Q.1 a. Define a Data warehouse. Compare OLTP and OLAP systems.

Answer:

Data Warehouse: A data warehouse is a subject-oriented, integrated, time-variant, and Non volatile collection of data in support of management's decision making process. To support business decision-making activities, including 2

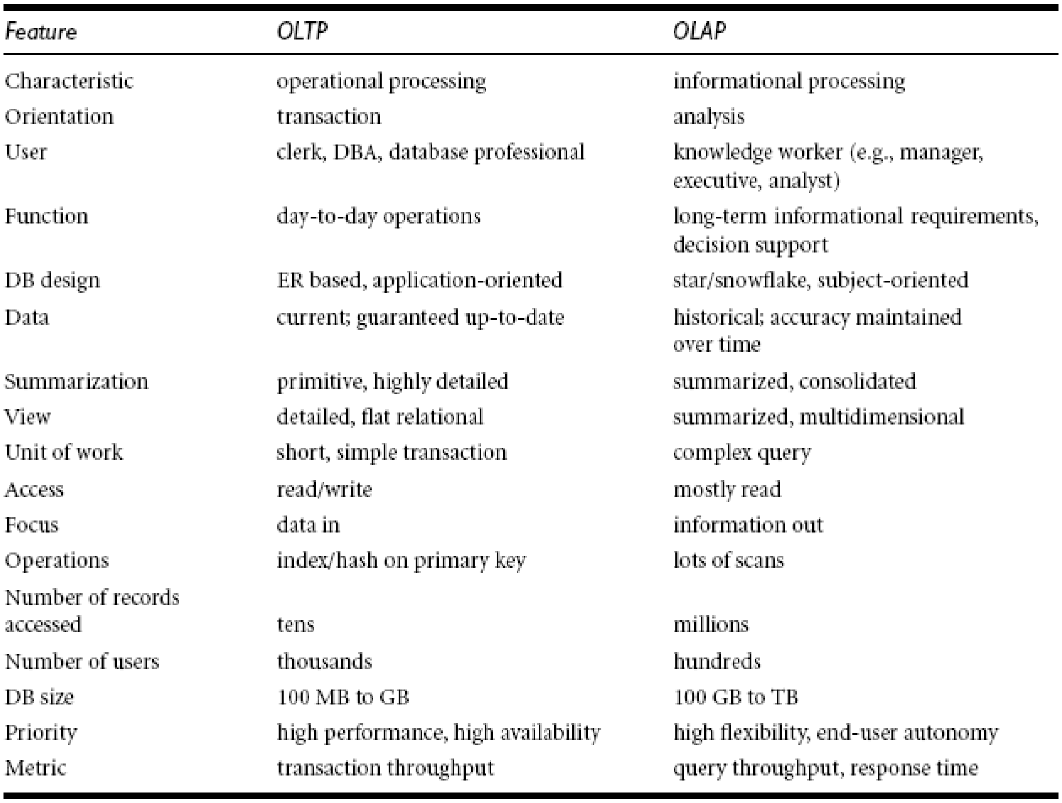
(1) increasing customer focus, which includes the analysis of customer buying patterns (such as buying preference, buying time, budget cycles, and appetites for spending);

(2) repositioning products and managing product portfolios by comparing the performance of sales by quarter, by year, and by geographic regions in order to fine tune production strategies;

(3) analyzing operations and looking for sources of profit; and

(4) managing the customer relationships, making environmental corrections, and managing the cost of corporate assets.

Comparison between OLTP and OLAP are n as follows: 5



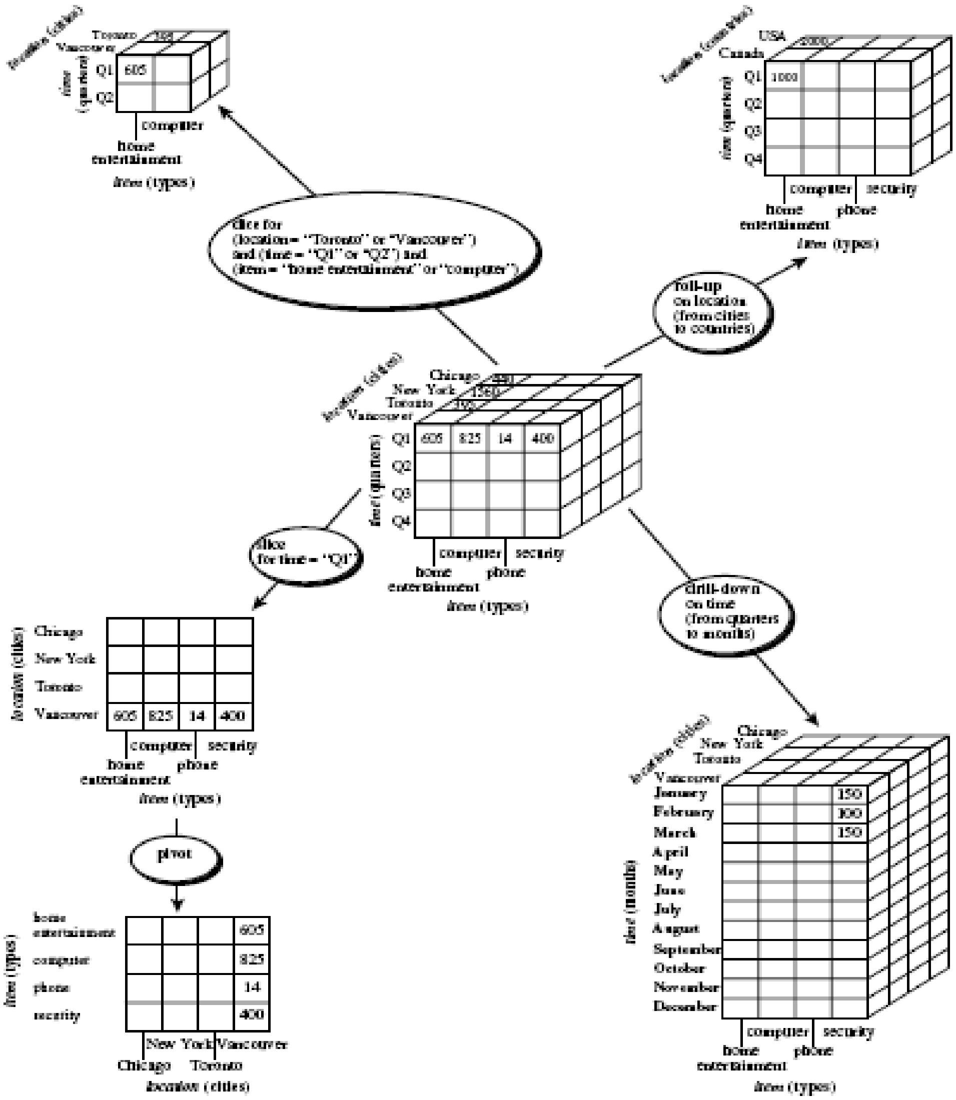
Feature	*OLTP*	*OLAP*
Characteristic	operational processing	informational processing
Orientation	transaction	analysis
User	clerk, DBA, database professional	knowledge worker (e.g., manager, executive, analyst)
Function	day-to-day operations	long-term informational requirements, decision support
DB design	ER based, application-oriented	star/snowflake, subject-oriented
Data	current; guaranteed up-to-date	historical; accuracy maintained over time
Summarization	primitive, highly detailed	summarized, consolidated
View	detailed, flat relational	summarized, multidimensional
Unit of work	short, simple transaction	complex query
Access	read/write	mostly read
Focus	data in	information out
Operations	index/hash on primary key	lots of scans
Number of records accessed	tens	millions
Number of users	thousands	hundreds
DB size	100 MB to GB	100 GB to TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

b. Discuss OLAP operations in the multidimensional data model.

Answer:

OLAP operations:

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



3.5

Explanation of OLAP operations include **rollup, drill-(down, across, through), slice-and-dice, pivot (rotate),** as well as statistical operations such as ranking and computing moving averages and growth rates. OLAP operations can be implemented efficiently using the data cube structure. **4**

c. Briefly explain agglomerative and divisive Hierarchical clustering methods.

Answer: Book 2, Page 355, 815.1

d.“Data mining tools may improve telecommunication services”. Explain.

Answer: Book 2, Page 456, 10.1.4

e. What do you understand by Noisy data? Briefly, explain any two smoothing techniques.

Answer: Book 2, Page 110, 3.22

f. State any four criteria on which classification and prediction methods can be compared and evaluated.

Answer: Book 2, Page 283, 7.2.2

g. Briefly explain Apriori algorithm. **(7×4)**

Answer: Book 2, Page 231, 6.2.1

Q.2 Briefly compare the following concepts. You may use an example to explain your point.
(i) *Snowflake* schema, *fact constellation*, *starnet* query model
(ii) Data cleaning, data transformation, data refresh
(iii) Enterprise warehouse, data mart, virtual warehouse **(3×6)**

Answer:

The most popular data model for a data warehouse is a multidimensional model. Such a model can exist in the form of a star schema, a snowflake schema, or a fact constellation schema.

Star schema: The most common modeling paradigm is the star schema, in which the data warehouse contains (1) a large central table (fact table) containing the bulk of the data, with no redundancy, and (2) a set of smaller attendant tables (dimension tables), one for each dimension. The schema graph resembles a starburst, with the dimension tables displayed in a radial pattern around the central fact table. **2**

Snowflake schema: The snowflake schema is a variant of the star schema model, where some dimension tables are normalized, thereby further splitting the data into additional tables. The resulting schema graph forms a shape similar to a snowflake. **2**

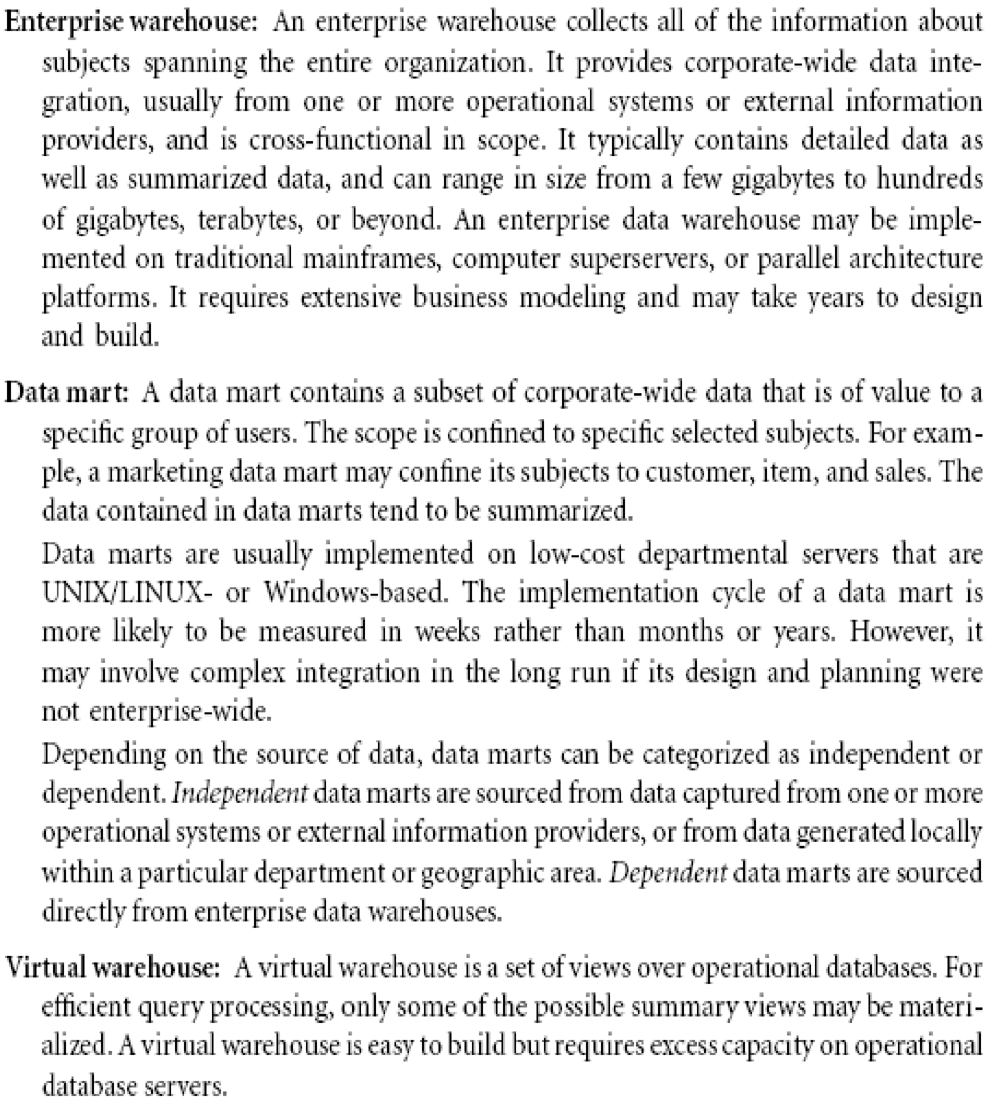
The major difference between the snowflake and star schema models is that the dimension tables of the snowflake model may be kept in normalized form to reduce redundancies. Such a table is easy to maintain and saves storage space. However, this saving of space is negligible in comparison to the typical magnitude of the fact table. Furthermore, the snowflake structure can reduce the effectiveness of browsing, since more joins will be needed to execute a query. Consequently, the system performance may be adversely impacted. Hence, although the snowflake schema reduces redundancy, it is not as popular as the star schema in data warehouse design

The querying of multidimensional databases can be based on a **starnet model.** A starnet model consists of radial lines emanating froma central point, where each line represents a concept hierarchy for a dimension. Each abstraction level in the hierarchy is called a footprint. These represent the granularities available for use by OLAP operations such as drill-down and roll-up. **2**

Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. Data cleaning is usually performed as an iterative two-step process consisting of discrepancy detection and data transformation. **2**

Data transformation routines convert the data into appropriate forms for mining. For example, attribute data may be normalized so as to fall between a small range, such as 0:0 to 1:0. 2

Data reduction techniques such as data cube aggregation, attribute subset selection, dimensionality reduction, numerosity reduction, and discretization can be used to obtain a reduced representation of the data while minimizing the loss of information content. 2



Enterprise warehouse: An enterprise warehouse collects all of the information about subjects spanning the entire organization. It provides corporate-wide data integration, usually from one or more operational systems or external information providers, and is cross-functional in scope. It typically contains detailed data as well as summarized data, and can range in size from a few gigabytes to hundreds of gigabytes, terabytes, or beyond. An enterprise data warehouse may be implemented on traditional mainframes, computer superservers, or parallel architecture platforms. It requires extensive business modeling and may take years to design and build. 2

Data mart: A data mart contains a subset of corporate-wide data that is of value to a specific group of users. The scope is confined to specific selected subjects. For example, a marketing data mart may confine its subjects to customer, item, and sales. The data contained in data marts tend to be summarized. 2

Data marts are usually implemented on low-cost departmental servers that are UNIX/LINUX- or Windows-based. The implementation cycle of a data mart is more likely to be measured in weeks rather than months or years. However, it may involve complex integration in the long run if its design and planning were not enterprise-wide.

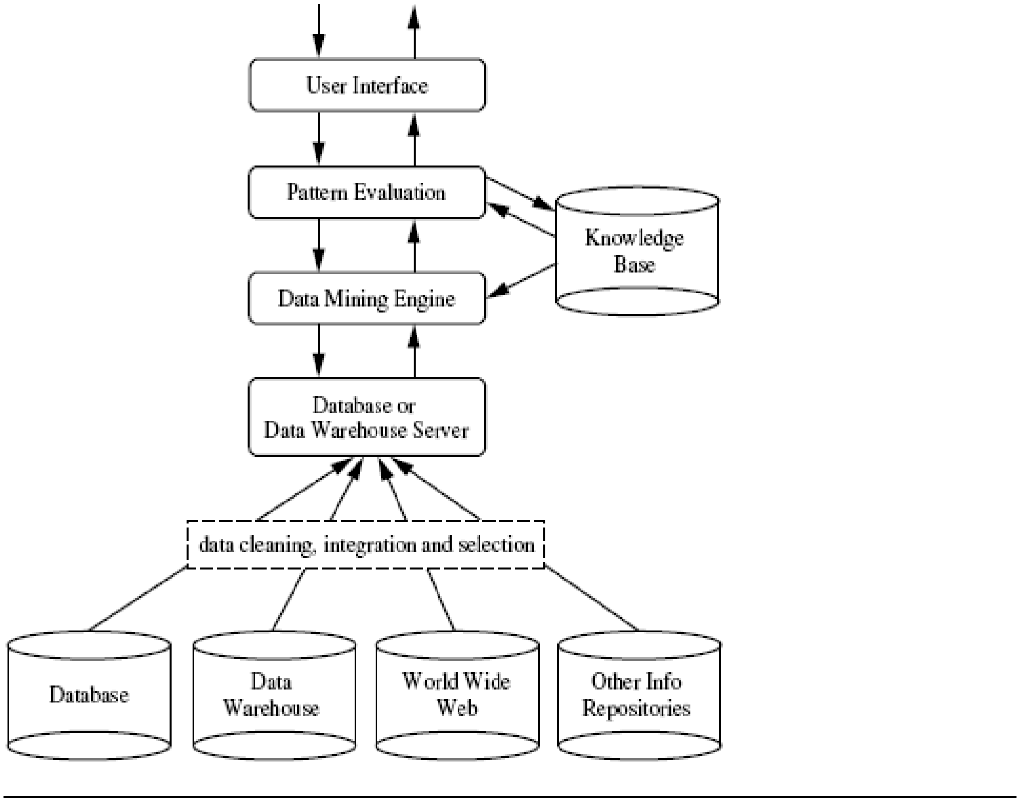
Depending on the source of data, data marts can be categorized as independent or dependent. *Independent* data marts are sourced from data captured from one or more operational systems or external information providers, or from data generated locally within a particular department or geographic area. *Dependent* data marts are sourced directly from enterprise data warehouses.

Virtual warehouse: A virtual warehouse is a set of views over operational databases. For efficient query processing, only some of the possible summary views may be materialized. A virtual warehouse is easy to build but requires excess capacity on operational database servers. 2

Q.3 a. Define Data mining. Draw and explain the architecture of typical data mining system. (10)

Answer:

Data mining: data mining refers to extracting or "mining" knowledge from large amounts of data 1



2

5 Architecture of a typical data mining system.

Database, data warehouse, WorldWideWeb, or other information repository: This is one or a set of databases, data warehouses, spreadsheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.

Database or data warehouse server: The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.

Knowledge base: This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns. Such knowledge can include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction.

Data mining engine: This is essential to the data mining systemand ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis.

Pattern evaluation module: This component typically employs interestingness measures and interacts with the data mining modules so as to focus the search toward interesting patterns.

User interface: This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results.

b. Describe three challenges to data mining regarding data mining methodology and User-interaction issues. (8)

Answer:

Mining methodology and user interaction issues: These reflect the kinds of knowledge mined, the ability to mine knowledge at multiple granularities, the use of domain knowledge, ad hoc mining, and knowledge visualization. **8**

Mining different kinds of knowledge in databases:
Interactive mining of knowledge at multiple levels of abstraction:
Incorporation of background knowledge:
Data mining query languages and ad hoc data mining:
Presentation and visualization of data mining results:
Handling noisy or incomplete data:
Pattern evaluation

Q.4 a. How does data mining relate to information processing and online analytical processing? (10)

Answer:

"How does data mining relate to information processing and on-line analytical processing?"
Explanation with Justification: **10**

Information processing, based on queries, can find useful information. However, answers to such queries reflect the information directly stored in databases or computable by aggregate functions. They do not reflect sophisticated patterns or regularities buried in the database. Therefore, information processing is not data mining. On-line analytical processing comes a step closer to data mining because it can derive information summarized at multiple granularities from user-specified subsets of a data warehouse.

b. What is metadata? Explain metadata repository. (8)

Answer:

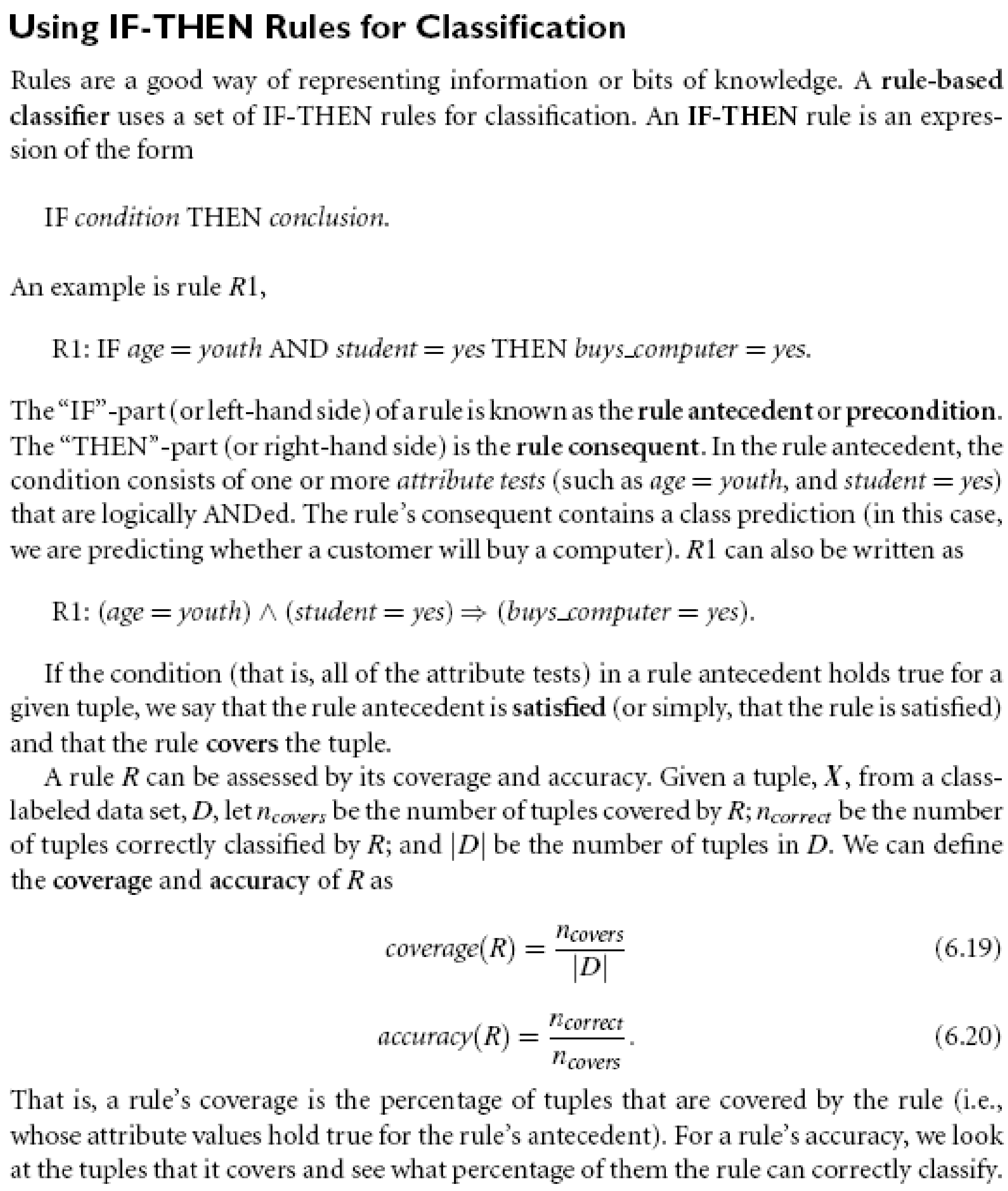
Metadata are data about data. When used in a data warehouse, metadata are the data that define warehouse objects. **1**

A metadata repository should contain the following:

Description of the **structure of the data warehouse**, which includes the warehouse schema, view, dimensions, hierarchies, and derived data definitions, as well as data mart locations and contents **7**
Operational metadata, which include data lineage (history of migrated data and the sequence of transformations applied to it), currency of data (active, archived, or purged), and monitoring information (warehouse usage statistics, error reports, and audit trails)

The **algorithms used for summarization**, which include measure and dimension definition algorithms, data on granularity, partitions, subject areas, aggregation, summarization, and predefined queries and reports
The **mapping from the operational environment to the data warehouse**, which includes
source databases and their contents, gateway descriptions, data partitions, data extraction, cleaning, transformation rules and defaults, data refresh and purging rules, and security (user authorization and access control)
Data related to system performance, which include indices and profiles that improve data access and retrieval performance, in addition to rules for the timing and scheduling of refresh, update, and replication cycles
Business metadata, which include business terms and definitions, data ownership information, and charging policie

Q.5 a. Explain how Rule-based classification is used in data mining. (10)
Answer:
Rule-based classifiers, where the learned model is represented as a set of IF-THEN rules: **10**



Using IF-THEN Rules for Classification

Rules are a good way of representing information or bits of knowledge. A **rule-based classifier** uses a set of IF-THEN rules for classification. An **IF-THEN** rule is an expression of the form

IF *condition* THEN *conclusion*.

An example is rule $R1$,

R1: IF $age = youth$ AND $student = yes$ THEN $buys\_computer = yes$.

The "IF"-part (or left-hand side) of a rule is known as the **rule antecedent** or **precondition**. The "THEN"-part (or right-hand side) is the **rule consequent**. In the rule antecedent, the condition consists of one or more *attribute tests* (such as $age = youth$, and $student = yes$) that are logically ANDed. The rule's consequent contains a class prediction (in this case, we are predicting whether a customer will buy a computer). $R1$ can also be written as

R1: $(age = youth) \wedge (student = yes) \Rightarrow (buys\_computer = yes)$.

If the condition (that is, all of the attribute tests) in a rule antecedent holds true for a given tuple, we say that the rule antecedent is **satisfied** (or simply, that the rule is satisfied) and that the rule **covers** the tuple.

A rule R can be assessed by its coverage and accuracy. Given a tuple, X, from a class-labeled data set, D, let n_{covers} be the number of tuples covered by R; $n_{correct}$ be the number of tuples correctly classified by R; and $|D|$ be the number of tuples in D. We can define the **coverage** and **accuracy** of R as

$$coverage(R) = \frac{n_{covers}}{|D|} \tag{6.19}$$

$$accuracy(R) = \frac{n_{correct}}{n_{covers}}. \tag{6.20}$$

That is, a rule's coverage is the percentage of tuples that are covered by the rule (i.e., whose attribute values hold true for the rule's antecedent). For a rule's accuracy, we look at the tuples that it covers and see what percentage of them the rule can correctly classify.

b. Describe Classification based on Multiple Association Rules (CMAR) with an example. (8)

Answer:

Associative classification, where association rules are generated and analyzed for use in classification. One of the earliest and simplest algorithms for associative classification is CBA (Classification-Based Association). CBA uses an iterative approach to frequent item set mining. **8**

CMAR (Classification based on Multiple Association Rules) differs from CBA in its strategy for frequent itemset mining and its construction of the classifier. It also employs several rule pruning strategies with the help of a tree structure for efficient storage and retrieval of rules. CMAR adopts a variant of the FP-growth algorithm to find the

complete set of rules satisfying the minimum confidence and minimum support thresholds.
FP-growth uses a tree structure, called an FP-tree, to register all of the frequent itemset information contained in the given data set, D. This requires only two scans of D. The frequent itemsets are then mined from the FP-tree. CMAR uses an enhanced FP-tree that maintains the distribution of class labels among tuples satisfying each frequent itemset. In this way, it is able to combine rule generation together with frequent itemset mining in a single step.

CMAR employs another tree structure to store and retrieve rules efficiently and to prune rules based on confidence, correlation, and database coverage. Rule pruning strategies are triggered whenever a rule is inserted into the tree. For example, given two rules, R1 and R2, if the antecedent of R1 is more general than that of R2 and conf(R1) _ conf(R2), then R2 is pruned. The rationale is that highly specialized rules with low confidence can be pruned if a more generalized version with higher confidence exists. CMAR also prunes rules for which the rule antecedent and class are not positively correlated, based on a c2 test of statistical significance.

Q.6 a. What is cluster analysis? Explain major categorization of clustering Methods. (10)
Answer:
Cluster Analysis:
The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. 2

Cluster analysis has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing. In business, clustering can help marketers discover distinct groups in their customer bases and characterize customer groups based on purchasing patterns.

Clustering can be categorized into **partitioning methods, hierarchical methods, density-based methods, grid-based methods, model-based methods, methods for high-dimensional data (including frequent pattern–based methods), and constraintbased methods.**

A **partitioning method** first creates an initial set of k partitions, where parameter k is the number of partitions to construct. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. Typical partitioning methods include k-means, k-medoids, CLARANS, and their improvements. 2
A **hierarchical method** creates a hierarchical decomposition of the given set of data objects. The method can be classified as being either agglomerative (bottom-up) or 2 divisive (top-down), based on how the hierarchical decomposition is formed. To compensate for the rigidity of merge or split, the quality of hierarchical agglomeration can be improved by analyzing object linkages at each hierarchical partitioning (such as in ROCK and Chameleon), or by first performing microclustering (that is, grouping objects

into “microclusters”) and then operating on the microclusters with other clustering techniques, such as iterative relocation (as in BIRCH).

A **density-based method** clusters objects based on the notion of density. It either grows clusters according to the density of neighborhood objects (such as in DBSCAN) or according to some density function (such as in DENCLUE). OPTICS is a densitybased method that generates an augmented ordering of the clustering structure of the data. **2**

Clustering high-dimensional data is of crucial importance, because in many advanced applications, data objects such as text documents and microarray data are high-dimensional in nature. There are three typical methods to handle highdimensional data sets: dimension-growth subspace clustering, represented by CLIQUE, dimension-reduction projected clustering, represented by PROCLUS, and frequent pattern–based clustering, represented by pCluster. **2**

b. What do you mean by a web-enabled data-warehouse? Describe three of its functional features **(8)**

Answer:

Book 1, Page 58 - 60

Q.7 Write short note on **(3×6)**

(i) Data Mining for the Retail Industry

(ii) Statistical Data Mining

(iii) Strategies of data reduction

Answer:

(i) The retail industry is a major application area for data mining, since it collects huge amounts of data on sales, customer shopping history, goods transportation, consumption, and service. The quantity of data collected continues to expand rapidly, especially due to the increasing ease, availability, and popularity of business conducted on theWeb, or e-commerce. Today, many stores also have websites where customers can make purchases on-line. Some businesses, such as Amazon.com (www.amazon.com), exist solely on-line, without any brick-and-mortar (i.e., physical) store locations. Retail data provide a rich source for data mining. **6**

A few examples of data mining in the retail industry are outlined as follows:

Design and construction of data warehouses based on the benefits of data mining
Multidimensional analysis of sales, customers, products, time, and region
Analysis of the effectiveness of sales campaigns
Customer retention—analysis of customer loyalty
Product recommendation and cross-referencing of items:

(ii) The efficient handling of huge amounts of data that are typically multidimensional and possibly of various complex types. There are, however, many well-established statistical techniques for data analysis, particularly for numeric data. These techniques have been applied extensively to some types of scientific data (e.g., dataFrom experiments in physics, engineering, manufacturing, psychology, and medicine), as well as to data from economics and the social sciences. **6**

Major statistical methods for data analysis are: **Regression, Generalized linear models, Analysis of variance, Factor analysis, Time series analysis etc.**

(iii) 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. **6**

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 rawdata 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.

TEXT BOOKS

1. J Han and M Kamber, "Data Mining: Concepts and Techniques "2e, Elservier, 2006

2. I.H Witten and E Frank, "Data Mining: Practical Machine Learning Tools and Techniques, "2e, Elsevier, 2005