analysis.**	**4 M**
		Market basket analysis is a technique that discovers relationships between pairs of products purchased together. The technique can be used to uncover interesting cross-sells and related products. The idea behind market basket analysis is simple. Simply examine your orders for products have been purchased together. For example using market basket analysis you might un cover the fact that customers tend to buy hot dogs and buns together. Using this information you might organize the store so that hot dogs and buns are next to each other. In an e-commerce environment you might create a cross-sell rule to offer the shopper buns whenever they place hot dogs in their shopping cart. There are a couple of measures we use when doing market basket analysis and they are described here. The first measure is the frequency. The frequency is defined as the number of times two products were purchased together. If hot dogs and buns were found together in 820 baskets this would be its frequency. Frequency by itself doesn’t tell the whole story. For instance if I told you hot dogs and buns were purchased 820 times together you wouldn’t know if that was relevant or not. Therefore we introduce two other measures called support and confidence to help with the analysis. If you divide the frequency by the total number of orders you get the percentage of order containing the pair. This is called the support. Another way to thinking about support is as the probability of the pair being purchased. Now if 820 hot dogs and buns were purchased together and your store took 1000 orders the support for this would be calculated as (820 / 1000) = 82.0% .We can extend this even further by	**(Problem: 2 marks, Technique for remove problem any 2,1 mark each)**

defining a calculation called confidence. Confidence compares the number of times the pair was purchased to the number of times one of the items in the pair was purchased. In probability terms this is referred to as the conditional probability of the pair. Sogoing back to our hot dogs example if hot dogs were purchases 900 times and out of those 900 purchases 820 contained buns we would have a confidence of (820 / 900) = 91.1%.Now that we've defined frequency, support and confidence we can talk a little about what a market basket analysis report might look like. The report would have the user select the product they are interested in performing the analysis on (i.e. hot dogs). Then it would list all the products that were purchased together with the selected products ranked by it frequency. It might look something like the following

e) Describe the method of data preprocessing with its block diagram. **4 M**

Ans:

(Method Explanation :2 marks, Block Diagram: 2 marks)

Data cleaning techniques Objectives

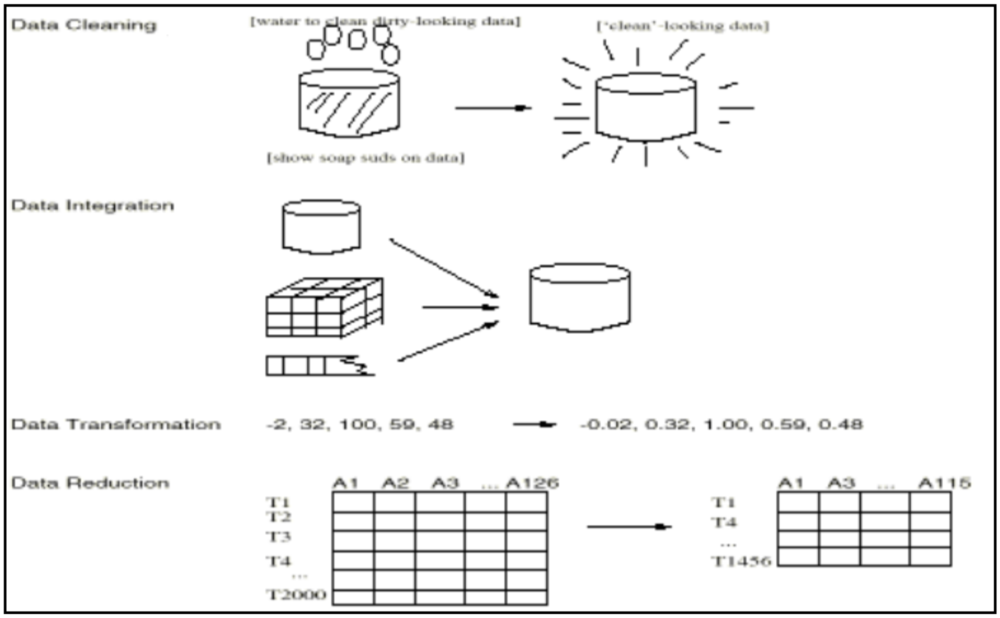
•Incomplete: Lacking attribute values, lacking certain attributes of interest, or containing only aggregate data: e.g., occupation=""
•Noisy: Containing errors or outliers e.g., Salary="-10"
•Inconsistent: Containing discrepancies in codes or names e.g., Age="42" Birthday= "03/07/1997" e.g., Was rating "1,2,3", now rating "A, B, C" e.g., discrepancy between duplicate records

Why Is Data Dirty?
•Incomplete data comes from n/a data value when collected Different consideration between the time when the data was collected and when it is analyzed.

- Human/hardware/software problems
- Noisy data comes from the process of data
- Collection Entry
- Transmission
- Inconsistent data comes from
- Different data sources
- Functional dependency violation

Major Tasks in Data Processing

•Data cleaning
• Fill in missing values, smooth noisy data, identify or remove outliers, and resolve Inconsistencies
•Data integration Integration of multiple databases, data cubes, or files
•Data transformation Normalization and aggregation
•Data reduction Obtains reduced representation in volume but produces the same or similar analytical results
•Data discretization Part of data reduction but with particular importance, especially for numerical data



4. a) Answer any three of the following : **(4x3=12) Marks**

i) Explain Categories and classes of DSSs **4 M**

Ans: DSS have been classified in different ways as the concept matured with time. As. and when the full potential and possibilities for the field emerged, different classification systems also emerged. Some of the well-known classification mode ls are given below. According to Donovan and Mad nick (1977) DSS can be classified as:

(Classes: 2 marks (any 2), Categories: 2 marks (any 2))

1). Institutional-when the DSS supports ongoing and recurring decisions
2). Ad hoc-when the DSS supports a one off-kind of decision. Hack thorn and Keen (1981)

Classified DSS as.
1). Personal DSS
2). Group DSS
3). Organizational DSS

Alter (1980) opined that decision support systems could be classified into seven types based on their generic nature of operations. He described the seven types as.

1). File drawer systems. This type of DSS primarily provides access to data stores/data

		related items. Examples--ATM Machine, Use the balance to make transfer of funds decisions 2). Data analysis systems. This type of DSS supports the manipulation of data through the use of specific or generic computerized settings or tools. Examples: Airline Reservation system, use the info to make flight plans 3). Analysis information systems. This type of DSS provides access to sets of decision oriented databases and simple small models. 4). Accounting and financial models. This type of DSS can perform 'what if analysis' and calculate the outcomes of different decision paths. Examples: calculate production cost.	
	ii)	**Explain Mining Text Databases.**	**4 M**
	Ans:	A substantial portion of the available information is stored in text databases (or document databases), which consist of large collections of documents from various sources, such as news articles, research papers, books, digital libraries, e-mail messages, and Web pages. Text databases are rapidly growing due to the increasing amount of information available in electronic form, such as electronic publications, various kinds of electronic documents, e-mail, and the World Wide Web (which can also be viewed as a huge, interconnected, dynamic text database). Nowadays most of the information in government, industry, business, and other institutions are stored electronically, in the form of text databases. Data stored in most text databases are semi structured data in that they are neither completely unstructured nor completely structured. For example, a document may contain structured fields, such as title, authors, publication date, and category, and so on, but also contain some largely unstructured text components, such as abstract and contents. There have been a great deal of studies on the modelling and implementation of semi structured data in recent database research. Moreover, information retrieval techniques, such as text indexing methods, have been developed to handle unstructured documents. Traditional information retrieval techniques become inadequate for the increasingly vast amounts of text data. Typically, only a small fraction of the many available documents will be relevant to a given individual user. Without knowing what could be in the documents, it is difficult to formulate effective queries for analyzing and extracting useful information from the data. Users need tools to compare different documents, rank the importance and relevance of the documents, or find patterns and trends across multiple documents. Thus, text mining has become an increasingly popular and essential theme in data mining. Text Mining Approaches: There are many approaches to text mining, which can be classified from different perspectives, based on the inputs taken in the text mining system and the data mining tasks to be performed. In general, the major approaches, based on the kinds of data they take as input, are:	**(Explanation : 4 marks)**

1. The keyword-based approach, where the input is a set of keywords or terms in the documents,
2. The tagging approach, where the input is a set of tags, and
3. The information-extraction approach, which inputs semantic information, such as events, facts, or entities uncovered by information extraction.

iii) Draw block diagram of data warehouse architecture and list its components. **4 M**

Ans:

(Block Diagram: 2 marks, List: 2 marks)

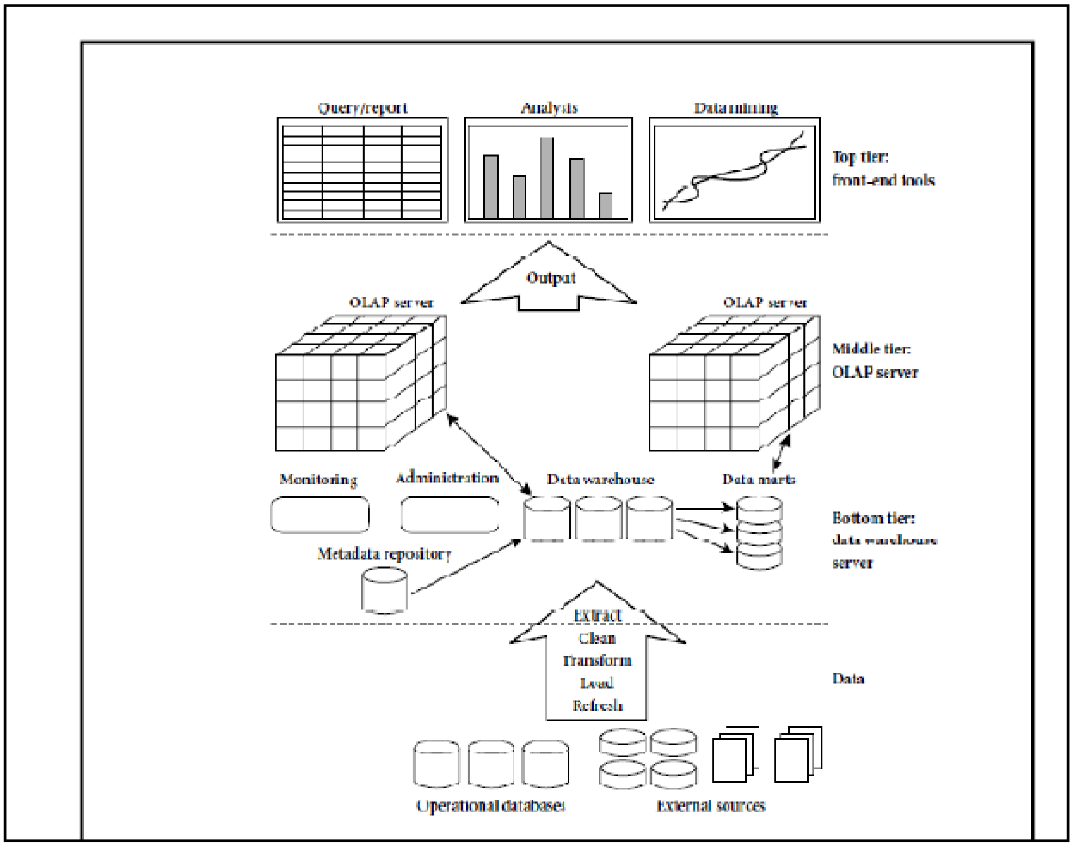
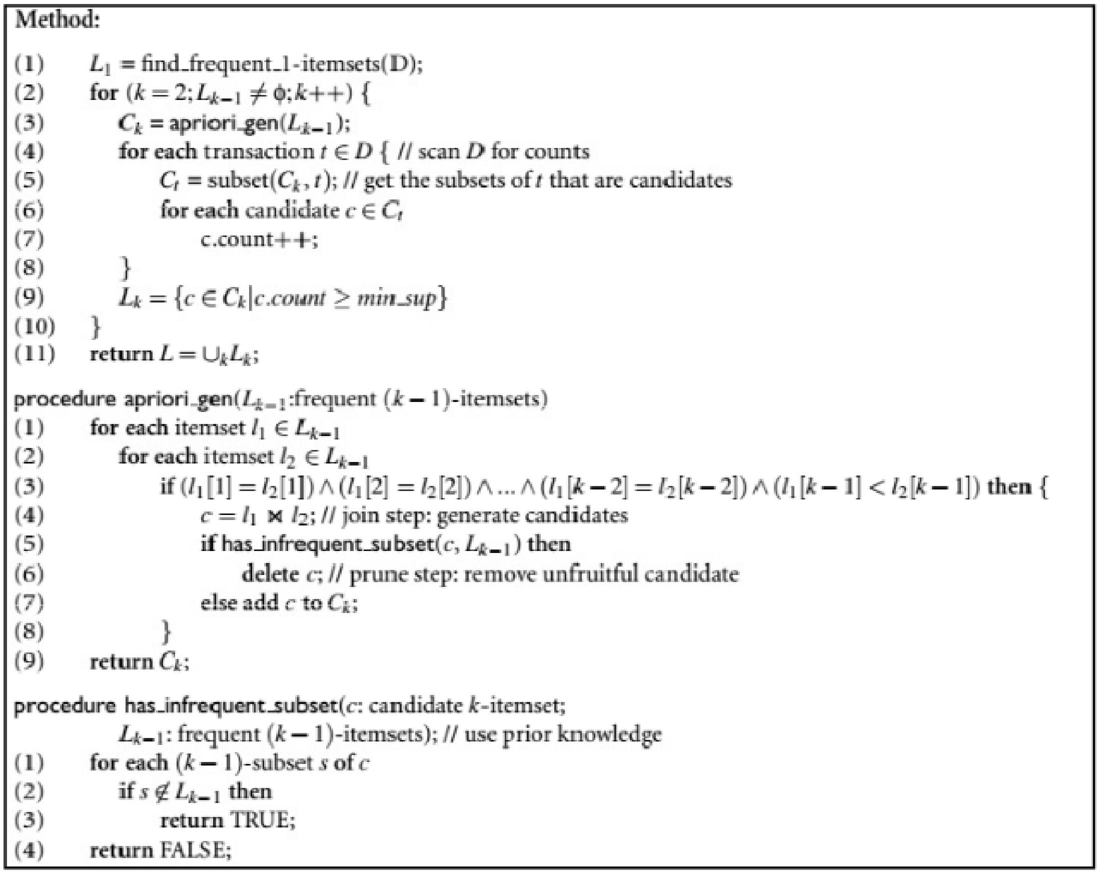


Fig: A three-tire data warehousing architecture.

1). The bottom tier is a warehouse database server that is almost always a relational database system. Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources (such as customer profile information provided by external consultants).These tools and utilities perform data extraction, cleaning, and transformation (e.g., to merge similar data from different Sources into a unified format), as well as load and refresh functions to update the data warehouse. The data are extracted using application program interfaces known as gateways. A gateway is supported by the underlying DBMS and allows client programs to generate SQL code to be executed at a server. Examples of gateways include ODBC

		(Open Database Connection) and OLEDB (Open Linking And Embedding for Databases) by Microsoft and JDBC (Java Database Connection).This tier also contains a metadata repository, which stores information about the data warehouse and its contents. The middle tier is an OLAP server that is typically implemented using either (1) a relational OLAP (ROLAP) model, that is, an extended relational DBMS that maps operations on multidimensional data to standard relational operations; or (2) A multidimensional OLAP (MOLAP) model, that is, a special-purpose server that directly implements multidimensional data and operations. The top tier is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools (e.g., trend analysis, prediction, and so on).	
	iv)	**Describe data integration with an example.**	**4 M**
		Data analysis task will involve data integration, which combines data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes, or flat files. There are a number of issues to consider during data integration. Schema integration and object matching can be tricky. How can equivalent real-world entities from multiple data sources be matched up? This is referred to as the entity identification problem. **Example:** For example, how can the data analyst or the computer be sure that customer id in one database and cust number in another refers to the same attribute? Examples of metadata for each attribute include the name, meaning, data type, and range of values permitted for the attribute, and null rules for handling blank, zero, or null values. Such metadata can be used to help avoid errors in schema integration. The metadata may also be used to help transform the data (e.g., where data codes for pay type in one database may be "H" and "S", and 1 and 2 in another). Hence, this step also relates to data cleaning, Redundancy is another important issue. An attribute (such as annual revenue, for instance) may be redundant if it can be "derived" from another attribute or set of attributes. Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set.	**(Explanation: 2 marks, Example: 2 marks)**

	b)	**Attempt any one of the following :**	**(6x1=6) Marks**
	i)	**Describe the Apriori algorithm.**	**6 M**
	Ans:	Apriori is a seminal algorithm for mining frequent item sets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent Item set properties Apriori employs an iterative approach known as a level-wise search, where k item sets are used to explore (k + 1)-item sets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted L1. Next, L1 is used to find L2 the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-item sets can be found. The finding of each Lk requires one full scan of the database. Once the frequent item sets from transactions in a database D have been found, it is straightforward to generate strong association rules from them (where strong association rules satisfy both minimum support and minimum confidence). This can be done using Equation for confidence, which we show again here for completeness: $\boxed{confidence(A \Rightarrow B) = P(B\|A) = \frac{support\_count(A \cup B)}{support\_count(A)}.}$ Input: D, a database of transactions; Min_sup, the minimum support count threshold Output: L, frequent itemsets in D	**(Description :3 marks, Algorithm :3 marks)**

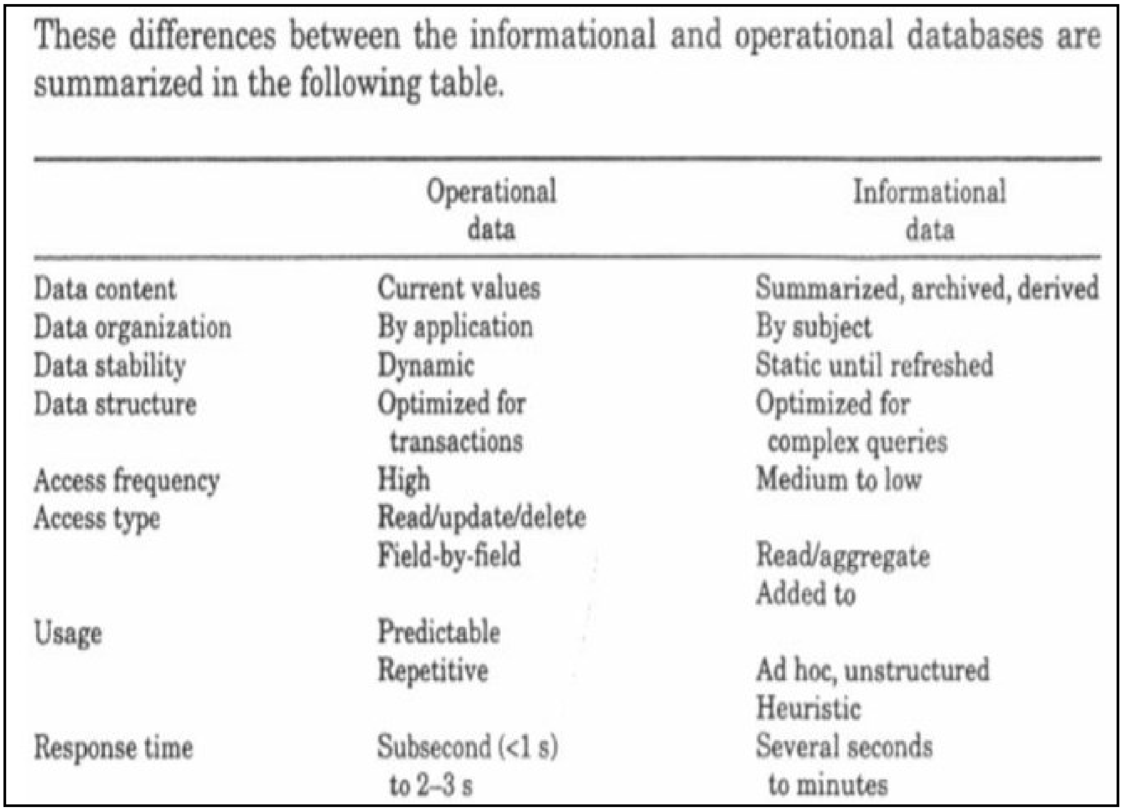


```
Method:
L_1 = find_frequent_1-itemsets(D);
for ($k = 2; L_{k-1} \neq \phi; k$++) {
    $C_k$ = apriori_gen($L_{k-1}$);
    for each transaction $t \in D$ { // scan $D$ for counts
        $C_t$ = subset($C_k, t$); // get the subsets of $t$ that are candidates
        for each candidate $c \in C_t$
            c.count++;
    }
    $L_k = \{c \in C_k | c.count \geq min\_sup\}$
}
return $L = \cup_k L_k$;

procedure apriori_gen($L_{k-1}$:frequent $(k-1)$-itemsets)
    for each itemset $l_1 \in L_{k-1}$
        for each itemset $l_2 \in L_{k-1}$
            if $(l_1[1] = l_2[1]) \wedge (l_1[2] = l_2[2]) \wedge ... \wedge (l_1[k-2] = l_2[k-2]) \wedge (l_1[k-1] < l_2[k-1])$ then {
                $c = l_1 \bowtie l_2$; // join step: generate candidates
                if has_infrequent_subset($c, L_{k-1}$) then
                    delete $c$; // prune step: remove unfruitful candidate
                else add $c$ to $C_k$;
            }
    return $C_k$;

procedure has_infrequent_subset($c$: candidate $k$-itemset;
            $L_{k-1}$: frequent $(k-1)$-itemsets); // use prior knowledge
    for each $(k-1)$-subset $s$ of $c$
        if $s \notin L_{k-1}$ then
            return TRUE;
    return FALSE;
```

	ii)	**Explain operational and informational data.**	**6 M**
	Ans:	Operational Data & Informational Data: Operational Data: • Focusing on transactional function such as bank card withdrawals and deposits • Detailed Updateable • Reflects current data Informational Data: • Focusing on providing answers to problems posed by decision makers • Summarized • Non updateable	**(Explanation :6 marks)**



These differences between the informational and operational databases are summarized in the following table.

	Operational data	Informational data
Data content	Current values	Summarized, archived, derived
Data organization	By application	By subject
Data stability	Dynamic	Static until refreshed
Data structure	Optimized for transactions	Optimized for complex queries
Access frequency	High	Medium to low
Access type	Read/update/delete Field-by-field	Read/aggregate Added to
Usage	Predictable Repetitive	Ad hoc, unstructured Heuristic
Response time	Subsecond (<1 s) to 2–3 s	Several seconds to minutes

5. **Attempt any tow of the following :** **(8x2=16) Marks**

a) Describe four different OLAP operation in the multidimensional model with neat diagram. **8 M**

Ans: In the multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by concept hierarchies. This organization provides users with the flexibility to view data from different perspectives. A number of OLAP data cube operations exist to materialize these different views, allowing interactive querying and analysis of the data at hand. Hence, OLAP provides a user-friendly environment for interactive data analysis.

(Diagram: 2 marks, Points: 2 marks, Explanation :4 marks)

Example: OLAP operations: Each of the operations described below is illustrated in below Figure. At the center of the figure is a data cube for All Electronics sales. The cube contains the dimensions location, time, and item, where location is aggregated with

respect to city values,
time is aggregated with respect to quarters, and item is aggregated with respect to item types.

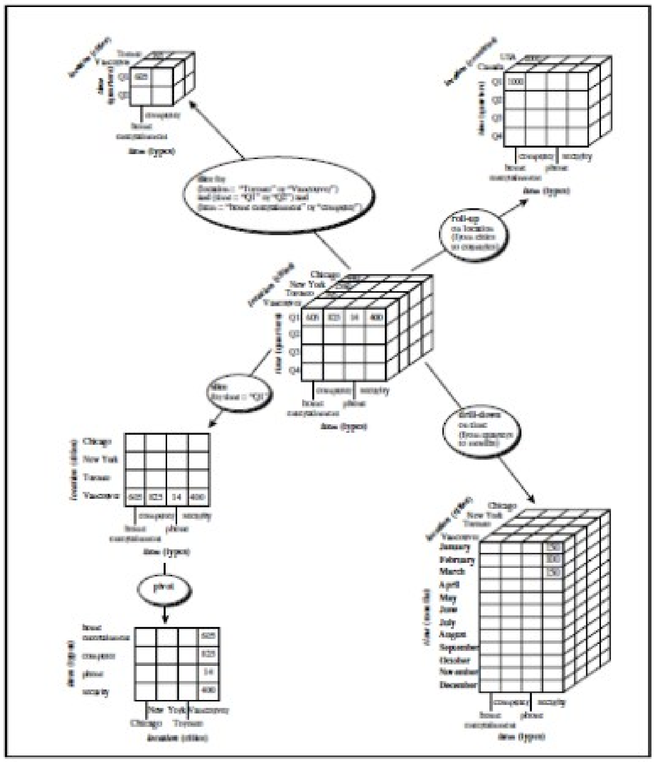


Fig: Example of typical OLAP operation on Multidimensional data

This cube is referred to as the central cube. The measure displayed is dollars sold (in thousands).The data examined are for the cities Chicago, New York, Toronto, and Vancouver. Roll-up: The roll-up operation (also called the drill-up operation by some vendors) performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction. Figure shows the result of a roll-up operation performed on the central cube by climbing up the concept hierarchy for location given in Figure. This hierarchy was defined as the total order “street< city < province or state < country.” The roll-up operation shown aggregates the data by ascending the location hierarchy from the level of city to the level of country. In other words, rather than grouping the data by city, the resulting cube groups the data by country. When roll-up is performed by dimension reduction, one or more dimensions are removed from the given cube. For example, consider a sales data cube containing only the two dimensions location and time. Roll-up may be performed by removing, say, the time dimension,

		resulting in an aggregation of the total sales by location, rather than by location and by time. Drill-down: Drill-down is the reverse of roll-up. It navigates from less detailed data to more detailed data. Drill-down can be realized by either stepping down a concept hierarchy for a dimension or introducing additional dimensions. Figure shows the result of a drill-down operation performed on the central cube by stepping down a concept hierarchy for time defined as “day < month < quarter < year.” Drill-down occurs by descending the time hierarchy from the level of quarter to the more detailed level of month. The resulting data cube details the total sales per month rather than summarizing them by quarter. Because a drill-down adds more detail to the given data, it can also be performed by adding new dimensions to a cube. For example, a drill-down on the central cube of Figure can occur by introducing an additional dimension, such as customer group. Slice and dice: The slice operation performs a selection on one dimension of the given cube, resulting in a subcube. Figure shows a slice operation where the sales data are selected from the central cube for the dimension time using the criterion time = “Q1”. The dice operation defines a sub cube by performing a selection on two or more dimensions. Figure shows a dice operation on the central cube based on the following selection criteria that involve three dimensions: (location = “Toronto” or “Vancouver”) and (time = “Q1” or “Q2”) and (item =“home entertainment” or “computer”). Pivot (rotate):Pivot (also called rotate) is a visualization operation that rotates the data axes in view in order to provide an alternative presentation of the data. Figure shows a pivot operation where the item and location axes in a 2-D slice are rotated. Other examples include rotating the axes in a 3-D cube, or transforming a 3-D cube into a series of 2-D planes. Other OLAP operations: Some OLAP systems offer additional drilling operations. For example, drill-across executes queries involving (i.e., across) more than one fact table. The drill-through operation uses relational SQL facilities to drill through the bottom level of a data cube down to its back-end relational tables. Other OLAP operations may include ranking the top N or bottom N items in lists, as well as computing moving averages, growth rates, and interests, internal rates of return, depreciation, currency conversions, and statistical functions. OLAP offers analytical modelling capabilities, including a calculation engine for deriving ratios, variance, and so on, and for computing measures across multiple dimensions. It can generate summarizations, aggregations, and hierarchies at each granularity level and at every dimension intersection. OLAP also supports functional models for forecasting, trend analysis, and statistical analysis. In this context, an OLAP engine is a powerful data analysis tool.	
	b)	**State the following mining techniques.** **i) Constraint based association mining** **ii) Sequential mining**	**8 M**
	Ans:	**i) Constraint based association mining:** A data mining process may uncover thousands of rules from a given set of data, most of which end up being unrelated or uninteresting to the users. Often, users have a good sense of which “direction” of mining may lead to interesting patterns and the “form” of	**(4 Marks each)**

the patterns or rules they would like to find. Thus, a good heuristic is to have the users specify such intuition or expectations as constraints to confine the search space. This strategy is known as constraint-based mining. The constraints can include the following: Knowledge type constraints: These specify the type of knowledge to be mined, such as association or correlation. Data constraints: These specify the set of task-relevant data. Dimension/level constraints: These specify the desired dimensions (or attributes) of the data, or levels of the concept hierarchies, to be used in mining. Interestingness constraints: These specify thresholds on statistical measures of rule interestingness, such as support, confidence, and correlation. Rule constraints: These specify the form of rules to be mined. Such constraints may be expressed as Meta rules (rule templates), as the maximum or minimum number of predicates that can occur in the rule antecedent or consequent, or as relationships among attributes, attribute values, and/or aggregates. The above constraints can be specified using a high-level declarative data mining query language and user interface. The first four of the above types of constraints have already been addressed in earlier parts of this book and chapter. In this section, we discuss the use of rule constraints to focus the mining task. This form of constraint-based mining allows users to describe the rules that they would like to uncover, thereby making the data mining process more effective. In addition, a sophisticated mining query optimizer can be used to exploit the constraints specified by the user, thereby making the mining process more efficient. Constraint-based mining encourages interactive exploratory mining and analysis.

ii) Sequential mining:

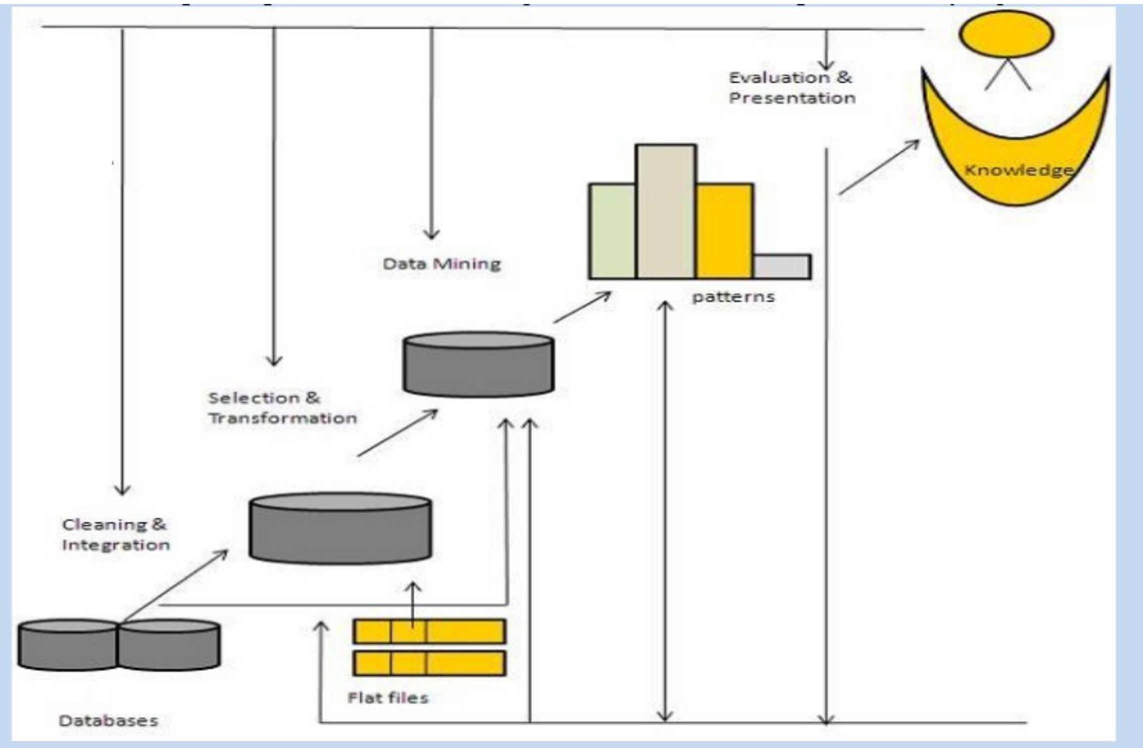
A sequence database consists of sequences of ordered elements or events, recorded with or without a concrete notion of time. There are many applications involving sequence data. Typical examples include customer shopping sequences, Web click streams, biological sequences, sequences of events in science and engineering, and in natural and social developments. In this section, we study sequential pattern mining in transactional databases. Concepts and Primitives: Sequential pattern mining is the mining of frequently occurring ordered events or subsequences as patterns. An example of a sequential pattern is “Customers who buy a Canon digital camera are likely to buy an HP color printer within a month.” For retail data, sequential patterns are useful for shelf placement and promotions. This industry, as well as telecommunications and other businesses, may also use sequential patterns for targeted marketing, customer retention, and many other tasks. Other areas in which sequential patterns can be applied include Web access pattern analysis, weather prediction, production processes, and network intrusion detection. Most of the studies of sequential pattern mining concentrate on categorical (or symbolic) patterns, whereas numerical curve analysis usually belongs to the scope of trend analysis and forecasting in statistical time-series analysis. The sequential pattern mining problem was first introduced by Agrawal and Srikant in 1995 [AS95] based on their study of customer purchase sequences, as follows:

“Given a set of sequences, where each sequence consists of a list of events (or elements) and each event consists of a set of items, and given a user-specified minimum support threshold of min sup, sequential pattern mining finds all frequent subsequences, that is,

		the subsequences whose occurrence frequency in the set of sequences is no less than min sup.” Let I = fI1, I2, : : : , Ipg be the set of all items. An itemset is a nonempty set of items. A sequence is an ordered list of events. A sequences is denoted he1e2e3 _ _ _eli, where event e1 occurs before e2, which occurs before e3, and so on. Event e j is also called an element of s. In the case of customer purchase data, an event refers to a shopping trip in which a customer bought items at acertain store. The event is thus an itemset, that is, an unordered list of items that the customer purchased during the trip. The itemset (or event) is denoted (x1x2 _ _ _xq), where xk is an item. For brevity, the brackets are omitted if an element has only one item, that is, element (x) is written as x. Suppose that a customer made several shopping trips to the store. These ordered events form a sequence for the customer. That is, the customer first bought the items in s1, then later bought the items in s2, and so on. An item can occur at most once in an event of a sequence, but can occur multiple times in different events of a sequence. The number of instances of items in a sequence is called the length of the sequence. A sequence with length l is called an l-sequence. A sequence a = ha1a2 _ _ _ani is called a subsequence of another sequence b = hb1b2 _ _ _bmi, and b is a supersequence of a, denoted as a v b, if there exist integers 1 _ j1 < j2 < _ _ _ < jn _ m such that a1 _ bj1 , a2 _ bj2 , . . . , an _ bjn . For example, if a h(ab), di and b = h(abc), (de)i, where a, b, c, d, and e are items, then a is a subsequence of b and b is a supersequence of a. A sequence database, S, is a set of tuples, hSID, si, where SID is a sequence ID and s is a sequence. For our example, S contains sequences for all customers of the store. A tuple hSID, si is said to contain a sequence a, if a is a subsequence of s. The support of a sequence a in a sequence database S is the number of tuples in the database containing a, that is, supportS(a) = j fh SID, sij(hSID, si 2 S)^(a v s)g j. It can be denoted as support (a) if the sequence database is clear from the context. Given a positive integer min sup as the minimum support threshold, a sequence a is frequent in sequence database S if supportS(a)_min sup. That is, for sequence a to be frequent, it must occur at least min sup times in S. A frequent sequence is called a sequential pattern. A sequential pattern with length l is called an l-pattern.	
	c)	**Define knowledge discovery and describe any six innovative technique for knowledge discovery.**	**8 M**
	Ans:	What is Knowledge Discovery Knowledge discovery in databases (KDD) is the non-trivial process of identifying valid, potentially useful and ultimately understandable patterns in data process of extracting previously unknown, valid, and actionable (understandable) information from large databases Data mining is a step in the KDD process of applying data analysis and discovery algorithms Machine learning, pattern recognition, statistics, databases, data visualization. Knowledge Discovery and Data Mining (KDD) is an interdisciplinary area focusing upon methodologies for extracting useful knowledge from data. The ongoing rapid growth of online data due to the Internet and the widespread use of databases have created an immense need for KDD methodologies. The challenge of extracting knowledge from data draws upon research in statistics, databases, pattern recognition, machine learning, data visualization, optimization, and high-performance computing, to deliver advanced business intelligence and web discovery solutions. Some people treat data mining same as	**(Explanation :8 marks)**

Knowledge discovery while some people view data mining essential step in process of knowledge discovery. Here is the 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.



6. **Answer any four of the following :** **(4x4=16) Marks**

a) **Describe Mining in World Wide Web.** **4 M**

Ans: Data mining refers to extracting or "mining" knowledge from large amounts of data. Mining can be broadly defined as the discovery and analysis of useful information from the World Wide Web. The World Wide Web contains the huge information such as hyperlink information, web page access info, education etc that provide rich source for data mining. The basic structure of the webpage is based on Document Object Model (DOM). The DOM structure refers to a tree like structure. In this structure the HTML tag in the page corresponds to a node in the DOM tree. We can segment the web page by

(Describtion: 4 marks)

		using predefined tags in HTML. The HTML syntax is flexible therefore; the web pages do not follow the W3C specifications. Not following the specifications of W3C may cause error in DOM tree structure. The DOM structure was initially introduced for presentation in the browser not for description of semantic structure of the web page. The DOM structure cannot correctly identify the semantic relationship between different parts of a web page.	
	b)	**Describe the significant role of meta data.**	**4 M**
	Ans:	Metadata are data about data. When used in a data warehouse, metadata are the data that define warehouse objects. Figure 3.12 showed a metadata repository within the bottom tier of the data warehousing architecture. Metadata are created for the data names and definitions of the given warehouse. Additional metadata are created and captured for time stamping any extracted data, the source of the extracted data, and missing fields that have been added by data cleaning or integration processes. A metadata repository should contain the following: •A description of the structure of the data warehouse, which includes the warehouse schema, view, dimensions, hierarchies, and derived data definitions, as well as data mart locations and contents •Operational metadata, which include data lineage (history of migrated data and the sequence of transformations applied to it), currency of data (active, archived, or purged), and monitoring information (warehouse usage statistics, error reports, and audit trails) •The algorithms used for summarization, which include measure and dimension definition •algorithms, data on granularity, partitions, subject areas, aggregation, summarization, and predefined queries and reports •The mapping from the operational environment to the data warehouse, which includes source databases and their contents, gateway descriptions, data partitions, data extraction, cleaning, transformation rules and defaults, data refresh and purging rules, and security (user authorization and access control) •Data related to system performance, which include indices and profiles that improve data access and retrieval performance, in addition to rules for the timing and scheduling of refresh, update, and replication cycles •Business metadata, which include business terms and definitions, data ownership information, and charging policies	**(Describtion: 4 marks)**
	c)	**Describe concept of hierarchy with an example.**	**4 M**
	Ans:	A concept hierarchy for a given numerical attribute defines a discretization of the attribute. Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts (such as numerical values for the attribute age) with higher-level concepts (such as youth, middle-aged, or senior). Although detail is lost by such data generalization, the generalized data may be more meaningful and easier to interpret. This contributes to a consistent representation of data mining results among multiple mining tasks, which is a common requirement. In addition, mining on a reduced data set	**(Describtion:4 marks)**

		require fewer input/output operations and is more efficient than mining on a larger, un-generalized data set. Because of these benefits, discretization techniques and concept hierarchies are typically applied before data mining as a preprocessing step, rather than during mining. Concept hierarchies for numeric attributes can be constructed automatically based on data distribution analysis. Five methods for concept hierarchy generation are defined below- binning histogram analysis entropy-based discretization and data segmentation by • **Natural partitioning Binning**: Attribute values can be discretized by distributing the values into bin and replacing each bin by the mean bin value or bin median value. These technique can be applied recursively to the resulting partitions in order to generate concept hierarchies. • **Histogram Analysis**: Histograms can also be used for discretization. Partitioning rules can be applied to define range of values. The histogram analyses algorithm can be applied recursively to each partition in order to automatically generate a multilevel concept hierarchy, with the procedure terminating once a pre specified number of concept levels have been reached. A Minimum interval size can be used per level to control the recursive procedure. This specifies the minimum width of the partition, or the minimum member of partitions at each level. • **Cluster Analysis**: A clustering algorithm can be applied to partition data into clusters or groups. Each cluster forms a node of a concept hierarchy, where all noses are at the same conceptual level. Each cluster may be further decomposed into sub-clusters, forming a lower level in the hierarchy. Clusters may also be grouped together to form a higher-level concept hierarchy. Segmentation by natural partitioning: Breaking up annual salaries in the range of into ranges like ($50,000-$100,000) are often more desirable than ranges like ($51, 263, 89-$60,765.3) arrived at by cluster analysis. The 3-4-5 rule can be used to segment numeric data into relatively uniform • **Natural intervals**. In general the rule partitions a give range of data into 3,4,or 5 equinity intervals, recursively level by level based on value range at the most significant digit. The rule can be recursively applied to each interval creating a concept hierarchy for the given numeric attribute.	
	d)	**Describe mining descriptive statistical measure in large databases.**	**4 M**
	Ans:	Class description can be explained with respect to the terms of popular Measures, such as count, sum, and average. Relational database systems provide five Built-in aggregate functions: count ( ), sum ( ), max ( ), and min ( ). These Functions can also be computed efficiently (in incremental and distributed manners) in data cubes. Thus, there is no problem in including these aggregate functions as basic measures in the descriptive mining of multidimensional data. For many data mining tasks, however, users would like to learn more data characteristics regarding both central tendency and data dispersion. Measures of central tendency include mean, median, mode, and midrange, while measures of data dispersion include quartiles, outliers, and variance. These descriptive statistics are of great help in Understanding the distribution of the data. Such	**(Describtion:4 marks)**

		measures have been studied extensively In the statistical literature. From the data mining point of view, we need to examine how They can be computed efficiently in large multidimensional databases.	
	e)	**Describe issues regarding classification and predication.**	**4 M**
	Ans:	**Preparing the Data for Classification and Prediction:** The following preprocessing steps may be applied to the data in order to help improve the accuracy, efficiency, and scalability of the classification or prediction process. **Data Cleaning:** This refers to the preprocessing of data in order to remove or reduce noise (by applying smoothing techniques) and the treatment of missing values (e.g., by replacing a missing value with the most commonly occurring value for that attribute, or with the most probable value based on statistics.) Although most classification algorithms have some mechanisms for handling noisy or missing data, this step can help reduce confusion during learning. **Relevance Analysis:** Many of the attributes in the data may be irrelevant to the classification or prediction task. For example, data recording the day of the week on which a bank loan application was filed is unlikely to be relevant to the success of the application. Furthermore, other attributes may be redundant. Hence, relevance analysis may be performed on the data with the aim of removing any irrelevant or redundant attributes from the learning process. In machine learning, this step is known as feature selection. Including such attributes may otherwise slow down, and possibly mislead, the learning step. Ideally, the time spent on relevance analysis, when added to the time spent on learning from the resulting “reduced” feature subset should be less than the time that would have been spent on learning from the original set of features. Hence, such analysis can help improve classification efficiency and scalability. **Data Transformation:** The data can be generalized to higher – level concepts. Concept hierarchies may be used for this purpose. This is particularly useful for continuous – valued attributes. For example, numeric values for the attribute income may be generalized to discrete ranges such as low, medium, and high. Similarly, nominal – valued attributes like street, can be generalized to higher – level concepts, like city. Since generalization compresses the original training data, fewer input / output operations may be involved during learning. The data may also be normalized, particularly when neural networks or methods involving distance measurements are used in the learning step. Normalization involves scaling all values for a given attribute so that they fall within a small specified range, such as – 1.0 to 1.0, or 0.0 to 1.0. In methods that use distance measurements, for example, this would prevent attributes with initially large ranges (like, say, income) from outweighing attributes with initially smaller ranges (such as binary attributes). **Comparing Classification Methods:** Classification and prediction methods can be compared and evaluated according to the following criteria:	**(4 marks)**

		Predictive Accuracy: This refers to the ability of the model to correctly predict the class label of new or previously unseen data. **Speed:** This refers to the computation costs involved in generating and using the model. Robustness: This is the ability of the model to make correct predictions given noisy data or data with missing values. **Scalability:** This refers to the ability to construct the model efficiently given large amount of data. **Interpretability:** This refers to the level of understanding and insight that is provided by the model.	