

# DR. SUBHASH UNIVERSITY

## School of Engineering and Technology

### Department of CSE / IT

Q.1 What is Data Mining? Explain Steps for Data mining or KDD process.

Data Mining:

Data mining refers to extracting or “mining” knowledge from large amounts of data. Many people treat data mining as a synonym for another popularly used term, knowledge Discovery from Data, or KDD.

Many other terms carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging.

Steps for KDD (Knowledge Discovery from Data):

1. Data cleaning: to remove noise and inconsistent data
2. Data integration: where multiple data sources may be combined
3. Data selection: where data relevant to analysis task are retrieved from the database
4. Data transformation: where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregate operation
5. Data mining: an essential process where intelligent methods are applied in order to extract data patterns
6. Pattern evaluation: to identify the truly increasing patterns representing knowledge based on some interestingness measures
7. Knowledge presentation: where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

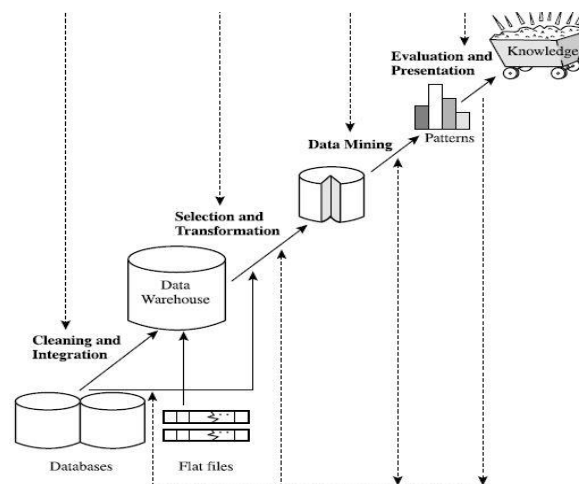
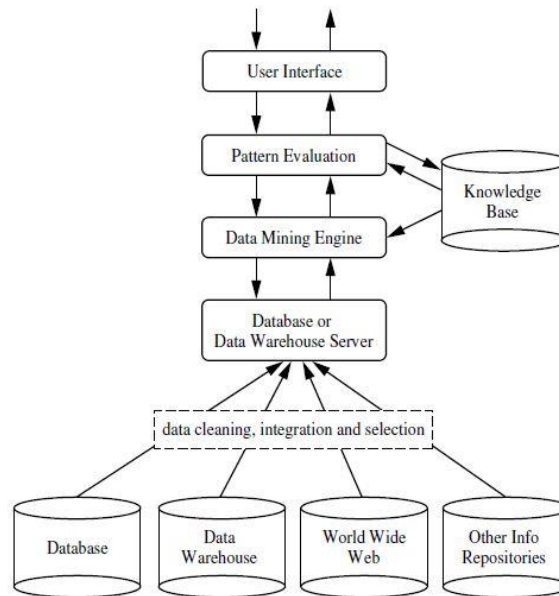


Fig. Steps For KDD.

**DR. SUBHASH UNIVERSITY**  
**School of Engineering and Technology**  
**Department of CSE / IT**

Q.2 Data mining Architecture.



- ✓ Database, data warehouse, Worldwide Web, 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.
- ✓ Data mining engine: This is essential to the data mining system and 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.

**DR. SUBHASH UNIVERSITY**  
**School of Engineering and Technology**  
**Department of CSE / IT**

- ✓ 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. It may use interestingness thresholds to filter out discovered patterns. Alternatively, the pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining method used.
- ✓ 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. In addition, this component allows the user to browse database and data warehouse schemas or data Structures, evaluate mined patterns, and visualize the patterns in different forms.

**Q.3 Data Mining—On What Kind of Data?**

- ✓ Flat File: Flat files are actually the most common data source for data mining algorithm, especially at the research level. Flat files are simple data files in text or binary format with a structure known by the data mining algorithm to be applied. The Data in these files can be transactions, time-series data, and scientific measurements.
- ✓ Relational Databases: A relational database consists of a set of tables containing either values of entity attributes, or values of attributes from entity relationships. Tables have columns and rows, where columns represent attributes and rows represent tuples. A tuple in a relational table correspond to either an object or a relationship between objects and is identified by a set of attribute values representing a unique key.
- ✓ Data Warehouses: A data warehouse as a storehouse, is a repository of data collected from multiple data sources (often heterogeneous) and is intended to be used as a whole under the same unified schema. A data warehouse gives the option to analyze data form different source sunder the same roof. Data from the different stores would be loaded, cleaned, transformed and integrated together.
- ✓ Multimedia Databases: Multimedia databases include video, images, audio, and text Media. They can be stored on extended object-relational or object-oriented databases, or simply on a file system. Multimedia is characterized by its high dimensionality, which makes data mining even more challenging. Data mining from multimedia

# DR. SUBHASH UNIVERSITY

## School of Engineering and Technology

### Department of CSE / IT

repositories may require computer vision, computer graphics, image interpretation, and natural language processing methodologies.

- ✓ **Spatial Databases:** Spatial databases are databases that in addition to usual data, store geographical information like maps, and global or regional positioning.
- ✓ **Time-Series Databases:** Time-series databases contain time related data such stock market data or logged activities. These databases usually have a continuous flow of new data coming in, which sometimes causes the need for a challenging real time analysis.
- ✓ **World Wide Web:** The World Wide Web is the most heterogeneous and dynamic repository available. Data in the World Wide Web is organized in interconnected documents. These documents can be text, audio, video, raw data and even applications. Conceptually, the World Wide Web is comprised of three major components: The content of the Web, which encompasses documents available, the structure of the Web, which covers the hyperlink and the relationships between documents; and the usage of the web, describing how and when the resources are accessed. A fourth dimension can be added relating the dynamic nature or evolution of the documents. Data mining in the World Wide Web or Web mining, tries to address all these issues and is often divided into web content mining, web structure mining and web usage mining.

#### Q.4 Data Mining Functionalities—What Kinds of Patterns Can Be Mined?

We have observed various types of databases and information repositories on which data Mining can be performed. Let us now examine the kinds of data patterns that can be mined.

In general, data mining tasks can be classified into two categories: descriptive and predictive.

- ✓ **Descriptive mining:** tasks characterize the general properties of the data in the database.
- ✓ **Predictive mining:** tasks perform inference on the current data in order to make predictions.

**DR. SUBHASH UNIVERSITY**  
**School of Engineering and Technology**  
**Department of CSE / IT**

1. Data Characterization And Data Discrimination:

- ✓ Data characterization is a summarization of the general characteristics or features of a target class of data. For example, the characteristics of students can be produced, generating a profile of all the University result year computer science students, which may include such information as a high GPA and large number of courses taken.
- ✓ Data discrimination is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes. For example, the general features of students with high GPA's may be compared with the general features of students with low GPA's. The resulting description could be a general comparative profile of the students such as 75% of the students with high GPA's are fourth-year computing science students while 65% of the students with low GPA's are not.

2. Mining Frequent Patterns, Associations, and Correlations.

- ✓ Frequent patterns, as the name suggests, are patterns that occur frequently in data. There are many kinds of frequent patterns, including item sets, subsequences, and substructures. A frequent item set typically refers to a set of items that frequently appear together in a transactional data set, such as milk and bread.
- ✓ A substructure can refer to different structural forms, such as graphs, trees, or lattices, which may be combined with item sets or subsequences. If a substructure occurs frequently, it is called a (frequent) structured pattern. Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

$\text{buys}(X; \text{"computer"}) \rightarrow \text{buys}(X; \text{"software"})$  [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.

# DR. SUBHASH UNIVERSITY

## School of Engineering and Technology

### Department of CSE / IT

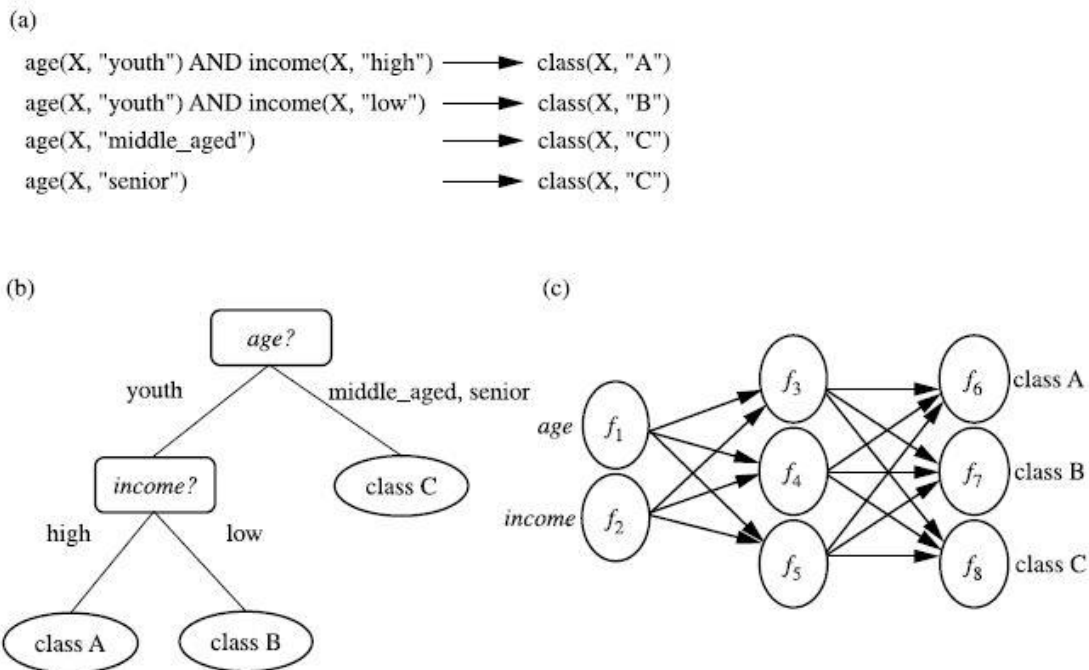
Single Dimensional Association rule.  
Multi-Dimensional Association rule.

### 3. Classification and Prediction.

- ✓ 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)

#### Classification Methods

- Classification Rule.
- Decision Tree.
- Neural Network or mathematical formula.

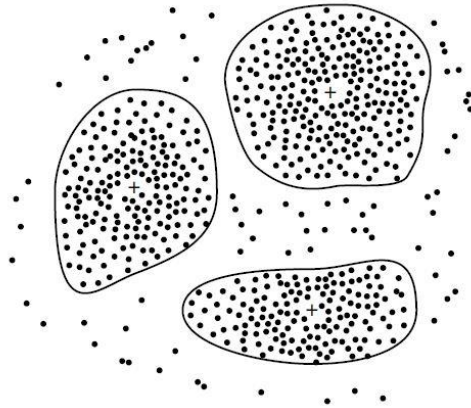


The prediction may refer to both numeric prediction and class label prediction. Regression analysis is a statistical methodology that is most often used for numeric prediction, although other methods exist as well. Prediction also encompasses the

identification of distribution trends based on the available data.

#### 4. Cluster Analysis

Unlike classification and prediction, which analyze class-labeled data objects, clustering analyzes data objects without consulting a known class label.



#### 5. 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.

#### 6. 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.

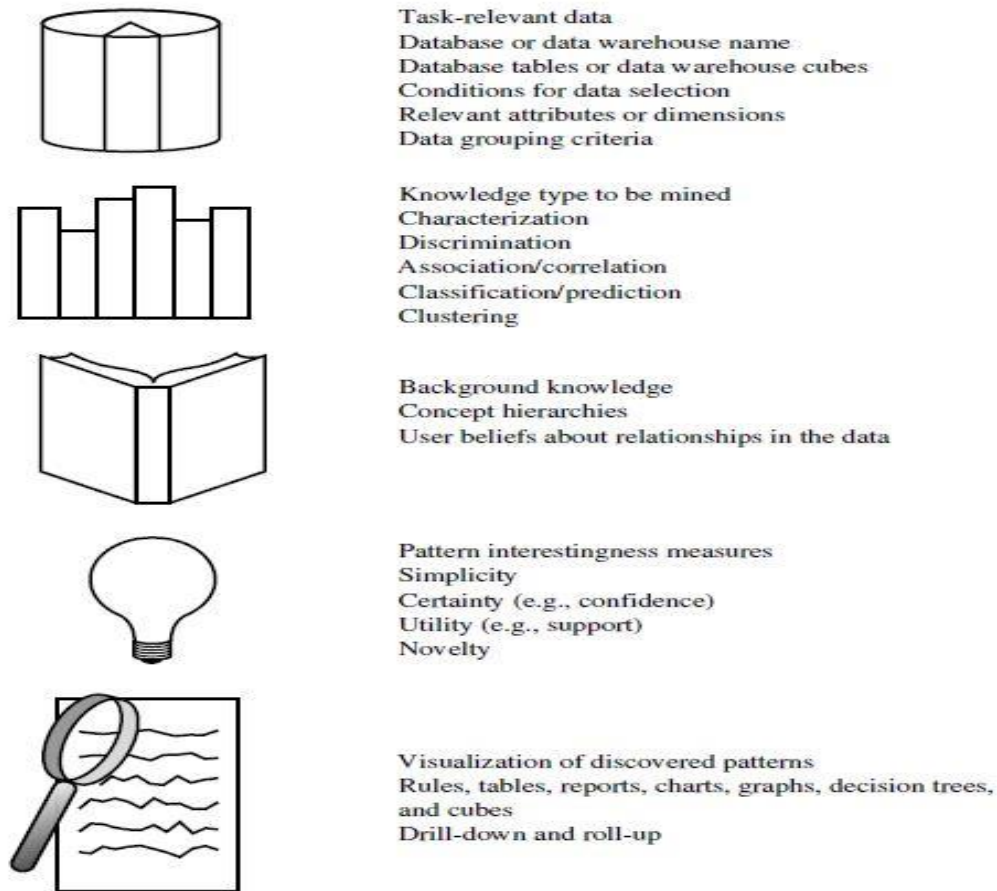
#### Q.5 Data Mining Task Primitives

A data mining task can be specified in the form of a data mining query, which is input to the data mining system. A data mining query is defined in terms of data mining task primitives.

- The set of task-relevant data to be mined: This specifies the portions of the database or the set of data in which the user is interested.
- The kind of knowledge to be mined: This specifies the data mining functions to be performed, such as characterization, discrimination, association or correlation analysis, Classification, prediction, clustering, outlier analysis, or evolution analysis.
- The background knowledge to be used in the discovery process: This knowledge about the domain to be mined is useful for guiding the knowledge discovery process and for evaluating the patterns found.
- The interestingness measures and thresholds for pattern evaluation: They may be used to guide the mining process or, after discovery, to evaluate the discovered patterns.
- The expected representation for visualizing the discovered patterns: This refers to the Form in which discovered patterns are to be displayed, which may include rules, tables, charts, graphs, decision trees, and cubes.



**DR. SUBHASH UNIVERSITY**  
**School of Engineering and Technology**  
**Department of CSE / IT**



**Q.6 Major Issues in Data Mining**

1. Efficiency and Scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed, stream, and incremental mining methods
2. Diversity of data types
  - Handling complex types of data
  - Mining dynamic, networked, and global data repositories
3. Data mining and society
  - Social impacts of data mining
  - Privacy-preserving data mining
  - Invisible data mining

**Q.7** What is noise? Describe the possible reasons for noisy data. Explain the different techniques to remove the noise from data.

**DR. SUBHASH UNIVERSITY**  
**School of Engineering and Technology**  
**Department of CSE / IT**

Noise is a random error or variance in a measured variable. Noise is any disturbance that interferes with data transmission and corrupts the quality of the signal. Let's look at the following data smoothing techniques.

Example:

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

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15

Bin 2: 21, 21, 24

Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9

Bin 2: 22, 22, 22

Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15

Bin 2: 21, 21, 24

Bin 3: 25, 25, 34

1. Binning: Binning methods smooth a sorted data value by consulting its "neighbourhood," that is, the values around it. The sorted values are distributed into a number of "buckets," or *bins*. Because binning methods consult the neighbourhood of values, they perform *local* smoothing.

In this example, the data for *price* are first sorted and then partitioned into *equal frequency* bins of size 3.

In smoothing by bin means, each value in a bin is replaced by the mean value of the bin. For example, the mean of the values 4, 8, and 15 in Bin 1 is 9. Therefore, each original value in this bin is replaced by the value 9. Similarly, smoothing by bin medians can be employed, in which each bin value is replaced by the bin median. In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the *bin boundaries*. Each bin value is then replaced by the closest boundary value.

**DR. SUBHASH UNIVERSITY**  
**School of Engineering and Technology**  
**Department of CSE / IT**

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 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.

Q.8 List and describe the methods for handling the missing values in data cleaning.

1. Ignore the tuple: This is usually done when the class label is missing (assuming the mining task involves classification). This method is not very effective, unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.
2. Fill in the missing value manually: In general, this approach is time-consuming and may 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 the missing 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 value with the average income value for customers in the same credit risk category as that of the given tuple.
6. Use the most probable value to fill in the missing value: This may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree induction. For example, using the other customer attributes in your data set, you may construct a decision tree to predict the missing values for income.

**DR. SUBHASH UNIVERSITY**  
**School of Engineering and Technology**  
**Department of CSE / IT**

Q.9 Explain data transformation in data mining.

In data transformation, the data are transformed or consolidated into forms appropriate for mining. Data transformation can involve the following:

Smoothing, which works to remove noise from the data. Such techniques include binning, regression, and clustering.

Aggregation, where summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts. This step is typically used in constructing a data cube for analysis of the data at multiple granularities.

Generalization, of the data, where low-level or “primitive” (raw) data are replaced by higher level concepts through the use of concept hierarchies. For example, categorical attributes, like *street*, can be generalized to higher-level concepts, like *city* or *country*. Similarly, values for numerical attributes, like *age*, may be mapped to higher-level concepts, like *youth*, *middle-aged*, and *senior*.

Normalization, where the attribute data are scaled so as to fall within a small specified range, such as 1:0 to 1:0, or 0:0 to 1:0.

Attribute construction, (or *feature construction*), where new attributes are constructed and added from the given set of attributes to help the mining process.

Q.10 Write typical requirements of clustering in data mining.

Clustering is a challenging field of research in which its potential applications pose their own special requirements. The following are typical requirements of clustering in data mining:

Scalability, Many clustering algorithms work well on small data sets containing fewer than several hundred data objects; however, a large database may contain millions of Objects. Clustering on a *sample* of a given large data set may lead to biased results. Highly scalable clustering algorithms are needed.

Ability to deal with different types of attributes, Many algorithms are designed to cluster interval-based (numerical) data. However, applications may require clustering other types of data, such as binary, categorical (nominal), and ordinal data, or mixtures of

these data types.

Discovery of clusters with arbitrary shape, Many clustering algorithms determine clusters based on Euclidean or Manhattan distance measures. Algorithms based on such distance measures tend to find spherical clusters with similar size and density. However, a cluster could be of any shape. It is important to develop algorithms that can detect clusters of arbitrary shape.

Minimal requirements for domain knowledge to determine input parameters, Many clustering algorithms require users to input certain parameters in cluster analysis (such as the number of desired clusters). The clustering results can be quite sensitive to input parameters. Parameters are often difficult to determine, especially for data sets containing high dimensional objects. This not only burdens users, but it also makes the quality of clustering difficult to control.

Ability to deal with noisy data, Most real-world databases contain outliers or missing, unknown, or erroneous data. Some clustering algorithms are sensitive to such data and may lead to clusters of poor quality.

Incremental clustering and insensitivity to the order of input records, Some clustering algorithms cannot incorporate newly inserted data (i.e., database updates) into existing clustering structures and, instead, must determine a new clustering from scratch. Some clustering algorithms are sensitive to the order of input data. That is, given a set of data objects, such an algorithm may return dramatically different clustering's depending on the order of presentation of the input objects. It is important to develop incremental clustering algorithms and algorithms that are insensitive to the order of input.

High dimensionality, A database or a data warehouse can contain several dimensions or attributes. Many clustering algorithms are good at handling low-dimensional data, involving only two to three dimensions. Human eyes are good at judging the quality of clustering for up to three dimensions. Finding clusters of data objects in high dimensional space is challenging, especially considering that such data can be sparse and highly skewed.

Constraint-based clustering, Real-world applications may need to perform clustering under various kinds of constraints. Suppose that your job is to choose the

locations for a given number of new automatic banking machines (ATMs) in a city. To decide upon this, you may cluster households while considering constraints such as the city's rivers and highway networks, and the type and number of customers per cluster. A challenging task is to find groups of data with good clustering behaviour that satisfy specified constraints.

Interpretability and usability, Users expect clustering results to be interpretable, comprehensible, and usable. That is, clustering may need to be tied to specific semantic Interpretations and applications. It is important to study how an application goal may Influence the selection of clustering features and methods.

Q.11 Explain rule based classification and case based reasoning in details.

Rule based classification: A rule-based classifier uses a set of IF-THEN rules for classification. Rules can be extracted from a decision tree. Rules may also be generated directly from training data using sequential covering algorithms and associative classification algorithms.

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  $R_1$ ,

$R_1$ : 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).  $R_1$  can also be written as

$R_1: (age = youth) \wedge (student = yes)(buys\ computer = yes).$

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

**DR. SUBHASH UNIVERSITY**  
**School of Engineering and Technology**  
**Department of CSE / IT**

classable data set,  $D$ , let  $ncovers$  be the number of tuples covered by  $R$ ;  $ncorrect$  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

$$coverage(R) = ncovers / |D|$$

$$accuracy(R) = ncorrect / ncovers$$

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.

Case based reasoning: Case-based reasoning (CBR) classifiers use a database of problem solutions to solve new problems. Unlike nearest-neighbor classifiers, which store training tuples as points in Euclidean space, CBR stores the tuples or "cases" for problem solving as complex symbolic descriptions. Business applications of CBR include problem resolution for customer service help desks, where cases describe product-related diagnostic problems. CBR has also been applied to areas such as engineering and law, where cases are either technical designs or legal rulings, respectively. Medical education is another area for CBR, where patient case histories and treatments are used to help diagnose and treat new patients.

When given a new case to classify, a case-based reasoner will first check if an identical training case exists. If one is found, then the accompanying solution to that case is returned. If no identical case is found, then the case-based reasoner will search for training cases having components that are similar to those of the new case. Conceptually, these training cases may be considered as neighbours of the new case. If cases are represented as graphs, this involves searching for subgraphs that are similar to subgraphs within the new case. The case-based reasoner tries to combine the solutions of the neighbouring training cases in order to propose a solution for the new case. If incompatibilities arise with the individual solutions, then backtracking to search for other solutions may be necessary. The case-based reasoner may employ background knowledge and problem-solving strategies in order to propose a feasible combined solution.



**DR. SUBHASH UNIVERSITY**  
**School of Engineering and Technology**  
**Department of CSE / IT**

Q.12 Write the steps of the k-means clustering algorithm. Also state its limitations.

K-means algorithm.

K-means clustering is a method of cluster analysis which aims to partition  $n$  observations into  $k$  clusters in which each observation belongs to the cluster with the nearest mean. It is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The k-means algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters observations into  $k$  groups, where  $k$  is provided as an input parameter. It then assigns each observation to clusters based upon the observation's proximity to the mean of the cluster. The cluster's mean is then recomputed and the process begins again.

Here's how the algorithm works:

1. The algorithm arbitrarily selects  $k$  points as the initial cluster centers ("means").
2. Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.
3. Each cluster center is recomputed as the average of the points in that cluster.
4. Steps 2 and 3 repeat until the clusters converge. Convergence may be defined differently depending upon the implementation, but it normally means that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters.

Limitations in K-means algorithm:

**Handling Empty Clusters:** One of the problems with the basic K-means algorithm given earlier is that empty clusters can be obtained if no points are allocated to a cluster during the assignment step. If this happens, then a strategy is needed to choose a replacement centroid, since otherwise, the squared error will be larger than necessary.

**Outliers:** When outliers are present, the resulting cluster centroids (prototypes) may not be as representative as they otherwise would be and thus, the SSE will be higher as well.

**Reducing the SSE with Post processing:** In k-means to get better clustering we have to reduce the SSE that is most difficult task. There are various types of clustering methods available which reduces the SSE.

**Difficult to measure the nod of clusters:** The user has to choose the value of  $K$ , the number of clusters. Although for 2D data this choice can easily be made by visual inspection, it is not so for higher dimension data, and there are usually no clues as to what number of clusters might be appropriate.



**DR. SUBHASH UNIVERSITY**  
**School of Engineering and Technology**  
**Department of CSE / IT**

Q.13 Explain Data Warehouse.

Data warehouse states that “it is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process”

**Subject-oriented:** A data warehouse is organized around major subjects, such as customer, supplier, product, and sales. Rather than concentrating on the day-to-day operations and transaction processing of an organization, a data warehouse focuses on the modeling and analysis of data for decision makers. Hence, data warehouses typically provide a simple and concise view around particular subject issues by excluding data That are not useful in the decision support process.

**Integrated:** A data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and on-line transaction records. Data cleaning and data integration techniques are applied to ensure consistency in naming conventions, encoding structures, attribute measures, and so on.

**Time-variant:** Data are stored to provide information from a historical perspective (e.g., the past 5–10 years). Every key structure in the data warehouse contains, either implicitly or explicitly, an element of time.

**Nonvolatile:** A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment. Due to this separation, a data warehouse does not require transaction processing, recovery, and concurrency control mechanisms. It usually requires only two operations in data accessing: initial loading of data and access of data.

# DR. SUBHASH UNIVERSITY

## School of Engineering and Technology

### Department of CSE / IT

#### Q.13 Difference between OLAP v/s OLTP

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

#### Q.14 Data Cube, Cuboid, Dimensions, Dimension Table, Fact Table

**Dimension:** In general terms, dimensions are the perspectives or entities with respect to which an organization wants to keep records. For example, AllElectronics may create a sales data warehouse in order to keep records of the store's sales with respect to the dimensions time, item, branch, and location.

**Data Cube:** Data warehouses and OLAP tools are based on a multidimensional data model. This model views data in the form of a data cube. A data cube allows data to be modeled and viewed in multiple Dimensions. It is defined by dimensions and facts.

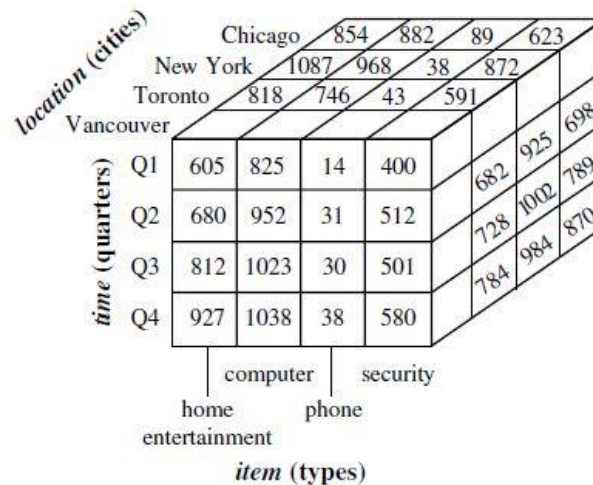
**Dimension Table:** Each dimension may have a table associated with it, called a dimension table, which further describes the dimension.

# DR. SUBHASH UNIVERSITY

## School of Engineering and Technology

### Department of CSE / IT

Fact Table: Facts are numerical measures. Think of the facts as the quantities by which we want to analyze relationships between dimensions. Examples of facts for a sales data warehouse include dollars sold (sales amount in dollars), units sold (number of units sold), and amount budgeted.

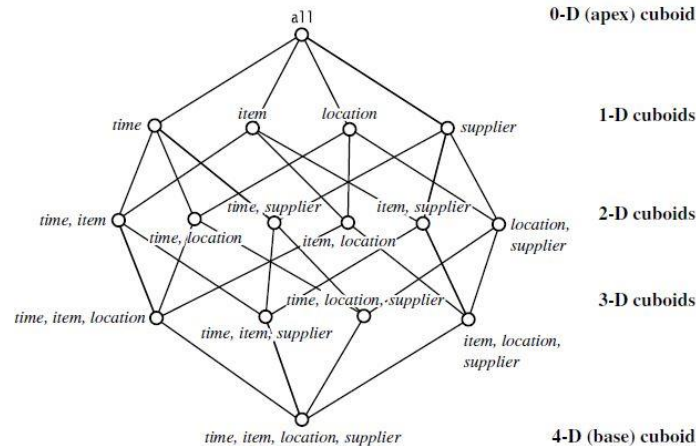


		location (cities)			
		Chicago	New York	Toronto	Vancouver
time (quarters)	Q1	605	825	14	400
	Q2	680	952	31	512
	Q3	812	1023	30	501
	Q4	927	1038	38	580

		item (types)			
		computer	home entertainment	phone	security
location (cities)	Chicago	854	882	89	623
	New York	1087	968	38	872
	Toronto	818	746	43	591
	Vancouver	682	925	728	1002

Cuboid: Cuboid represent the data at a different level of summarization, or group by.



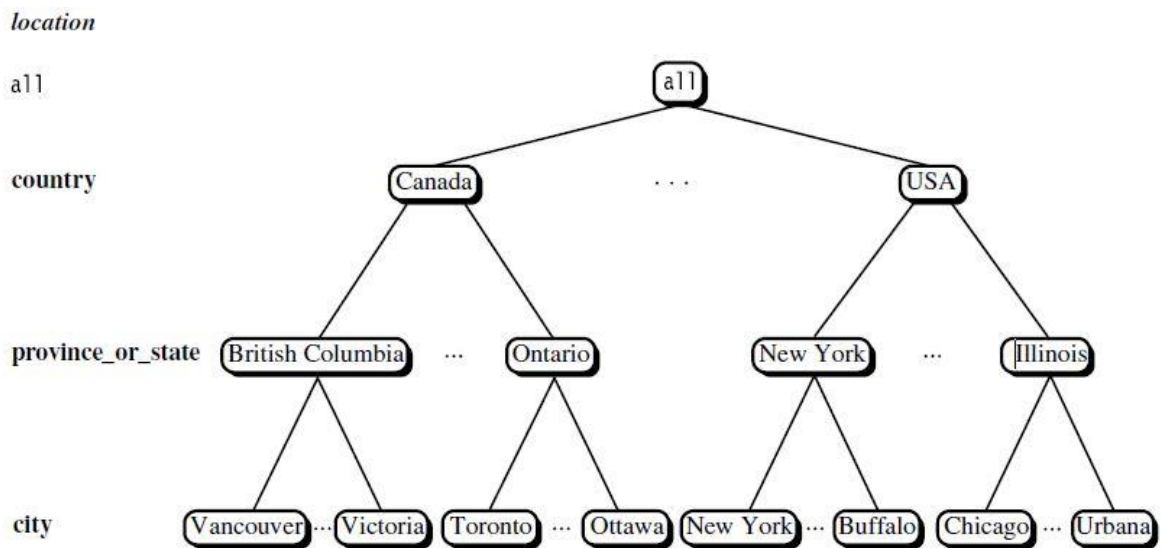
# DR. SUBHASH UNIVERSITY

## School of Engineering and Technology

### Department of CSE / IT

Q.15 Explain Concept Hierarchies.

A concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts. Consider a concept hierarchy for the dimension location. City values for location include Vancouver, Toronto, New York, and Chicago. Each city, however, can be mapped to the province or state to which it belongs. For example, Vancouver can be mapped to British Columbia, and Chicago to Illinois. The provinces and states can in turn be mapped to the country to which they belong, such as Canada or the USA. These mappings form a concept hierarchy for the dimension location, mapping a set of low-level concepts (i.e., cities) to higher-level, more general concepts (i.e., countries). The concept hierarchy described above is illustrated as below.



Q.16 Explain OLAP Operations.

**Roll-up:** The roll-up operation (also called the drill-up operation) performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction.

**Drill-down:** Drill-down is the reverse of roll-up. It navigates from less detailed data to more detailed data. Drill-down can be realized by either stepping down a concept hierarchy for a dimension or introducing additional dimensions.

**Slice and dice:** The slice operation performs a selection on one dimension of the given cube, resulting in a sub cube. The dice operation defines a sub cube by performing a selection on two or more dimensions.

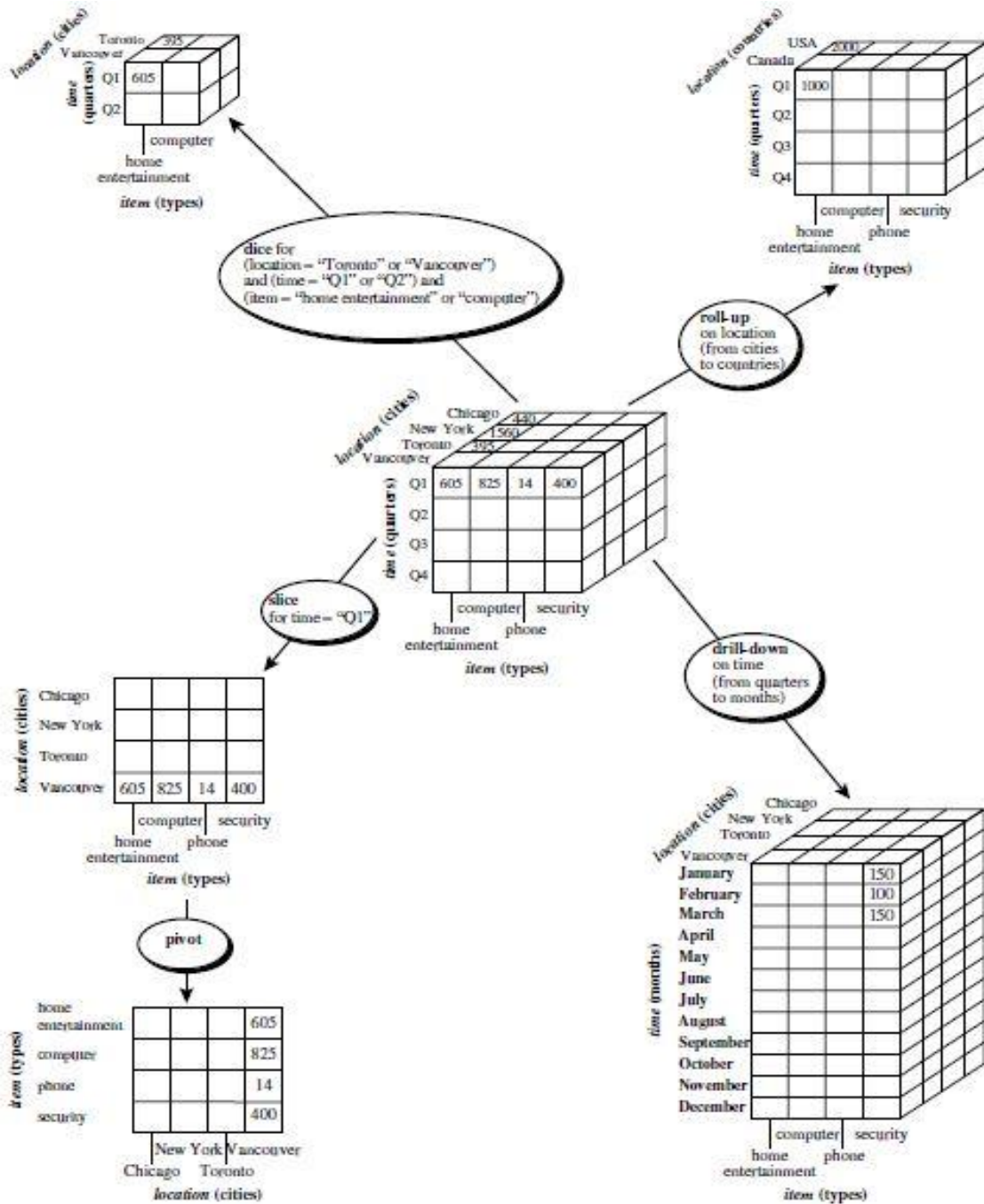
**Pivot (rotate):** Pivot (also called rotate) is a visualization operation that rotates the

# DR. SUBHASH UNIVERSITY

## School of Engineering and Technology

### Department of CSE / IT

data axes in view in order to provide an alternative presentation of the data.

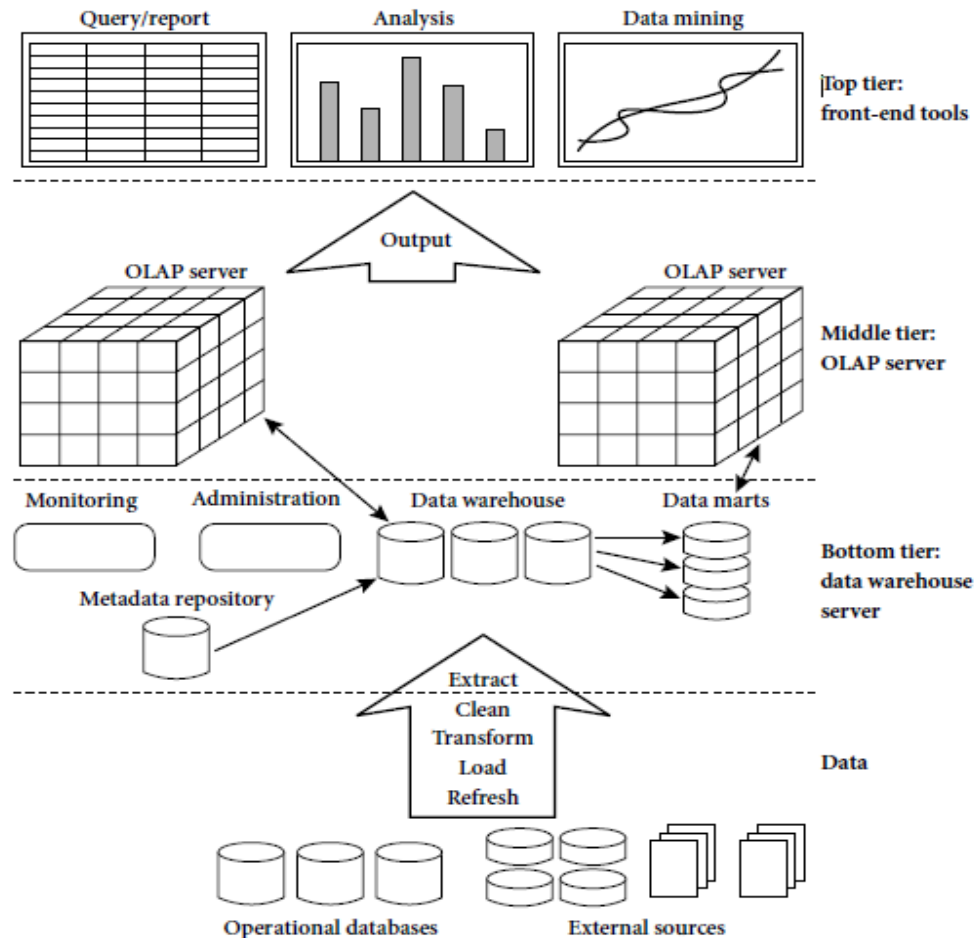


# DR. SUBHASH UNIVERSITY

## School of Engineering and Technology

### Department of CSE / IT

Q. 17 Three-tier Architecture of data warehousing.



The bottom tier is a warehouse database server that is almost always a relational database system. Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources (such as customer profile information provided by external consultants). These tools and utilities perform data extraction, cleaning, and transformation (e.g., to merge similar data from different sources into a unified format), as well as load and refresh functions to update the data warehouse.

The middle tier is an OLAP server that is typically implemented using either (1) a relational OLAP (ROLAP) model, that is, an extended relational DBMS that maps operations on multidimensional data to standard relational operations; or (2) a multidimensional OLAP (MOLAP) model, that is, a special-purpose server that directly implements multidimensional data and operations.

The top tier is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools (e.g., trend analysis, prediction, and so on).



**DR. SUBHASH UNIVERSITY**  
**School of Engineering and Technology**  
**Department of CSE / IT**

Q.18 Explain Different Data Warehouse model.

**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.

**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.

**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.

Q.19 Explain Metadata Repository.

**Metadata** are data about data. When used in a data warehouse, metadata are the data that define warehouse objects. Metadata are created for the data names and definitions of the given warehouse. Additional metadata are created and captured for time stamping any extracted data, the source of the extracted data, and missing fields that have been added by data cleaning or integration processes.

A description of the structure of the data warehouse, which includes the warehouse schema, view, dimensions, hierarchies, and derived data definitions, as well as data mart locations and contents.

**Operational metadata**, which include data lineage (history of migrated data and the sequence of transformations applied to it), currency of data (active, archived, or purged), and monitoring information (warehouse usage statistics, error reports, and audit trails)

The algorithms used for summarization, which include measure and dimension definition algorithms, data on granularity, partitions, subject areas, aggregation, summarization, and predefined queries and reports.

The mapping from the operational environment to the data warehouse, which includes source databases and their contents, gateway descriptions, data partitions, data extraction, cleaning, transformation rules and defaults, data refresh and purging rules, and security (user authorization and access control).

Data related to system performance, which include indices and profiles that improve data access and retrieval performance, in addition to rules for the timing and scheduling of refresh, update, and replication cycles.

**Business metadata**, which include business terms and definitions, data ownership information, and charging policies.