UNIT I - PART A

1. Define data warehouse. [Dec 2013][May 2012]

A data warehouse is a subject oriented, integrated, time-variant and nonvolatile collection of data in support of management's decision making process. Data warehouse refers to a database that is maintained separately from an organization's operational databases. They support information processing by providing a solid platform of consolidated historical data for analysis.

2. What are the components of data warehousing?

1. Sourcing, Acquisition, cleanup, and transformation tools 2.Repository 3.Data warehouse database,4. Data mart 5.Applications and tools 6.Management platform 7.Information delivery systems

3.Define data mart. [Dec 2011] [May 2013]

A data mart is a department subset of the data warehouse that focuses on selected subjects and thus it is department-wide. For data marts star schema or snowflake schema is used since both are geared towards modeling single subjects although star schema is more popular and efficient.

4. **How is a data warehouse different from a database? How are they similar?[May 2012]**

Data warehouse	**Database**
Online Analytical Processing	Online Transaction Processing
Data analysis and decision making	Day-to-day operations purchasing, inventory, banking,payroll, registration etc.
Structure for corporate view of data	Structure to suite departmental view of data
Up-to-date driven	Query driven

Similarity

Both store data and information. Data warehouse is a repository of multiple heterogeneous data sources, organized under a unified schema at a single site in order to facilitate management decision-making.

Database consists of a collection of interrelated data.

5.List the three important issues that have to be considered during data integration.[Nov 2011]

The three important issues to be addresses during integration are

- Schema integration and object matching
- Redundancy
- Detection and resolution of data value conflict.

6. why data transformation is essential in the process of knowledge discovery? [June 2012][May 2011]

In this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations. Methods here include dimension reduction (such as feature selection and extraction and record sampling), and attribute transformation (such as discretization of numerical attributes and functional transformation). This step can be crucial for the success of the entire KDD project, and it is usually very project-specific. For example, in medical examinations, the quotient of attributes may often be the most important factor, and not each one by itself.

7.State one of the biggest challenges when designing a data warehouse.[June 2013]

The biggest challenge when designing a data warehouse is the data placement and distribution. The data placement and distribution should consider factors like

- Subject area
- Location
- Time

Distribution solves many problems but creates network related problems.

8. What is a Meta data?Give an example.[May 2011] [Dec 2013][Nov 2014] or Mention the role of meta data in a data warehouse.[June 2012]

Metadata are data about data. When used in data Warehouse, metadata are data that define warehouse objects. Meta data are created for the data names and definitions of the given warehouse. Metadata summarizes basic information about data, which can make finding and working with particular instances of data easier. For example, author, date created and date modified and file size are examples of very basic document metadata. Having the abilty to filter through that metadata makes it much easier for someone to locate a specific document.

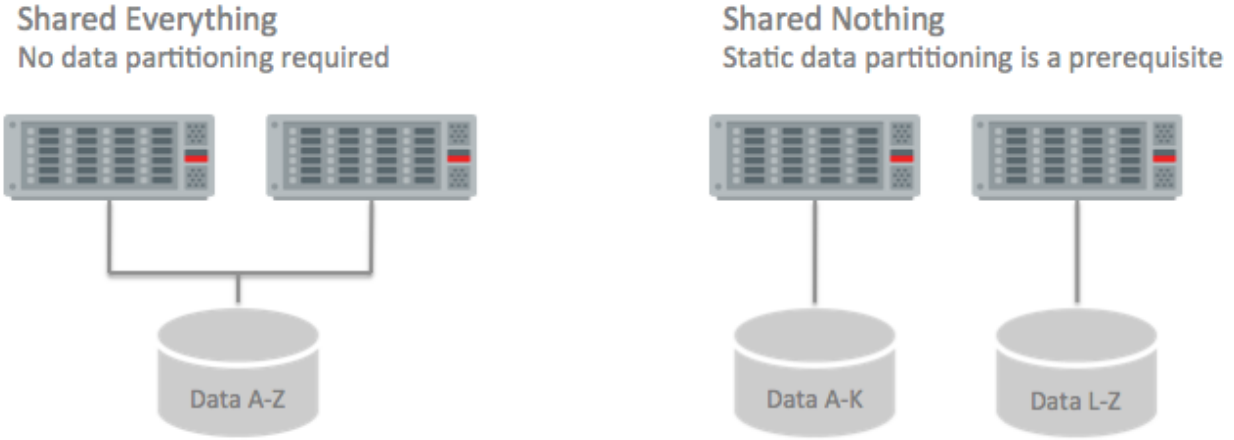
The role of meta data is to specify data about data which will be like structure of data warehouse, background data. It is calssified into technical metadata and business metadata.

9. Compare data mart and data warehouse.

Data mart is a department subset of a data warehouse. It focuses on selected subjects and thus its scope is department wide. On the other hand data warehouse collects information about subjects that span an entire organization and thus its scope is department wide.

10.List the two ways the parallel execution of the tasks within SQL statements can be done.[Dec 2012]

The "partitioning of data" for parallel processing is commonly done in two basic, but fundamentally different ways. The main differentiation is whether or not physical data partitioning (placement) is used as a foundation – and therefore as static prerequisite – for parallelizing the work.
These fundamental conceptually different approaches are known as shared everything architecture and shared nothing architecture respectively.



11. Define data cube. [May 2013]

A data cube allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.

12. What is a dimension table? [Dec 2013]

Dimensions are perspectives or entities with respect to whish an organization wants to keep records. Each dimension may have a table associated with it called dimension table which further describes the dimension.

13. Define facts.

A multidimensional data model is typically organized around a central theme and the theme is represented by a fact table. Facts are numerical measures. Fact table contains the names of the facts and keys to each of the related dimension tables.

14. Define star schema and snowflake schema.[Nov 2014]

In star schema data warehouse contains 1) a large central table (fact table) containing the bulk of the data containing no redundancy and 2) set of smaller attendant tables (dimension tables) one for each dimension. The schema graph resembles a star burst with the denormalised dimension table displayed in a radial pattern around a central fact table. It may have any number of dimension tables and many-to-one relationship between the fact table and each dimension table.Snowflake schema is the further splitting of star schema dimension tables into one or more multiple normalized table thereby reducing the redundancy. A snowflake schema can have any number of dimensions and each dimension can have any number of levels.

15. Give the advantages and disadvantages of snowflake schema.

Advantage: Dimension table are kept in a normalized form and thus it is easy to maintain and saves the storage space.

Disadvantage: It reduces the effectives of browsing since more join is needed to execute a query.

16. Define fact constellation.

Sophisticated applications may require multiple fact tables to share dimension tables. This kind of schema can be viewed as a collection of stars and hence it is called as galaxy schema or fact constellation.

17. List the steps involved in Warehouse design process

- Choose a business process to model
- Choose the grain of the business process
- Choose the dimensions
- Choose the measures.

18. What are the approaches for building a data warehouse? Or How can we design a data warehouse?

A data warehouse can be designed using a top-down approach, bottom-up approach, or a combination of both. In combined approach an organization can exploit the planned strategic nature of top-down approach while retaining the rapid implementation and opportunistic approach of the bottom-up approach.

The top-down approach: meaning that an organization has developed an enterprise data model, collected enterprise wide business requirements, and decided to build an enterprise data warehouse with subset data marts. The top-down approach starts with overall design and planning.

The bottom-up approach: implying that the business priorities resulted in developing individual data marts, which are then integrated into the enterprise data warehouse. The bottom-up approach starts with experiments and prototypes.

19. What are the types of data partitioning? [May 2013]

1. Hash partitioning 2.Kay range partitioning 3. Schema partitioning 4. Use defined partitioning

20. Give the major features of data warehouse. [April/May 2010]

Subject-oriented, integrated, time-variant and nonvolatile.

21. What are the advantages of dimensional modeling?[June 2014]

1. Single version of the truth 2)Data integration 3)Analyze on the fly 4)Drill up or drill down to any level of detail contained in the data 5)Maximize flexibility and scalability. 6)Optimize the end-user experience.

22. What are the data design issues and technical isues in building data warehouse? [May 2013]

- Heterogeneity of data sources, which affects data conversion, quality, timeliness.
- Use of historical data, which implies that data may be "old"
- Tendency of databases to grow very large
- End user requirements and data sources will change.

Technical issues

- The hardware platform that would house the data warehouse
- The DBMS that supports the warehouse database
- The communications infrastructure that connects the warehouse, data marts, operational systems, and end users.
- The hardware platform and software to support the metadata repository.

23. What a metadata repository should contain?

A metadata repository should contain the following:

- A description of the structure of the data warehouse.
- Operational Meta data.
- Algorithms used for summarization
- Mapping from operational environment to data warehouses.
- Data related to system performance.
- Business Meta data.

24. What is the need for back end process in data warehouse design?[June 2014]or What is ETL process?Give its significance.[Dec 2013]

Extract – Transformation – Load(ETL) processes, which take place in the back stage of the data warehouse environment are data intensive, complex, and costly. The functionality of these processes includes: (a) the identification of relevant information at the source side; (b) the extraction of this information; (c) the transportation of this information from the sources to an intermediate place called Data Staging Area (DSA); (d) the customization and integration of the information coming from multiple sources into a common format; (e) the cleaning of the resulting data set, on the basis of database and business rules; and (f) the propagation of the homogenized and cleansed data to the data warehouse and/or data marts.

25. Give the data warehouse applications.[May 2008]

Information processing, analytical processing, data mining,decision mining

26. What are the nine steps involed in the design of a data warehouse?

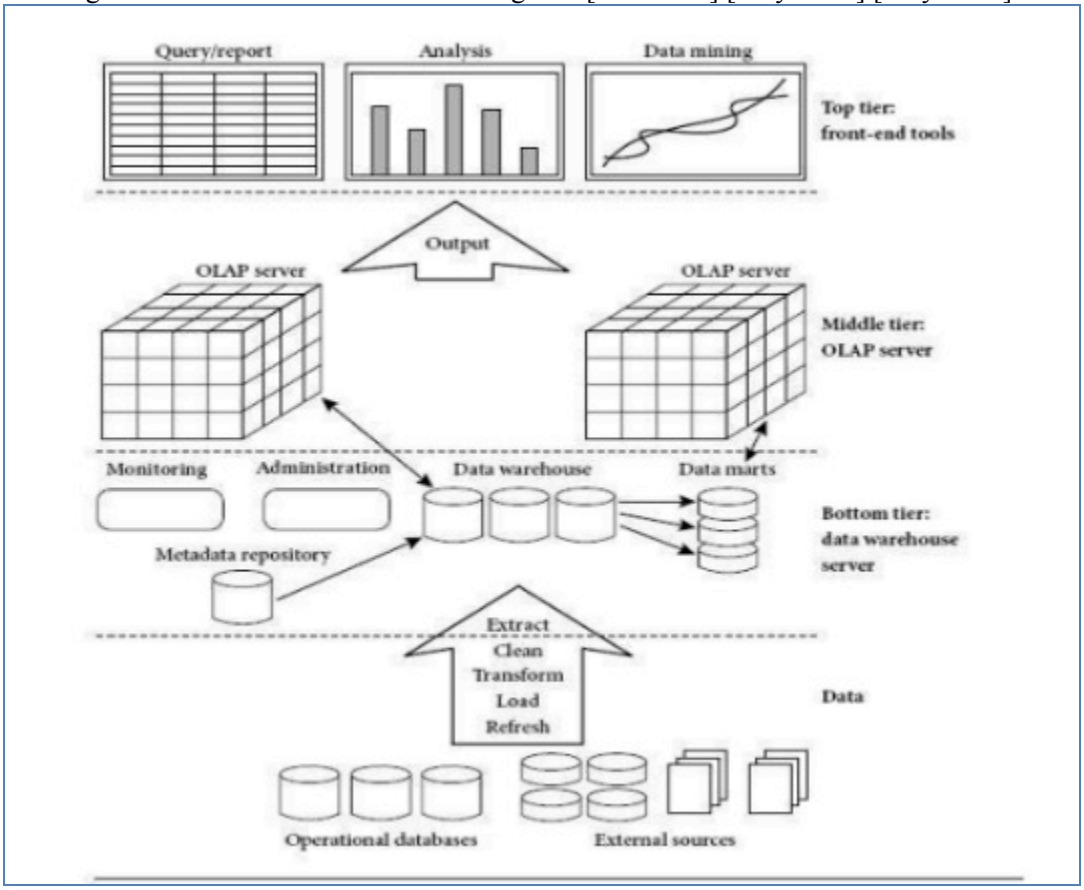
1. Choosing the subject matter 2.deciding what a fact table represents 3.Identofying and confirming the dimensions 4.choosing the facts 5.storing precalculations in the table 6.rounding out the dimension table 7.choosing the duration of the table 8.the need to track slowly changing dimensions 9.deciding the query priorities and the query modes.

27. Define data transformation. [May 2011]

Data transformation from one format to another on the basis of possible differences between the source and the target platforms. Ex: calculating age from the date of birth, replacing a possible numeric gender code with a more meaningful "male" and "female".

UNIT I - PARTB

1.i) Define data warehouse. Explain its features. Diagrammatically illustrate and discuss the data warehouses architecture [May 2011] [Dec 2011][Nov 2014] or Explain the multi-tier architecture suitable for evolving a data warehouse with suitable diagram. [Dec 2013] [May 2012] [May 2010]



BOTTOM TIER**:** It is a warehouse database server. Data is fed using Back end tools and utilities. Data extracted using programs called gateways. It also contains Meta data repository.

MIDDLE TIER: 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.

TOP TIER: The top tier is a front-end client layer, which containsquery and reporting tools, analysis tools, and/or datamining tools.

A common way of introducing data warehousing is to refer to the characteristics of a data warehouse.

- Subject Oriented
- Integrated
- Nonvolatile
- Time Variant

Subject Oriented:

Data warehouses are designed to help you analyze data. For example, to learn more about your company's sales data, you can build a warehouse that concentrates on sales. Using this warehouse, you can answer questions like "Who was our best customer for this item last year?" This ability to define a data warehouse by subject matter, sales in this case, makes the data warehouse subject oriented.

Integrated:

Integration is closely related to subject orientation. Data warehouses must put data from disparate sources into a consistent format. They must resolve such problems as naming conflicts and inconsistencies among units of measure. When they achieve this, they are said to be integrated.

Nonvolatile:

Nonvolatile means that, once entered into the warehouse, data should not change. This is logical because the purpose of a warehouse is to enable you to analyze what has occurred.

Time Variant:

In order to discover trends in business, analysts need large amounts of data. This is very much in contrast to online transaction processing (OLTP) systems, where performance requirements demand that historical data be moved to an archive. A data warehouse's focus on change over time is what is meant by the term time variant.

ii) **Explain the different types of data repositories on which mining can be performed?[**Nov 2014]

The different types of data repositories on which mining can be performed are:

- Relational Databases
- DataWarehouses
- Transactional Databases
- Advanced Databases
- Flat files
- World Wide Web

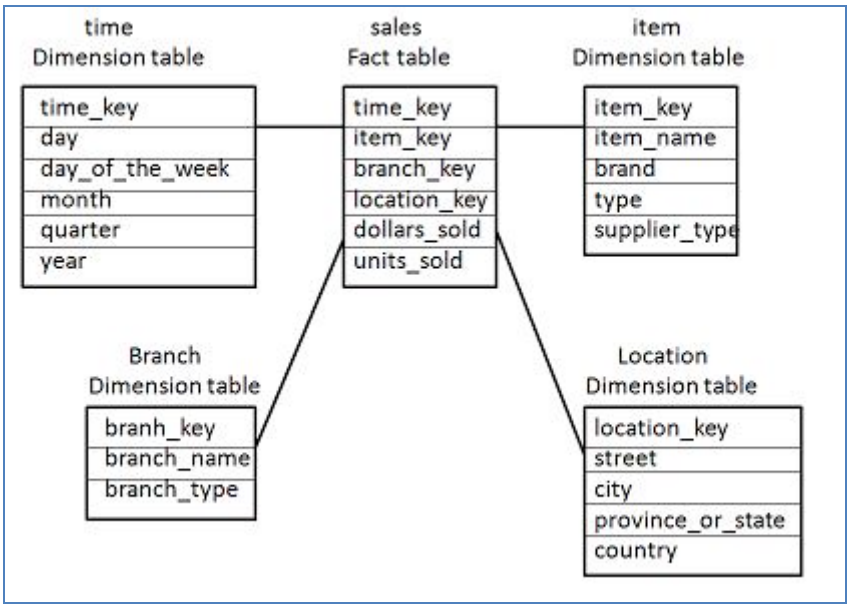
2. **Briefly describe star snowflake and fact constellations schemas with examples? (Or) Giving suitable examples, describe the various multi-dimensional schema. [Dec 2011] [May 2012] [May 2013] or Explain in detail the DBMS Schema for decision support**

A star schema is organized around a central fact table that is joined to some dimension tables using foreign key references. The fact table contains data like price, discount values, number of units sold, and dollar value of sales. The fact table usually has some summarized and aggregated data and it is usually very large in terms of both fields and records. The basic premise of a star schema is that information can be classified into two groups: facts and dimensions. Facts are the core data elements one is analyzing. A snowflake schema is an expansion and extension of a star schema to additional secondary dimensional tables. The database uses the relational model on the other hand the data warehouse uses the Stars, snowflake and fact constellation schema. In this chapter we will discuss the schemas used in data warehouse.

Star Schema

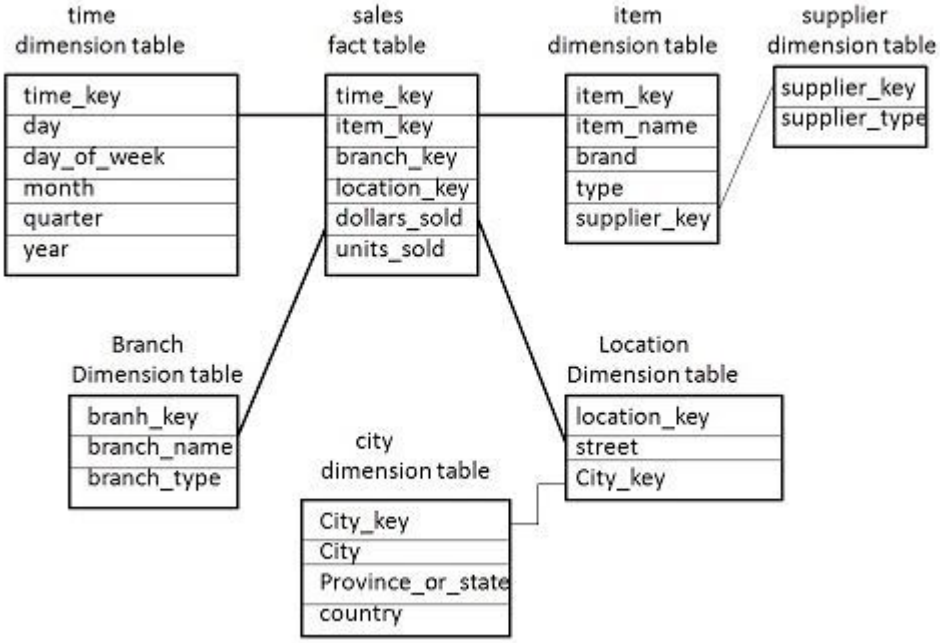
- In star schema each dimension is represented with only one dimension table.
- This dimension table contains the set of attributes.
- In the following diagram we have shown the sales data of a company with respect to the four dimensions namely, time, item, branch and location.\
- There is a fact table at the centre. This fact table contains the keys to each of four dimensions.
- The fact table also contain the attributes namely, dollars sold and units sold.

Note: Each dimension has only one dimension table and each table holds a set of attributes. For example the location dimension table contains the attribute set location_key,street,city,province_or_state,country}. This constraint may cause data redundancy. For example the "Vancouver" and "Victoria" both cities are both in Canadian province of British Columbia. The entries for such cities may cause data redundancy along the attributes province_or_state and country.



Snowflake Schema

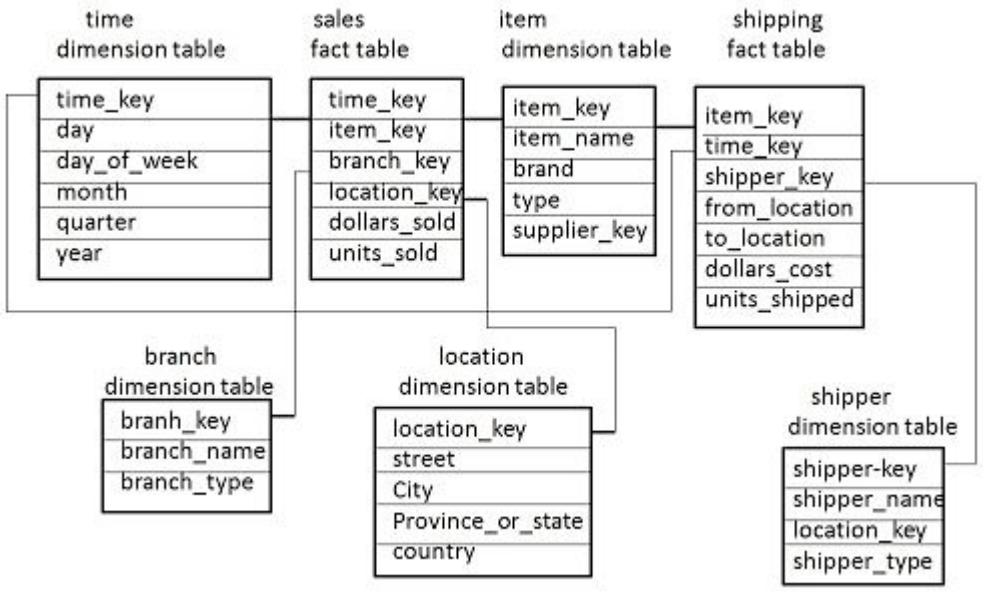
- In Snowflake schema some dimension tables are normalized.
- The normalization split up the data into additional tables.
- Unlike Star schema the dimensions table in snowflake schema are normalized for example the item dimension table in star schema is normalized and split into two dimension tables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



Therefore now the item dimension table contains the attributes item_key, item_name, type, brand, and supplier-key.The supplier key is linked to supplier dimension table. The supplier dimension table contains the attributes supplier_key, and supplier_type.Note: Due to normalization in Snowflake schema the redundancy is reduced therefore it becomes easy to maintain and save storage space.

Fact Constellation Schema

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



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

Explain in detail the DBMS Schema for decision support

- Multidimensional data model
- Star Schema
  - DBA Viewpoint
  - Performance problems of star schema
  - Solutions to the Problems
- Star join and Star index
- Bitmap indexing

3. **Explain the role played by sourcing, acquisition, cleanup, and transformation tools in building a data warehouse.** [May 2013]

The data sourcing, cleanup, transformation and migration tools perform all of the conversions, summarizations, key changes, structural changes and condensations needed to transform disparate data into information that can be used by the decision support tool. Data Integration includes combining several source records into a single record to be loaded into the warehouse. Data transformation means converting data from one format to another on the basis of possible differences between the source and the target platforms.A significant portion of the implementation effort is spent extracting data from operational systems and putting it in a format suitable for informational applications that run off the data warehouse.

The data sourcing, cleanup, transformation and migration tools perform all of the conversions, summarizations, key changes, structural changes and condensations needed to transform disparate data into information that can be used by the decision support tool. They produce the programs and control statements, including the COBOL programs, MVS job-control language (JCL), UNIX scripts, and SQL data definition language (DDL) needed to move data into the data warehouse for multiple operational systems. These tools also maintain the meta data. The functionality includes:

- Removing unwanted data from operational databases
- Converting to common data names and definitions
- Establishing defaults for missing data
- Accommodating source data definition changes

The data sourcing, cleanup, extract, transformation and migration tools have to deal with some significant issues including:

- Database heterogeneity. DBMSs are very different in data models, data access language, data navigation, operations, concurrency, integrity, recovery etc.
- Data heterogeneity. This is the difference in the way data is defined and used in different models - homonyms, synonyms, unit compatibility (U.S. vs metric), different attributes for the same entity and different ways of modeling the same fact.

These tools can save a considerable amount of time and effort. However, significant shortcomings do exist. For example, many available tools are generally useful for simpler data extracts. Frequently, customized extract routines need to be developed for the more complicated data extraction procedures.

4.**List and Discuss the steps involved in mapping the data warehouse to a multiprocessor architecture.** [May 2011] [Dec 2011][Nov 2014]

The goals of linear performance and scalability can be satisfied by parallel hardware architectures, parallel operating systems, and parallel DBMSs. Parallel hardware architectures are based on Multi-processor systems designed as a Shared-memory model, Shared-disk model or distributed-memory model.

Parallelism can be achieved in three different ways:1.Horizontal Parallelism (Database is partitioned across different disks) 2.Vertical Parallelism (occurs among different tasks – all components query operations i.e. scans, join, sort) 3.Data Partitioning

Shared-memory Architecture- multiple processors share the main memory space, as well as mass storage (e.g. hard disk drives) **Shared Disk Architecture** - each node has its own main memory, but all nodes share mass storage, usually a storage area network **Shared-nothing Architecture** - each node has its own mass storage as well as main memory

5.i) **Describe the steps involved in the design and construction of data warehouses.[**June 2012][Dec 2012]

In general, building any data warehouse consists of the following steps:

1. Extracting the transactional data from the data sources into a staging area
2. Transforming the transactional data
3. Loading the transformed data into a dimensional database
4. Building pre-calculated summary values to speed up report generation
5. Building (or purchasing) a front-end reporting tool

ii) **Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order):13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.**

1.Use smoothing by bin means to smooth the above data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.

Step 1: Sort the data (this step is not required here as the data is already sorted.)
Step 2: Partition the data into equidepth bins of depth 3
Bin 1: 13, 15, 16 Bin 2: 16,19,20 Bin 3: 20, 21, 22
Bin 4: 22, 25, 25 Bin 5: 25,25,30 Bin 6: 33, 33, 35
Bin 7: 35, 35, 35 Bin 8: 36, 40, 45 Bin 9: 46, 52, 70
Step 3: calculate the arithmetic mean of each bin
Step 4: Replace each of the values in each bin by the arithmetic mean calculated for the bin.

Bin 1: 44/3, 44/3, 44/3 Bin 2: 55/3, 55/3, 55/3 Bin 3: 21, 21, 21
Bin 4: 24, 24, 24 Bin 5: 80/3, 80/3, 80/3 Bin 6: 101/3, 101/3, 101/3
Bin 7: 35, 35, 35 Bin 8: 121/3, 121/3, 121/3 Bin 9: 56, 56, 56

2. **How might you determine outliers in the data**?

Outliers in the data may be detected by clustering, where similar values are organized into groups, or "clusters". Values that fall outside of the set of groups may be considered outliers. Alternatively, a combination of computer and human inspection can be used where a predetermined data distribution is implemented to allow the computer to identify possible outliers. These possible outliers can then be verified by human inspection with much less effort than would be required to verify the entire data set.

7.i) **How do datawarehousing and OLAP relate to data mining? Explain.[Dec 2012]**

OLAP and Data Mining Comparison

OLAP and data mining are used to solve different kinds of analytic problems:

- OLAP provides summary data and generates rich calculations. For example, OLAP answers questions like "How do sales of mutual funds in North America for this quarter compare with sales a year ago? What can we predict for sales next quarter? What is the trend as measured by percent change?"
- Data mining discovers hidden patterns in data. Data mining operates at a detail level instead of a summary level. Data mining answers questions like "Who is likely to buy a mutual fund in the next six months, and what are the characteristics of these likely buyers?"

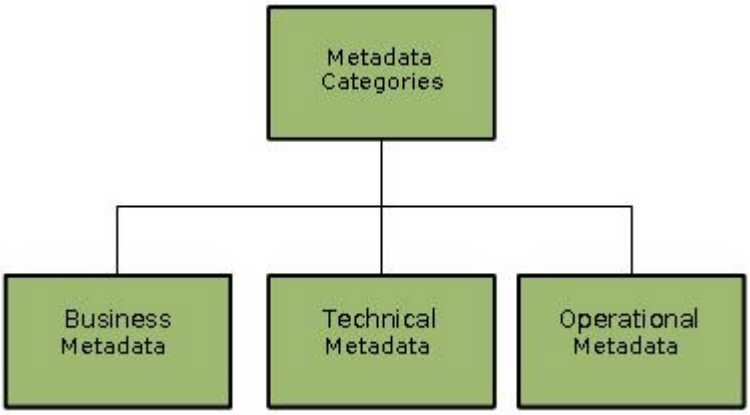
OLAP and data mining can complement each other. For example, OLAP might pinpoint problems with sales of mutual funds in a certain region. Data mining could then be used to gain insight about the behavior of individual customers in the region. Finally, after data mining predicts something like a 5% increase in sales, OLAP can be used to track the net income. Or, Data Mining might be used to identify the most important attributes concerning sales of mutual funds, and those attributes could be used to design the data model in OLAP.

ii) **Explain metadata in detail. Classify metadata and explain the same. [May 2013]**

Metadata is simply defined as data about data. The data that are used to represent other data is known as metadata.

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

Categories of Metadata: The metadata can be broadly categorized into three categories:



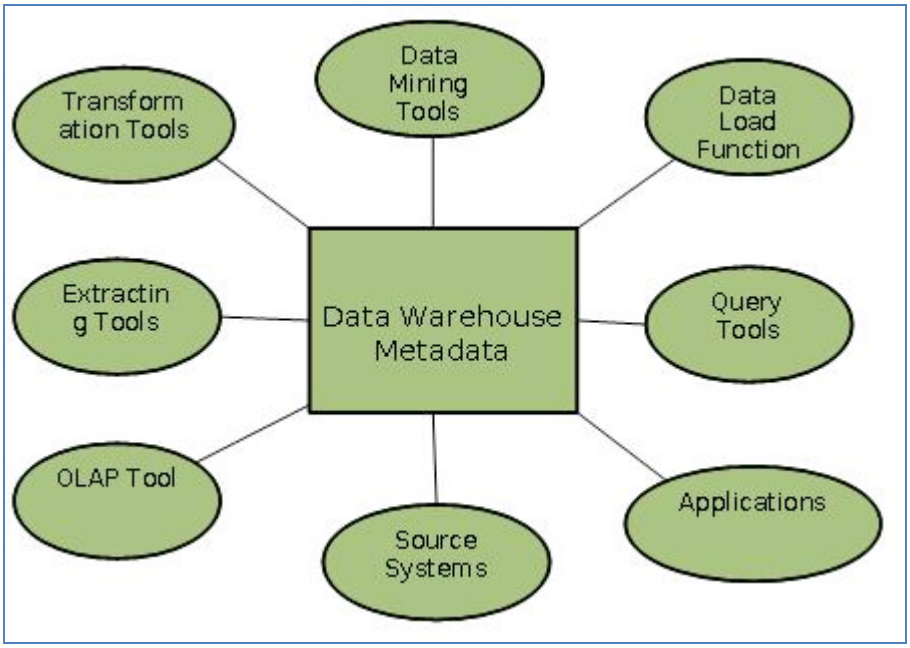
1. **Business Metadata** - This metadata has the data ownership information, business definition and changing policies.
2. **Technical Metadata** - Technical metadata includes database system names, table and column names and sizes, data types and allowed values. Technical metadata also includes structural information such as primary and foreign key attributes and indices.
3. **Operational Metadata** - This metadata includes currency of data and data lineage.Currency of data means whether data is active, archived or purged. Lineage of data means history of data migrated and transformation applied on it.

Role of Metadata

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

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

Diagram to understand role of Metadata.



Metadata Respiratory

The Metadata Respiratory is an integral part of data warehouse system. The Metadata Respiratory has the following metadata:

- Definition of data warehouse - This includes the description of structure of data warehouse. The description is defined by schema, view, hierarchies, derived data definitions, and data mart locations and contents.
- Business Metadata - This metadata has the data ownership information, business definition and changing policies.
- Operational Metadata - This metadata includes currency of data and data lineage. Currency of data means whether data is active, archived or purged. Lineage of data means history of data migrated and transformation applied on it.
- Data for mapping from operational environment to data warehouse - This metadata includes source databases and their contents, data extraction,data partition cleaning, transformation rules, data refresh and purging rules.
- The algorithms for summarization - This includes dimension algorithms, data on granularity, aggregation, summarizing etc.

Challenges for Metadata Management

The importance of metadata can not be overstated. Metadata helps in driving the accuracy of reports, validates data transformation and ensures the accuracy of calculations. The metadata also enforces the consistent definition of business terms to business end users. With all these uses of Metadata it also has challenges for metadata management. The some of the challenges are discussed below.

- The Metadata in a big organization is scattered across the organization. This metadata is spreaded in spreadsheets, databases, and applications.
- The metadata could present in text file or multimedia file. To use this data for information management solution, this data need to be correctly defined.
- There are no industry wide accepted standards. The data management solution vendors have narrow focus.
- There is no easy and accepted methods of passing metadata.

8.**Explain the potential performance problems with star schema.Give Examples[June 2013]**

The star schema suffers the following performance problems.

Indexing

Multipart key presents some problems in the star schema model.
(day->week-> month-> quarter-> year)

- It requires multiple metadata definition(one for each component) to design a single table.
- Since the fact table must carry all key components as part of its primary key, addition or deletion of levels in the hierarchy will require physical modification of the affected table, which is time-consuming processed that limits flexibility.
- Carrying all the segments of the compound dimensional key in the fact table increases the size of the index, thus impacting both performance and scalability.
- **Level Indicator**
- The dimension table design includes a level of hierarchy indicator for every record.
- Every query that is retrieving detail records from a table that stores details and aggregates must use this indicator as an additional constraint to obtain a correct result.
- The user is not and aware of the level indicator, or its values are in correct, the otherwise valid query may result in a totally invalid answer.
- Alternative to using the level indicator is the **snowflake** schema
- Aggregate fact tables are created separately from detail tables
- Snowflake schema contains separate fact tables for each level of aggregation
- **Other problems with the star schema design**
- ***Pairwise Join Problem***
- 5 tables require joining first two tables, the result of this join with third table and so on.
- The intermediate result of every join operation is used to join with the next table.
- Selecting the best order of pairwise joins rarely can be solve in a reasonable amount of time.
- Five-table query has 5!=120 combinations
- This problem is so serious that some databases will not run a query that tries to join too many tables

9. **Draw any two multidimensional schemas suitable for representing weather data and give their advantages and disadvantages.[**Dec 2013]

A number of data models have been proposed to conceptually model the multi-dimensional data maintained in the warehouse. These include the star schema, the snowflake schema, and the fact constellation schema. Since our data model, the cascaded star model, is an extension of the star model, in the following, we present these three models with examples, and bring out the limitations of these models in representing the data in our spatial data warehouse.

The Star Schema

Perhaps, star schema, first introduced by Ralph Kimball, is the earliest schema used to model the data warehouse implemented as a relational databases. In this schema, the data warehouse contains a large central table (fact table) containing the bulk of data (dimensions) with no redundancy, and a set of smaller attendant tables (dimension tables) with one for each

dimension. The schema graph resembles a starburst, with the dimension tables displayed in a radial pattern around the central fact table, as shown in Figure 4, where A is the fact table, and b, c, d, e and f are dimensions and represented by dimensional tables.

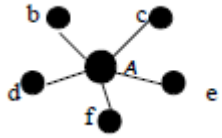


Figure 4: The Star Model

Note that in the star schema, only one dimension table represents each dimension, and each dimension table contains a set of attributes and joins with fact table by common keys when implemented as a relational database. Moreover, the attributes within a dimension table may form either a hierarchy (total order) or a lattice (partial order). Currently, most traditional data strong support for OLAP operations.

To illustrate, in the following, we provide an example of the implementation in star schema [8]. Suppose the multi-dimensional data for the weather in northeast region in USA consists of four dimensions: temperature, precipitation, time, and region_name, and three measures: region_map, area, and count, where region_map is a spatial measure which represents a collection of spatial pointers pointing to corresponding regions, area is a numerical measure which represents the sum of the total areas of the corresponding spatial objects, and count is a numerical measure which represents the total number of base regions accumulated in the corresponding cell.

The following figure illustrates the implementation for a star model in this case:

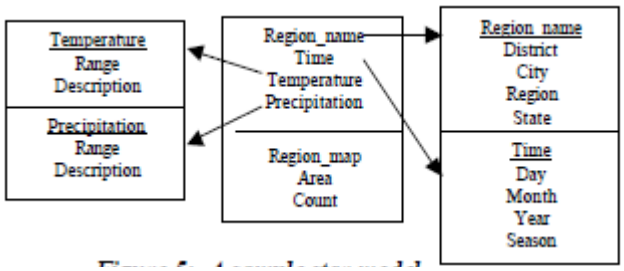


Figure A sample star model

The following tables show some sample data set that maybe collected from a number of weather districts tested in northeast of USA.

Region_name	Time	Temperature	Precipitatio n	…
A111	02/23/01	33	1.4	…
B111	02/24/01	41	1.5	…
…	…	…	…	…

Region_name	District	City	Region	State
A111	A	Flushing	111	NY
B111	B	Edison	111	NJ
…	…	…	…	…

Time	Day	Month	Year	Season
02/23/01	23	February	2001	Winter
02/24/01	24	February	2001	Winter
…	…	…	…	…

Temperature	Range	Description
33	11	Chilly
41	12	Mild cold
…	…	…

Precipitation	Range	Description
1.4	21	Middle
1.5	22	Middle
...	...	...

From this sample, we can see that a star model consists of a fact table with multiple dimension tables, and the fact table joins the dimension tables with different keys. In this example, all attributes in each dimension table are only one-dimensional and can be expressed completely in one table. Our question is: if some or all of the attributes in the dimension tables are also multidimensional,i.e., one attribute in one dimension table has multiple attributes associated with it,how can we implement it in this model? The answer is impossible.

UNIT II - PART A

1. What are the categories of decision support tools?[Nov 2014]

1. Reporting
2. Managed query
3. Executive information systems
4. On-line analytical processing
5. Data mining

2. What is the use of reporting tools? [May 2013]

Reporting tools can be divided into production reporting tools and desktop report writers. Production reporting tools will let companies generate regular operational reports or support high volume batch jobs, such as calculating and printing paychecks. Desktop tools designed for end users.

3. Define metalayer.

Managed query tools shield end users from the complexities of SQL and database structures by inserting a matalayer between users and the database. Metalayer is the software that provides subject-oriented views of a database and supports point-and-click creation of SQL.

4. Define EIS.

Executive Information Systems (EIS) tools predate report writers and managed query tools; they were first deployed on mainframes. EIS tools allow developers to build customized, graphical decision support application. That gives managers and executives a high-level view of the business and access to external resources.

5. What is the use of OLAP tools?

Online Analytical Processing (OLAP) tools provide an intuitive way to view corporate data. These tools aggregate data along common business subjects or dimensions and then let users navigate through the hierarchies and dimensions with the click of a mouse button. Users can drill down, across, or up levels in each dimension or pivot and swap out dimensions to change their view of the data.

6. What is the use of Cognos Impromptu tool?

Impromptu is the database reporting tool that exploits the power of the database, while offering complete control over all reporting within the enterprise. Users access Impromptu through its easy-to-use graphical user interface. Impromptu has been well received by users because querying and reporting are unified in one interface, and the users can get meaningful views of corporate data quickly and easily.

7. What are the special reporting options in Impromptu tool?

- Picklists and prompts
- Custom templates
- Exception reporting
- Interactive reporting
- Frames

8. What are the features of Impromptu tool?

- Unified query and reporting interface
- Object0oriented architecture
- Complete integration with powerplay
- Scalability
- Security and control
- Frame based reporting ,Database independent catalogs

9. What is the need of OLAP?

Modern business problems need query centric database schemas that are array oriented and multidimensional in nature. The characteristics of such problems are:

i)Need to retrieve large number of records from very large datasets

ii) summarize the data on the fly.

To solve these problems OLAP is needed. Eg:Solving modern business problems such as market analysis and financial forecasting requires query-centric database schemes that are array-oriented and multidimensional in nature. These business problems need OLAP to retrieve large number of records. from very large data sets and summarize them.

10.List the OLAP operations used in multidimensional data model.

Roll-up, drill-down, slice and dice, pivot (rotate)

11. List the categories of OLAP tools. [May 2011][Dec 2013]

- MOLAP (Multidimensional OLAP)
- ROLAP (Relational OLAP).
- Hybrid OLAP (HOLAP)
- Web OLAP

12. Differentiate MOLAP and OLAP. [Dec 2013]

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

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

13. What are the OLAP guidelines? [Dec 2013]

1. Multidimensional conceptual view: The OLAP should provide an appropriate multidimensional Business model that suits the Business problems and Requirements.
2. Transparency: The OLAP tool should provide transparency to the input data for the users.
3. Accessibility: The OLAP tool should only access the data required only to the analysis needed.
4. Consistent reporting performance: The Size of the database should not affect in any way the performance.
5. Client/server architecture: The OLAP tool should use the client server architecture to ensure better performance and flexibility.

14. Differentiate multidimensional and multirelational OLAP.

Relational implementations of multidimensional database systems are referred to as multirelational database systems. To achieve the required speed, these products use the star or snowflake schemas – specially optimized and denormalized data models that involve data restructuring and aggregation.

15. List out the two different types of reporting tools.[June 2014]

a)production reporting tools b)Desktop report writer

16. Define Data cube.[June 2013]

A data cube refers is a three-dimensional (3D) (or higher) range of values that are generally used to explain the time sequence of an image's data. It is a data abstraction to evaluate aggregated data from a variety of viewpoints. A data cube can also be described as the multidimensional extensions of two-dimensional tables. It can be viewed as a collection of identical 2-D tables stacked upon one another. Data cubes are used to represent data that is too complex to be described by a table of columns and rows. As such, data cubes can go far beyond 3-D to include many more dimensions.

17. Define OLAP.[June 2014]

OLAP can be defined as computer-based techniques used to analyze trends and perform business analysis using multidimensional views of business data. **OLAP** (online analytical processing) enables a user to easily and selectively extract and view data from different points of view.

18. What are the features of first generation web enabled data access.

Web sites used a static distribution model, in which clients access static HTML pages via web browsers. In this model, the decision support reports were stored as HTML documents and delivered to users on request.

19. What are the features of second generation and third generation web enabled data access.

Second generation

Web sites support interactive database queries by utilizing a multitiered architecture in which a web client submits a query in the form of HTML encoded request to a web server, which in turn transforms the request for structured data into a CGI scripts.

Third generation

Web sites replace HTML gateways with web based application servers. These servers can download java applets or ActiveX applications that can execute on clients, or interact with corresponding applets running on the server.

20.What is Virtual Warehouse?[Dec 2014]

The view over an operational data warehouse is known as a virtual warehouse. It is easy to build a virtual warehouse. Building a virtual warehouse requires excess capacity on operational database servers.

21. How the web is related with OLAP tool?

- The internet is a virtually free resource which provides a universal connectivity within and between companies.
- The web eases complex administrative tasks of managing distributed environments.
- The web allows companies to store and manage both data and applications on servers that can be centrally managed, maintained, and updated, thus eliminating problems with software and data currency.

22. Name some OLAP tools. [Dec 2013]

Arbor'sEssbase, Oracle Express, Planning Sciences' Gentia, Kenan Technologies' Acumate ES.

23. What are the various approaches for deploying OLAP tools on the web?

- HTML publishing
- Helper applications
- Plug-ins
- Server-centric components
- Java and ActiveX applications

24. Define OLTP systems.

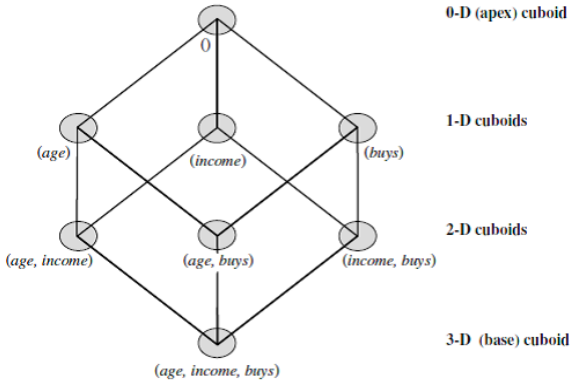
The major task of online operational database system is to perform online transaction and query processing. These systems are called On Line Transaction Processing (OLTP) systems. They cover most of the day-to-day operations of an organization such as purchasing, inventory, manufacturing and banking.

25. What is the need of tools for applications?

- Easy-to-use
- Point-and-click tools accept SQL or generate SQL statements to query relational data stored in the warehouse
- Tools can format the retrieved data into easy-to-read reports

26. What is apex cuboid?Give Example.[May 2011][Dec 2011]

The 0-D cuboid which holds the highest level of summarization is called the apex cuboid. The apex cuboid is typically denoted by all.



27.What is multidimensional database?[Dec 2011]

A multidimensional database (MDB) is a type of database that is optimized for data warehouse and online analytical processing (OLAP) applications. Multidimensional databases are frequently created using input from existing relational databases.

28.What is time series analysis?[June 2012]

Time series analysis comprises methods for analyzing **time series** data in order to extract meaningful statistics and other characteristics of the data. **Time series** forecasting is the use of a model to predict future values based on previously observed values.

29.What is a reporting tool?[Dec 2012]

Reporting is one of the most important part of performance management process. It's data presentation performed to allow target make more efficient decisions. Reporting tool enable business users without technical knowledge to extract corporate data, analyze it and assemble reports.

30.Give examples for managed query tools.[Dec 2012]

IQ Software's IQ Objects, Andyne Computing Ltd's GQL,IBM's Decision server,Oracle Corp's Discoverer/2000, Speedware Corp's Esperant.

31.What are production reporting tools? Give examples[June 2013]

Third Generation Language COBOL, Fourth generation language Information Builder, Focus, Client server tools MITI's SQR.

UNIT II -PART B

1.i) **List out the OLAP operations in multidimensional data model and explain with an example. [Dec 2009] [May 2013][Nov 2014]**

In the multidimensional data model, data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by concept hierarchies. This organization provides users with the flexibility to view data from different perspectives.

Ex: At the centre of the figure is a data cube for AllElectronics sales. The cube contains 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. The measure displayed is dollars_sold (in thousands).

Roll-up: The roll-up operation performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction. The fig shows the result of a roll-up operation performed on the central cube by climbing up the concept hierarchy for the location from the level of city to the level of country. In other words, rather than grouping the data by city, the resulting cube groups the data by country.

Drill-down: It is the reverse of roll-up. It can be realized by either stepping down a concept hierarchy for a dimension or introducing additional dimensions. In the fig. drill-down occurs by descending the time hierarchy from the level of quarter to the more detailed level of month. The resulting data cube details the total sales per month rather than summarized by quarter.

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

more dimensions. The dice operation in the fig. involves on three dimensions: (location = Toronto or Vancouver) and (time = Q1 or Q2) and (item=home entertainment or computer)
Pivot: Pivot is a visualization operation that rotates the data axes in view in order to provide an alternative presentation of the data. The fig. shows a pivot operation where the item and location axes in a 2-D slice are rotated.

ii) **Differentiate OLTP and OLAP. [Nov 2014] [June 2012][Dec 2011].**

OLTP stands for On Line Transaction Processing and is a data modeling approach typically used to facilitate and manage usual business applications. Most of applications you see and use are OLTP based. OLTP technology used to perform updates on operational or transactional systems (e.g., point of sale systems)
OLAP stands for On Line Analytic Processing and is an approach to answer multi-dimensional queries. OLAP was conceived for Management Information Systems and Decision Support Systems. OLAP technology used to perform complex analysis of the data in a data warehouse.

The following table summarizes the major differences between OLTP and OLAP system design.

	OLTP System Online Transaction Processing (Operational System)	OLAP System Online Analytical Processing (Data Warehouse)
Source of data	Operational data; OLTPs are the original source of the data.	Consolidation data; OLAP data comes from the various OLTP Databases
Purpose of data	To control and run fundamental business tasks	To help with planning, problem solving, and decision support
What the data	Reveals a snapshot of ongoing business processes	Multi-dimensional views of various kinds of business activities
Inserts and Updates	Short and fast inserts and updates initiated by end users	Periodic long-running batch jobs refresh the data
Queries	Relatively standardized and simple queries Returning relatively few records	Often complex queries involving aggregations
Processing Speed	Typically very fast	Depends on the amount of data involved; batch data refreshes and complex queries may take many hours; query speed can be improved by creating indexes
Space Requirements	Can be relatively small if historical data is archived	Larger due to the existence of aggregation structures and history data; requires more indexes than OLTP
Database Design	Highly normalized with many tables	Typically de-normalized with fewer tables; use of star and/or snowflake schemas
Backup and Recovery	Backup religiously; operational data is critical to run the business, data loss is likely to entail significant monetary loss and legal liability	Instead of regular backups, some environments may consider simply reloading the OLTP data as a recovery method

2.i)**Explain the categories of OLAP tools. [May 2013]or Discuss the architecture of MOLAP and ROLAP.[DEC 2012] or compare multidimensional OLAP(MOLAP) and multirelational OLAP(ROLAP)[June 2014] or With relevant examples discuss multidimensional online analytical processing(MOLAP) and multirelational online analytical processing(ROLAP). [May 2011]**

MOLAP

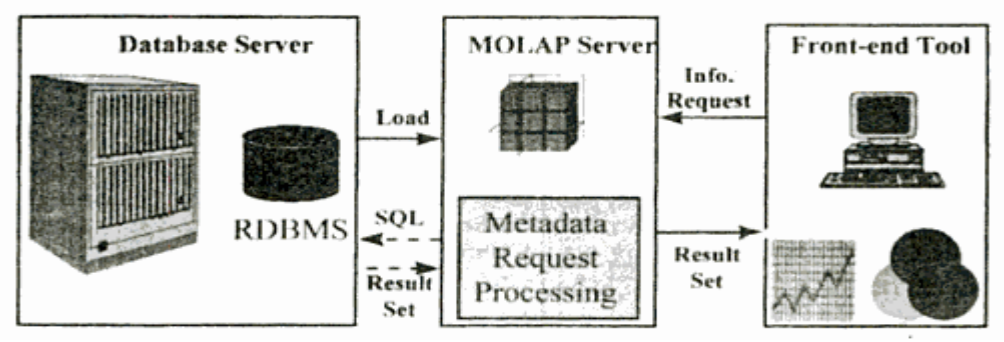
This is the more traditional way of OLAP analysis. In MOLAP, data is stored in a multidimensional cube. The storage is not in the relational database, but in proprietary formats. That is, data stored in array-based structures.

Advantages:

- Excellent performance: MOLAP cubes are built for fast data retrieval, and are optimal for slicing and dicing operations.
- Can perform complex calculations: All calculations have been pre-generated when the cube is created. Hence, complex calculations are not only doable, but they return quickly.

Disadvantages:

- Limited in the amount of data it can handle: Because all calculations are performed when the cube is built, it is not possible to include a large amount of data in the cube itself. This is not to say that the data in the cube cannot be derived from a large amount of data. Indeed, this is possible. But in this case, only summary-level information will be included in the cube itself.
- Requires additional investment: Cube technology are often proprietary and do not already exist in the organization. Therefore, to adopt MOLAP technology, chances are additional investments in human and capital resources are needed.



Examples: Hyperion Essbase, Fusion (Information Builders)

ROLAP

This methodology relies on manipulating the data stored in the relational database to give the appearance of traditional OLAP's slicing and dicing functionality. In essence, each action of slicing and dicing is equivalent to adding a "WHERE" clause in the SQL statement. Data stored in relational tables

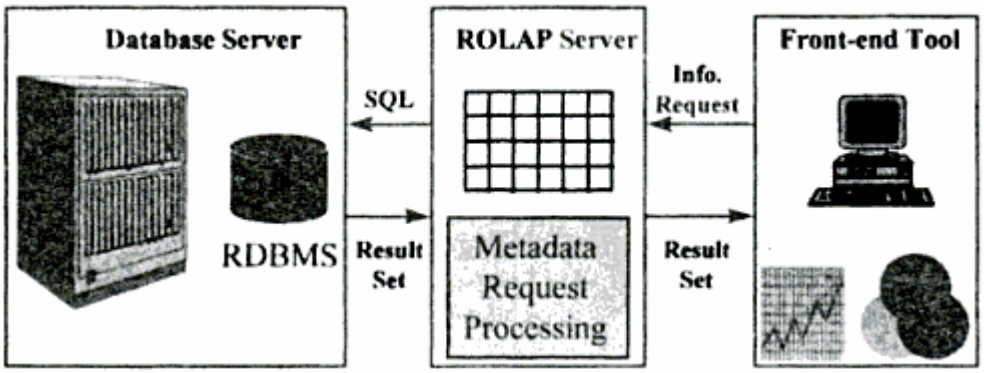
Advantages:

- Can handle large amounts of data: The data size limitation of ROLAP technology is the limitation on data size of the underlying relational database. In other words, ROLAP itself places no limitation on data amount.

- Can leverage functionalities inherent in the relational database: Often, relational database already comes with a host of functionalities. ROLAP technologies, since they sit on top of the relational database, can therefore leverage these functionalities.

Disadvantages:

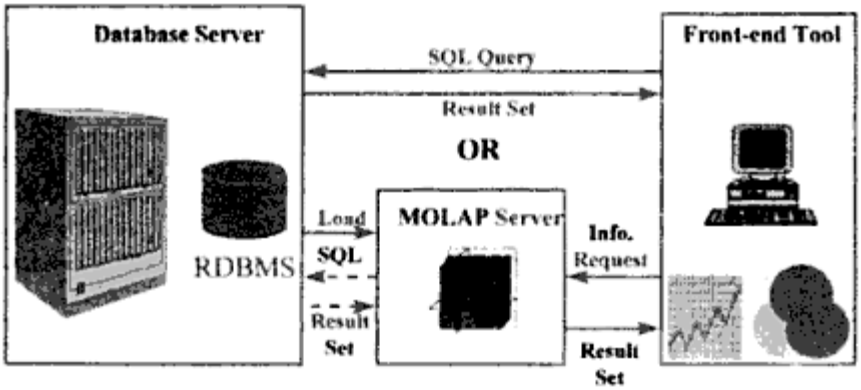
- Performance can be slow: Because each ROLAP report is essentially a SQL query (or multiple SQL queries) in the relational database, the query time can be long if the underlying data size is large.
- Limited by SQL functionalities: Because ROLAP technology mainly relies on generating SQL statements to query the relational database, and SQL statements do not fit all needs (for example, it is difficult to perform complex calculations using SQL), ROLAP technologies are therefore traditionally limited by what SQL can do. ROLAP vendors have mitigated this risk by building into the tool out-of-the-box complex functions as well as the ability to allow users to define their own functions.



Examples: Microstrategy Intelligence Server, MetaCube (Informix/IBM)

HOLAP (MQE: Managed Query Environment)

HOLAP technologies attempt to combine the advantages of MOLAP and ROLAP. For summary-type information, HOLAP leverages cube technology for faster performance. It stores only the indexes and aggregations in the multidimensional form while the rest of the data is stored in the relational database.



Examples: PowerPlay (Cognos), Brio, Microsoft Analysis Services, Oracle Advanced Analytic Services

Compare multidimensional OLAP(MOLAP) and multirelational OLAP(ROLAP)[June 2014]

MOLAP vs ROLAP

Sr.No.	MOLAP	ROLAP
1	Information retrieval is fast.	Information retrieval is comparatively slow.
2	Uses sparse array to store data-sets.	Uses relational table.
3	MOLAP is best suited for inexperienced users, since it is very easy to use.	ROLAP is best suited for experienced users.
4	Maintains a separate database for data cubes.	It may not require space other than available in the Data warehouse.
5	DBMS facility is weak.	DBMS facility is strong.

3.i)**Explain the OLAP tool with the internet.**

Web sites used a static distribution model, in which clients access static HTML pages via web browsers. In this model, the decision support reports were stored as HTML documents and delivered to users on request. Web sites support interactive database queries by utilizing a multitiered architecture in which a web client submits a query in the form of HTML encoded request to a web server, which in turn transforms the request for structured data into a CGI scripts. Web sites replace HTML gateways with web based application servers. These servers can download java applets or ActiveX applications that can execute on clients, or interact with corresponding applets running on the server.

ii) **Explain OLAP guidelines.**

Dr. E.F. Codd the "father" of the relational model, created a list of rules to deal with the OLAP systems. Users should priorities these rules according to their needs to match their business requirements (reference 3). These rules are:

1) Multidimensional conceptual view: The OLAP should provide an appropriate multidimensional Business model that suits the Business problems and Requirements.
2) Transparency: The OLAP tool should provide transparency to the input data for the users.
3) Accessibility: The OLAP tool should only access the data required only to the analysis needed.
4) Consistent reporting performance: The Size of the database should not affect in any way the performance.
5) Client/server architecture: The OLAP tool should use the client server architecture to ensure better performance and flexibility.
6) Generic dimensionality: Data entered should be equivalent to the structure and operation requirements.
7) Dynamic sparse matrix handling: The OLAP too should be able to manage the sparse matrix and so maintain the level of performance.
8) Multi-user support: The OLAP should allow several users working concurrently to work together.
9) Unrestricted cross-dimensional operations: The OLAP tool should be able to perform operations across the dimensions of the cube.

10) Intuitive data manipulation. "Consolidation path re-orientation, drilling down across columns or rows, zooming out, and other manipulation inherent in the consolidation path outlines should be accomplished via direct action upon the cells of the analytical model, and should neither require the use of a menu nor multiple trips across the user interface."(Reference 4)
11) Flexible reporting: It is the ability of the tool to present the rows and column in a manner suitable to be analyzed.
12) Unlimited dimensions and aggregation levels: This depends on the kind of Business, where multiple dimensions and defining hierarchies can be made.

iii)**How to reduce the size of the fact table? Explain with an example[Dec 2014]**

Each data warehouse or data mart includes one or more fact tables. Central to a star or snowflake schema, a fact table captures the data that measures the organization's business operations. A fact table might contain business sales events such as cash register transactions or the contributions and expenditures of a nonprofit organization. Fact tables usually contain large numbers of rows, sometimes in the hundreds of millions of records when they contain one or more years of history for a large organization.

A key characteristic of a fact table is that it contains numerical data (facts) that can be summarized to provide information about the history of the operation of the organization. Each fact table also includes a multipart index that contains as foreign keys the primary keys of related dimension tables, which contain the attributes of the fact records. Fact tables should not contain descriptive information or any data other than the numerical measurement fields and the index fields that relate the facts to corresponding entries in the dimension tables.

In the **FoodMart 2000** sample database provided with Microsoft® SQL Server™ 2000 Analysis Services, one fact table,**sales_fact_1998**, contains the following columns.

Column	Description
product_id	Foreign key for dimension table **product**.
time_id	Foreign key for dimension table **time_by_day**.
customer_id	Foreign key for dimension table **customer**.
promotion_id	Foreign key for dimension table **promotion**.
store_id	Foreign key for dimension table **store**.
store_sales	Currency column containing the value of the sale.

store_cost	Currency column containing the cost to the store of the sale.
unit_sales	Numeric column containing the quantity sold.

In this fact table, each entry represents the sale of a specific product on a specific day to a specific customer in accordance with a specific promotion at a specific store. The business measurements captured are the value of the sale, the cost to the store, and the quantity sold.

The most useful measures to include in a fact table are numbers that are additive. Additive measures allow summary information to be obtained by adding various quantities of the measure, such as the sales of a specific item at a group of stores for a particular time period. Nonadditive measures such as inventory quantity-on-hand values can also be used in fact tables, but different summarization techniques must then be used.

Aggregation in Fact Tables

Aggregation is the process of calculating summary data from detail records. It is often tempting to reduce the size of fact tables by aggregating data into summary records when the fact table is created. However, when data is summarized in the fact table, detailed information is no longer directly available to the analyst. If detailed information is needed, the detail rows that were summarized will have to be identified and located, possibly in the source system that provided the data. Fact table data should be maintained at the finest granularity possible. Aggregating data in the fact table should only be done after considering the consequences.

Mixing aggregated and detailed data in the fact table can cause issues and complications when using the data warehouse. For example, a sales order often contains several line items and may contain a discount, tax, or shipping cost that is applied to the order total instead of individual line items, yet the quantities and item identification are recorded at the line item level. Summarization queries become more complex in this situation, and tools such as Analysis Services often require the creation of special filters to deal with the mixture of granularity.

There are two approaches that can be used in this situation. One approach is to allocate the order level values to line items based on value, quantity, or shipping weight. Another approach is to create two fact tables, one containing data at the line item level, the other containing the order level information. The order identification key should be carried in the detail fact table so the two tables can be related. The order table can then be used as a dimension table to the detail table, with the order-level values considered as attributes of the order level in the dimension hierarchy.

4.**List and discuss the basic features that are provided by reporting and query tools used for business analysis. [May 2011] [Dec 2011]**

Production reporting tool used to generate regular operational reports, Desktop report writer are inexpensive desktop tools designed for end users.

Application development tools:This is a graphical data access environment which integrates OLAP tools with data warehouse and can be used to access all db systems *OLAP Tools:* are used to analyze the data in multi dimensional and complex views. To enable multidimensional properties it uses MDDB and MRDB where MDDB refers multi dimensional data base and MRDB refers multi relational data bases.

Data mining tools:are used to discover knowledge from the data warehouse data also can be used for data visualization and data correction purposes.

Reporting features:

- Topline results
- Simple cross tables

- Interlocking cross breaks(e.g. age by gender - male/18-24)
- View data in tabular and graphical formats (i.e. as tables or charts)
- Report exportable to Excel and/or PowerPoint
- Multiple reports exportable to Excel
- Charts exportable to Powerpoint

Report features

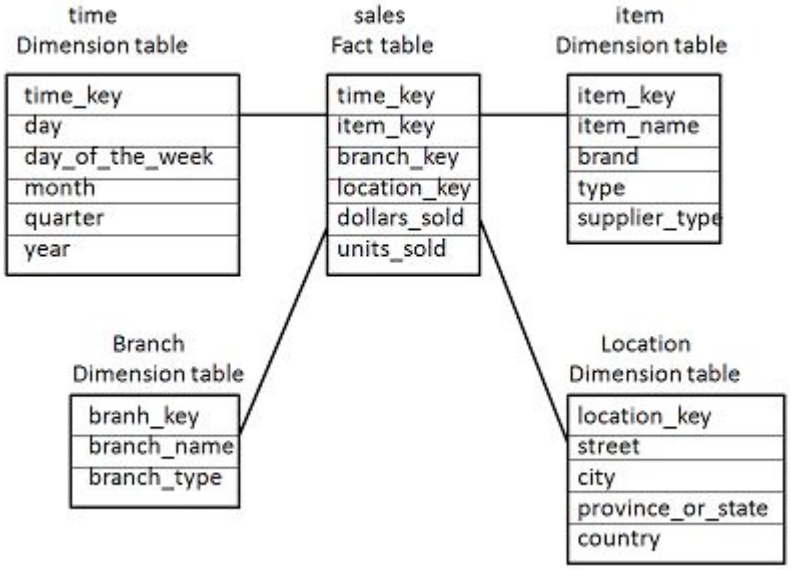
- Statistical tests
- Score values can be applied to variable to compute mean scores
- T-tests (significance testing)
- Report formatting
- Hide blank rows columns
- Show row column percentages
- Show or hide counts
- Show indices

5.**Giving the suitable examples , describe the various multi-dimensional schema**[Dec 2011].

Schema is a logical description of the entire database. It includes the name and description of records of all record types including all associated data-items and aggregates. Much like a database, a data warehouse also requires to maintain a schema. A database uses relational model, while a data warehouse uses Star, Snowflake, and Fact Constellation schema. In this chapter, we will discuss the schemas used in a data warehouse.

Star Schema

- Each dimension in a star schema is represented with only one-dimension table.
- This dimension table contains the set of attributes.
- The following diagram shows the sales data of a company with respect to the four dimensions, namely time, item, branch, and location.

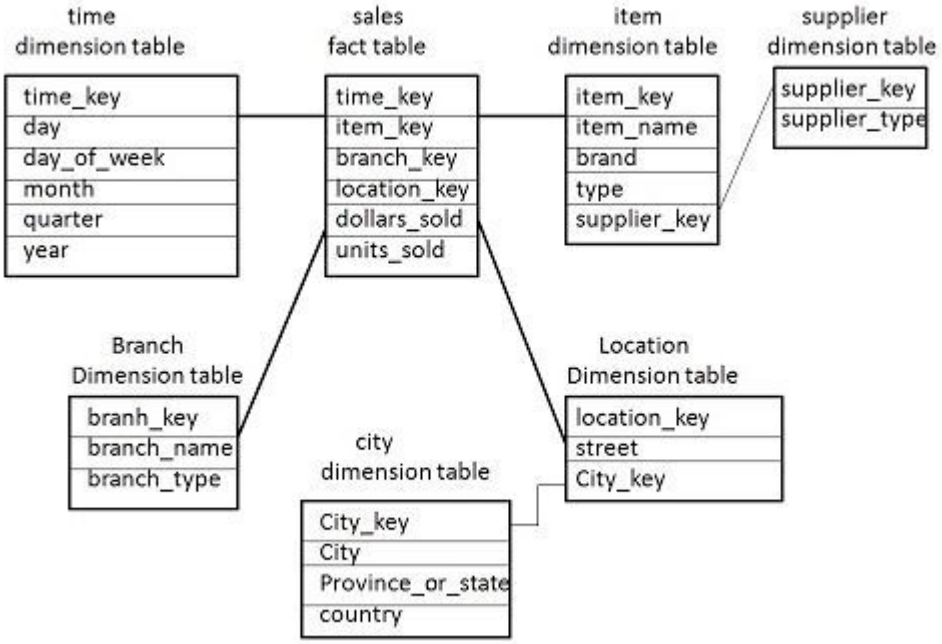


- There is a fact table at the center. It contains the keys to each of four dimensions.
- The fact table also contains the attributes, namely dollars sold and units sold.

Note: Each dimension has only one dimension table and each table holds a set of attributes. For example, the location dimension table contains the attribute set {location_key, street, city, province_or_state,country}. This constraint may cause data redundancy. For example, "Vancouver" and "Victoria" both the cities are in the Canadian province of British Columbia. The entries for such cities may cause data redundancy along the attributes province_or_state and country.

Snowflake Schema

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

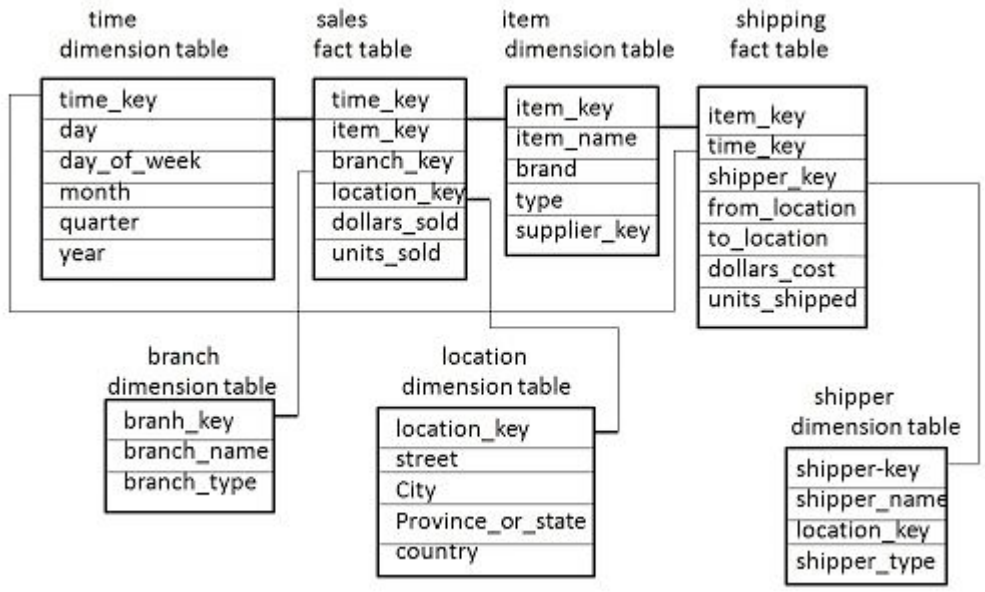


- Now the item dimension table contains the attributes item_key, item_name, type, brand, and supplier-key.
- The supplier key is linked to the supplier dimension table. The supplier dimension table contains the attributes supplier_key and supplier_type.

<b< style="box-sizing: border-box;">Note: Due to normalization in the Snowflake schema, the redundancy is reduced and therefore, it becomes easy to maintain and the save storage space.</b<>

Fact Constellation Schema

- A fact constellation has multiple fact tables. It is also known as galaxy schema.
- The following diagram shows two fact tables, namely sales and shipping.



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

Schema Definition

Multidimensional schema is defined using Data Mining Query Language (DMQL). The two primitives, cube definition and dimension definition, can be used for defining the data warehouses and data marts.

Syntax for Cube Definition

define cube < cube_name > [< dimension-list > }: < measure_list >

Syntax for Dimension Definition

define dimension < dimension_name > as (< attribute_or_dimension_list >)

Star Schema Definition

The star schema that we have discussed can be defined using Data Mining Query Language (DMQL) as follows:

define cube sales star [time, item, branch, location]:

dollars sold = sum(sales in dollars), units sold = count(*)

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)

Snowflake Schema Definition

Snowflake schema can be defined using DMQL as follows:

define cube sales snowflake [time, item, branch, location]:

dollars sold = sum(sales in dollars), units sold = count(*)

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

Fact Constellation Schema Definition

Fact constellation schema can be defined using DMQL as follows:

define cube sales [time, item, branch, location]:

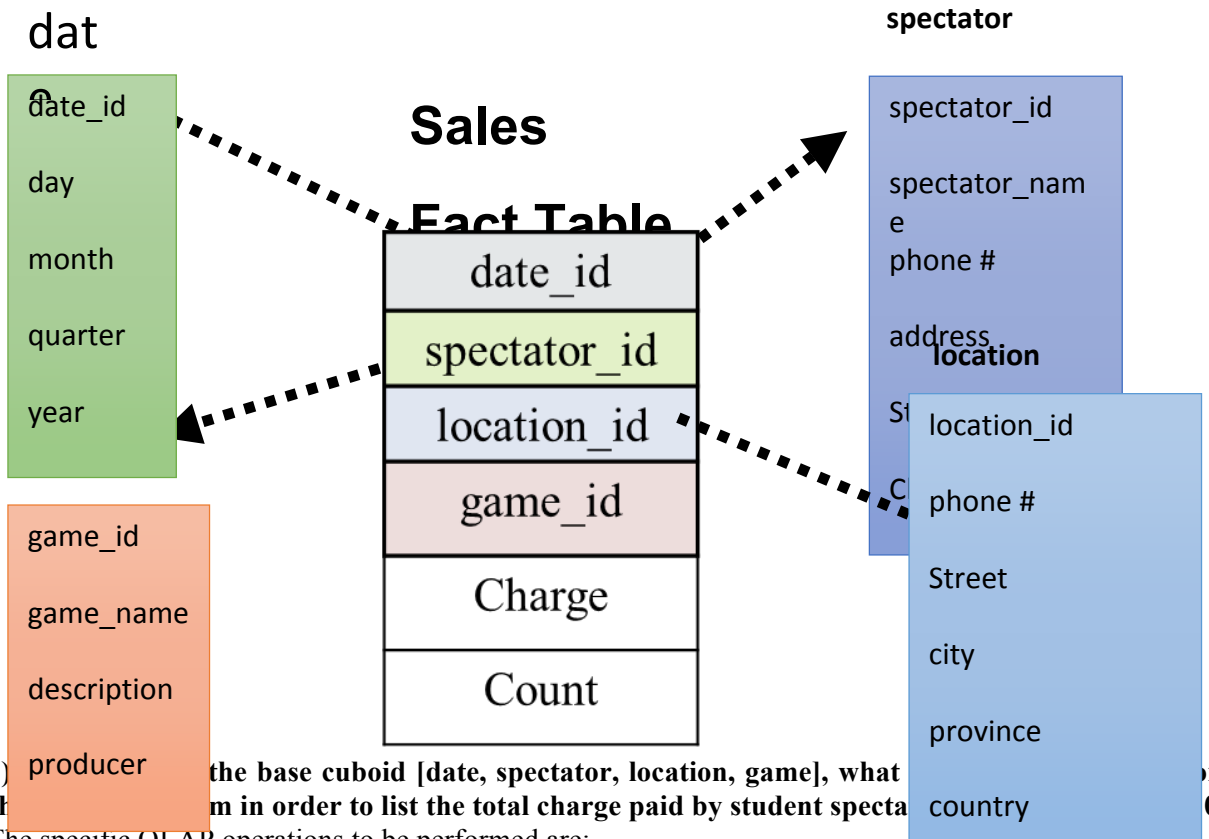
dollars sold = sum(sales in dollars), units sold = count(*)

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)
define cube shipping [time, item, shipper, from location, to location]:

dollars cost = sum(cost in dollars), units shipped = count(*)

define dimension time as time in cube sales
define dimension item as item in cube sales
define dimension shipper as (shipper key, shipper name, location as location in cube sales, shipper type)
define dimension from location as location in cube sales
define dimension to location as location in cube sales

6i)compare the concepts :discovery driven cube, multifeature cube,and virtual warehouse.
[June 2012]
ii) Suppose that a data warehouse consists of the four dimensions, date, spectator, location, and game, and the two measures, count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate.[June 2012]
1)Draw a star schema diagram for the data warehouse.



2) ...the base cuboid [date, spectator, location, game], what ...ons sh...m in order to list the total charge paid by student specta...0?

The specific OLAP operations to be performed are:

- Roll-up on date from date id to year.
- Roll-up on game from game id to all.
- Roll-up on location from location id to location name.
- Roll-up on spectator from spectator id to status.
- Dice with status="students", location name="GM Place", and year = 2010.

7. Explain the features of the reporting and query tool COGNOS IMPROMPTU. [Dec 2013][May 2011][Dec 2012]

Impromptu is an interactive database reporting tool. It allows Power Users to query data without programming knowledge. When using the Impromptu tool, no data is written or changed in the database. It is only capable of reading the data.

Impromptu's main features includes,

- Interactive reporting capability
- Enterprise-wide scalability
- Superior user interface
- Fastest time to result
- Lowest cost of ownership

A report can include a prompt that asks you to select a product type from a list of those available in the database. Only the products belonging to the product type you select are retrieved and displayed in your report. Reports are created by choosing fields from the catalog folders.

8. i)**Describe multidimensional data model in detail.[June 2013]**

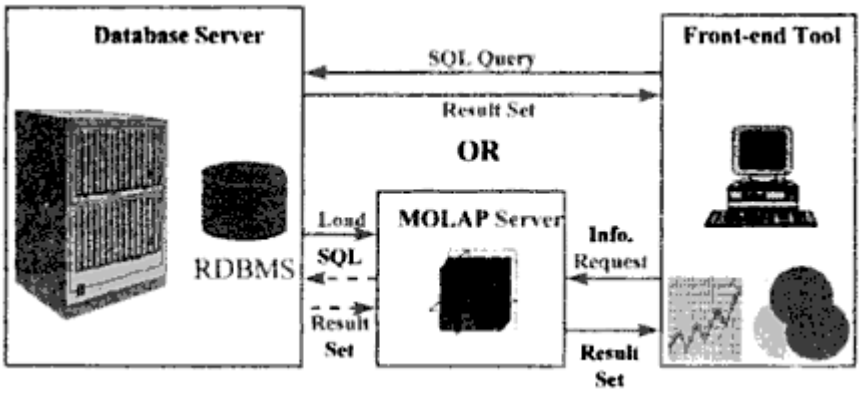
The multidimensional data model is an integral part of On-Line Analytical Processing, or OLAP. Because OLAP is on-line, it must provide answers quickly; analysts pose iterative queries duringinteractive sessions, not in batch jobs that run overnight. And because OLAP is also analytic, the queries are complex. The multidimensional data model is designed to solve complex queries in real time. Multidimensional data model is to view it as a cube. The cable at the left contains detailed sales data by product, market and time. The cube on the right associates sales number (unit sold) with dimensions-product type, market and time with the unit variables organized as cell in an array.This cube can be

expended to include another array-price-which can be associates with all or only some dimensions. As number of dimensions increases number of cubes cell increase exponentially. Dimensions are hierarchical in nature i.e. time dimension may contain hierarchies for years, quarters, months, weak and day.

ii) **Explain with diagrammatic illustration managed query environment (MQE) architecture.[June 2013**]

HOLAP (MQE: Managed Query Environment)

HOLAP technologies attempt to combine the advantages of MOLAP and ROLAP. For summary-type information, HOLAP leverages cube technology for faster performance. It stores only the indexes and aggregations in the multidimensional form while the rest of the data is stored in the relational database.



Examples: PowerPlay (Cognos), Brio, Microsoft Analysis Services, Oracle Advanced Analytic Services

Client Runtime Steps:

– Fetch data from MOLAP Server, or RDBMS directly

– Build memory-based data structures, as required

– Execute the analysis application

• Advantages:

– Distributes workload to the clients, offloading the servers

– Simple to install, and maintain => reduced cost

• Disadvantages:

– Provides limited analysis capability (i.e., client is less powerful than a server)

– Lots of redundant data stored on the client systems

– Client-defined and cached datacubes can cause inconsistent data

– Uses lots of network bandwidth9.i)**what are the differences between the three main types of data warehouse usage: information processing, Analytical processing, data mining. Page Number 146 Topic 3.5 From Data Warehousing to Data Mining**

Three kinds of data warehouse applications

» **Information processing**

• supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts and graphs

» **Analytical processing**

• multidimensional analysis of data warehouse data • supports basic OLAP operations, slice-dice, drilling, pivoting

» **Data mining**

• knowledge discovery from hidden patterns • supports associations, constructing analytical models,performing classification and prediction, and presenting the mining results using visualization tools

online analytical processing uses basic operations such as slice and dice drilldown and roll up on historical data in order to provide multidimensional analysis of data.

Data mining uses knowledge discovery to find out hidden patterns and association constructing analytical models and presenting mining results with visualization tools.

Sr.No.	Data Warehouse (OLAP)	Operational Database(OLTP)
1	It involves historical processing of information.	It involves day-to-day processing.
2	OLAP systems are used by knowledge workers such as executives, managers, and analysts.	OLTP systems are used by clerks, DBAs, or database professionals.
3	It is used to analyze the business.	It is used to run the business.
4	It focuses on Information out.	It focuses on Data in.
5	It is based on Star Schema, Snowflake Schema, and Fact Constellation Schema.	It is based on Entity Relationship Model.
6	It focuses on Information out.	It is application oriented.
7	It contains historical data.	It contains current data.
8	It provides summarized and consolidated data.	It provides primitive and highly detailed data.
9	It provides summarized and multidimensional view of data.	It provides detailed and flat relational view of data.
10	The number of users is in hundreds.	The number of users is in thousands.
11	The number of records accessed is in millions.	The number of records accessed is in tens.
12	The database size is from 100GB to 100 TB.	The database size is from 100 MB to 100 GB.
13	These are highly flexible.	It provides high performance.

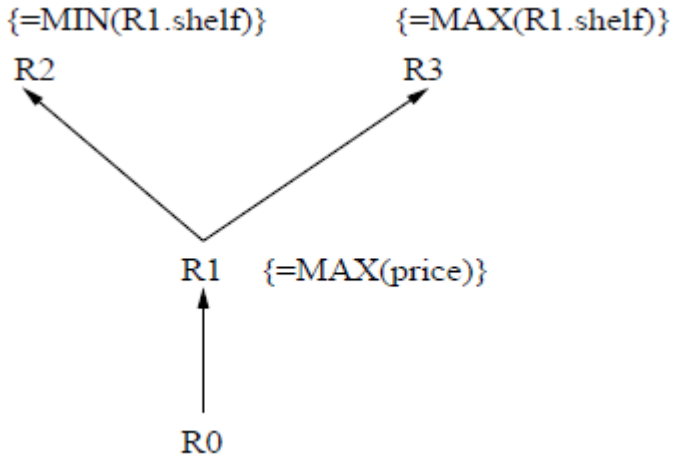
- **Information Processing** - A data warehouse allows to process the data stored in it. The data can be processed by means of querying, basic statistical analysis, reporting using crosstabs, tables, charts, or graphs.

- **Analytical Processing** - A data warehouse supports analytical processing of the information stored in it. The data can be analyzed by means of basic OLAP operations, including slice-and-dice, drill down, drill up, and pivoting.

- **Data Mining** - Data mining supports knowledge discovery by finding hidden patterns and associations, constructing analytical models, performing classification and prediction. These mining results can be presented using the visualization tools.

ii)consider the following multi-feature cube query: Grouping by all subsets of (item, region, month)
Find the maximum price in 1997 for each group. Among the maximum price tuples, find the minimum and maximum item shelf lives. Also find the fraction of the total sales due to tuples that have minimum shelf life within the set of all maximum price tuples, and the fraction of the total sales due to tuples that have maximum shelf life within the set of all maximum price tuples.

1)Draw a multi-feature cube graph for the query



2)Express the query in extended SQL

```
select   item, region, month, MAX(price), MIN(R1.shelf), MAX(R1.shelf),
         SUM(R1.sales), SUM(R2.sales), SUM(R3.sales)
from     Purchases
where    year = 1997
cube by  item, region, month: R1, R2, R3
such that R1.price = MAX(price) and
         R2 in R1 and R2.shelf = MIN(R1.shelf) and
         R3 in R1 and R3.shelf = MAX(R1.shelf)
```

3)Is this a distributive multi-feature cube?Justify

Intuitively, Query 2 is a distributive multifeature cube since we can distribute its computation by incrementally generating the output of the cube at a higher level granularity using only the output of the cube at a lower level granularity. Similarly, Query 3 is also distributive. Some multifeature cubes that are not distributive may be "converted" by adding aggregates to the `select` clause so that the resulting cube is distributive. For example, suppose that the `select` clause for a given multifeature cube has `AVG(sales)`, but neither `COUNT(sales)` nor `SUM(sales)`. By adding `SUM(sales)` to the `select` clause, the resulting data cube is distributive. The original cube is therefore algebraic. In the new distributive cube, the average sales at a higher level granularity can be computed from the average and total sales at lower level granularities. A cube that is neither distributive nor algebraic is holistic.

11. **Discuss how datawarehousing is used in retail and telecommunication industry.[Dec 2013]**

Uses of Data Warehousing in Telecommunications

- Churn
  - Differentiate between the propensity to churn and actual churn
  - Differentiate between product church and customer churn
- Fraud Detection
  - Data mining tools can predict fraud by spotting patterns in consolidated customer information and call detail records
- Product Packaging and Custom Pricing
  - Using knowledge discover and modeling, companies can tell which products will see well together, as well as which customers or customer segments are most likely to buy them
    - Packaging of vertical features
      - Voice products such as caller ID, call waiting
  - Employ price elasticity models to determine the new package's optimal price
- Network Feature Management
  - By monitoring call patterns and traffic routing, a carrier can install a switch or cell in a location where it is liable to route the maximum amount of calls
- Historical activity analysis can help telecommunications companies predict equipment outages before they occur
- Call Detail Analysis
  - Analysis of specific call records
  - Helps provide powerful information about origin and destination patterns that could spur additional sales to important customers
- Customer Satisfaction

12. **Suppose that a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit.[Dec 2014]**

(a) Enumerate three classes of schemas that are popularly used for modeling data warehouses.

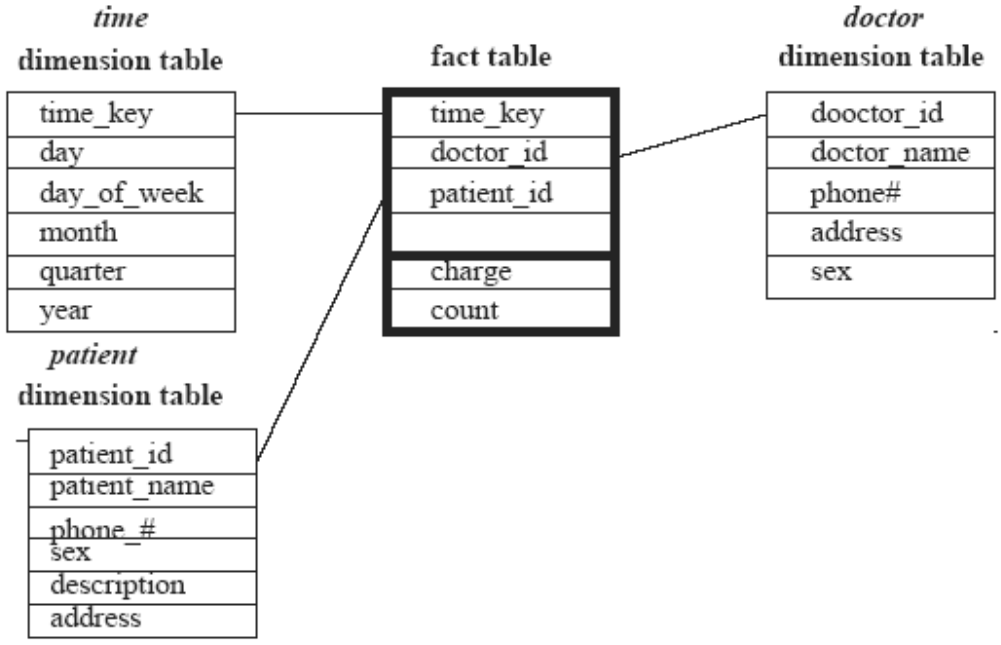
(a) star schema: a fact table in the middle connected to a set of dimension tables
snowflake schema: a refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake.

Fact constellations: multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation.

(b) Draw a schema diagram for the above data warehouse using one of the schema classes listed in (a).

(c) Starting with the base cuboid [**day, doctor, patient**], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2004?

(d) To obtain the same list, write an SQL query assuming the data are stored in a relational database with the schema fee (**day, month, year, doctor, hospital, patient, count, charge**).



c.Starting with the base cuboid [**day, doctor, patient**], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2004?

1. roll up from day to month to year
2. slice for year = "2004"
3. roll up on patient from individual patient to all
4. slice for patient = "all"
4. get the list of total fee collected by each doctor in 2004

d.

```
Select doctor, Sum(charge)
From fee
Where year = 2004
Group by doctor
```

UNIT III - PART A

1. **What is Data mining?**

Data mining refers to extracting or "mining" knowledge from large amount of data. It is considered as a synonym for another popularly used term Knowledge Discovery in Databases or KDD.

2. Give the steps involved in KDD. [Dec 2013]

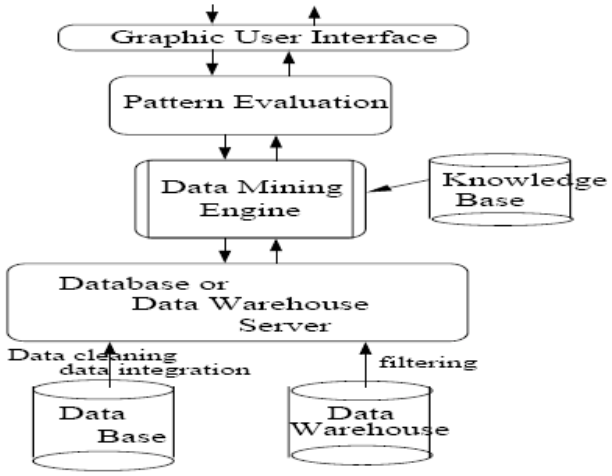
KDD consists of the iterative sequence of the following steps:

- Data cleaning
- Data integration.
- Data selection
- Data transformation
- Data mining

- Pattern Evaluation
- Knowledge Presentation

3. Give the architecture of a typical data mining system.

The architecture of a typical data mining system consists of the following components:



Architecture of a typical data mining system.

4. Give the classification of data mining tasks.[June 2014]

Descriptive – Characterizes the general property of the data in the database.

Predictive – perform inference on the current data in order to make predictions

5. Give the classification of data mining systems.

Data mining systems can be categorized according to various criteria:

- Classification according to the kinds of databases mined.
- Classification according to the kinds of knowledge mined.
- Classification according to the kinds of techniques utilized.
- Classification according to the kinds of the application adapted.

6. Define pattern interestingness. What makes a pattern interesting? Or What is Pattern evaluation.[June 2013][Dec 2011]

Pattern evaluation is used to identify the truly interesting patterns representing knowledge based on some interestingness measures.A pattern is interesting if it validates a hypothesis that the user sought to confirm. An interesting pattern represents knowledge. A pattern is interesting if it is 1). Easily understood by humans 2).valid on new or test data with some degree of certainty 3).potentially useful 4).Novel.

7. What kind of data can be mined?

Kinds of data are Database data, data warehouses, transactional data, other kinds of data like time related data, data streams, spatial data, engineering design data, multimedia data and web data.

8. State why preprocessing data an important issue for data mining and data warehousing. Or Why preprocess the data? [May 2011] [Dec 2013] [June 2012]

Data that is to be analyzed by data mining techniques are incomplete, noisy, and inconsistent. These are the common place properties of large real world databases and data warehouses. To remove all these errors data must be preprocessed.

9. What are the data preprocessing techniques?

Data preprocessing techniques are

- Data cleaning-removes noise and correct inconsistencies in the data.
- Data integration-merges data from multiple sources into a coherent data store such as data warehouse or a data cube.
- Data transformations-such as normalization improve the accuracy and efficiency of mining algorithms involving distance measurements.
- Data reduction-reduces the data size by aggregating, eliminating redundant features, or clustering.

10. Define Database management system.

A database system also called database management system consists of a collection of interrelated data known as a database and a set of software programs to manage and access the data.

11. Give the various data smoothing techniques.

Binning, clustering, combined computer and human inspection, regression.

12. Define data integration.

Data integration combines data from multiple sources into a coherent data store. These sources may include multiple databases, data cubes or flat files.

13. Give the issues to be considered during data integration. [Dec 2011]

Schema integration, Redundancy, detection and resolution of data value conflicts.

14. What is the significance of task relevant data.[June 2012]

It specifies only the concept and relevant data for a task in a summarative & consise manner.

15. What are the types of data?[Nov 2014]

- Qualitative data
- Quantitative data

16. What are the data mining functionalities? Or What kind of patterns can be mined? [May 2011]

It is used to specify the kinds of patterns or knowledge to be found in data mining tasks. It includes

- Concept description: Characterization and
- discrimination
- Generalize, summarize, and contrast data
- characteristics, e.g., dry vs. wet regions
- Classification and Prediction
- Clusters and outliers

17. Define data discrimination. [Dec 2013]

It is a comparison of the general features of target data objects with the general features of objects from one or a set of contrasting classes. The target and contrasting classes are specified by the user and the corresponding data objects retrieved through database queries.

18.Define data generalization.

Data generalization is a process that abstracts a large set of task-relevant data in a database from a relatively low conceptual level to higher conceptual levels. Methods for generalization can be categorized according to two approaches: 1) data cube approach and 2) attribute-oriented induction approach.

19. Define data reduction.

It is used to obtain a reduced representation of the data set that is much smaller in volume yet closely maintains the integrity of the original data. I-e mining on the reduced set should be more efficient yet produce the same analytical results.

20. Give the strategies used for data reduction.

Data cube aggregation, dimension reduction, data compression, numerosity reduction, and discretization and concept hierarchy generation.

21. What is data cube aggregation?

Data cube store multidimensional aggregated information. Each cell holds an aggregate data value data value, corresponding to the data point in multidimensional space. Concept hierarchies may exist for each attribute allowing the analysis of data at multiple levels of abstraction. Data cubes provide fast access to precomputed, summarized data, thereby benefiting on-line analytical processing as well as data mining.

22. What is dimensionality reduction?

Dimensionality reduction reduces the data set size by removing such attributes from it. Mining on a reduced set of attributes reduces the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand.

23. Define data characterization. [April/May 2010] [Dec 2013][Dec 2012]

It is a summarization of the general characteristics or feature of a target class of data. The data corresponding to the user-specified class are typically collected by a database query.

24. Give the output forms of data characterization.

Pie charts, bar charts, curves, multidimensional data cubes and multidimensional tables including cross tabs. The resulting descriptions can also be presented as generalized relations or in rule form called characteristic rule.

25. State the need of data cleaning. [Dec 2011] [May 2013]

Data cleaning removes noise and correct inconsistencies in the data. It cleans the data by filling in missing values smoothing noisy data, identifying or removing outliers and resolving inconsistencies.

26.List the primitives that specify a data mining task.[June 2012]

1. The set of task-relevant data to be mined
2. The kind of knowledge to be mined
3. The background knowledge to be used in the discovery process.

27.Mention the steps involved in the class comparison procedure[June 2012]

It is basically the comparison of the general features of objects between two classes referred to as the target class and the contrasting class. For example, one may want to compare the general characteristics of the customers who rented more than 30 movies in the last year with those whose rental account is lower than 5. The techniques used for data discrimination are very similar to the techniques used for data characterization with the exception that data discrimination results include comparative measures.

28. State why concept hierarchies are useful in data mining[Dec 2012]

The concept hierarchies are induced by a partial order over the values of a given attribute. Usually data can be abstracted at different conceptual levels. The raw data in a database is called at its primitive level and the knowledge is said to be at a primitive level if it is discovered by using raw data only. Abstracting raw data to a higher conceptual level and discovering and expressing knowledge at higher abstraction levels have superior advantage over data mining at a primitive level.

29.State the need for data pre-processing.[Dec 2013]

Real world data are generally 1)Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data 2)Noisy: containing errors or outliers 3)Inconsistent: containing discrepancies in codes or names. So to remove all these data preprocessing is needed.

30.Differentiate between data characterization and discrimination.[Dec 2013]

Characterization: provides a concise and succinct summarization of the given collection of data

Discrimination or Comparison: provides descriptions comparing two or more collections of data

31.What is a legacy database?[June 2014]

It is a group of heterogeneous databases that combines different kinds of data systems such as relational or object oriented databases, hierarchical databases, spreadsheets, network databases, multimedia databases or file systems. The heterogeous databases in legact database may be connected by inter or intra computer networks.

32. What is meta learning.[Dec 2014]

1. Meta learning is a subfield of Machine learning where automaticlearning algorithms are applied on meta-data about machine learningexperiments.

UNIT III – PART B

1.i) **Discuss the various issues that have to be addressed during data integration. [May 2012]**

- Schema Integration
- Redundancy
- Detection and Resolution of data value conflict

ii)**What is attribute –oriented induction?Describe how this is implemented.[May 2012]**

The Attribute-Oriented Induction (AOI) approach to data generalization and summarization-based characterization was first proposed in 1989, a few years prior to the introduction of the data cube approach. The data cube approach can be considered as a data warehouse-based, precomputation-oriented, materialized view approach. It performs online aggregation before an OLAP or data mining query is submitted for processing. On the other hand, the attribute-oriented induction approach, at least in its initial proposal, is a relational database query-oriented, generalization- based, on-line data analysis technique. However, there is no inherent barrier distinguishing the two approaches based on on-line aggregation versus online precomputation. Some aggregations in the data cube can be computed on-line, while online precomputation of multidimensional space can speed up attribute-oriented induction as well. In fact, data mining

systems based on attribute-oriented induction, such as DBMiner, have been optimized to include such online precomputation.

The general idea of attribute-oriented induction is to first collect the task-relevant data using a relational database query and then perform generalization based on the examination of the number of distinct values of each attribute in the relevant set of data. The generalization is performed by either attribute removal or attribute generalization (also known as concept hierarchy ascension). Aggregation is performed by merging identical, generalized tuples, and accumulating their respective counts. This reduces the size of the generalized data set. The resulting generalized relation can be mapped into di_erent forms for presentation to the user, such as charts or rules.

2. **Define data mining. Describe the steps involved in data mining when viewed as a process of knowledge discovery. Explain the architecture of the data mining system? [May 2010] [May 2013]or What is the use of data mining task?What are the basic types of data mining tasks?Explain with example.[June 2014].**

In contrast with traditional data analysis, the KDD process is interactive and iterative. One has to make several decisions in the process of KDD.

- Selection: selecting a data set, or focusing on a subset of variables, or data samples
- Preprocessing: strategies for handling missing value and noise
- Transformation: dimensionality reduction or transformation
- Data Mining: searching for patterns
- Interpretation: interpreting rules and consolidating discovered knowledge

3.**i)Explain with diagrammatic illustration the primitives for specifying a data mining task.[June 2013]** or **Explain the data mining task primitives. [Dec 2013]**

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. This includes the database attributes or data warehouse dimensions of interest (referred to as the relevant attributes or dimensions).

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. Concept hierarchies are a popular form of background knowledge, which allow data to be mined at multiple levels of abstraction.

ii)**Explain the evolution of database technology.[Dec 2014]**

Evolutionary Step	**Business Question**	**Enabling Technologies**	**Product Providers**	**Characteristics**
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels

Support (1990s)		warehouses		
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	Prospective, proactive information delivery

Steps in the Evolution of Data Mining.

4. i)**Describe the different types of data repositories on which data mining can be performed?**

As a general technology, data mining can be applied to any kind of data as long as the data are meaningful for a target application. The most basic forms of data for mining applications are database data, data warehouse data, and transactional data. The concepts and techniques presented in this book focus on such data. Data mining can also be applied to other forms of data (e.g., data streams, ordered/sequence data, graph or networked data, spatial data, text data, multimedia data, and the WWW).Relational databases,Datawarehouses,Transactionaldatabases,Object oriented databases,Spatialdatabase,Time series databases,Textdatabase,Multi mediadatabase,WWW.

ii)**Briefly explain the kinds of patterns that can be mined? What kind of data can be mined? (Or)Explain the data mining functionalities. [Dec 2012][Dec 2014]**

Data mining functionalities, and the kinds of patterns are, Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions,Presentation: decision-tree, classification rule, neural network,Prediction: Predict some unknown or missing numerical values,Outlier analysis

1. Concept/class description
2. Association analysis
3. Classication and prediction
4. Clustering analysis
5. Evolution and deviation analysis

5. **How data mining systems are classified? Discuss each classification with an example. (Or) Give the classification of data mining system. Describe the issues related to data mining. [May 2011] [Dec 2011] [Dec 2013] [May 2012][Dec 2014]**

a. Classification according to the kinds of databases mined.

A data mining system can be classied according to the kinds of databases mined. Database systems themselves can be classified according to different criteria (such as data models, or the types of data or applications involved), each of which may require its own data mining technique. Data mining systems can therefore be classified accordingly.For instance, if classifying according to data models, we may have a relational, transactional, object-oriented, object-relational, or data warehouse mining system. If classifying according to the special types of data handled, we may have a spatial, time-series, text, or multimedia data mining system, or a World-Wide Web mining system. Other system types include heterogeneous data mining systems, and legacy data mining systems.

b. Classifiion according to the kinds of knowledge mined.

Data mining systems can be categorized according to the kinds of knowledge they mine, i.e., based on data mining functionalities, such as characterization, discrimination, association, classification, clustering, trend and evolution analysis, deviation analysis, similarity analysis, etc. A comprehensive data mining system usually provides multiple and/or integrated data mining functionalities.

Moreover, data mining systems can also be distinguished based on the granularity or levels of abstraction of the knowledge mined, including generalized knowledge (at a high level of abstraction), primitive-level knowledge(at a raw data level), or knowledge at multiple levels (considering several levels of

abstraction). An advanced data mining system should facilitate the discovery of knowledge at multiple levels of abstraction.

c. Classification according to the kinds of techniques utilized.

Data mining systems can also be categorized according to the underlying data mining techniques employed. These techniques can be described according to the degree of user interaction involved (e.g., autonomous systems, interactive exploratory systems, query-driven systems), or the methods of data analysis employed (e.g., database-oriented or data warehouse-oriented techniques, machine learning, statistics, visualization, pattern recognition, neural networks, and so on). A sophisticated data mining system will often adopt multiple data mining techniques or work out aneffective, integrated technique which combines the merits of a few individual approaches.

6. **Explain in detail data cleaning and data integration process in detail. [Dec 2012][June 2014]**

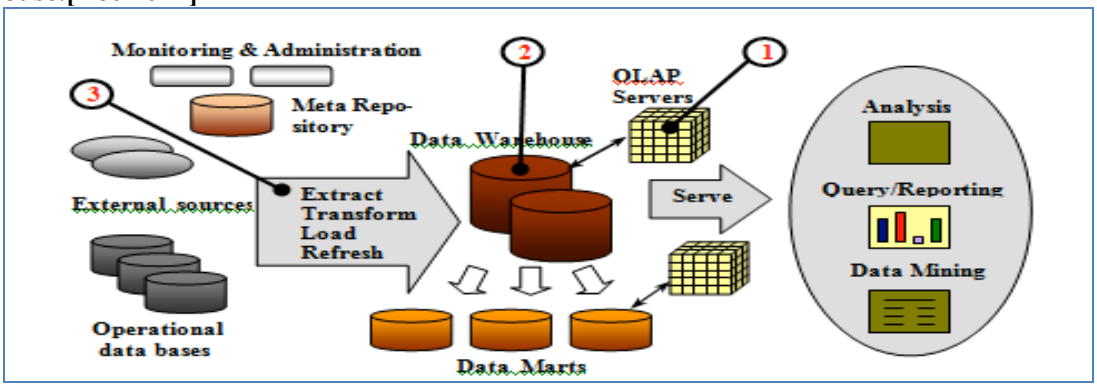
Data integration is the process of standardising the data definitions and data structures of multiple data sources by using a common schema thereby providing a unified view of the data for enterprise-level planning and decision making.

Design: The data integration initiative within a company must be an initiative of business, not IT. There should be a champion who understands the data assets of the enterprise and will be able to lead the discussion about the long-term data integration initiative in order to make it consistent, successful and benefitial.

Implementation: The larger enterprise or the enterprises which already have started other projects of data integration are in an easier position as they already have experience and can extend the existing system and exploit the existing knowledge to implement the system more effectively.Datacleansing is the process of detecting, correcting or removing incomplete, incorrect, inaccurate, irrelevant, out-of-date, corrupt, redundant, incorrectly formatted, duplicate, inconsistent, etc. records from a record set, table or database.

Validity: The degree to which the measures conform to defined business rules or constraints. Data-Type Constraints – e.g., values in a particular column must be of a particular datatype, e.g., Boolean, numeric (integer or real), date, etc. Range Constraints: typically, numbers or dates should fall within a certain range. That is, they have minimum and/or maximum permissible values. Mandatory Constraints: Certain columns cannot be empty.**Decleansing**is detecting errors and syntactically removing them for better programming.

7. **How a data mining system can be integrated with a data warehouse? Discuss with an example. [May 2011]or List and discuss the steps for integrating a data mining system with data warehouse.[Dec 2011]**



1. Integration on the front end level combining On-Line Analytical Processing (OLAP) and data mining tools into a homogeneous Graphic User Interface;
2. Integration on the database level adding of data mining components directly in DBMS;
3. Interaction on back end level – the usage of data mining techniques during the data warehouse design process.

Most data mining tools need to work o n integrated, consistent, and cleaned data, which requires costly data cleaning, data transformation, and data integration as preprocessing steps. A data warehouse constructed by such preprocessing serves as a valuable source of high quality of data for OLAP as well as for data mining. Effective data mining needs exploratory data analysis.

8.i)**List the challenges and issues in implementation of data mining systems. [Dec 2011]**

One of the key issues raised by data mining technology is not a business or technological one, but a social one. It is the issue of individual privacy. Data mining makes it possible to analyze routine business transactions and glean a significant amount of information about individuals buying habits and preferences. Another issue is that of data integrity. Clearly, data analysis can only be as good as the data that is being analyzed. A key implementation challenge is integrating conflicting or redundant data from different sources. For example, a bank may maintain credit cards accounts on several different databases. A hotly debated technical issue is whether it is better to set up a relational database structure or a multidimensional one. In a relational structure, data is stored in tables, permitting ad hoc queries. In a multidimensional structure, on the other hand, sets of cubes are arranged in arrays, with subsets created according to category. Finally, there is the issue of cost. While system hardware costs have dropped dramatically within the past five years, data mining and data warehousing tend to be self-reinforcing.

Challenges:

A) improving the scalability of data mining algorithms, B) mining non-vector data, C) mining distributed data, D) improving the ease of use of data mining systems and environments, and E) privacy and security issues for data mining.

ii)**What is the significance of interestingness measures in data mining system? Give examples.[Dec 2011]**

A pattern is interesting if,

(1) It is easily understood by humans,
(2) Valid on new or test data with some degree of certainty,
(3) Potentially useful, and
(4) Novel.

A pattern is also interesting if it validates a hypothesis that the user sought to confirm. An interesting pattern represents knowledge.

- **Objective vs. subjective interestingness measures**
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, actionability, etc.

9. **What is evolution analysis? Give example. [May 2013]**

- Evolution Analysis: Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time.
- Example: Time-series data. If the stock market data (time-series) of the last several years available from the New York Stock exchange and one would like to invest in shares of high tech industrial companies. A data mining study of stock exchange data may identify stock evolution regularities for overall stocks and for the stocks of particular companies. Such regularities may help predict future trends in stock market prices, contributing to one's decision making regarding stock investments.

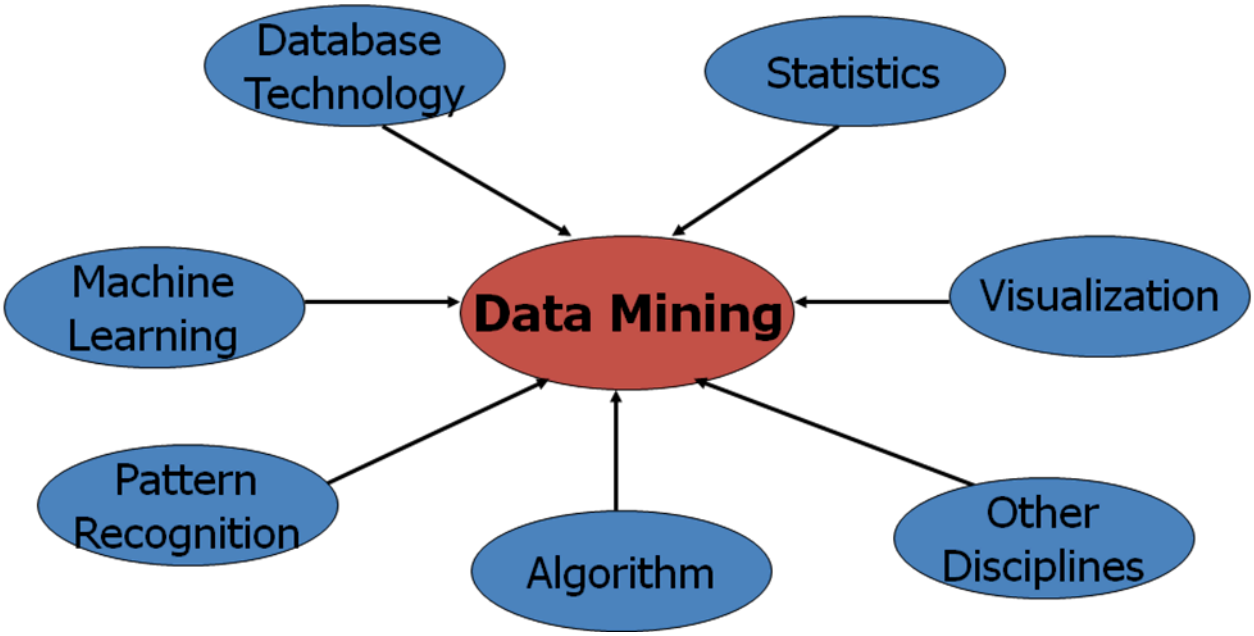
10.i) **Explain the issues in integration of data mining with data warehouse. [Dec 2009]**

Large volumes of data from multiple sources are involved; there is a high probability of errors and anomalies in the data. Real-world data tend to be incomplete, noisy and inconsistence. Data cleansing is a non-trivial task in data warehouse environments. The main focus is the identification of missing or incorrect data (noise) and conflicts between data of different sources and the correction of these problems.data analysis can only be as good as the data that is being analyzed. A key implementation challenge is integrating conflicting or redundant data from different sources. For example, a bank may maintain credit cards accounts on several different databases. The addresses (or even the names) of a single cardholder may be different in each. Software must translate data from one system to another and select the address most recently entered.

ii) Explain the various data preprocessing tasks in data mining. [Dec 2013] [May 2012]

- Data Cleaning: Missing values,Noisydata,Inconsistent data
- Data Integration
- Data transformation
- Data Reduction: Data cube aggregation,Dimensionreduction,Data compression, Numerosity reduction,
- Discretization and concept hierarchy

11**. With diagrammatic illustration discuss data mining as a confluence of multiple disciplines.[Dec 2012][June 2013]**



12.**Discuss the following schemes used for integration of a data mining system with a database or data warehouse system: Page number 35 Topic 1.8 Integration of a Data mining system with database or data warehouse system.**

i)No coupling ii)Loose Coupling iii)semitight coupling iv)Tight coupling

No coupling—flat file processing, not recommended

ii) Loose coupling

- Fetching data from DB/DW

iii)Semi-tight coupling—enhanced DM performance

- Provide efficient implement a few data mining primitives in a DB/DW system, e.g., sorting, indexing, aggregation, histogram analysis, multiway join, precomputation of some stat functions

iv)Tight coupling—A uniform information processing environment

- DM is smoothly integrated into a DB/DW system, mining query is optimized based on mining query, indexing, etc.

UNIT IV - PART A

1. Define classification. [May 2012]

Data classification is a two step process. In the first step a model is built describing a predetermined set of data classes are concepts. The model is constructed by analyzing database tuples described b attributes. Each tuple is assumed to belong to the predefined class as determined by one of the attributes called class label attribute. In the second step the modal is used for classification.

2. Define training data set.

The data tuples analyzed to build the model collectively form the training data set. Individual tuples making up the training set are referred to as training samples and a randomly selected from the sample population.

3. State the need for pruning phase in decision tree construction.[May 2013]

Pruning methods can improve the generalization performance of a decision tree, especially in noisy domains.Another key motivation of pruning is "trading accuracy for simplicity". When the goal is to produce a sufficiently accurate compact concept description, pruning is highly useful. Within this

process, the initial decision tree is seen as a completely accurate one. Thus the accuracy of a pruned decision tree indicates how close it is to the initial tree.

4. Define prediction.

Prediction can be viewed as the construction and use of a model to assess the class of an unlabeled sample or to assess the value or value ranges of an attribute that a given sample is likely to have. Classification and regression are the two major types of prediction.

5. Differentiate classification and prediction.[June 2014]

Classification is used to predict discrete or nominal values whereas prediction is used to predict continuous values. Classification is also known as supervised learning whereas prediction is also known as unsupervised learning.

6. List the applications of classification and prediction.

Applications include credit approval, medical diagnosis, performance prediction, and selective marketing.

7. How do you choose best split while constructing a decision tree?[June 2014]

The choice of best split test condition is determined by comparing the impurity of child nodes and also depends on which impurity measurement is used. After building the decision tree, a tree-prunning step can be performed to reduce the size of decision tree. Decision trees that are too large are susceptible to a phenomenon known as overfitting. Pruning helps by trimming the branches of the initail tree in a way that improves the generalization capability of the decision tree.

8.What is market basket analysis?[June 2013]

Market Basket Analysis is a modelling technique based upon the theory that if a customer buys a certain group of items, they are more (or less) likely to buy another group of items. For example, if a customer buys bread, they are more likely to buy butter.

9. Define tree pruning. [May 2013]

When decision trees are built many of the branches may reflect noise or outliers in training data. Tree pruning attempts to identify and remove such branches with the goal of improving classification accuracy on unseen data.

10. Define information gain.

The information gain measure is used to select the test attribute at each node in the tree. Such a measure is referred to as an attribute selection measure or a measure of the goodness of split. The attribute with the highest information gain is chosen as the test attribute for the current node. This attribute minimizes the information needed to classify the samples in the resulting partitions and reflects the least randomness is “impurity” in the partitions.

11. List the two common approaches for tree pruning.[Dec 2014]

Prepruning approach – a tree is “Pruned” by halting its construction early. Upon halting the node becomes a leaf. The leaf may hold the most frequent class among the subsets samples or the probability distribution of the samples.

Post pruning approach – removes branches from a “fully grown” tree. A tree node is pruned by removing its branches the lowest unpruned node becomes the leaf and is labeled by the most frequent class among its former branches.

12. List the problems in decision tree induction and how it can be prevented.

Fragmentation, repetition, and replication. Attribute construction is an approach for preventing these problems, where the limited representation of the given attributes is improved by creating new attributes based on the existing ones.

13. What are Bayesian classifiers? [May 2012]

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class. Bayesian classification is based on bayes theorem. Bayesian classifiers exhibit high accuracy and speed when applied to large databases. Bayesian classifier also known as naïve Bayesian classifiers is comparable in performance with decision tree and neural network classifiers.

14. Define Bayesian belief networks.

Bayesian belief networks are graphical models which allow the representation of dependencies among subsets of attributes. It can also be used for classification.

15. Define rule based classification.Give Example [Dec 2011]

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

Examples of classification rules:

- (Blood Type=Warm) □ (Lay Eggs=Yes) □ Birds
- (Taxable Income < 50K) □ (Refund=Yes) □ Evade=No

16. Define support vector machine. [May 2011]

A support vector machine is an algorithm for the classification of both linear and nonlinear data. It transforms the original data into a higher dimension, from where it can find a hyperplane for data separation using essential training tuples called support vectors.

17. Define backpropagation.

Backpropagation is a neural algorithm fro classification that employs a method of gradient descent. It searches for a set of weights that can model the data so as to minimize the mean-squared distance between the network's class prediction and the actual class lable of data tuples.

18. List the methods used for classification based on concepts from association rule mining.

ARCS (Association Rule Clustering System), Associative classification, CAEP (Classification by Aggregating Emerging Patterns).

19. Define single dimensional association rule. [Dec 2013]

Buys(X, "IBM desktop computer") => buys(X, "Sony b/w printer")

The above rule is said to be single dimensional rule since it contains a single distinct predicate (eg buys) with multiple occurrences (i.e., the predicate occurs more than once within the rule. It is also known as intra dimension association rule.

20. Define multi dimensional association rules.

Association rules that involve two or more dimensions or predicates can be referred to as multi dimensional associational rules.

Age(X, "20…29") ^ occupation (X, "Student") => buys (X,"Laptop")

The above rule contains three predicates (age, occupation, buys) each of which occurs only once in the rule. There are no repeated predicates in the above rule. Multi dimensional association rules with no repeated predicates are called interdimension association rules.

21. List some of the other classification methods.

Other classification methods are K – nearest neighbor classification, case based reasoning, genetic algorithms, rough set and fuzzy set approaches.

22. What is K - nearest neighbor classifiers?

Nearest Neighbor classifiers are based on learning by analogy. The training samples are described by n - dimensional numeric attributes. Each Sample represents a point in an n – dimensional space. In this way all of the training samples are stored in an n - dimensional pattern space. When given an unknown sample a K – nearest neighbor classifier searches the pattern space for the K – training samples that are closes to the unknown sample. These K training samples are the K – nearest neighbors of the unknown sample.

23. Define support in association rule mining

The rule A => B holds in the transaction set D with support s where s is the percentage of transactions in D that contain A U B i.e., both A & B. This is taken to be the probability, P (A U B).

24. What is decision tree induction?[Dec 2012]

The decision tree induction algorithm works by recursively selecting the best attribute to split the data and expanding the leaf nodes of the tree until the stopping cirterion is met. Decision Tree learning is one of the most widely used and practical methods for inductive inference over supervised data. A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.

25. List the two step process involved in Apriori algorithm.

- Join Step
- Prune Step

26. Define correlation analysis with an example. [May 2011] [Dec 2011] [May 2012]

Inconsistencies in attribute or dimension naming can cause redundancies in the resulting data set. Redundancies can be detected by correlation analysis. Given two attributes, such analysis can measure how strongly one attribute implies the other, based on available data. For nominal data chi-Square test can be used.

27. Define frequent itemset. [Dec 2013]

The number of transactions required for the item set to satisfy minimum support is therefore referred to as minimum support count. If an item set satisfies minimum support then it is a frequent item. set.

28.What is support Vector machine.[May 2011]

1. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.

29.What is naïve Bayesian classification? How is it differ from Bayesian classification?[June 2012]

Bayesian classifiers use Bayes theorem, which says

p(cj | d) = p(d | cj) p(cj)/p(d)

p(cj | d) = probability of instance d being in class cj, p(d | cj) = probability of generating instance d given class cj, p(cj) = probability of occurrence of class cj, p(d) = probability of instance d occurring.

Naïve Bayesian classifiers assume attributes have independent distributions, and thereby estimate

p(d|cj) = p(d1|cj) * p(d2|cj) * ….* p(dn|cj)

p(d|cj) =The probability of class cj generating instance d, equals…

p(d1|cj)=The probability of class cj generating the observed value for feature 1, multiplied by..

30.List the two interesting measures of an association rule.[Dec 2012]

Support,Confidence

31.Define Lazy learners. Or Differentiate lazy learners.[Dec 2014]

Lazy learning (e.g., instance-based learning): Simply stores training data (or only minor processing) and waits until it is given a test tuple

Eager learning (the above discussed methods): Given a set of training set, constructs a classification model before receiving new (e.g., test) data to classify.

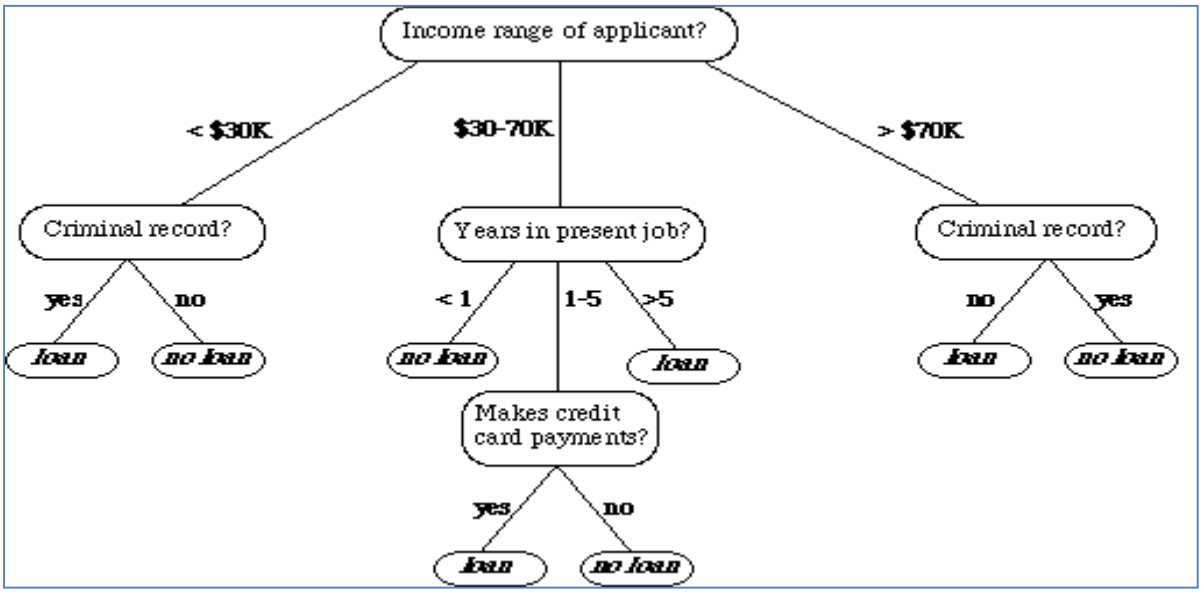
Lazy: less time in training but more time in predicting

UNIT IV - PARTB

1. Discuss about mining association rules using the apriori algorithm in detail.[Dec 2013] [May 2012][May 2011][Dec 2011][Dec 2013][Dec 2014]

- Find the frequent itemsets: the sets of items that have minimum support
  - A subset of a frequent itemset must also be a frequent itemset
- • i.e., if {AB} is a frequent itemset, both {A} and {B} should be a frequent itemset
  - Iteratively find frequent itemsets with cardinality from 1 to k (k-itemset)
- Use the frequent itemsets to generate association rules
- The Apriori Algorithm : Pseudocode
- Join Step: Ck is generated by joining Lk-1 with itself
- Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset

2. **Explain classification by decision tree induction in detail. [May 2012] or what is decision tree? Explain how classification is done using decision tree induction.[June 2012]or Develop an algorithm for classification using decision trees. Illustrate the algorithm with a relevant example.[Dec 2012]**



3. Explain Bayesian classification in detail. [May 2008][May 2012] (Or) Develop an algorithm for classification using Bayesian classification. Illustrate the algorithm with a relevant example. [May 2011] [May 2013]

Bayesian approaches are a fundamentally important DM technique. Given the probability distribution, Bayes classifier can provably achieve the optimal result. Bayesian method is based on the probability theory. Bayes Rule is applied here to calculate the posterior from the prior and the likelihood, because the later two is generally easier to be calculated from a probability model. One limitation that the Bayesian approaches can not cross is the need of the probability estimation from the training dataset. Bayesian classification provides practical learning algorithms and prior knowledge andobserved data can be combined. Bayesian Classification provides a useful perspective forunderstanding and evaluating many learning algrithms. It calculates explicit probabilities forhypothesis and it is robust to noise in input data.

the probability model for a classifier is a conditional model

$$p(C|F_1,\ldots,F_n)$$

over a dependent class variable $C$with a small number of outcomes or *classes*, conditional on several feature variables F_1through F_n. The problem is that if the number of features $n$is large or when a feature can take on a large number of values, then basing such a model on probability tables is infeasible. We therefore reformulate the model to make it more tractable.

Using Bayes' theorem, this can be written

$$p(C|F_1,\ldots,F_n) = \frac{p(C)\ p(F_1,\ldots,F_n|C)}{p(F_1,\ldots,F_n)}.$$

In plain English, using Bayesian Probability terminology, the above equation can be written as

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}.$$

4. **Explain constraint based association mining. [May 2008] [May 2012]**

A good heuristic is to have the users specify such intuition or expectations as constraints to confine the search space. The constraint can include the following:

- Key type constraints
- Data constraints
- Dimensional constraints
- Interestingness constraints
- Rule constraints

5. **What is classification? With an example explain how support vector machine can be used for classification. [Dec 2011]**

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilisticbinarylinear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.

6. With an example explain various attribution selection measures in classification. [May 2014]

Attribute selection measures

- Information Gain
- Gain Ratio

7. **Write short notes on the following a. Compare classification and prediction.b. Issues in classification and prediction. [Dec 2011]**

a.Comparison of Classification and Prediction Methods:Here is the criteria for comparing methods of Classification and Prediction:

Accuracy - Accuracy of classifier refers to ability of classifier predict the class label correctly and the accuracy of predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.

Speed - This refers to the computational cost in generating and using the classifier or predictor.

Robustness - It refers to the ability of classifier or predictor to make correct predictions from given noisy data.

Scalability - Scalability refers to ability to construct the classifier or predictor efficiently given large amount of data.

Interpretability - This refers to the to what extent the classifier or predictor understand.

b.The major issue is preparing the data for Classification and Prediction. preparing the data involves the following activities:

Data Cleaning - Data cleaning involves removing the noise and treatment of missing values. The noise is removed by applying smoothing techniques and the problem of missing values is solved by replacing a missing value with most commonly occurring value for that attribute.

Relevance Analysis - Database may also have the irrelevant attributes. Correlation analysis is used to know whether any two given attributes are related.

Data Transformation and reduction - The data can be transformed by any of the following methods.

Normalization - The data is transformed using normalization. Normalization involves scaling all values for given attribute in order to make them fall within a small specified range. Normalization is used when in the learning step, the neural networks or the methods involving measurements are used.

Generalization -The data can also be transformed by generalizing it to the higher concept. For this purpose we can use the concept hierarchies.

8. **Explain Bayesian classification and rule based classification. Give example for any one classification and explain in detail.[Dec 2014]**

Classfication rules:

"if…then…" rules

(Blood Type=Warm) ∧(Lay Eggs=Yes) → Birds

(Taxable_Income< 50K) ∧(Refund=Yes) → Evade=No

Rule: (Condition) → y

where Condition is a **conjunction of attribute tests**(A1 = v1) **and** (A2 = v2) **and** … **and** (An = vn) and y is the **class label**

– LHS: rule antecedent or condition

– RHS: rule consequent

Ex:A rule *r* **covers** an instance **x** if the attributes of the instance satisfy the condition (LHS) of the rule

R1: (Give Birth = no) □ (Can Fly = yes) → Birds

R2: (Give Birth = no) □ (Live in Water = yes) → Fishes

9.Explain about classification and prediction techniques. [Dec 2012][Dec 2011]

Major method for prediction: regression

– Many variants of regression analysis in statistics

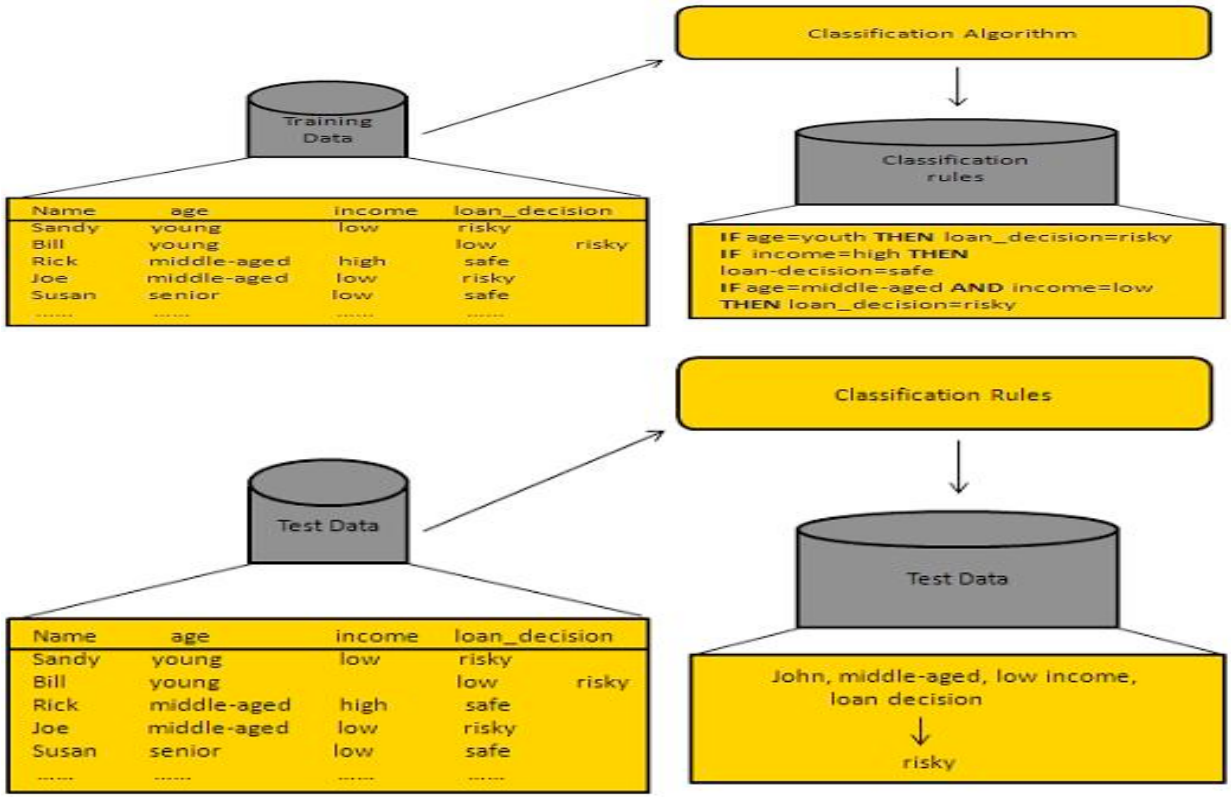
The Data Classification process includes the two steps:Building the Classifier or Model,Using Classifier for Classification

a. Building the Classifier or Model: This step is the learning step or the learning phase. In this step the classification algorithms build the classifier. The classifier is built from the training set made up of database tuples and their associated class labels. Each tuple that constitutes the training set is referred to as a category or class. These tuples can also be referred to as sample, object or data points.

b.Using Classifier for Classification

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

Figure a and b are given below



10. **Discuss the apriori algorithm for discovering frequent itemsets. Apply apriori algorithm to the following data set. Use 0.3 for the minimum support value. [May 2011] [Dec 2011] [May 2013][Dec 2012][June 2013][June 2014]**

Trans ID	Item purchased
101	strawberry, litchi, oranges
102	strawberry, butterfruit
103	butterfruit, vanilla
104	strawberry, litchi, oranges
105	banana, oranges
106	Banana
107	banana, butterfruit
108	strawberry, litchi, apple, oranges
109	apple, vanilla
110	strawberry, litchi,

The set of item is {strawberry, litchi, oranges, butterfruit, vanilla, Banana, apple}. Use 0.3 for the minimum support value.

Algorithm:

```
Apriori(T, ε)
    L1 ← {large 1 — itemsets}
    k ← 2
        while Lk−1 ≠ ∅
            Ck ← {a ∪ {b} | a ∈ Lk−1 ∧ b ∈ ⋃ Lk−1 ∧ b ∉ a}
            for transactions t ∈ T
                Ct ← {c | c ∈ Ck ∧ c ⊆ t}
                for candidates c ∈ Ct
                    count[c] ← count[c] + 1
            Lk ← {c | c ∈ Ck ∧ count[c] ≥ ε}
            k ← k + 1
        return ⋃k Lk
```

The pseudo code for the algorithm is given below for a transaction database T, and a support threshold of ϵ. Usual set theoretic notation is employed, though note that T is a multiset. $C_k$is the candidate set for level k. Generate() algorithm is assumed to generate the candidate sets from the large item sets of the preceding level, heeding the downward closure lemma. $count[c]$accesses a field of the data structure that represents candidate set C, which is initially assumed to be zero. Many details are omitted below, usually the most important part of the implementation is the data structure used for storing the candidate sets, and counting their frequencies

11.**Explain how the Bayesian belief networks are trained to perform classification.[June 2012]**

A belief network is defined by two components:

- A directed acyclic graph
- Conditional probability table

Training Bayesian belief network

12. Explain as to how neural networks are used for classification of data.[Dec 2013].

Backpropagation is a neural network learning algorithm. The neural networks field was originally kindled by psychologists and neurobiologists who sought to develop and test computational analogs of neurons. Roughly speaking, a **neural network** is a set of connected input/output units in which each connection has a weight associated with it. During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of the input tuples. Neural network learning is also referred to as **connectionist learning** due to the connections between units.

```
(1)  Initialize all weights and biases in network;
(2)  while terminating condition is not satisfied {
(3)      for each training tuple X in D {
(4)          // Propagate the inputs forward:
(5)          for each input layer unit j {
(6)              Oj = Ij; // output of an input unit is its actual input value
(7)          for each hidden or output layer unit j {
(8)              Ij = Σi wij Oi + θj; //compute the net input of unit j with respect to
                     the previous layer, i
(9)              Oj = 1/(1+e^−Ij); } // compute the output of each unit j
(10)         // Backpropagate the errors:
(11)         for each unit j in the output layer
(12)             Errj = Oj(1 − Oj)(Tj − Oj); // compute the error
(13)         for each unit j in the hidden layers, from the last to the first hidden layer
(14)             Errj = Oj(1 − Oj) Σk Errk wjk; // compute the error with respect to
                       the next higher layer, k
(15)         for each weight wij in network {
(16)             Δwij = (l)Errj Oi; // weight increment
(17)             wij = wij + Δwij; } // weight update
(18)         for each bias θj in network {
(19)             Δθj = (l)Errj; // bias increment
(20)             θj = θj + Δθj; } // bias update
```

UNIT V - PART A

1. What is cluster analysis?

Clustering analyses data objects without consulting a known class label. Class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity.

2. Define Clustering. [Dec 2011] [May 2013]

A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to objects in other clusters. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering.

3. How is the quality of a cluster represented?

Quality of a cluster mat is represented by its diameter, the maximum distance between any two objects in the cluster. Centroid distance is another alternative measure of cluster quality and is defined by the average distance of each distance object from the cluster centroid.

4. Give the categorization of major clustering methods. [May 2012] [May 2013]

The major clustering methods are partitioning methods, hierarchical methods, density based methods, grid based methods and model based methods.

5. List the commonly used partitioning methods.

K- Means and K-Medoids

6. What are the requirements of clustering in data mining?

- Scalability
- Ability to deal with different types of attributes
- Discover of clusters with arbitrary shape
- Requirements for domain knowledge to determine input parameters
- Ability to deal with noisy data
- Interpretability and usability

7. What are the characteristics of partitioning methods?

- Find mutually exclusive clusters of spherical shape
- Distance-based
- May use mean-medoid to represent cluster center
- Effective for small to medium-size data sets

8.What are the characteristics of Hierarchical methods?

- Clustering is a hierarchical decomposition
- Cannot correct erroneous merges or splits
- May incorporate other technique like micro clustering

9.What are the characteristics of Density based methods?

- Can find arbitrary shaped clusters
- Clusters are dense regions of objects in space that are separated by low-density regions
- Custer density-Each point must have a minimum number of points within its “neighborhood”.
- May filter out outliers

10. What are the characteristics of Grid based methods?

- Use a multiresolution grid data structure
- Fast processing time

11. What are the applications of cluster analysis?

Applications are Business Intelligence, Image pattern recognition, web search, biology and security. Cluster analysis can be used as a standalone data mining tool to gain insight into the data distribution, or as a preprocessing step for other data mining algorithms operating on the detected clusters.

12. What is the concept of partitioning methods?

It 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 that partitioning by moving objects from one group to another.

13. Define hierarchical method in clustering.

It creates a hierarchical decomposition of the given set of data objects. The method is classified as being either agglomerative (bottom-up) or divisive (top-down), based on how the hierarchical decomposition is firmed.

14. Define density-based method in clustering.

A density-based method clusters objects based on the notion of density. It grows clusters either according to the density of neighborhood objects or according to a density function.

15. What is an outlier? Give example[May 2011] [Dec 2011][May 2013][June 2013]

Some data objects do not comply with the general behavior or model of the data. Such data objects which are grossly different from or inconsistent with the remaining set of data are called outliers. Outliers can be caused by measurement or execution error. Outlier detection and analysis is referred to as outlier mining.

Identify the outlier of the data set. 216, 403, 203, 210, 227, 220, 190, 194

Correct answer is 403

Solution:

Step 1: An outlier is an element of a data set that distinctly stands out from the rest of the data.
Step 2: In the given data set, 403 is far apart from the remaining data values.
Step 3: So, the outlier of the data set is 403.

16. Define grid-based method in clustering.

A grid-based method first quantizes the object space into a finite number of cells that form a grid structure, and then performs clustering on the grid structure. Ex: STING

17. What are the types of clustering methods for high-dimensional data?

1. Subspace clustering methods - search for clusters in subspaces of the original space
2. Dimensionality reduction – creates a new space of lower dimensionality and search for clusters there.

18. What is STING?[June 2014]

STING (STatistical INformation Grid approach). The spatial area area is divided into rectangular cells.There are several levels of cells corresponding to different levels of resolution.Each cell at a high level is partitioned into a number of smaller cells in the next lower level.Statistical info of each cell is calculated and stored beforehand and is used to answer queries.Parameters of higher level cells can be easily calculated from parameters of lower level cell.Use a top-down approach to answer spatial data queries

19. What are the types of constraints with clustering?

1. Constraints on instances
2. Constraints on clusters
3. Constraints on similarity measurement

20. What are types of outliers?

Global outliers, contextual (or conditional) outliers and collective outliers.

21. List the methods for outlier detection.

Statistical approach, proximity-based methods, and clustering-based methods

22. What is distance based outlier?

An object o in a data set S is a distance based outlier with parameters p and d i.e., DB (p, d) if at least a fraction p of the objects in S lie at a distance greater than d from o.

23. List the algorithms for mining distance based outliers.

Index based algorithm, Nested Loop Algorithm, Cell based Algorithm.

24. Give the two techniques used for deviation based outlier detection.

Sequential exception technique, OLAP data cube technique.

25. What are the data mining applications? [May 2013] [May 2011]

- Financial data analysis
- Retail and telecommunication industries
- Science and Engineering
- Intrusion detection and prevention
- Data mining and Recommender systems
- Applications are business intelligence, web search, bioinformatics, health informatics, finance, digital libraries, and digital governments.

26.Mention the applications of outlier.[Dec 2011][Dec 2013]

Fraud Detection ,Intrusion Detection mFault/ Damage Detection ,Crime Investigation,Medical Informatics.

27.Distinguish between classification and clustering.[June 2012][Dec 2012]

Classification is supervised learning technique used to assign per-defined tag to instance on the basis of features. So classification algorithm requires training data. Classification model is created from training data, then classification model is used to classify new instances.

Clustering is unsupervised technique used to group similar instances on the basis of features. Clustering does not require training data. Clustering does not assign per-defined label to each and every group.

28.What is outlier analysis?[Dec 2012]

Data objects which are grossly different from or inconsistent with the remaining set of data are called as outliers. The outliers may be of particular interest, such as in the case of fraud detection, where outliers may indicate fraudulent activity. Thus outlier detection and analysis is an interesting data mining tasks referred to as outlier mining or outlier analysis.

29.Classify hierarchical clustering methods.[June 2013][Dec 2014]

Hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy

30.How is the goodness of clusters is measured.[Dec 2013]

Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types.

1. **External Index**: Used to measure the extent to which cluster labels match externally supplied class labels.
    a. Entropy
2. **Internal Index**: Used to measure the goodness of a clustering structure without respect to external information.
    a. Sum of Squared Error (SSE)
3. **Relative Index**: Used to compare two different clusterings or clusters.
    a.Often an external or internal index is used for this function, e.g., SSE or entropy

31.Define Wave Cluster.[June 2014]

It is a grid based multi resolution clustering method. In this method all the objects are represented by a multidimensional grid structure and a wavelet transformation is applied for finding the dense region. Each grid cell contains the information of the group of objects that map into a cell. A wavelet transformation is a process of signaling that produces the signal of various frequency sub bands.

UNIT V - PARTB

1.**Explain K-Means partitioning algorithm in detail. [May 2011] [Dec 2011] [Dec 2012] [May 2013]**

***k*-means clustering** is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *k*-means clustering aims to partition*n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

Given a set of observations ($\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$), where each observation is a *d*-dimensional real vector, *k*-means clustering aims to partition the *n* observations into *k* sets ($k \leq n$) $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS):

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

where $\boldsymbol{\mu}_i$ is the mean of points in S_i.

The most common algorithm uses an iterative refinement technique. Due to its ubiquity it is often called the ***k*-means algorithm**; it is also referred to as **Lloyd's algorithm**, particularly in the computer science community.

Assignment step: Assign each observation to the cluster whose mean yields the least within-cluster sum of squares (WCSS). Since the sum of squares is the squared Euclidean distance, this is intuitively the "nearest" mean.[8] (Mathematically, this means partitioning the observations according to the Voronoi diagram generated by the means).

$$S_i^{(t)} = \{x_p : \|x_p - m_i^{(t)}\|^2 \leq \|x_p - m_j^{(t)}\|^2 \ \forall j, 1 \leq j \leq k\},$$

where each x_p is assigned to exactly one $S^{(t)}$, even if it could be is assigned to two or more of them.

Update step: Calculate the new means to be the centroids of the observations in the new clusters.

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

Since the arithmetic mean is a least-squares estimator, this also minimizes the within-cluster sum of squares (WCSS) objective.

2. i)Explain outlier analysis in detail with an example. Discuss the use of outlier analysis.[June 2014] [Dec 2013][May 2012] [Dec 2014]

"An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism." Outliers are also referred to as abnormalities, discordants, deviants, or anomalies in the data mining and statistics literature. The recognition of such unusual characteristics provides useful application-specific insights. Some examples are as follows:**Intrusion Detection Systems**: In many host-based or networked computer systems, different kinds of data are collected about the operating system calls, network traffic, or other activity in the system**Credit Card Fraud:** Credit card fraud is quite prevalent, because of the ease with which sensitive information such as a credit card number may be compromised**Interesting Sensor Events:** The sudden changes in the underlying patterns may represent events of interest. Event detection is one of the primary motivating applications in the field of sensor networks.**Medical Diagnosis:** In many medical applications the data is collected from a variety of devices such as MRI scans, PET scans or ECG time-series. Unusual patterns in such data typically reflect disease conditions.**Law Enforcement:** Outlier detection finds numerous applications to law enforcement, especially in cases, where unusual patterns can only be discovered over time through multiple actions of an entity. **Earth Science:** A significant amount of spatiotemporal data about weather patterns, climate changes, or land cover patterns is collected through a variety of mechanisms such as satellites or remote sensing.

ii)DBSCAN[Dec 2013]

The DBSCAN algorithm

The DBSCAN algorithm can identify clusters in large spatial data sets by looking at the local density of database elements, using only one input parameter. Furthermore, the user gets a suggestion on which parameter value that would be suitable. Therefore, minimal knowledge of the domain is required. The DBSCAN can also determine what information should be classified as noise or outliers. In spite of this, its working process is quick and scales very well with the size of the database – almost linearly. By using the density distribution of nodes in the database, DBSCAN can categorize these nodes into separate clusters that define the different classes. DBSCAN can find clusters of arbitrary shape, as can be seen in figure 1 [1]. However, clusters that lie close to each other tend to belong to the same class.

Figure 1. The node distribution of three different databases, taken from SEQUOIA 2000 benchmark database.

The following section will describe further how the DBSCAN algorithm works. Its computing process is based on six rules or definitions, creating two lemmas.

Definition 1: *(The Eps-neighborhood of a point)*

$N_{Eps}(p) = \{q \in D | dist(p,q) < Eps\}$

For a point to belong to a cluster it needs to have at least one other point that lies closer to it than thedistance *Eps*.

Definition 2: *(Directly density-reachable)*

There are two kinds of points belonging to a cluster; there are border points and core points, as can be seen in *figure 2* .

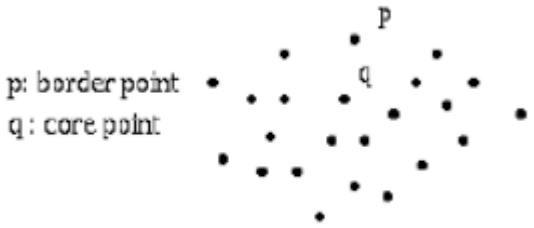


Figure 2. Border- and core points

"The *Eps-neighborhood* of a border point tends to have significantly less points than the *Epsneighborhood* of a core point."**[1]**. The border points will still be a part of the cluster and in order to include these points, they must belong to the *Eps-neighborhood* of a core point *q* as seen in *figure 3*

1) $p \in N_{Eps}(q)$

In order for point *q* to be a core point it needs to have a minimum number of points within its *Epsneighborhood*.

2) $|N_{Eps}(q)| \geq$ MinPts (core point condition)

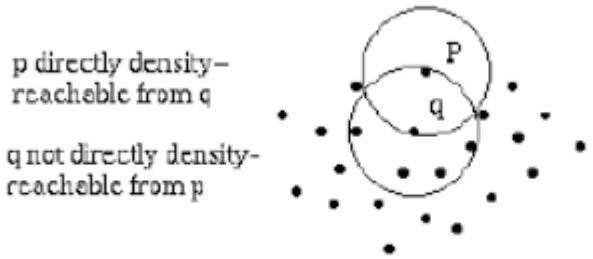


Figure 3. Point *p* is directly density-reachable from point *q* but not vice versa.

Definition 3: *(Density-reachable)*
"A point *p* is *density-reachable* from a point *q* with respect to *Eps* and *MinPts* if there is a chain of points $p_1 \ldots, p_n$, $p_1=q$, $p_n=p$ such that p_{i+1} is *directly density-reachable* from p_i." **[1]** *Figure 4* **[1]** shows an illustration of a density-reachable point.

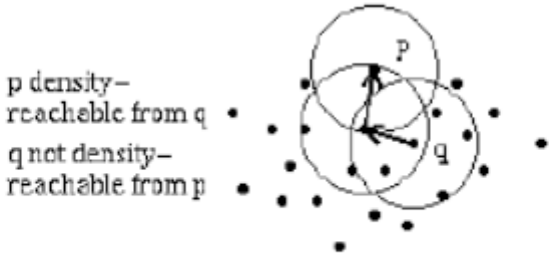


Figure 4. Point *p* is density-reachable from point *q* and not vice versa.

There are cases when two border points will belong to the same cluster but where the two border points don't share a specific core point. In these situations the points will not be *density-reachable* from each other. There must however be a core point *q* from which they are both *density-reachable*. *Figure 5* shows how density connectivity works.
"A point p is *density-connected* to a point q with respect to *Eps* and *MinPts* if there is a point *o* such that both, *p* and *q* are density-reachable from *o* with respect to *Eps* and *MinPts*."

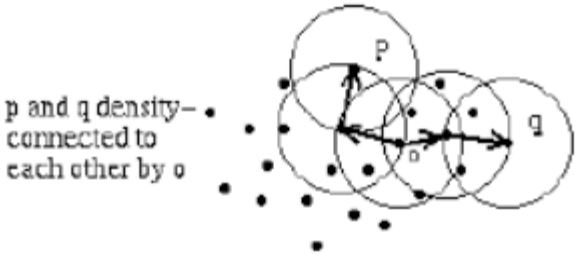


Figure 5. Density connectivity.

Definition 5: *(cluster)*
If point *p* is a part of a cluster *C* and point *q* is *density-reachable* from point *p* with respect to a given distance and a minimum number of points within that distance, then *q* is also a part of cluster *C*.
1) "∀ *p*, *q*: if $p \in C$ and *q* is *density-reachable* from *p* with respect to *Eps* and *MinPts*, then $q \in C$.
Two points belongs to the same cluster *C*, is the same as saying that *p* is density-connected to *q* with respect to the given distance and the number of points within that given distance.
2) "∀ *p*, $q \in C$: *p* is *density-connected* to *q* with respect to *Eps* and *MinPts*.

Definition 6: *(noise)*
Noise is the set of points, in the database, that don't belong to any of the clusters.

Lemma 1:
A cluster can be formed from any of its core points and will always have the same shape.

Lemma 2:
Let p be a core point in cluster C with a given minimum distance (Eps) and a minimum number of points within that distance ($MinPts$). If the set O is *density-reachable* from p with respect to the same Eps and $MinPts$, then C is equal to the set O.
"To find a cluster, DBSCAN starts with an arbitrary point p and retrieves all points density-reachable from p with respect to Eps and $MinPts$. If p is a core point, this procedure yields a cluster with respect to Eps and $MinPts$ (see Lemma 2). If p is a border point then no points are density-reachable from p and DBSCAN visits the next point of the database."

3. Applications

An example of software program that has the DBSCAN algorithm implemented is WEKA. The following of this section gives some examples of practical application of the DBSCAN algorithm.

Satellites images

A lot of data is received from satellites all around the world and this data have to be translated into comprehensible information, for instance, classifying areas of the satellite-taken images according to forest, water and mountains. Before the DBSCAN algorithm can classify these three elements in the database, some work have to be done with image processing. Once the image processing is done, the data appears as spatial data where the DBSCAN can classify the clusters as desired.

X-ray crystallography

X-ray crystallography is another practical application that locates all atoms within a crystal, which results in a large amount of data. The DBSCAN algorithm can be used to find and classify the atoms in the data.

Anomaly Detection in Temperature Data

This kind of application focuses on pattern anomalies in data, which is important in several cases, e.g.credit fraud, health condition etc. This application measures anomalies in temperatures , which is relevant due to the environmental changes (global warming). It can also discover equipment errors and so forth. These unusual patterns need to be detected and examined to get control over the situation. The DBSCAN algorithm has the capability to discover such patterns in the data.

3. Explain model based clustering.

Observe characteristics of some objects{**x**1, …,**x**N}N objects
•An object belongs to one of M clusters, but you don't know which{z1, …, zN}cluster memberships, numbers from 1..M
•Some clusters are more likely than othersP(zk=m) = πm(πm= frequency cluster m occurs)
•Within a cluster, objects' characteristics are generated by the same distribution, which has free parametersP(**x**k|zk=m) = f(**x**k,**λ**m)(**λ**m= parameters of cluster m),f doesn't have to be Gaussian
Now you have a model connecting the observations to the cluster memberships and parametersP(**x**k) = Σm=1..MP(**x**k|zk=m) P(zk=m)
= Σm=1..Mf(**x**k,**λ**m)πm, P(**x**1… **x**N) = Πk=1..NP(**x**k)(assuming **x**'sare independent)
1.Find the values of the parameters by maximizing the likelihood (usually the log of the likelihood) of the observationsmax log P(**x**1… **x**N) over **λ**1… **λ**Mand π1… πM

4. **Explain with an example density based clustering methods.[Dec 2014]**

Is a set of density-connected objects that is maximal with respect to density-reachability. Every object not contained in any cluster is considered to be noise. That is, for each data point within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of points. Such an algorithm can be used to filter out noise (outliers) and discover clusters of arbitrary shape. Clustering based on density (local cluster criterion), such as density-connected points or based on an explicitly constructed density function
Major features:Discover clusters of arbitrary shape,Handlenoise,Onescan,Need density parameters

Local outliers: Outliers comparing to their local neighborhoods, instead of the global data distribution

•In Fig., o1 and o2 are local outliers to C1, o3 is a global outlier, but o4 is not an outlier. However, proximity-based clustering cannot find o1 and o2 are outlier (e.g., comparing with O4).

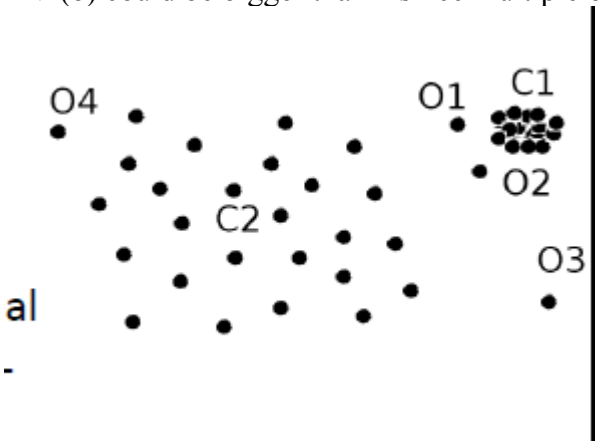
nIntuition (density-based outlier detection): The density around an outlier object is significantly different from the density around its neighbors

nMethod: Use the relative density of an object against its neighbors as the indicator of the degree of the object being outliers

n*k-distance* of an object o, distk(o): distance between o and its k-th NN

n*k-distance neighborhood* of o, Nk(o) = {o'| o' in D, dist(o, o') ≤ distk(o)}

nNk(o) could be bigger than k since multiple objects may have identical distance to o

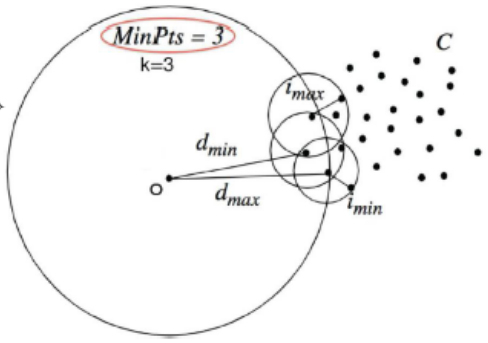


Local Outlier Factor: LOF

- Reachability distance from *o'* to *o*:

$$reachdist_k(o \leftarrow o') = \max\{dist_k(o), dist(o, o')\}$$

  - where k is a user-specified parameter



- Local reachability density of *o*:

$$lrd_k(o) = \frac{\|N_k(o)\|}{\sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o)}$$

- LOF (Local outlier factor) of an object o is the average of the ratio of local reachability of *o* and those of *o*'s k-nearest neighbors

$$LOF_k(o) = \frac{\sum_{o' \in N_k(o)} \frac{lrd_k(o')}{lrd_k(o)}}{\|N_k(o)\|} = \sum_{o' \in N_k(o)} lrd_k(o') \cdot \sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o)$$

- The lower the local reachability density of o, and the higher the local reachability density of the kNN of o, the higher LOF
- This captures a local outlier whose local density is relatively low comparing to the local densities of its kNN

LOF(Local Outlier Factor) Example

•Consider the following 4 data points: a(0, 0), b(0, 1), c(1, 1), d(3, 0) Calculate the LOF for each point and show the top 1 outlier, set k = 2 and use Manhattan Distance.

Step 1: calculate all the distances between each two data points

•There are 4 data points:

a(0, 0), b(0, 1), c(1, 1), d(3, 0)

(Manhattan Distance here)

dist(a, b) = 1

dist(a, c) = 2

dist(a, d) = 3

dist(b, c) = 1

dist(b, d) = 3+1=4

dist(c, d) = 2+1=3

Step 2: calculate all the dist2(o)
•distk(o): distance between o and its k-th NN(k-th nearest neighbor)

dist2(a) = **dist(a, c) = 2 (c is the 2nd nearest neighbor)**
dist2(b) = **dist(b, a) = 1 (a/c is the 2nd nearest neighbor)**
dist2(c) = **dist(c, a) = 2 (a is the 2nd nearest neighbor)**
dist2(d) = **dist(d, a) = 3 (a/c is the 2nd nearest neighbor)**

Step 3: calculate all the Nk(o)
n*k-distance neighborhood* of o, **Nk(o) = {o'| o' in D, dist(o, o') ≤ distk(o)}**

N2(a) = {b, c}
N2(b) = {a, c}
N2(c) = {b, a}
N2(d) = {a, c}

Step 4: calculate all the $lrd_k(o)$

$$reachdist_k(o \leftarrow o') = \max\{dist_k(o), dist(o, o')\}$$

$$reachdist_2(b \leftarrow a) = max\{dist_2(b), dist(b, a)\}$$

$$= \max\{1, 1\} = 1$$

$$reachdist_2(c \leftarrow a) = max\{dist_2(c), dist(c, a)\}$$

$$= \max\{2, 2\} = 2$$

Thus, $lrd_2(a)$

$$= \frac{|| N_2(a) ||}{reachdist_2(b \leftarrow a) + reachdist_2(c \leftarrow a)} = 2/(1+2) = 0.667$$

Step 4: calculate all the $lrd_k(o)$

- $lrd_k(o)$: Local Reachability Density of o

$$lrd_k(o) = \frac{\|N_k(o)\|}{\sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o)}$$

$$reachdist_k(o \leftarrow o') = \max\{dist_k(o), dist(o, o')\}$$

|| $N_k(o)$ || means the number of objects in $N_k(o)$, For example: || $N_2(a)$ || = || {b, c} || = 2

$$lrd_k(a) = \frac{\| N_2(a) \|}{reachdist_2(b \leftarrow a) + reachdist_2(c \leftarrow a)}$$

Step 4: calculate all the $lrd_k(o)$

Similarly,

$$lrd_2(b) = \frac{\| N_2(b) \|}{reachdist_2(a \leftarrow b) + reachdist_2(c \leftarrow b)} = 2/(2+2) = 0.5$$

$$lrd_2(c) = \frac{\| N_2(c) \|}{reachdist_2(b \leftarrow c) + reachdist_2(a \leftarrow c)} = 2/(1+2) = 0.667$$

$$lrd_2(d) = \frac{\| N_2(b) \|}{reachdist_2(a \leftarrow d) + reachdist_2(c \leftarrow d)} = 2/(3+3) = 0.33$$

Step 5: calculate all the $LOF_k(o)$

$$LOF_k(o) = \frac{\sum_{o' \in N_k(o)} \frac{lrd_k(o')}{lrd_k(o)}}{\|N_k(o)\|} = \sum_{o' \in N_k(o)} lrd_k(o') \cdot \sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o)$$

$LOF_2(a)$ =

$(lrd_2(b) + lrd_2(c)) * (reachdist_2(b \leftarrow a) + reachdist_2(c \leftarrow a))$

= (0.5+0.667) * (1+2) = 3.501

$LOF_2(b)$ =

$(lrd_2(a) + lrd_2(c)) * (reachdist_2(a \leftarrow b) + reachdist_2(c \leftarrow b))$

= (0.667+0.667) * (2+2) = 5.336

Step 6: Sort all the $LOF_k(o)$

The sorted order is:

$LOF_2(\mathbf{d})$ = 8.004

$LOF_2(\mathbf{b})$ = 5.336

$LOF_2(\mathbf{a})$ = 3.501

$LOF_2(\mathbf{c})$ = 3.501

Obviously, top 1 outlier is point d.

5. Explain about data mining applications.or. Expalin how data mining is used for retail industry.[June 2014]

Here is the list of areas where data mining is widely used:Financial Data Analysis,RetailIndustry,TelecommunicationIndustry,Biological DataAnalysis,OtherScientific Applications,Intrusion Detection. Data Mining in Telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Here is the list examples for which data mining improve telecommunication services:Multidimensional Analysis of Telecommunication data,Fraudulent pattern analysis,Identification of unusual patterns,Multidimensional association and sequential patterns analysis,Mobile Telecommunication services,Use of visualization tools in telecommunication data analysis.

6. With relevant example discuss constraint based cluster analysis. [May 2011]

In computer science, **constrained clustering** is a class of semi-supervised learning algorithms. Typically, constrained clustering incorporates either a set of must-link constraints, cannot-link constraints, or both, with a Data clustering algorithm. Both a must-link and a cannot-link constraint define a relationship between two data instances. A must-link constraint is used to specify that the two instances in the must-link relation should be associated with the same cluster. A cannot-link constraint is used to specify that the two instances in the cannot-link relation should *not* be associated with the same cluster. These sets of constraints acts as a guide for which a constrained clustering algorithm will attempt to find clusters in a data set which satisfy the specified must-link and cannot-link constraints. Some constrained clustering algorithms will abort if no such clustering exists which satisfies the specified constraints. Others will try to minimize the amount of constraint violation should it be impossible to find a clustering which satisfies the constraints.Examples of constrained clustering algorithms include:COP K-means, PCKmeans,CMWK-Means

7. Describe cluster analysis in detail.

Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters.Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning,pattern recognition, image analysis, information retrieval, and bioinformatics.

Use of clustering: Business, Biology,Statistics,Data Mining

8. What is grid based clustering? With an example explain an algorithm for grid based clustering. [Dec 2011]

1. Define a set of grid-cells
2. Assign ojects to the appropriate grid cell and compute the density of each cell.
3. Eliminate cells, whose density is below a certain threshold t.
4. Form clusters from contiguous (adjacent) groups of dense cells (usually minimizing a given objective function)

Advantages:fast:No distance computations, Clustering is performed on summaries and not individual objects; complexity is usually O(#-populated-grid-cells) and not O(#objects), Easy to determine which clusters are neighboring

- Shapes are limited to union of grid-cells
- Using multi-resolution grid data structure
- Clustering complexity depends on the number of populated grid cells and not on the number of objects in the dataset
    - Several interesting methods (in addition to the basic grid-based algorithm):STINGand CLIQUE

9. Explain hierarchical clustering techniques stating their pros and cons. [Dec 2013]or what is hierarchical clustering? With an example discuss dendegram representation for hierarchical clustering of data objects.[Dec 2012][Dec 2014]

In data mining, **hierarchical clustering** is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types: **Agglomerative**: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.**Divisive**: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.In order to decide which clusters should be combined (for agglomerative), or where a cluster should be split (for divisive), a measure of dissimilarity between sets of observations is required. In most methods of hierarchical clustering, this is achieved by use of an appropriate metric (a measure of distance between pairs of observations), and a linkage criterion which specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets.

A hierarchical clustering of distances in kilometers between some Italian cities. The method used is single-linkage.

Input distance matrix (L = 0 for all the clusters):

	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

The nearest pair of cities is MI and TO, at distance 138. These are merged into a single cluster called "MI/TO". The level of the new cluster is L(MI/TO) = 138 and the new sequence number is m = 1.
Then we compute the distance from this new compound object to all other objects. In single link clustering the rule is that the distance from the compound object to another object is equal to the shortest distance from any member of the cluster to the outside object. So the distance from "MI/TO" to RM is chosen to be 564, which is the distance from MI to RM, and so on.
After merging MI with TO we obtain the following matrix:

	BA	FI	MI/TO	NA	RM
BA	0	662	877	255	412
FI	662	0	295	468	268
MI/TO	877	295	0	754	564
NA	255	468	754	0	219
RM	412	268	564	219	0

min d(i,j) = d(NA,RM) = 219 => merge NA and RM into a new cluster called NA/RM
L(NA/RM) = 219
m = 2

	BA	FI	MI/TO	NA/RM
BA	0	662	877	255
FI	662	0	295	268
MI/TO	877	295	0	564
NA/RM	255	268	564	0

min d(i,j) = d(BA,NA/RM) = 255 => merge BA and NA/RM into a new cluster called

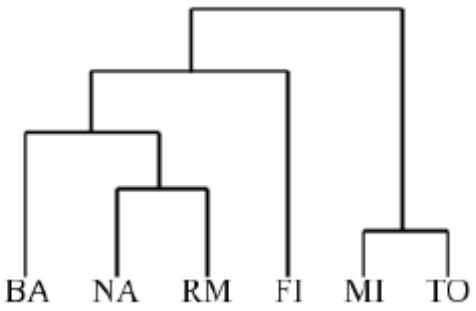
BA/NA/RM
L(BA/NA/RM) = 255
m = 3

	BA/NA/RM	FI	MI/TO
BA/NA/RM	0	268	564
FI	268	0	295
MI/TO	564	295	0

min d(i,j) = d(BA/NA/RM,FI) = 268 => merge BA/NA/RM and FI into a new cluster called BA/FI/NA/RM
L(BA/FI/NA/RM) = 268
m = 4

	BA/FI/NA/RM	MI/TO
BA/FI/NA/RM	0	295
MI/TO	295	0

Finally, we merge the last two clusters at level 295.
The process is summarized by the following hierarchical tree:



10. Explain high dimensional data clustering.

Clustering high-dimensional data is the cluster analysis of data with anywhere from a few dozen to many thousands of dimensions. Such high-dimensional data spaces are often encountered in areas such as medicine, where DNA microarray technology can produce a large number of measurements at once, and the clustering of text documents, where, if a word-frequency vector is used, the number of dimensions equals the size of the vocabulary.Multiple dimensions are hard to think in, impossible to visualize, and, due to the exponential growth of the number of possible values with each dimension, complete enumeration of all subspaces becomes intractable with increasing dimensionality. This problem is known as the curse of dimensionality.A cluster is intended to group objects that are related, based on observations of their attribute's values.

11. What is K-Means algorithm? Cluster the following eight points (with (x, y) representing locations) into three clusters A1(2, 10) A2(2, 5) A3(8, 4) A4(5, 8) A5(7, 5) A6(6, 4) A7(1, 2) A8(4, 9). Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2). The distance function between two points a=(x1,

y1) and b=(x2, y2) is defined as: $\rho(a, b) = |x2 - x1| + |y2 - y1|$. Use k-means algorithm to find the three cluster centers after the first iteration and the final three clusters.[June 2012]

12.i)Explain the different types of data used in cluster analysis.[June 2014][Dec 2014]

- Interval Scaled Variables
- Binary Variables
- Ratio-Scaled Variable
- Variables of mixed types

ii)write the difference between CLARA and CLARANS

- *CLARA* (Kaufmann and Rousseeuw in 1990)
  - Built in statistical analysis packages, such as S+
- It draws *multiple samples* of the data set, applies *PAM* on each sample, and gives the best clustering as the output
- Strength: deals with larger data sets than *PAM*
- Weakness:
  - Efficiency depends on the sample size
  - A good clustering based on samples will not necessarily represent a good clustering of the whole data set if the sample is biased

- *CLARANS* (A Clustering Algorithm based on Randomized Search) (Ng and Han'94)
  - CLARANS draws sample of neighbors dynamically
  - The clustering process can be presented as searching a graph where every node is a potential solution, that is, a set of *k* medoids
  - If the local optimum is found, *CLARANS* starts with new randomly selected node in search for a new local optimum
  - It is more efficient and scalable than both *PAM* and *CLARA*