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 judgment 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.

1. a) Attempt any THREE of the following: Marks12

(i) Describe decision support system.

(Explanation -3 Marks, Types -1 Mark)

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.

(ii) Explain why preprocessing data?

(Explanation - 4 Marks)

Ans: Incomplete, inconsistent and noisy data are commonplace properties of large real-world databases. Attributes of interest may not always be available and other data was included just because it was considered to be important at the time of entity. Relevant data may not sometimes be recorded. Furthermore, the recording of the modifications to the data may not have been done. There are many possible reasons for noisy data (incorrect attribute values). They could have been human as well as computer errors that occurred during data entry. There could be inconsistent in the naming conventions adopted. Sometimes duplicate tuples may occur.

Data cleaning routines work to "clean" the data by filling in the missing values, smoothing noisy data, identifying and remo0ving outliers, and resolving inconsistencies in the data. Although mining routines have some form of handling noisy data, they are always not robust. If you would like to include files from many sources in your analysis then requires data integration. Naming inconsistencies may occur in this context. A large amount of redundant data may confuse or slow down the knowledge discovery process. In addition to data cleaning steps must taken to remove redundancies in the data.

Sometimes data would have to be normalized so that it scaled to a specific range e.g. [0.0, 1.0] in order to work data mining algorithms such as neural-networks, or clustering. Furthermore, you would require aggregating data e.g. as sales per region-something the t is not part of the data transformation methods need to be applied to the data.

Data reduction obtains a reduced representation of the data set that is much smaller in volume, yet produces the same or almost the same analytical results. There are a number of strategies for data reduction-data compression, numerosity reduction, generalization; data reduction

(iii) Describe multi dimensional data model.

(Explanation - 3 Marks, Example -1 Mark)

Ans: The model views data in the form of a data cube. OLAP tools are based on multidimensional data model. Data cubes usually model n-dimensional data.

From Tables Spreadsheets to Data Cubes

A data cube allows data to be modeled and viewed in multiple dimensions. Dimensions are facts that define a data cube. Dimensions are the perspective or entities with respect to which organizations would like to keep records. For example National Bank may create a customer warehouse in order to keep records of the bank's customers with respect to the dimension time, transaction, branch and location. These dimensions allow the bank to keep track of things like monthly transactions, branches and locations where the transactions were made. Each dimension may have a table associated with it, called the dimension table. For example the dimension tables for a transaction might include amount, type of transaction etc.

A multidimensional data model is typically organized around a central; theme like transactions. A fact table represents this theme where facts are numerical measures. Facts are usually quantities, which are used for analyzing relationship between dimensions. The fact table contains then names of facts, or measures, as well as keys related dimensions.

Although we hand to visualize data cubes three-dimensional geometric structures in the data warehouse the data cube inn n-dimensional.

(iv) What is concept description?

(Explanation - 4 Marks)

Ans: 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 class 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 ABCompany 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.

b) Attempt any ONE of the following: Marks 06

(i) Explain constraints based association mining.

(Explanation -2 Marks, Types - 4 Marks)

Ans: 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 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 maybe expressed as metarules (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.

(ii) Describe OLAP operations in the multi dimensional data models.

(Any 3 operation - 2 Marks each)

Ans: OLAP Operations

As we know that the OLAP server is based on the multidimensional view of data hence we will discuss the OLAP operations in multidimensional data.

Here is the list of OLAP operations.

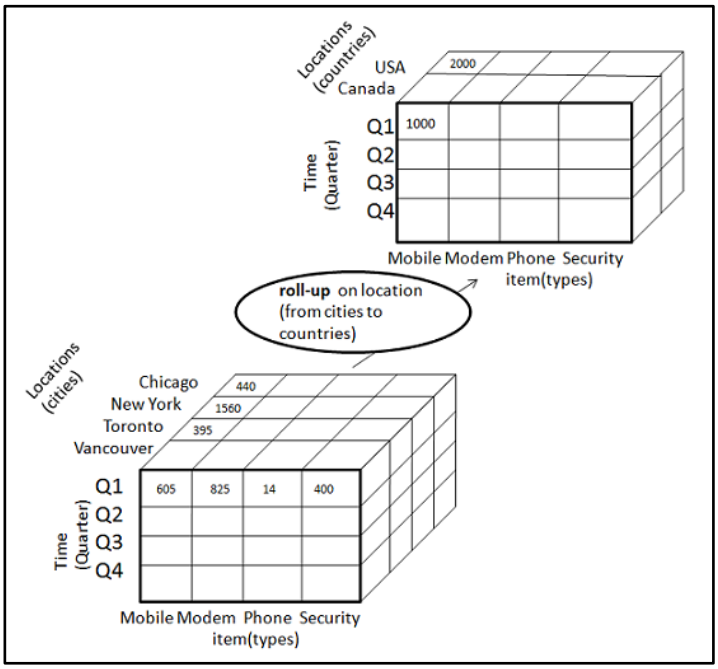
- **Roll-up**
- **Drill-down**
- **Slice and dice**
- **Pivot (rotate**)

ROLL-UP

This operation performs aggregation on a data cube in any of the following way:

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

Consider the following diagram showing the roll-up operation.



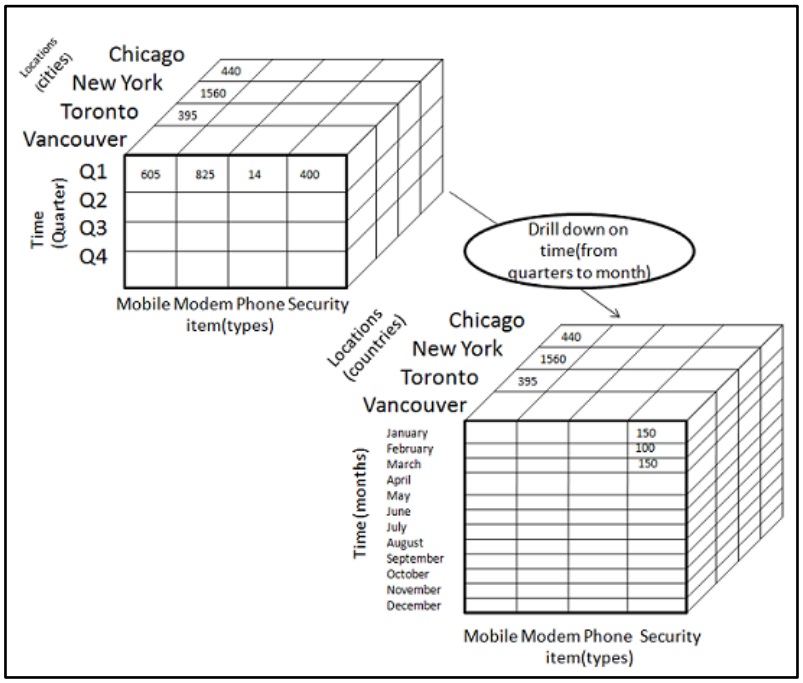
- The roll-up operation is performed by climbing up a concept hierarchy for the dimension location.
- Initially the concept hierarchy was "street < city < province < country".
- On rolling up the data is aggregated by ascending the location hierarchy from the level of city to level of country.
- The data is grouped into cities rather than countries.
- When roll-up operation is performed then one or more dimensions from the data cube are removed.

DRILL-DOWN

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

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

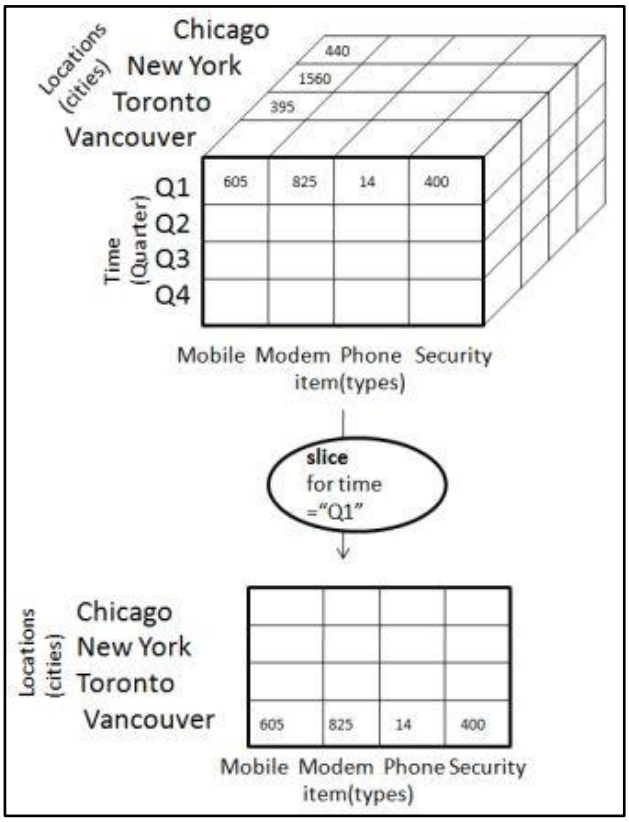
Consider the following diagram showing the drill-down operation:



- The drill-down operation is performed by stepping down a concept hierarchy for the dimension time.
- Initially the concept hierarchy was "day < month < quarter < year."
- On drill-up the time dimension is descended from the level quarter to the level of month.
- When drill-down operation is performed then one or more dimensions from the data cube are added.
- It navigates the data from less detailed data to highly detailed data.

SLICE

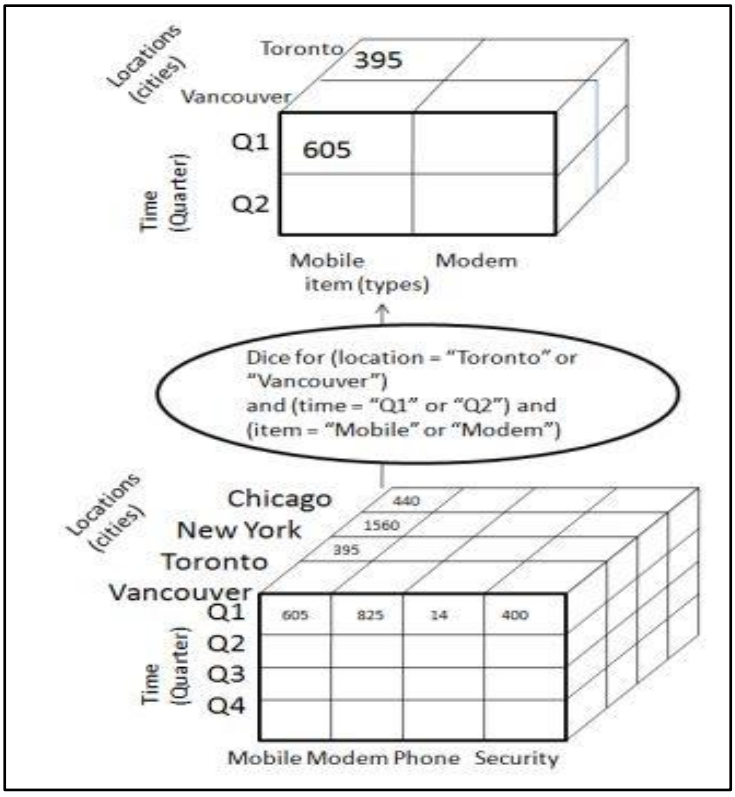
The slice operation performs selection of one dimension on a given cube and gives us a new sub cube. Consider the following diagram showing the slice operation.



- The Slice operation is performed for the dimension time using the criterion time ="Q1".
- It will form a new sub cube by selecting one or more dimensions.

DICE

The Dice operation performs selection of two or more dimension on a given cube and gives us a new subcube. Consider the following diagram showing the dice operation:

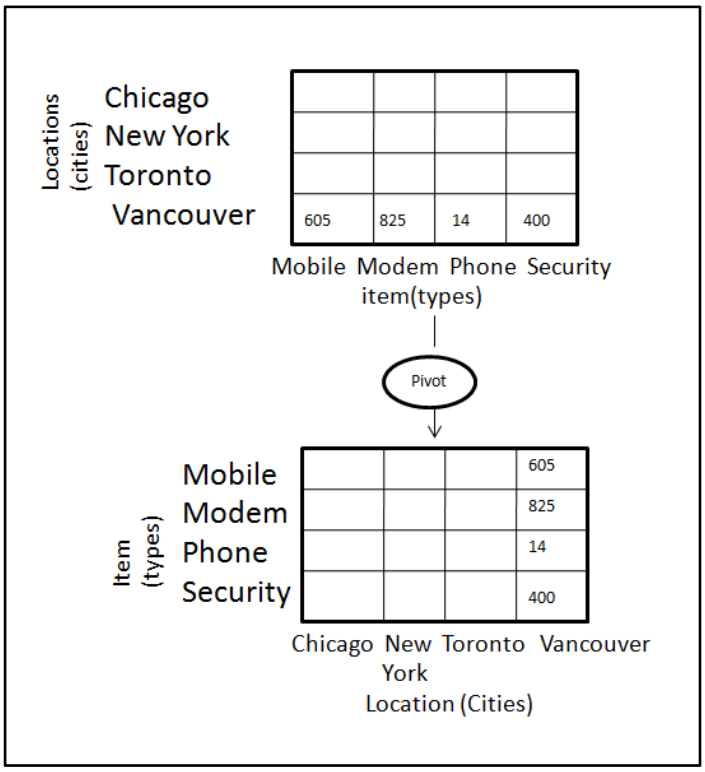


The dice operation on the cube based on the following selection criteria that involve three dimensions.

- (location = "Toronto" or "Vancouver")
- (time = "Q1" or "Q2")
- (item =" Mobile" or "Modem").

PIVOT

The pivot operation is also known as rotation. It rotates the data axes in view in order to provide an alternative presentation of data. Consider the following diagram showing the pivot operation.



2. Attempt any TWO of the following: Marks 16

a) Describe discretization and concept hierarchy generation for numeric and categorial data.
(For Numeric data -4 Marks and For Categorized data - 4 Marks)

Ans: Discretization and Concept Hierarchy Generation for Numeric Data:

It is difficult and laborious for to specify concept hierarchies for numeric attributes due to the wide diversity of possible data ranges and the frequent updates if data values. Manual specification also could be arbitrary. Concept hierarchies for numeric attributes

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.

Discretization and Concept Hierarchy Generation for Categorical Data:

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.

Specification of only of partial set of attributes:

Sometimes a user can be sloppy when defining a hierarchy, or may have only a vague idea about what should be included in a hierarchy. Consequently the user may have included only a small subset of the relevant attributes for the location, the user may have only specified street and city. To handle such partially specified hierarchies, it is important to embed data semantics in the database schema so that attributes with tight semantic connections can be pinned together.

b) Explain significant role of data with example.

(Role of Meta data -6 Marks and any example -2 Marks)

Ans: Metadata is simply defined as data about data. The data that are used to represent other data is known as metadata. For example the index of a book serves as metadata for the contents in the book. In other words we can say that metadata is the summarized data that leads us to the detailed data. In terms of data warehouse we can define metadata as following.

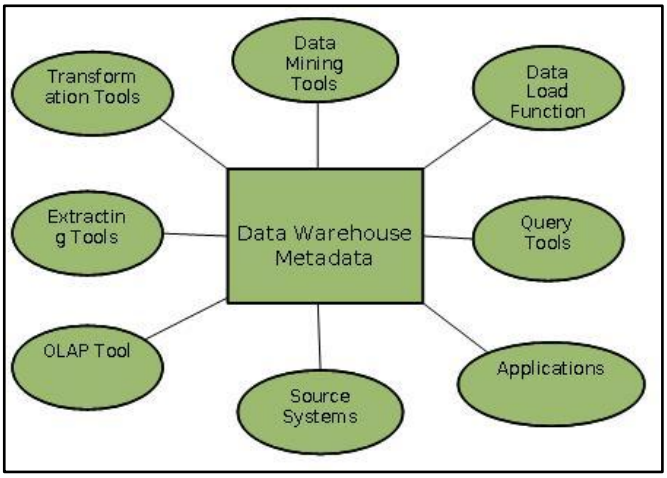
- Metadata is a road map to data warehouse.
- Metadata in data warehouse define the warehouse objects.
- The metadata act as a directory. This directory helps the decision support system to locate the contents of data warehouse.

Role of Meta data

Metadata has very important role in data warehouse. The role of metadata in warehouse is different from the warehouse data yet it has very important role. The various roles of metadata are explained below.

- The metadata act as a directory.
- This directory helps the decision support system to locate the contents of data warehouse.

- Metadata helps in decision support system for mapping of data when data are transformed from operational environment to data warehouse environment.
- Metadata helps in summarization between current detailed data and highly summarized data.
- Metadata also helps in summarization between lightly detailed data and highly summarized data.
- Metadata are also used for query tools.
- Metadata are used in reporting tools.
- Metadata are used in extraction and cleansing tools.
- Metadata are used in transformation tools.
- Metadata also plays important role in loading functions.



c) Explain schemas for multi dimensional databases.

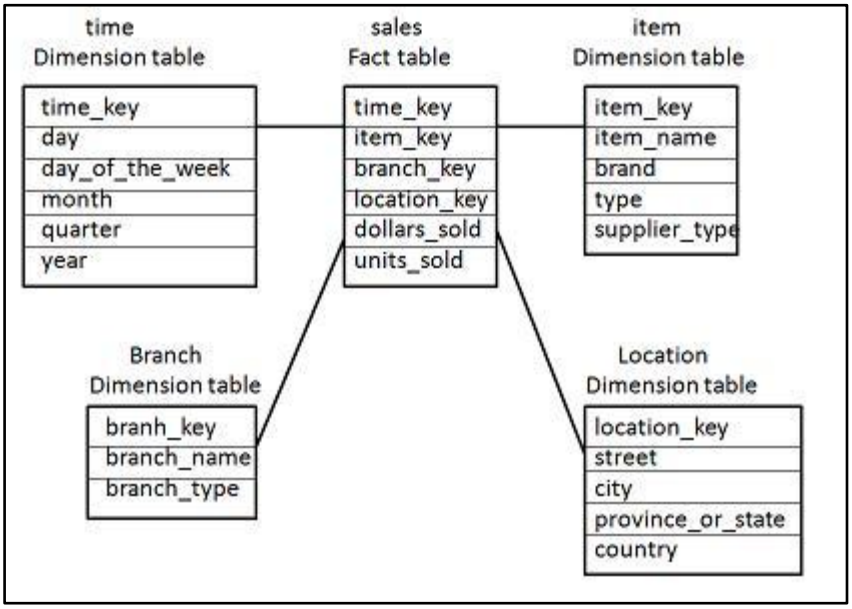
(List of Schema -2 Marks and for each schema – 2 Marks)

Ans: The schema is a logical description of the entire database. The schema includes the name and description of records of all record types including all associated data-items and aggregates. Likewise the database the data warehouse also requires the schema. The database uses the relational model on the other hand the data warehouse uses the Stars, snowflake and fact constellation schema. In this chapter we will discuss the schemas used in data warehouse.

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.

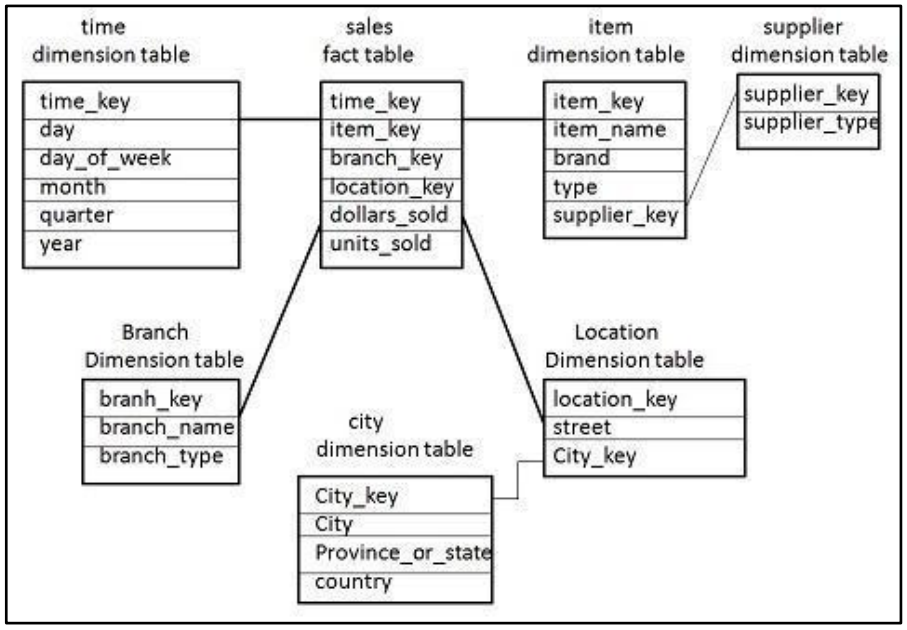


- There is a fact table at the centre. This fact table contains the keys to each of four dimensions.
- The fact table also contain the attributes namely, dollars sold and units sold.

.

Snowflake Schema

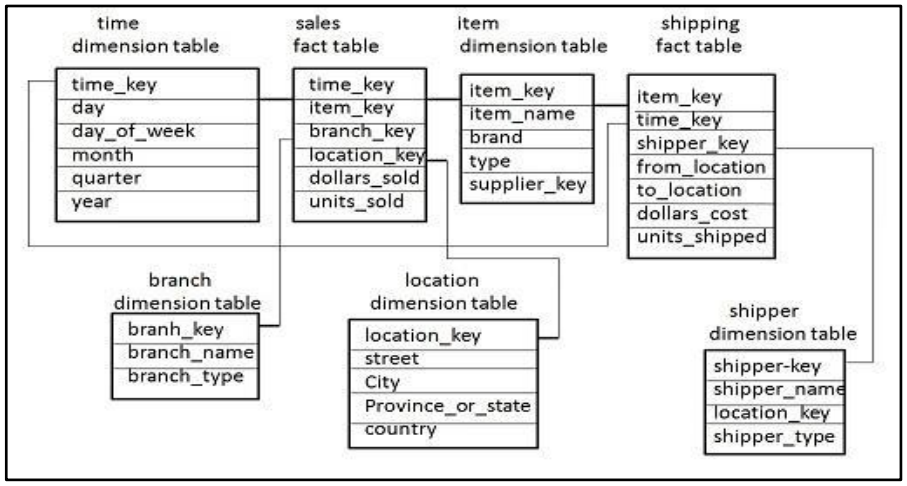
- In Snowflake schema some dimension tables are normalized.
- The normalization split up the data into additional tables.
- Unlike Star schema the dimensions table in snowflake schema is normalized for example the item dimension table in star schema is normalized and split into two dimension tables namely, item and supplier table.



- 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.

Fact Constellation Schema

- In fact Constellation there are multiple fact tables. This schema is also known as galaxy schema.
- In the following diagram we have two fact tables namely, sales and shipping.



- The sale fact table is same as that in star schema.
- The shipping fact table has the five dimensions namely, item_key, time_key, shipper-key, from-location.
- The shipping fact table also contains two measures namely, dollars sold and units sold.
- It is also possible for dimension table to share between fact tables. For example time, item and location dimension tables are shared between sales and shipping fact table.

3. Attempt any FOUR of the following: Marks 16

a) Describe data classification process.

(Steps – 1 Mark, Explanation-3 Marks)

Ans: The Data Classification process includes the two steps:

- Building the Classifier or Model
- Using Classifier for Classification

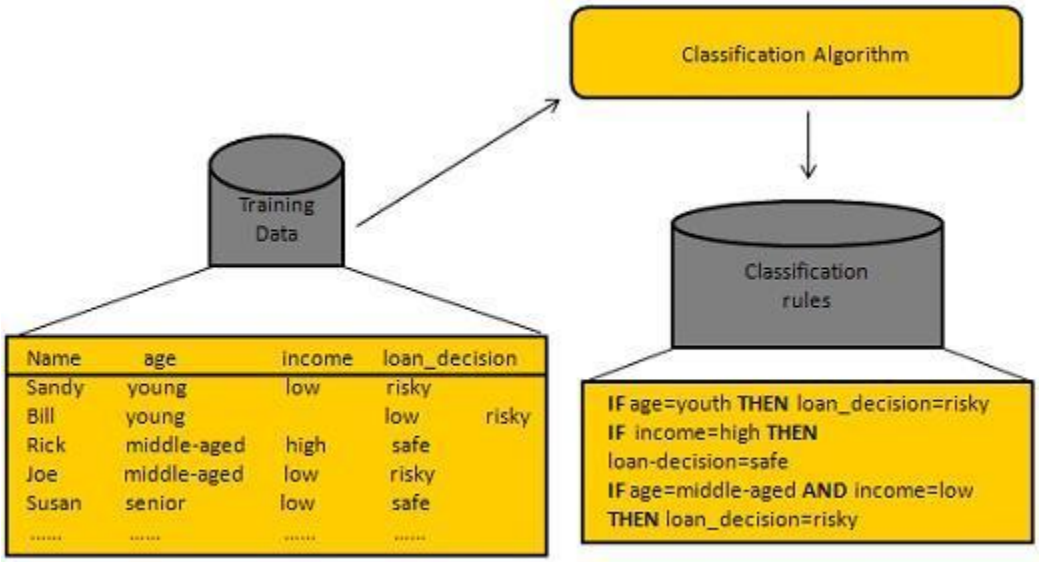
Building the classifier or model.

This step is the learning step or the learning phase.

In this step the classification algorithms build the classifier.

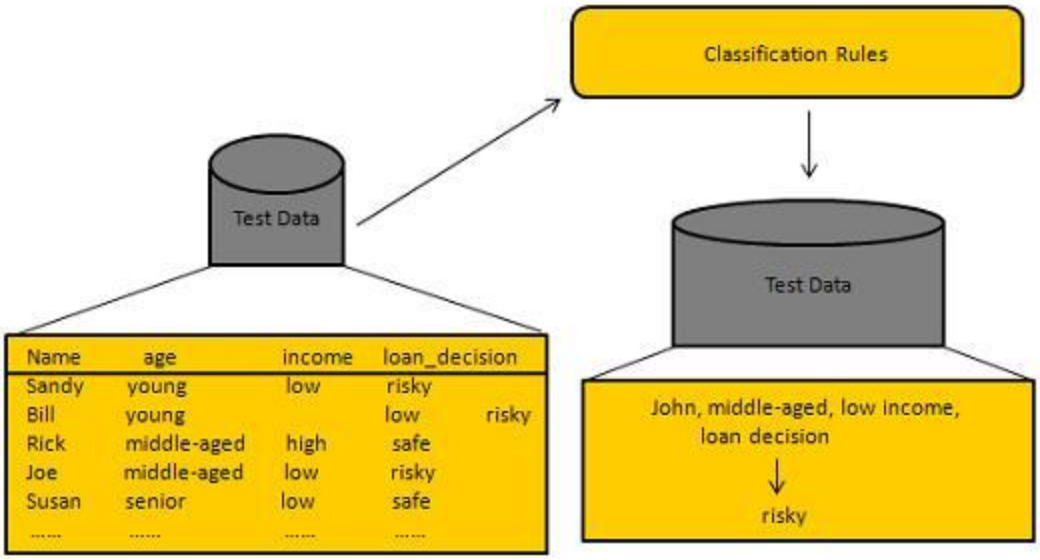
The classifier is built from the training set made up of database tuples and their associated class labels.

Each tuple that constitutes the training set is referred to as a category or class. These tuples can also be referred to as sample, object or data points.



Using classifier for classification

In this step the classifier is used for classification.Here the test data is used to estimate the accuracy of classification rules. The classification rules can be applied to the new data tuples if the accuracy is considered acceptable.



b) Explain sequential mining.

(Classification - 1 Mark, Explanation-3 Marks)

Ans: Sequential Pattern mining is a topic of data mining concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence. It is usually

presumed that the values are discrete, and thus time series mining is closely related, but usually considered a different activity. Sequential pattern mining is a special case of structured data mining.

There are several key traditional computational problems addressed within this field. These include building efficient databases and indexes for sequence information, extracting the frequently occurring patterns, comparing sequences for similarity, and recovering missing sequence members. In general, sequence mining problems can be classified as:

String mining which is typically based on string processing algorithms and Itemset mining which is typically based on association rule learning.

String Mining:

String mining typically deals with a limited alphabet for items that appear in a sequence, but the sequence itself may be typically very long. Examples of an alphabet can be those in the ASCII character set used in natural language text, nucleotide bases 'A', 'G', 'C' and 'T' in DNA sequences, or amino acids for protein sequences. In biology applications analysis of thearrangement of the alphabet in strings can be used to examine gene and protein sequences to determine their properties. Knowing the sequence of letters of a DNA a protein is not an ultimate goal in itself. Rather, the major task is to understand the sequence, in terms of its structure and biological function. This is typically achieved first by identifying individual regions or structural units within each sequence and then assigning a function to each structural unit. In many cases this requires comparing a given sequence with previously studied ones. The

Comparison between the strings becomes complicated

when insertions, deletions and mutations occur in a string.

ItemSet Mining:

Some problems in sequence mining lend themselves discovering frequent itemsets and the order they appear, for example, one is seeking rules of the form "if a {customer buys a car}, he or she is likely to {buy insurance} within 1 week", or in the context of stock prices, "if {Nokia up and Ericsson Up}, it is likely that {Motorola up and Samsung up} within 2 days". Traditionally, itemset mining is used in marketing applications for discovering regularities between frequently co-occurring items in large transactions. For example, by analysing transactions of customer

shopping baskets in a supermarket, one can produce a rule which reads "if a customer buys onions and potatoes together, he or she is likely to also buy hamburger meat in the same transaction".

The two common techniques that are applied to sequence databases for frequent itemset mining are the influential apriori algorithm and the more-recent FP-Growth technique.

c) Describe need for OLAP.

(Explanation- 4 Marks)

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".

d) Describe OLAP tools.

(List of Types- 1 Mark, Explanation – 3 Marks)

Ans: OLAP tools enable users to analyze multidimensional data interactively from multiple perspectives. OLAP consists of three basic analytical operations: consolidation (roll-up), drill-down, and slicing and dicing. Consolidation involves the aggregation of data that can be accumulated and computed in one or more dimensions. For example, all sales offices are rolled up to the sales department or sales division to anticipate sales trends. By contrast, the drill-down is a

technique that allows users to navigate through the details. For instance, users can view the sales by individual products that make up a region's sales. Slicing and dicing is a feature whereby users can take out (slicing) a specific set of data of the OLAP cube and view (dicing) the slices from different viewpoints.Designed for managers looking to make sense of their corporate data and related information, **OLAP tools** structure data hierarchically – the way managers think of their enterprises. But the best OLAP tools also allow business analysts to rotate their views on the information, changing the relationships in order to get more detailed insight into corporate trends and identify potential issues and opportunities.

Information Builders' WebFOCUS combines all the functionality of query tools, reporting tools, analytics and OLAP into a single powerful solution with one common interface so business analysts can slice and dice the data and see business processes in a new way.

The two types of OLAP tools are MOLAP (Multidimensional OLAP) and ROLAP (Relational OLAP).

1. MOLAP: In this type of OLAP, a cube is aggregated from the relational data source (data warehouse). When user generates a report request, the MOLAP tool can generate the results quickly because all data is already pre-aggregated within the cube.

2. ROLAP: In this type of OLAP, instead of pre-aggregating everything into a cube, the ROLAP engine essentially acts as a smart SQL generator. The ROLAP tool typically comes with a 'Designer' piece, where the data warehouse administrator can specify the relationship between the relational tables, as well as how dimensions, attributes, and hierarchies map to the underlying database tables.

e) Explain benefits of data warehousing.

(List of benefits - 1 Mark, Explanation - 3 Marks)

Ans: Benefits from a successful implementation of a data warehouse include:

• **Enhanced Business Intelligence**

Insights will be gained through improved information access. Managers and executives will be freed from making their decisions based on limited data. Decisions that affect the strategy and operations of organizations will be based upon credible facts and will be backed up with evidence and actual organizational data.

• **Increased Query and System Performance**

The data warehouse is built for analysis and retrieval of data rather than efficient upkeep of invidual records (i.e. transactions). Further, the data warehouse allows for a large system burden to be taken off the operational environment and effectively distributes system load across an entire organization's technology infrastructure.

• Business Intelligence from Multiple Sources

For many organizations, enterprise information systems are comprised of multiple subsystems, physically separated and built on different platforms. Moreover, merging data from multiple disparate data sources is a common need when conducting business intelligence. To solve this problem, the data warehouse performs integration of existing disparate data sources and makes them accessible in one place.

• Timely Access to Data

The data warehouse enables business users and decision makers to have access to data from many different sources as they need to have access to the data. Additionally, business users will spend little time in the data retrieval process. Scheduled data integration routines, known as ETL, are leveraged within a data warehouse environment.

• Enhanced Data Quality and Consistency

A data warehouse implementation typically includes the conversion of data from numerous source systems and data files and transformation of the disparate data into a common format. Data from the various business units and departments is standardized and the inconsistent nature of data from the unique source systems is removed. Moreover, individual business units and departments including sales, marketing, finance, and operations, will start to utilize the same data repository as the source system for their individual queries and reports. Thus each of these individual business units and departments will produce results that are consistent with the other business units within the organization.

• Historical Intelligence

Data warehouses generally contain many years worth of data that can neither be stored within nor reported from a transactional system. Typically transactional systems satisfy most operating reporting requirements for a given time-period but without the inclusion of historical data. In contrast, the data warehouse stores large amounts of historical data and can enable advanced business intelligence including time-period analysis, trend analysis, and trend prediction. The

advantage of the data warehouse is that it allows for advanced reporting and analysis of multiple time-periods.

• High Return on Investment

Return on investment (ROI) refers to the amount of increased revenue or decreased expenses a business will be able to realize from any project or investment of capital. Subsequently, implementations of data warehouses and complementary business intelligence systems have enabled business to generate higher amounts of revenue and provide substantial cost savings.

4. a) Attempt any THREE of the following: Marks 12

(i) How does data reduction technique helps to reduce size of data?

(List of Strategies-1 Mark, Explanation-3 Marks)

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.

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.

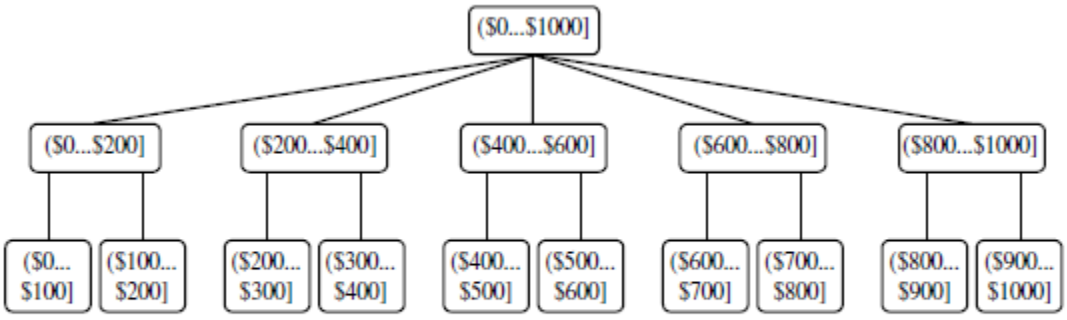
5. Discretization and concept hierarchy generation, where raw data values for attributes are replaced by ranges or higher conceptual levels. Data discretization is a form of numerosity reduction that is very useful for the automatic generation of concept hierarchies. Discretization and concept hierarchy generation are powerful tools for data mining, in that they allow the mining of data at multiple levels of abstraction

(ii) Describe concept hierarchies.

(Definition -1 Mark, Example -1 Mark, Explanation - 2 Marks)

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 requires fewer input/output operations and is more efficient than mining on a larger, ungeneralized 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.

An example of a concept hierarchy for the attribute *price* is given in Figure below .More than one concept hierarchy can be defined for the same attribute in order to accommodate the needs of various users.



A concept hierarchy for the attribute *price*, where an interval ($X ... $Y] denotes the range from $X (exclusive) to $Y (inclusive).

Figure

(iii) Explain association rule classification.

(Definition– 1 Mark, Explanation with respect to example-3 Marks)

Ans: Frequent patterns and their corresponding association or correlation rules characterize interesting relationships between attribute conditions and class labels, and thus havebeen recently used for effective classification. Association rules show strong associationsbetween attribute-value pairs (or items) that occur frequently in a given data set. Associationrules are commonly used to analyze the purchasing patterns of customers in astore. Such analysis is useful in many decision-making processes, such as product placement,catalog design, and cross-marketing. The discovery of association rules is based on frequent itemset mining. We can search for strong associations between frequentpatterns (conjunctions of attribute-value pairs) and class labels. Because association rules explore highly confident associations amongmultiple attributes, this approachmay overcomesome constraints introduced by decision-tree induction, which considers only oneattribute at a time.

Association rules aremined in a two-step process consisting of frequent itemset mining, followed by rule generation.

The first step searches for patterns of attribute-value pairs that occur repeatedly ina data set, where each attribute-value pair is considered an item. The resulting attribute value pairs form frequent itemsets. The second step analyzes the frequent itemsets in order to generate association rules. All association rules must satisfy certain criteria regardingtheir "accuracy" (or confidence) and the proportion of the data set that they actually represent(referred to as support). For example, the following is an association rule minedfrom a data set, D, shown with its confidence and support.

age = youth^credit = OK)buys computer = yes [support = 20%, confidence = 93%]

(3.3)

where "^" represents a logical "AND."

More formally, letDbe a data set of tuples. Each tuple inDis described by n attributes, A1, A2, : : : , An, and a class label attribute, Aclass. All continuous attributes are discretized and treated as categorical attributes. An item, p, is an attribute-value pair of the form (Ai, v), where Ai is an attribute taking a value, v. A data tuple X = (x1, x2, : : : , xn) satisfies an item, p = (Ai, v), if and only if xi = v, where xi is the value of the ith attribute of X. Association rules can have any number of items in the rule antecedent (left-hand side) and any number of items in the rule

consequent (right-hand side). However, when mining association rules for use in classification, we are only interested in association rules of the form p1 ^ p2 ^: : : pl) Aclass= C where the rule antecedent is a conjunction of items, p1, p2, : : : , pl(l _ n), associated with a class label, C. For a given rule, R, the percentage of tuples in D satisfying the rule antecedent that also have the class label C is called the confidence of R. From a classification point of view, this is akin to rule accuracy. For example, a confidence of 93% for Association Rule (3.3) means that 93% of the customers in D who are young and have an OK credit rating belong to the class buys computer = yes. The percentage of tuples inDsatisfying the rule antecedent and having class label C is called the support of R. A support of 20% for Association Rule means that 20% of the customers in D are young, have an OK credit rating, and belongto the class buys computer = yes.

(iv) Describe data generalization.

((Definition -1 Mark, Explanation - 3 Marks)

Ans: Data Generalization is the process of creating successive layers of summary data in an evolutional database. It is a process of zooming out to get a broader view of a problem, trend or situation. It is also known as rolling-up data.

There are millions and millions of data stored in the database and this number continues to increase everyday as a company heads for growth. In fact, a group of process of process called extract, transform, load (ETL) is periodically performed in order to manage data within the data warehouse.

A data warehouse is a rich repository of data, most of which are historical data from a company. But in modern data warehouses, data could come from other sources. Having data from several sources greatly helps in the overall business intelligence system of a company. With diverse data sources, the company can have a broader perspective not just about the trends and pattern within the organization but of the global industrial trends are well.

In order to get a view of trends and patterns based on the analytical outputs of the business intelligence system can be a daunting task. With those millions of data, most of which disparate (but of course ironed out by the ETL process), it may be difficult to generate reports.

Dealing alone with big volumes of data for consistent delivery of business critical applications can already affect the network management tools of a company. Many companies have found that

existing network management tools could hardly cope up with the great bulk of data required by the organization to monitor network and applications usage.

The existing tools could hardly capture, store and report on traffic with speed and granularity which are requirements for real network improvements. In order to keep the volume down to speed up network performance for effective delivery, some network tools discard the details. What they would do is convert some detailed data into hourly, daily or weekly summaries. This is the process called data generalization or as some database professionals call it, rolling up data. Ensuring network manageability is just one of the benefits of data generalization.

Data generalization can provide a great help in Online Analytical Processing (OLAP) technology. OLAP is used for providing quick answers to analytical queries which are by nature multidimensional. They are commonly used as part of a broader category of business intelligence. Since OLAP is used mostly for business reporting such as those for sales, marketing, management reporting, business process management and other related areas, having a better view of trends and patterns greatly speeds up these reports.

Data generalization is also especially beneficial in the implementation of an Online transaction processing (OLTP). OLTP refers to a class systems designed for managing and facilitating transaction oriented applications especially those involved with data entry and retrieval transaction processing. OLAP was created later than OLTP and had slight modifications from OLTP.

b) Attempt any ONE of the following: Marks 06

(i) Describe mining world wide web.

(Description – 2 Mark, Explanation-3 Marks, Diagram – 1 Mark)

Ans: 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 web page 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 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.

Vision-based page segmentation (VIPS)

The purpose of VIPS is to extract the semantic structure of a web page based on its visual presentation.

Such a semantic structure corresponds to tree structure. In this tree each node corresponds to a block.

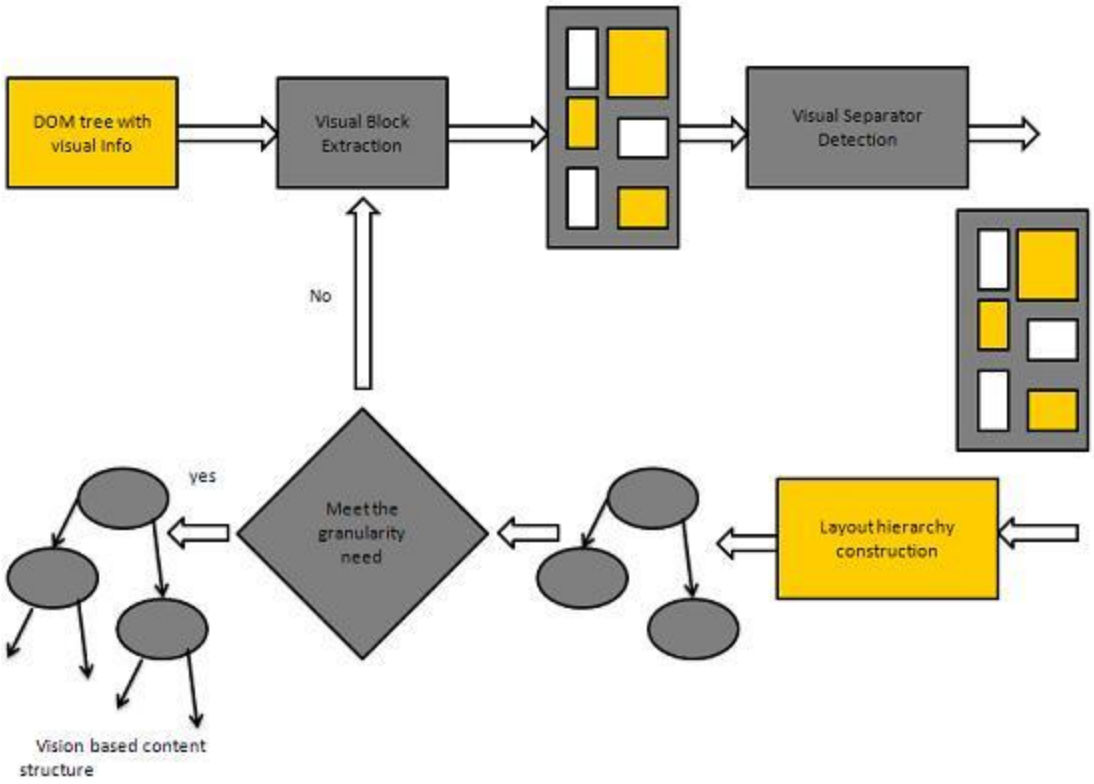
A value is assigned to each node. This value is called Degree of Coherence. This value is assigned to indicate how coherent is the content in the block based on visual perception.

The VIPS algorithm first extracts all the suitable blocks from the HTML DOM tree. After that it finds the separators between these blocks.

The separators refer to the horizontal or vertical lines in a web page that visually cross with no blocks.

The semantic of the web page is constructed on the basis of these blocks.

The following figure shows the procedure of VIPS algorithm:



(ii) Describe data cleaning techniques in data warehouse.

(Definition and Description-2 Marks, List of techniques- 1 Mark, Explanation-3 Marks)

Ans: Data cleaning is performed as data preprocessing step while preparing the data for a data warehouse. Real-world data tend to be incomplete, noisy, and inconsistent. Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifyingoutliers, and correct inconsistencies in the data.

Missing Values:

Imagine that you need to analyze AllElectronics sales and customer data. You note thatmany tuples have no recorded value for several attributes, such as customer income.Filling in the missing values of this attribute can be done using the following methods:

1. **Ignore the tuple**: This is usually done when the class label is missing (assuming themining task involves classification). This method is not very effective, unless the tuplecontains several attributes with missing values. It is especially poor when the percentageof missing values per attribute varies considerably.

2. **Fill in the missing value manually**: In general, this approach is time-consuming andmay not be feasible given a large data set with many missing values.

3. **Use a global constant to fill in the missing value**: Replace all missing attribute values by the same constant, such as a label like "Unknown" or ¥. If missing values are replaced by, say, "Unknown," then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common—that of "Unknown." Hence, although this method is simple, it is not foolproof.

4. **Use the attribute mean to fill in the missing value**: For example, suppose that the average income of All Electronics customers is $56,000. Use this value to replace themissing value for income.

5. **Use the attribute mean for all samples belonging to the same class as the given tuple**:

For example, if classifying customers according to credit risk, replace the missing valuewith the average income value for customers in the same credit risk category as thatof the given tuple.

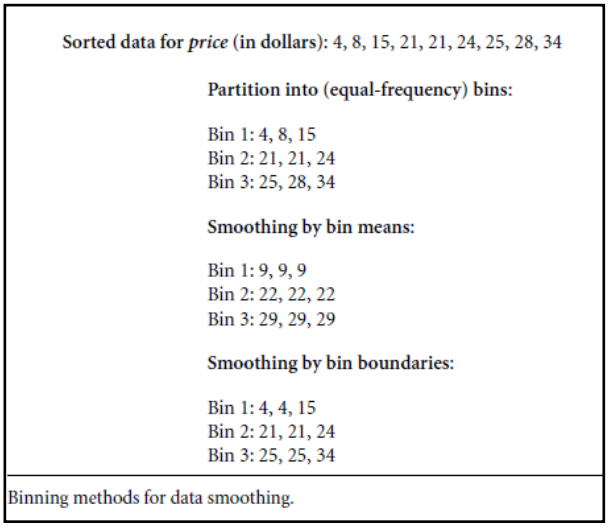
6. **Use the most probable value to fill in the missing value**: This may be determinedwith regression, inference-based tools using a Bayesian formalism, or decision treeinduction. For

example, using the other customer attributes in your data set, youmay construct a decision tree to predict the missing values for income.

Noisy Data:

Noise is a random error or variance in a measured variable. Given anumerical attribute such as, say, price, we can "smooth" out the data to remove thenoise by the following **data smoothing techniques**:

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 numberof "buckets," or bins. Because binning methods consult the neighborhood of values,they perform local smoothing. Figure illustrates some binning techniques. In thisexample, the data for price are first sorted and then partitioned into equal-frequencybins of size 3 (i.e., each bin contains three values). In smoothing by bin means, eachvalue in a bin is replaced by the mean value of the bin. For example, the mean of thevalues 4, 8, and 15 in Bin 1 is 9. Therefore, each original value in this bin is replacedby the value 9. Similarly, smoothing by bin medians can be employed, in which eachbin value is replaced by the bin median. In smoothing by bin boundaries, the minimumand maximum values in a given bin are identified as the bin boundaries. Eachbin value is then replaced by the closest boundary value. In general, the larger thewidth, the greater the effect of the smoothing. Alternatively, bins may be equal-width,where the interval range of values in each bin is constant.



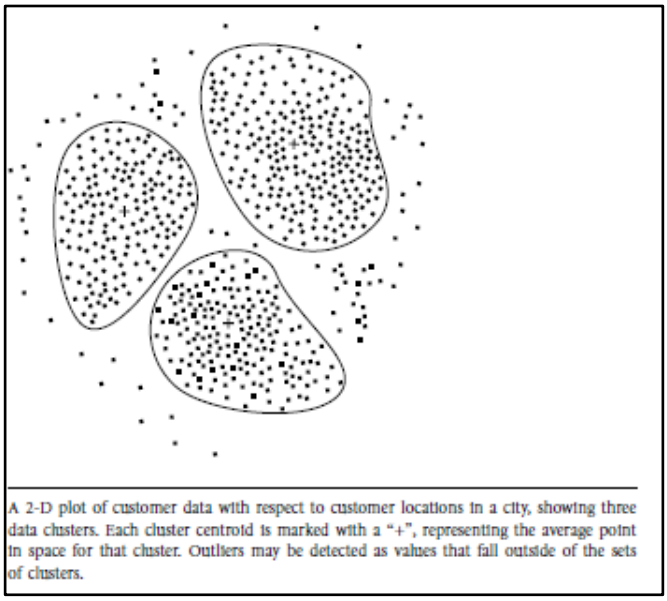
Binning methods for data smoothing.

Figure

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.



A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters. Each cluster centroid is marked with a "+", representing the average point in space for that cluster. Outliers may be detected as values that fall outside of the sets of clusters.

5. Attempt any TWO of the following: Marks 16

a) Explain innovative techniques for knowledge discovery, write application of these techniques.

(List of techniques - 2 Marks, Explanation of any two (2) technique - 2 Marks each 1 Mark for each application)

Ans: List of Knowledge discovery

Associations and correlations, classification, prediction, clustering, trend analysis, outlier and deviation analysis, and similarity analysis.

1. **Association analysis.** Suppose, as a marketing manager of All Electronics, you would like to determine which items are frequently purchased together within the same transactions. An example of such a rule, mined from the All Electronics transactional database, is buys(X , "computer") ⇒ buys(X , "software") [support = 1%, confidence = 50%] where X is a variable representing a customer. A confidence, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well. A 1% support means that 1% of all of the transactions under analysis showed that computer and software were purchased together. This association rule involves a single attribute or predicate (i.e., buys) that repeats. Association rules that contain a single predicate are referred to as single-dimensional association rules. Dropping the predicate notation, the above rule can be written simply as "computer⇒ software [1%, 50%]".Suppose, instead, that we are given the All Electronics relational database relating to purchases. A data mining system may find association rules like age(X , "20...29") ∧ income(X , "20K...29K") ⇒ buys(X , " CD player")[support = 2%, confidence = 60%] The rule indicates that of the AllElectronics customers under study, 2% are 20 to 29 years of age with an income of 20,000 to 29,000 and have purchased a CD player at AllElectronics. There is a 60% probability that a customer in this age and income group will purchase a CD player.

2. **Classification** is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known).

The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks.

Example: Suppose, as sales manager of AllElectronics, you would like to classify a large set of items in the store, based on three kinds of responses to a sales campaign: good response, mild response, and no response. You would like to derive a model for each of these three classes based on the descriptive features of the items, such as price, brand, place made, type, and category. The resulting classification should maximally distinguish each class from the others, presenting an organized picture of the data set.

3. **Clustering** analyzes data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Each cluster that is formed can be viewed as a class of objects, from which rules can be derived. Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together.

Example: Cluster analysis can be performed on AllElectronics customer data in order to identify homogeneous subpopulations of customers. These clusters may represent individual target groups for marketing.

4. Outlier Analysis A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. Most data mining methods discard outliers as noise or exceptions. However, in some applications such as fraud detection, the rare events can be more interesting than the more regularly occurring ones. Outliers may be detected using statistical tests that assume a distribution or probability model for the data, or using distance measures where objects that are a substantial distance from any other cluster are considered outliers. Rather than using statistical or distance measures, deviation-based methods identify outliers by examining differences in the main characteristics of objects in a group.

Example: Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of extremely large amounts for a given account number in comparison to regular charges incurred by the same account. Outlier values may also be detected with respect to the location and type of purchase, or the purchase frequency.

5. Evolution Analysis

Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time. Although this may include characterization, discrimination, association and correlation analysis, classification, prediction, or clustering of time related data, distinct features of such an analysis include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

Example

Suppose that you have the major stock market (time-series) data of the last several years available from the New York Stock Exchange and you would like to invest in shares of high-tech industrial companies. A data mining study of stock exchange data may identify stock evolution regularities for overall stocks and for the stocks of particular companies. Such regularities may help predict future trends in stock market prices, contributing to your decision making regarding stock investments.

b) Describe Apriori algorithm.

(Algorithm -4 Marks, Description of algorithm - 2 Marks, Example -2 Marks)

Ans: Apriori is a seminal algorithm for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties Apriori employs an iterative approach known as a level-wise search, where k-itemsets are used to explore (k + 1)-itemsets. 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 L_1. Next, L_1 is used to find L_2 the set of frequent 2-itemsets, which is used to find L_3, and so on, until no more frequent k-itemsets can be found. The finding of each L_k requires one full scan of the database.

Once the frequent itemsets 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:

$$confidence(A \Rightarrow B) = P(B|A) = \frac{support\_count(A \cup B)}{support\_count(A)}.$$

Input:

- *D*, a database of transactions;
- Min_sup, the minimum support count threshold

Output: *L*, frequent itemsets in *D*

Method:

Method:

```
(1)   L1 = find_frequent_1-itemsets(D);
(2)   for (k = 2; Lk−1 ≠ ϕ; k++) {
(3)      Ck = apriori_gen(Lk−1);
(4)      for each transaction t ∈ D { // scan D for counts
(5)         Ct = subset(Ck, t); // get the subsets of t that are candidates
(6)         for each candidate c ∈ Ct
(7)            c.count++;
(8)      }
(9)      Lk = {c ∈ Ck | c.count ≥ min_sup}
(10)  }
(11)  return L = ∪k Lk;

procedure apriori_gen(Lk−1:frequent (k − 1)-itemsets)
(1)   for each itemset l1 ∈ Lk−1
(2)      for each itemset l2 ∈ Lk−1
(3)         if (l1[1] = l2[1]) ∧ (l1[2] = l2[2]) ∧ ... ∧ (l1[k − 2] = l2[k − 2]) ∧ (l1[k − 1] < l2[k − 1]) then {
(4)            c = l1 ⋈ l2; // join step: generate candidates
(5)            if has_infrequent_subset(c, Lk−1) then
(6)               delete c; // prune step: remove unfruitful candidate
(7)            else add c to Ck;
(8)         }
(9)   return Ck;

procedure has_infrequent_subset(c: candidate k-itemset;
         Lk−1: frequent (k − 1)-itemsets); // use prior knowledge
(1)   for each (k − 1)-subset s of c
(2)      if s ∉ Lk−1 then
(3)         return TRUE;
(4)   return FALSE;
```

Example:

Assume that a large supermarket tracks sales data by stock-keeping unit (SKU) for each item: each item, such as "butter" or "bread", is identified by a numerical SKU. The supermarket has a database of transactions where each transaction is a set of SKUs that were bought together.

Let the database of transactions consist of following itemsets:

Itemsets

{1,2,3,4}

{1,2,4}

{1,2}

{2,3,4}

{2,3}

{3,4}

{2,4}

We will use Apriori to determine the frequent item sets of this database. To do so, we will say that an item set is frequent if it appears in at least 3 transactions of the database: the value 3 is the support threshold.

The first step of Apriori is to count up the number of occurrences, called the support, of each member item separately, by scanning the database a first time. We obtain the following result

Item	Support
{1}	3
{2}	6
{3}	4
{4}	5

All the itemsets of size 1 have a support of at least 3, so they are all frequent. The next step is to generate a list of all pairs of the frequent items:

Item	Support
{1,2}	3
{1,3}	1
{1,4}	2
{2,3}	3
{2,4}	4
{3,4}	3

The pairs {1,2}, {2,3}, {2,4}, and {3,4} all meet or exceed the minimum support of 3, so they are frequent. The pairs {1,3} and {1,4} are not. Now, because {1,3} and {1,4} are not frequent, any

larger set which contains {1,3} or {1,4} cannot be frequent. In this way, we can prune sets: we will now look for frequent triples in the database, but we can already exclude all the triples that contain one of these two pairs:

Item Support

{2,3,4}2

In the example, there are no frequent triplets -- {2,3,4} is below the minimal threshold, and the other triplets were excluded because they were super sets of pairs that were already below the threshold.

We have thus determined the frequent sets of items in the database, and illustrated how some items were not counted because one of their subsets was already known to be below the threshold.

c) Describe model management and user interface modes of DSS.

(Definition and characteristics of DSS - 2 Marks, Model management - 2 Marks, User interface- 2 Marks and 2 Marks from example)

Ans: A Decision Support System (DSS) is a computer-based information system that supports business or organizational decision-making activities. DSSs serve the management, operations, and planning levels of an organization (usually mid and higher management) and help to make decisions, which may be rapidly changing and not easily specified in advance (Unstructured and Semi-Structured decision problems). Decision support systems can be either fully computerized, human or a combination of both.

DSS as a tool to support decision making process, DSS users see DSS as a tool to facilitate organizational processes

DSS by its characteristics:

1. DSS tends to be aimed at the less well structured, underspecified problem that upper level managers typically face;
2. DSS attempts to combine the use of models or analytic techniques with traditional data access and retrieval functions;
3. DSS specifically focuses on features which make them easy to use by noncomputer people in an interactive mode; and
4. DSS emphasizes flexibility and adaptability to accommodate changes in the environment and the decision making approach of the user.

DSSs include knowledge-based systems. A properly designed DSS is an interactive software-based system intended to help decision makers compile useful information from a combination of raw data, documents, and personal knowledge, or business models to identify and solve problems and make decisions.

Components

Design of a drought mitigation decision support system three fundamental components of a DSS architecture are:

1. The database (or knowledge base),
2. The model (i.e., the decision context and user criteria), and
3. The user interface(allows you to communicate with the DSS).

Model Management Component

The model management component consists of both the Decision Support System models and the Decision Support System model management system. A model is a representation of some event, fact, or situation. As it is not always practical, or wise, to experiment with reality, people build models and use them for experimentation. Models can take various forms.

Businesses use models to represent variables and their relationships. For example, you would use a statistical model called analysis of variance to determine whether newspaper, TV, and billboard advertizing are equally effective in increasing sales.

Decision Support Systems help in various decision-making situations by utilizing models that allow you to analyze information in many different ways. The models you use in a Decision Support System depend on the decision you are making and, consequently, the kind of analysis you require. For example, you would use what-if analysis to see what effect the change of one or more variables will have on other variables, or optimization to find the most profitable solution given operating restrictions and limited resources. Spreadsheet software such as excel can be used as a Decision Support System for what-if analysis.

The model management system stores and maintains the Decision Support System's models. Its function of managing models is similar to that of a database management system. The model management component cannot select the best model for you to use for a particular problem that requires your expertise but it can help you create and manipulate models quickly and easily.

User Interface Management Component

It allows you to communicate with the Decision Support System. It consists of the user interface management system. This is the component that allows you to combine your know-how with the storage and processing capabilities of the computer.

The user interface is the part of the system you see through it when enter information, commands, and models. This is the only component of the system with which you have direct contract. If you have a Decision Support System with a poorly designed user interface, if it is too rigid or too cumbersome to use, you simply won't use it no matter what its capabilities. The best user interface uses your terminology and methods and is flexible, consistent, simple, and adaptable.

For an example of the components of a Decision Support System, let's consider the Decision Support System that Land's End has tens of millions of names in its customer database. It sells a wide range of women's, men's, and children's clothing, as well various household wares. To match the right customer with the catalog, land's end has identified 20 different specialty target markets. Customers in these target markets receive catalogs of merchandise that they are likely to buy, saving Lands' End the expense of sending catalogs of all products to all 20 million customers. To predict customer demand, lands' end needs to continuously monitor buying trends. And to meet that demand, lands' end must accurately forecast sales levels. To accomplish theses goals, it uses a Decision Support System which performs three tasks:

The user interface management component

- Data management: The Decision Support System stores customer and product information. In addition to this organizational information, Lands' End also needs external information, such as demographic information and industry and style trend information.
- Model management: The Decision Support System has to have models to analyze the information. The models create new information that decision makers need to plan product lines and inventory levels. For example, Lands' End uses a statistical model called regression analysis to determine trends in customer buying patterns and forecasting models to predict sales levels.
- User interface management: A user interface enables Lands' End decision makers to access information and to specify the models they want to use to create the information they need.

6. Attempt any FOUR of the following: Marks 16

a) Describe the needs of data warehousing.

(For each point-1Mark (any four points))

Ans:

- 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.
- 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.
- 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.
- 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 itsaccounting 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.
- 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.

- 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.
- 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) Explain operational and informational data.

(Operational Data - 2 Marks, Information Data -2 Marks)

Ans: Operational Data: **(2 Marks)**

- Focusing on transactional function such as bank card withdrawals and deposits
- Detailed
- Updateable
- Reflects current data

Informational Data: **(2 Marks)**

- Focusing on providing answers to problems posed by decision makers
- Summarized
- Non updateable

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

c) Describe mining descriptive statistical measures in large database.

Ans: ***(Listing -1 Mark)***

For many data mining tasks, however, users would like to learn more data characteristics regarding:

1) **Measuring the Central Tendency**

2) **Measuring the Dispersion of Data**

3) **Graph Displays of Basic Statistical Class Descriptions**

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.

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. From the data mining point of view, weneed to examine how they can be computed efficiently in large databases. In particular,it is necessary to introduce the notions of distributive measure, algebraic measure, andholistic measure. Knowing what kind of measure we are dealing with can help us chosen efficient implementation for it.

Measuring the Central Tendency (1 Mark)

In this section, we look at various ways to measure the central tendency of data. The most common and most effective numerical measure of the "center" of a set of data is the (arithmetic) mean. Let x1;x2; : :::;xNbe a set of N values or observations, such as forsome attribute, like salary. The mean of this set of values is

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}.$$

This corresponds to the built-in aggregate function, average (avg() in SQL), provided in relational database systems.

A distributive measure is a measure (i.e., function) that can be computed for a given data set by partitioning the data into smaller subsets, computing the measure for each subset, and then merging the results in order to arrive at the measure's value for the original (entire) data set. Both sum() and count() are distributive measures because they can be computed in this manner. Other examples include max() and min(). An algebraic measure is a measure that can be computed by applying an algebraic function to one or more distributive measures. Hence, average (or mean()) is an algebraic measure because it can be computed by sum()/count().When computing data cubes2, sum() and count() are typically saved in precomputation. Thus, the derivation of average for data cubes is straightforward. Sometimes, each value xi in a set may be associated with a weight wi, for i = 1; : : : ;N. The weights reflect the significance, importance, or occurrence frequency attached to their respective values. In this case, we can compute

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}.$$

This is called the weighted arithmetic mean or the weighted average. Note that the weighted average is another example of an algebraic measure.

Although the mean is the single most useful quantity for describing a data set, it is not always the best way of measuring the center of the data. A major problem with the mean is its sensitivity to extreme (e.g., outlier) values. Even a small number of extreme values can corrupt the mean. For example, the mean salary at a company may be substantially pushed up by that of a few highly

paid managers. Similarly, the average score of a class in an exam could be pulled down quite a bit by a few very low scores. To offset the effect caused by a small number of extreme values, we can instead use the trimmed mean, which is the mean obtained after chopping off values at the high and low extremes. For example, we can sort the values observed for salary and remove the top and bottom 2% before computing the mean. We should avoid trimming too large a portion (such as 20%) at both ends as this can result in the loss of valuable information.

For skewed (asymmetric) data, a better measure of the center of data is the median. Suppose that a given data set of N distinct values is sorted in numerical order. If N is odd, then the median is the middle value of the ordered set; otherwise (i.e., if N is even), the median is the average of the middle two values.

A holistic measure is a measure that must be computed on the entire data set as a whole. It cannot be computed by partitioning the given data into subsets and merging the values obtained for the measure in each subset. The median is an example of a holistic measure. Holistic measures are much more expensive to compute than distributive measures such as those listed above.

We can,however, easily approximate the median valueof a data set.Assumethat data are grouped in intervals according to their xi data values and that the frequency (i.e., numberof data values) of each interval is known. For example, people may be grouped according to their annual salary in intervals such as 10–20K, 20–30K, and so on. Let the interval that contains the median frequency be the median interval.We can approximate the median of the entire data set (e.g., the median salary) by interpolation using the formula:

$$median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}} \right) width,$$

Measuring the Dispersion of Data (1 Mark)

The degree to which numerical data tend to spread is called the dispersion, or variance of the data. The most common measures of data dispersion are range, the five-number summary (based on quartiles), the interquartile range, and the standard deviation. Boxplots can be plotted based on the five-number summary and are a useful tool for identifying outliers.

The K^{th} percentile of a set of data in numerical order is the value x having the property that K percent of the data entries lie at or below X. Values at or below the median M (discussed in the previous subsection) correspond to the 50th percentile.

The most commonly used percentiles other than the median are quartiles. The first quartile, denoted by Q1, is the 25th percentile; the third quartile, denoted by Q3, is the 75th percentile. The quartiles, including the median, give some indication of the center, spread, and shape of a distribution. The distance between the first and third quartiles is a sample measure of spread that gives the range covered by the middle half of the data. This distance is called the interquartile range(IQR) and is defined as

IQR=Q3-Q1

We should be aware that no single numerical measure of spread, such as IQR, is very useful for describing skewed distributions. The spreads of two sides of a skewed distribution are unequal. Therefore, it is more informative to also provide the two quartiles Q1 and Q3, along with the median, M. One common rule of thumb for identifying suspected outliers is to single out values falling at least 1:5 X IQR above the third quartile or below the first quartile.

Because Q,M, and Q3 contain no information about the endpoints (e.g., tails) of the data , a fuller summary of the shape of a distribution can be obtained by providing the highest and lowest data values as well. This is known as the five-number summary.

The five-number summary of ad distribution consists of the median M, the quartiles Q1 and Q3, and the smallest and the largest individual observations, written in the order Minimum,Q,M,Q3, Maximum. A popularly used visual representation of a distribution is the boxplot. In a boxplot: Typically, the ends of the box are at the quartiles, so that the box length is the Interquartile range, IQR.

• A line within the box marks the median

• Two lines (called whiskers) outside the box extend to the smallest(Minimum)and largest(Maximum)Observations.

When dealing with a moderate number of observations, it is worthwhile to plot potential outliers individually. To do this in a boxplot, the whiskers are extended to the extreme high and low observations only if these values are less than 1.5 X IQR beyond the Quartiles. Otherwise, the whiskers terminate at the most extreme observations occurring Within 1.5 X IQR of the quartiles

. The remaining cases are plotted individually . Boxplots can be used in the comparisons of several sets of compatible data.

Based on similar reasoning as in our analysis of the we can conclude that –Q1 and Q3 are holistic measures , as is IQR. The efficient computation of boxplots or even Approximate boxplots is interesting regarding the mining of large data sets.

Graph Displays of Basic Statistical Class Descriptions (1 Mark)

Aside from the bar charts, pie charts, and line graphs, there are also a few additional popularly used graphs for the display of data Summaries and distributions. These include histograms, quatile plots, q-q plots, scatter Plots, and curves.

Plotting histograms, or frequency histograms, is a graphical method for summarizing the distribution of a given attribute. A histogram for an attribute A partitions the data distribution of A into disjoint subsets, or buckets. Typically, the width of each bucket is uniform. Each bucket is represented by a rectangle whose height is equal to the count or relative frequency of the values at the bucket. If A is categoric, such as automobile model or item type, then one rectangle is drawn for each known value of A, and the resulting graph is more commonly referred to as a bar chart.

A quartile plot is a simple and effective way to have a first look at a unvaried Data distribution. First, it displays all of the data (allowing the user to assess both the Overallbehaviour and unusual occurrences). Second, it plots quartile information. The Mechanism used in this step is slightly different from the percentile computation. Let X(i), for(I=1 to n) be the data sorted in increasing order so that X(i) is the smallest Observation and X(n)is the largest. Each observation is paired with a percentage, which Indicates that approximately 100% of the data are below or equal to the value X(i).

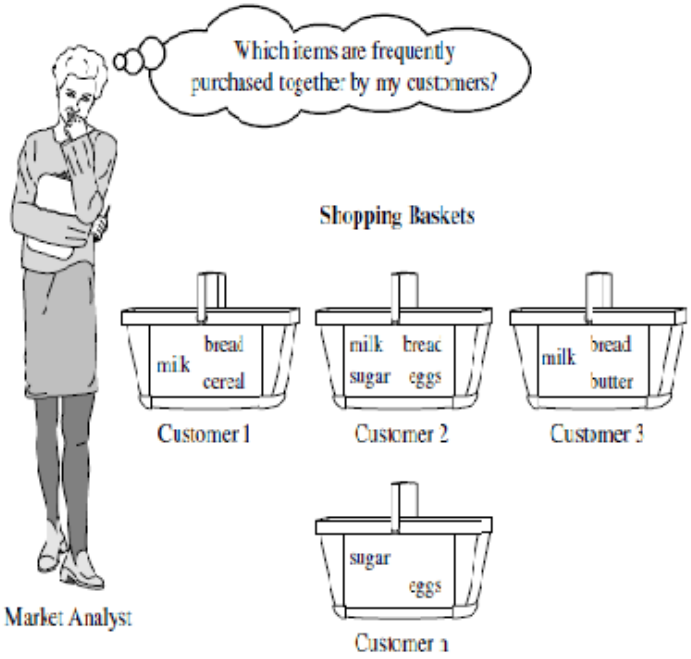
d) Describe basket analysis in association rule.

(Definition -1 Mark, Example-1 Mark, Explanation-2 Marks)

Ans: Frequent itemset mining leads to the discovery of associations and correlations among items in large transactional or relational data sets. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining such patterns from their databases. The discovery of interesting correlation relationships among huge amounts of business

transaction records can help in many business decision-making processes, such as catalog design, cross-marketing, and customershopping behavior analysis.

A typical example of frequent itemset mining is market basket analysis. This process analyzes customer buying habits by finding associations between the different items that customers place in their "shopping baskets" The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers. For instance, if customers are buying milk, how likely are they to also buy bread (and what kind of bread) on the same trip to the supermarket? Such information can lead to increased sales by helping retailers do selective marketing and plan their shelf space.



Market basket analysis

Market basket analysis Suppose, as manager of an AllElectronics branch, you would like to learn more about the buying habits of your customers. Specifically, you wonder, "Which groups or sets of items are customers likely to purchase on a given trip to the store?" To answer your question, market basket analysis may be performed on the retail data of customer transactions at your store. You can then use the results to plan marketing or advertising strategies, or in the design of a new catalog. For instance, market basket analysis may help you design different store

layouts. In one strategy, items that are frequently purchased together can be placed in proximity in order to further encourage the sale of such items together. If customers who purchase computers also tend to buy antivirus software at the same time, then placing the hardware display close to the software display may help increase the sales of both items. In an alternative strategy, placing hardware and software at opposite ends of the store may entice customers who purchase such items to pick up other items along the way. For instance, after deciding on an expensive computer, a customer may observe security systems for sale while heading toward the software display to purchase antivirus software and may decide to purchase a home security system as well. Market basket analysis can also help retailers plan which items to put on sale at reduced prices. If customers tend to purchase computers and printers together, then having a sale on printers may encourage the sale of printers as well as computers. If we think of the universe as the set of items available at the store, then each item has a Boolean variable representing the presence or absence of that item. Each basket can then be represented by a Boolean vector of values assigned to these variables. The Boolean vectors can be analyzed for buying patterns that reflect items that are frequently associated or purchased together. These patterns can be represented in the form of association rules. For example, the information that customers who purchase computers also tend to buy antivirus software at the same time is represented in

Association Rule below:

Computer=>antivirus software [support = 2%; confidence = 60%]

Rule support and confidence are two measures of rule interestingness. They respectively reflect the usefulness and certainty of discovered rules. A support of 2% for Association Rule means that 2% of all the transactions under analysis show that computer and antivirus software are purchased together. A confidence of 60% means that 60% of the customers who purchased a computer also bought the software. Typically, association rules are considered interesting if they satisfy both a minimum support threshold

and a minimum confidence threshold. Such thresholds can be set by users or domain experts. Additional analysis can be performed to uncover interesting statistical correlations between associated items.

Let τ = {i1, i2……im} be a set of items. Let D, the task-relevant data, be a set of database transactions

where each transaction T is a set of items such that T ⊂ τ each transaction is association with an identifier, called TID. Let A be a set of items. A transaction T is said to contain A if and only if A ⊆T . An association rule is an implication of the form A=>B, where A⊂ t, B⊂ t and A ∩ B = f. The rule A=> B holds in the transaction set D with support s, where s is the percentage of transaction in D that contains A ∪B (i.e both A and B). This is taken to be the probability, P(A∪B). The rule A=>B has confidence c in the transaction in the transaction set D if c is the percentage of transactions in D containing A that also contain B. This is taken to be the conditional probability P(B/A). That is

Support (A=>B)=P(A∪B)

Confidence (A=>B)=p(B/A)

Rules that satisfy both a minimum support threshold (min_sup) and a minimum confidence threshold (min_conf) are called strong. By convention, we write support and confidence value so as to occur between 0% and 100% rather than 0 to 1.0. A set of items is referred to as an itemset. An itemset that contains k item is a k-itemset. The set{computer, financial_management_software} is a 2-itemset. The occurrence frequency of an itemset is the number of transaction that contains the itemset. This is also known, simply, as the frequency, support count or count of the itemset. An itemset satisfies minimum support if the occurrence frequency of the itemset is greater than or equal to the product of min_sup and the total number of transactions in D. The number of transaction required for the itemset to satisfy minimum support is therefore referred to as the minimum support count. If an itemset satisfies minimum support, then it is a frequent itemset .The set of frequent K-itemsets is commonly denoted by LK.

Association rule mining is a two-step process:

1. Find all frequent itemsets: By definition, each of these itemsets will occur at least as frequently as a pre-determined minimum support count.

2.Generate strong association rules from the frequent itemsets: By definition, these rules must satisfy minimum support and minimum support and minimum confidence.Additional interestingness measures can be applied,if desired.The second step is the easiest,of the two.The overall performance of mining association rules is determined by the first step.

e) Explain categories and classes of DSSs.

(Categories -2 Marks, classes-2 Marks)

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 models are given below:

According to Donovan and Madnick (1977) DSS can be classified as,

1. Institutional-when the DSS supports ongoing and recurring decisions
2. Ad hoc-when the DSS supports a one off-kind of decision.

Hackathorn 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, make pricing decisions
5. Representational models. This type of DSS can also perform 'what if analysis' and calculate the outcomes of different decision paths, based on simulated models.
6. Optimization models. This kind of DSS provides solutions through the use of optimization models which have mathematical solutions.

7. Suggestion models. This kind of DSS works when the decision to be taken is based on well-structured tasks.

   Examples:Expert System• Applicant applies for personal loan

Modern classifications of DSS are,

1. Data driven DSS

These DSS has file drawer systems, data analysis systems, analysis information systems, data warehousing and emphasizes access to and manipulation of large databases of structured data

2. Model driven

The underlying model that drives the DSS can come from various disciplines or areas of specialty and might include accounting models, financial models, representation models, optimization models, etc. With model drive DSS the emphasize is on access to and manipulation of a model, rather than data, i.e. it uses data and parameters to aid decision makers in analyzing a situation. These systems usually are not data intensive and consequently are not linked to very large databases.

3. Knowledge driven

These systems provide recommendation and/or suggestion schemes which aids the user in selecting an appropriate alternative to a problem at hand. Knowledge driven DSS are often referred to as management expert systems or intelligent decision support systems. They focus on knowledge and recommends actions to managers based on an analysis of a certain knowledge base. Moreover, it has special problem solving expertise and are closely related to data mining i.e. sifting through large amounts of data to produce contend relationships.

4. Document driven

These systems help managers retrieve and mange unstructured documents and web pages by integrating a variety of storage and processing technologies to provide complete document retrieval and analysis. It also access documents such as company policies and procedures, product

specification, catalogs, corporate historical documents, minutes of meetings, important correspondence, corporate records, etc. and are usually driven by a task-specific search engine.[2]

5. Communication driven

This breed of DSS is often called group decision support systems (GDSS). They are a special type of hybrid DSS that emphasizes the use of communications and decision models intended to facilitate the solution of problems by decision makers working together as a group. GDSS supports electronic communication, scheduling, document sharing and other group productivity and decision enhancing activities and involves technologies such as two-way interactive video, bulletin boards, e-mail, etc.