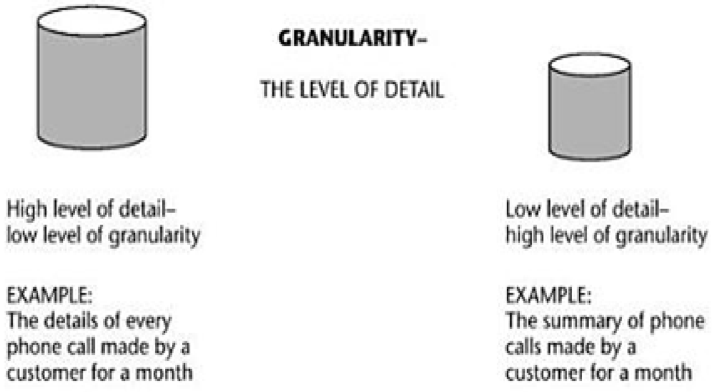
Q.1 a. Briefly explain “data granularity” with the help of example

Answer:

Data Granularity:

The single most important aspect and issue of the design of the data warehouse is the issue of granularity. It refers to the detail or summarization of the units of data in the data warehouse. The more detail there is, the lower the granularity level. The less detail there is, the higher the granularity level.

Granularity is a major design issue in the data warehouse as it profoundly affects the volume of data. The figure below shows the issue of granularity in a data warehouse.



3 + 1 for example

b. Differentiate between Data Warehouse and Data Mart.

Answer:

Data Warehouse and Data Mart:

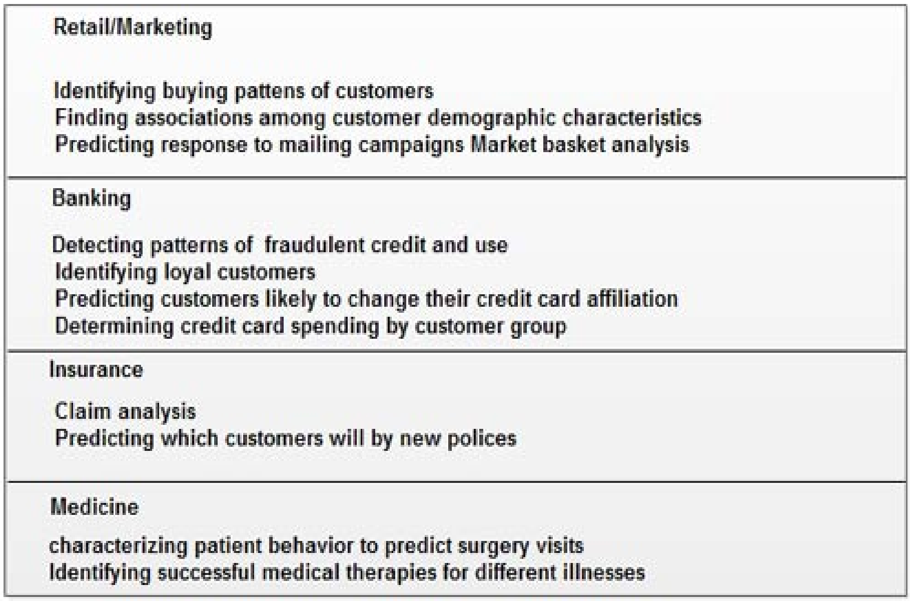
- Data mart and data warehousing are tools to assist management to come up with relevant information about the organization at any point of time
- While data marts are limited for use of a department only, data warehousing applies to an entire organization
- Data marts are easy to design and use while data warehousing is complex and difficult to manage
- Data warehousing is more useful as it can come up with information from any department
- Delaware house: It is a collection of data marts. Represents historical data.
- Data mart: It is a sub set of data ware housing. It can provide the data to analyze query reporting & analysis.

Any four differences can be Treated as answer

c. List the applications of Data Mining.

Answer:

Applications of Data Mining:



Retail/Marketing

Identifying buying pattens of customers
Finding associations among customer demographic characteristics
Predicting response to mailing campaigns Market basket analysis

Banking

Detecting patterns of fraudulent credit and use
Identifying loyal customers
Predicting customers likely to change their credit card affiliation
Determining credit card spending by customer group

Insurance

Claim analysis
Predicting which customers will by new polices

Medicine

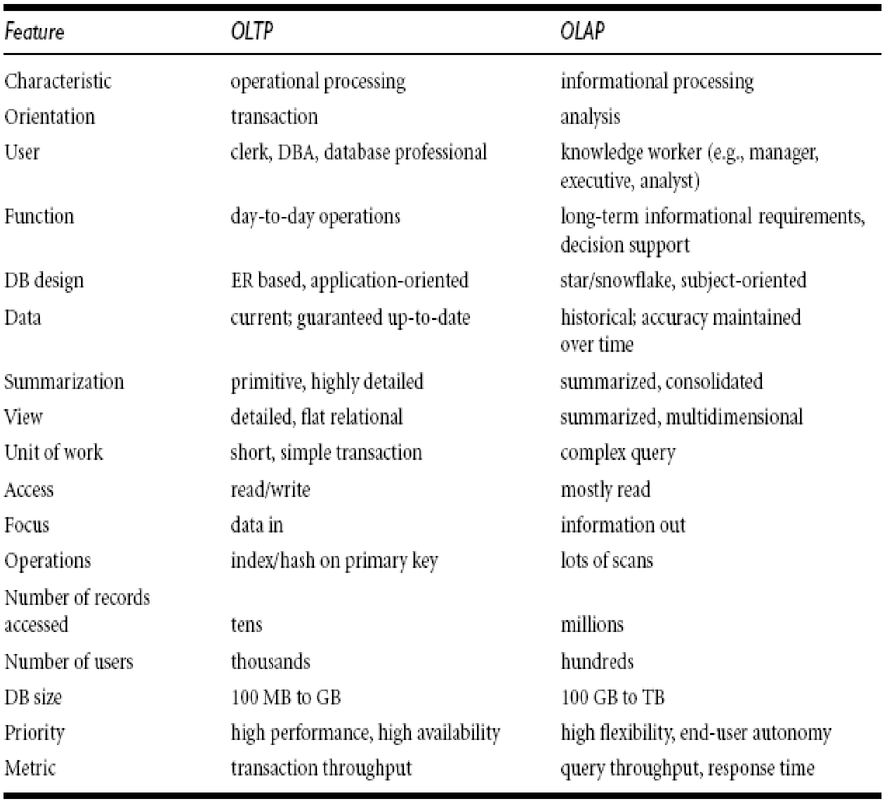
characterizing patient behavior to predict surgery visits
Identifying successful medical therapies for different illnesses

Any four field or Applications

d. Define a Data warehouse. Compare OLTP and OLAP systems.

Answer:

Compare OLTP and OLAP systems:



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

1 mark for definition of data mining + 3 marks for difference

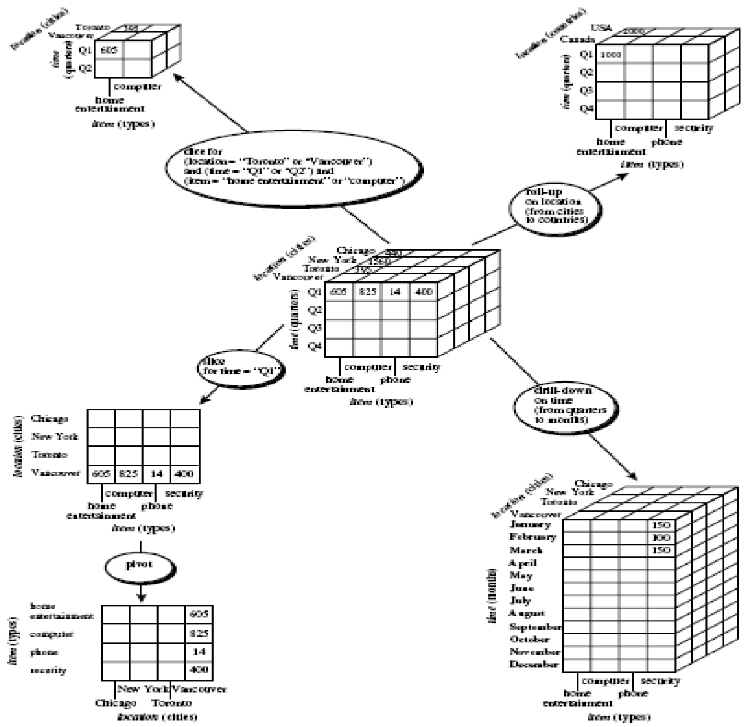
{Any SIX differences}

e. Discuss OLAP operations in the multidimensional data model.

Answer:

OLAP operations:

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



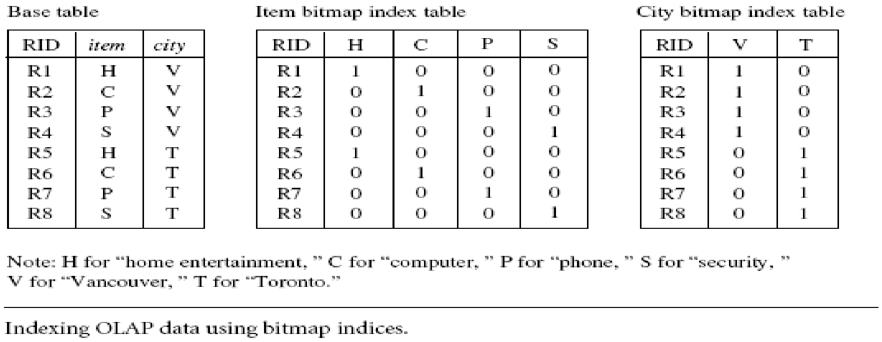
1 Mark for each operation

f. **Explain different indexing techniques for OLTP data**

Answer:

Indexing techniques for OLTP data:

In bitmap indexing, each attribute has its own bitmap index table. Bitmap indexing reduces join, aggregation, and comparison operations to bit arithmetic



Base table

RID	*item*	*city*
R1	H	V
R2	C	V
R3	P	V
R4	S	V
R5	H	T
R6	C	T
R7	P	T
R8	S	T

Item bitmap index table

RID	H	C	P	S
R1	1	0	0	0
R2	0	1	0	0
R3	0	0	1	0
R4	0	0	0	1
R5	1	0	0	0
R6	0	1	0	0
R7	0	0	1	0
R8	0	0	0	1

City bitmap index table

RID	V	T
R1	1	0
R2	1	0
R3	1	0
R4	1	0
R5	0	1
R6	0	1
R7	0	1
R8	0	1

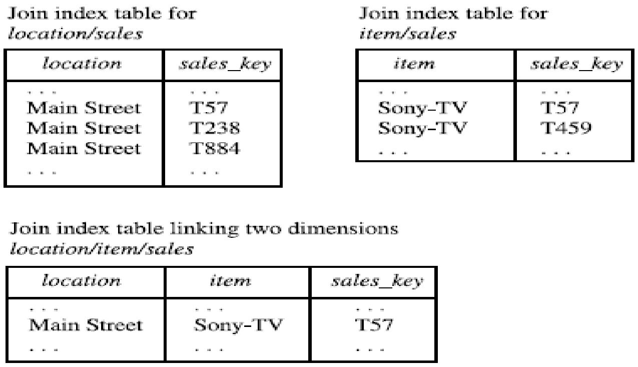
Note: H for "home entertainment, " C for "computer, " P for "phone, " S for "security, "
V for "Vancouver, " T for "Toronto."

Indexing OLAP data using bitmap indices.

2

Join indexing registers the joinable rows of two or more relations from a relational database, reducing the overall cost of OLAP join operations.

2



Join index table for *location/sales*

location	*sales_key*
. . .	. . .
Main Street	T57
Main Street	T238
Main Street	T884
. . .	. . .

Join index table for *item/sales*

item	*sales_key*
. . .	. . .
Sony-TV	T57
Sony-TV	T459
. . .	. . .

Join index table linking two dimensions *location/item/sales*

location	*item*	*sales_key*
. . .	. . .	. . .
Main Street	Sony-TV	T57
. . .	. . .	. . .

Join index tables based on the linkages between the *sales* fact table and dimension tables for *location* and *item* shown in Figure 3.16.

g. Explain different types of OLAP servers (7×4)

Answer:

Different types of OLAP servers:

A ROLAP server uses an extended relational DBMS that maps OLAP operations on multidimensional data to standard relational operations. These are the intermediate servers that stand in between a relational back-end server and client front-end tools. They use a relational or extended-relationalDBMS to store and manage warehouse data, and OLAP middleware to support missing pieces. ROLAP servers include optimization for each DBMS back end, implementation of aggregation navigation logic, and additional tools and services. ROLAP technology tends to have greater scalability than MOLAP technology. **1**

A MOLAP server maps multidimensional data views directly to array structures. These servers support multidimensional views of data through array-based multidimensional storage engines. They map multidimensional views directly to data cube array structures. The advantage of using a data cube is that it allows fast indexing to precomputed summarized data. Notice that with multidimensional data stores, the storage utilization may be low if the data set is sparse. **1**

The Hybrid OLAP (HOLAP) Servers approach combines ROLAP and MOLAP technology, benefiting from the greater scalability of ROLAP and the faster computation of MOLAP. For example, a HOLAP server may allow large volumes of detail data to be stored in a relational database, while aggregations are kept in a separate MOLAP store. **1**

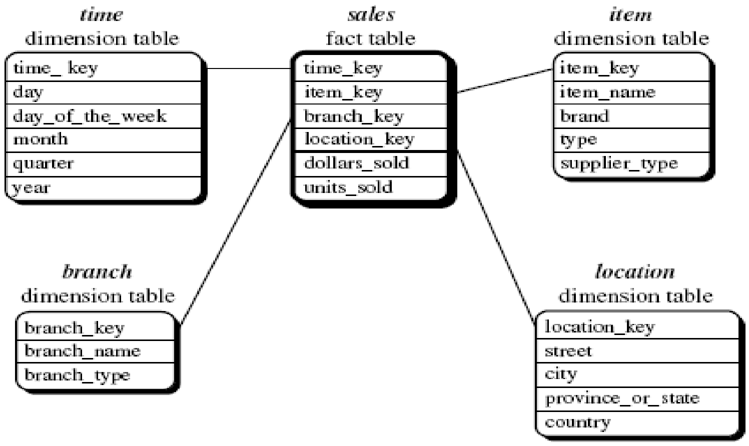
Specialized SQL servers: To meet the growing demand of OLAP processing in relational databases, some database system vendors implement specialized SQL servers that provide advanced query language and query processing support for SQL queries over star and snowflake schemas in a read-only environment **1**

Q.2 a. What are the various Star Schema Keys? Explain with the help of an example. (9)

Answer:

Star Schema Keys:

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. **3**



3

4 Star schema of a data warehouse for sales.

define cube sales star [time, item, branch, location]:
dollars sold = sum(sales in dollars), units sold = count(*)

3

define dimension time as (time key, day, day of week, month, quarter, year)
define dimension item as (item key, item name, brand, type, supplier type)
define dimension branch as (branch key, branch name, branch type)
define dimension location as (location key, street, city, province or state, country)

b. Explain Data cleaning, Data Integration and Data transformation. (9)

Answer:

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. 3

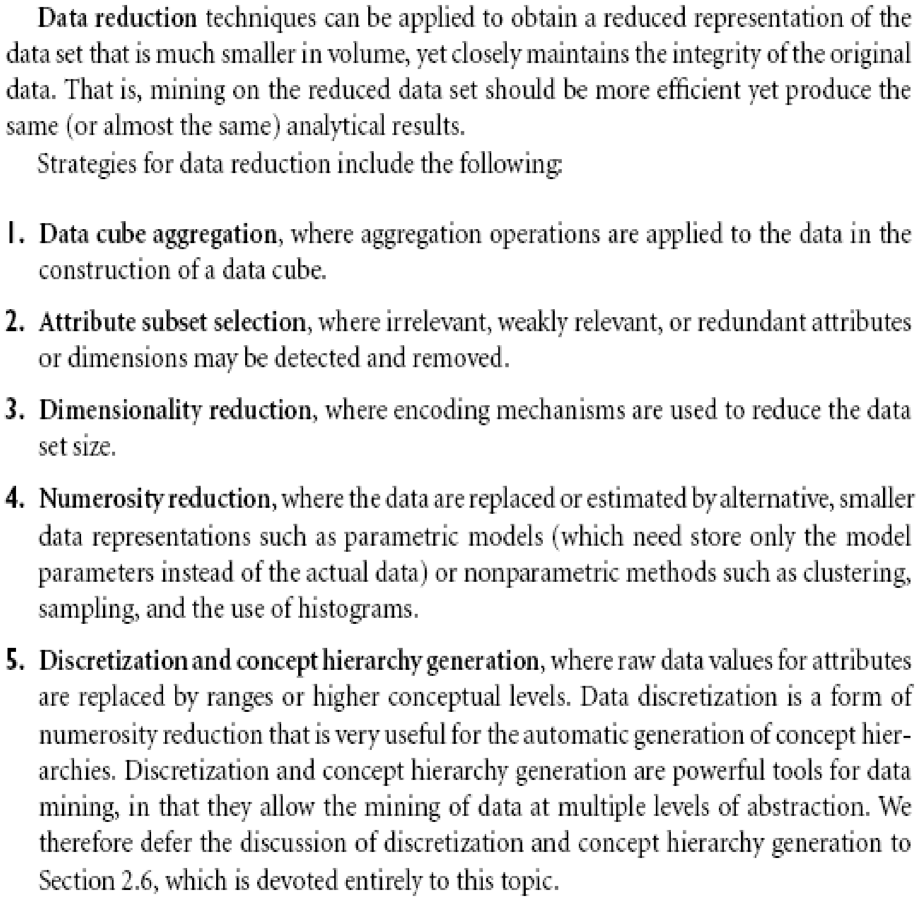
Data transformation routines convert the data into appropriate forms for mining. For example, attribute data may be normalized so as to fall between a small range, such as 0:0 to 1:0. 3

Data Integration: Combine data from multiple sources into a coherent data store as in data warehousing. These sources may include multiple databases, data cubes or flat files. 3

Q.3 a. What do you mean by data reduction? What are the strategies of the data reduction? (9)

Answer:

Data reduction:



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. We therefore defer the discussion of discretization and concept hierarchy generation to Section 2.6, which is devoted entirely to this topic.

2

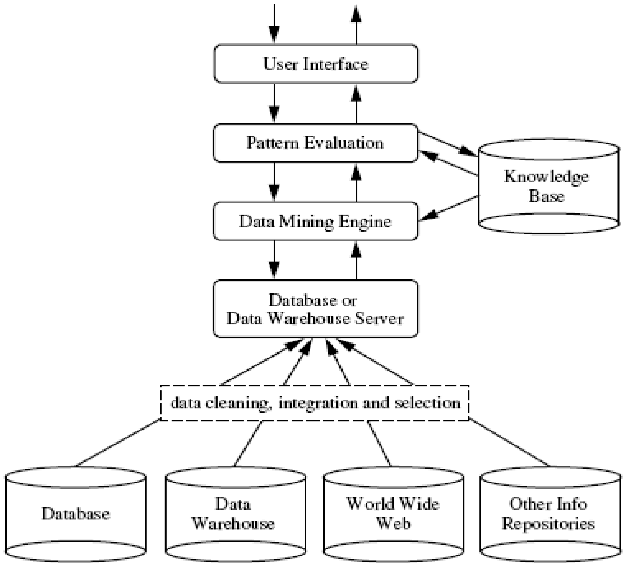
5

+ 2 marks for example if any

b. Define Data mining. Draw and explain the architecture of typical data mining system. (9)

Answer:

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



5 Architecture of a typical data mining system.

3 marks for definition
+
6 for figure and explanations of components

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.

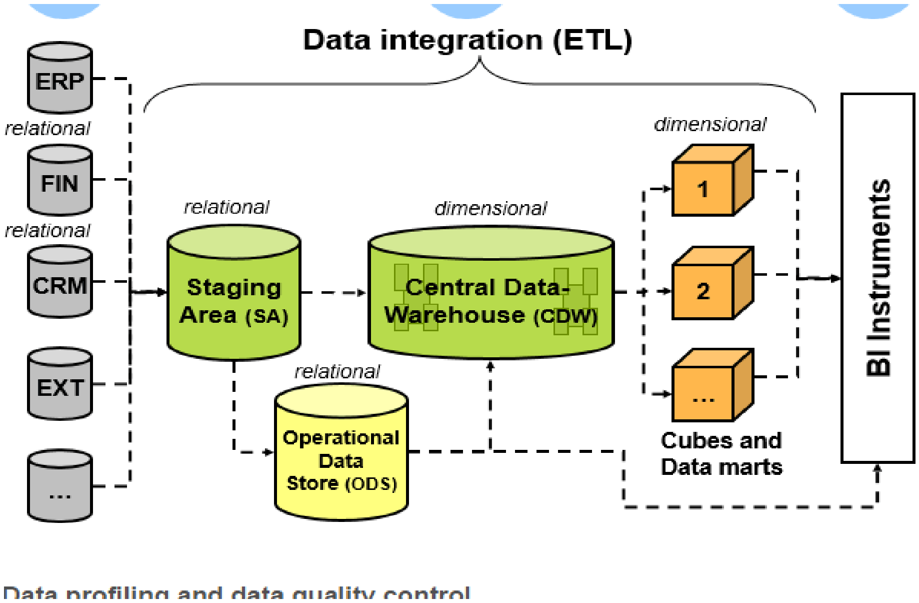
Database or data warehouse server: The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.

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.

Q.4 a. What is ETL (Extraction /Transformation /Loading) process? Discuss in brief.(9)
Answer:
ETL comes from Data Warehousing and stands for Extract-Transform-Load. ETL covers a process of how the data are loaded from the source system to the data warehouse. Currently, the ETL encompasses a cleaning step as a separate step. The sequence is then Extract-Clean-Transform-Load.



3

Extract

The Extract step covers the data extraction from the source system and makes it accessible for further processing. The main objective of the extract step is to retrieve all the required data from the source system with as little resources as possible. The extract step should be designed in a way that it does not negatively affect the source system in terms or performance, response time or any kind of locking. 2

Transform

The transform step applies a set of rules to transform the data from the source to the target. This includes converting any measured data to the same dimension (i.e. conformed dimension) using the same units so that they can later be joined. The transformation step also requires joining data from several sources, generating aggregates, generating surrogate keys, sorting, deriving new calculated values, and applying advanced validation rules. 2

Load

During the load step, it is necessary to ensure that the load is performed correctly and with as little resources as possible. The target of the Load process is often a database. In order to make the load process efficient, it is helpful to disable any constraints and indexes before the load and enable them back only after the load completes. The referential integrity needs to be maintained by ETL tool to ensure consistency. 2

b. How does data mining relate to information processing and online analytical processing? (9)

Answer:

Explanation with Justification:

More data required for 9 marks answer + examples of info processing & online A Processing

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.

Q.5 a. What is metadata? Explain metadata repository. (9)

Answer:

Metadata are data about data. When used in a data warehouse, metadata are the data that define warehouse objects. 3

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

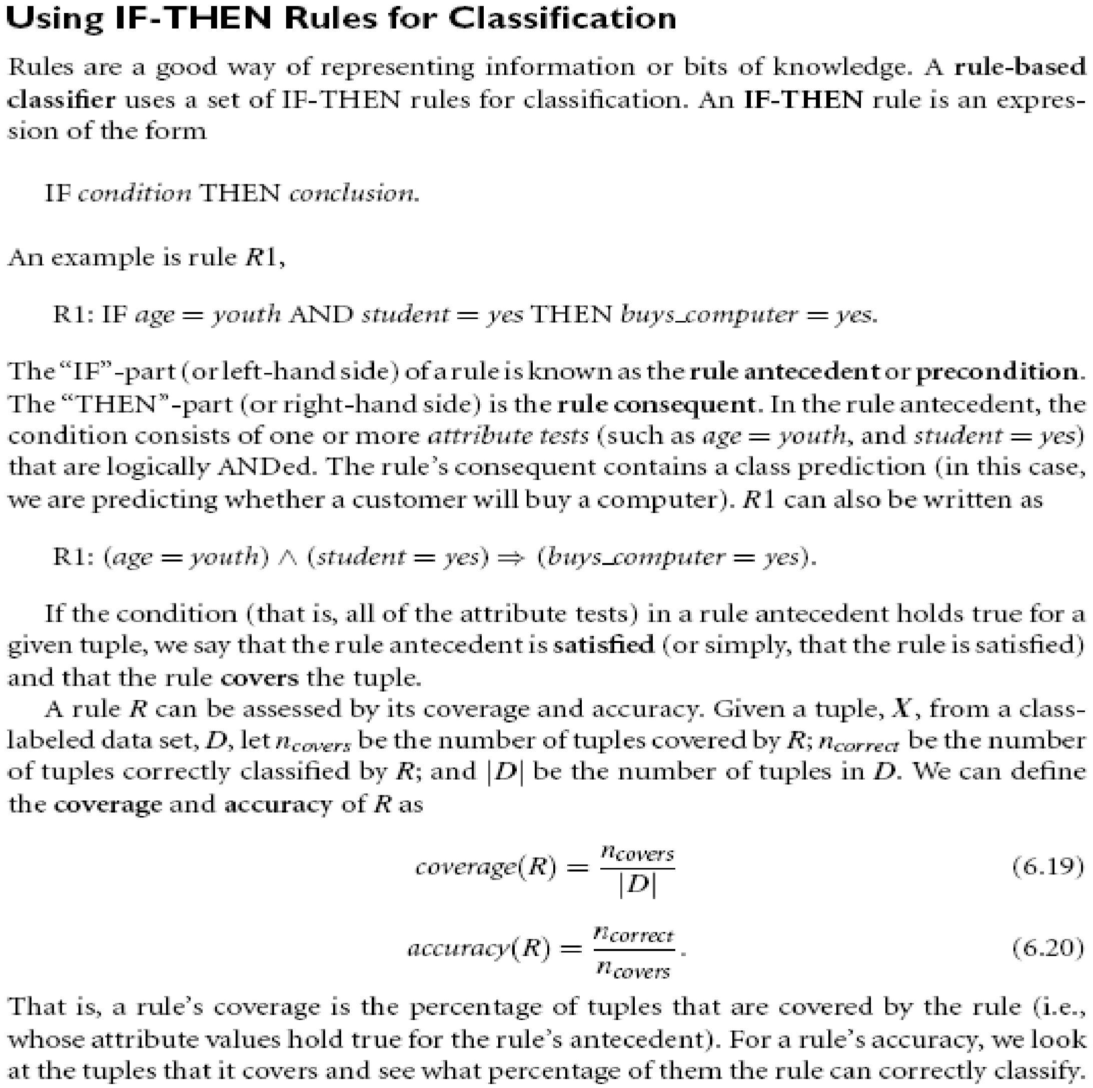
Operational metadata, which include data lineage (history of migrated data and the sequence of transformations applied to it), currency of data (active, archived, or

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 6
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

b. Explain how Rule-based classification is used in data mining? (9)

Answer:

Rule-based classifiers, where the learned model is represented as a set of IF-THEN rules:



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 $buys\_computer = yes$. 5

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 4

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

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

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.

Q.6 a. What is cluster analysis? Explain major categorization of clustering Methods. (9)

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.

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.

4

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.

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 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).

6 marks for different methods

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.

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.

b. How is data warehouse different from database? Justify your answer. (9)

Answer:

9 based on different points

Explanation with Justification:

A database designed to handle transactions isn’t designed to handle analytics. It isn’t structured to do analytics well. A data warehouse, on the other hand, is structured to make analytics fast and easy.

database” is one designed to make transactional systems run efficiently. Typically, this type of database is an OLTP (online transaction processing) database. an OLTP database is typically constrained to a single application.

The fact is that a transactional database doesn’t lend itself to analytics. To effectively perform analytics, you need a data warehouse. A data warehouse is a database of a different kind: an OLAP (online analytical processing) database. A data warehouse exists as a layer on top of another database or databases (usually OLTP databases). The data warehouse takes the data from all these databases and creates a layer optimized for and dedicated to analytics

Q.7 Write short note on :
(i) Data Mining for the Retail Industry
(ii) Statistical Data Mining
(iii) Constraint Based Association mining **(3×6)**

Answer:

(i) Data Mining for the Retail Industry:

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. **3**

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 **3**
- Customer retention—analysis of customer loyalty
- Product recommendation and cross-referencing of items:

(ii) Statistical Data Mining:

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., data **6**
From experiments in physics, engineering, manufacturing, psychology, and medicine), as well as to data from economics and the social sciences.
Major statistical methods for data analysis are: Regression, Generalized linear models, Analysis of variance, Factor analysis, Time series analysis etc.

(iii) Constraint-Based Association Mining:

Knowledge type constraints: These specify the type of knowledge to be mined, such
as association or correlation.
Data constraints: These specify the set of task-relevant data.
Dimension/level constraints: These specify the desired dimensions (or attributes) of

the data, or levels of the concept hierarchies, to be used in mining.
Interestingness constraints: These specify thresholds on statistical measures of rule interestingness, such as support, confidence, and correlation.
Rule constraints: These specify the form of rules to be mined. Such constraints may be expressed as 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.

6

<u>TEXT BOOK</u>

I. J Han and M Kamber, "Data Mining: Concepts and Techniques "2e,Elservier, 2006
II. I.H Witten and E Frank, " Data Mining: Practical Machine Learning Tools and Techniques, "2e, Elsevier, 2005