IMPORTANT QUESTIONS AND ANSWERS

Department of IT

SUBJECT CODE: **IT6702**

SUBJECT NAME: **DATA WAREHOUSING AND DATA MINING**

Regulation: 2013 **Semester and Year:IV/VII**

Verified by CLI, DLI and approved by the centralized monitoring team on

_____

IT6702 DATA WAREHOUSING AND DATA MINING L T P C
3 0 0 3

OBJECTIVES:

The student should be made to:

- Be familiar with the concepts of data warehouse and data mining,
- Be acquainted with the tools and techniques used for Knowledge Discovery in Databases.

UNIT I DATA WAREHOUSING 9

Data warehousing Components –Building a Data warehouse –- Mapping the Data Warehouse to a Multiprocessor Architecture – DBMS Schemas for Decision Support – Data Extraction, Cleanup, and Transformation Tools –Metadata.

UNIT II BUSINESS ANALYSIS 9

Reporting and Query tools and Applications – Tool Categories – The Need for Applications – Cognos Impromptu – Online Analytical Processing (OLAP) – Need – Multidimensional Data Model – OLAP Guidelines – Multidimensional versus Multirelational OLAP – Categories of Tools – OLAP Tools and the Internet.

UNIT III DATA MINING 9

Introduction – Data – Types of Data – Data Mining Functionalities – Interestingness of Patterns – Classification of Data Mining Systems – Data Mining Task Primitives – Integration of a Data Mining System with a Data Warehouse – Issues –Data Preprocessing.

UNIT IV ASSOCIATION RULE MINING AND CLASSIFICATION 9

Mining Frequent Patterns, Associations and Correlations – Mining Methods – Mining various Kinds of Association Rules – Correlation Analysis – Constraint Based Association Mining – Classification and Prediction - Basic Concepts - Decision Tree Induction - Bayesian Classification – Rule Based Classification – Classification by Back propagation – Support Vector Machines – Associative Classification – Lazy Learners – Other Classification Methods – Prediction.

UNIT V CLUSTERING AND TRENDS IN DATA MINING 9

Cluster Analysis - Types of Data – Categorization of Major Clustering Methods – K-means– Partitioning Methods – Hierarchical Methods - Density-Based Methods –Grid Based Methods – Model-Based Clustering Methods – Clustering High Dimensional Data - Constraint – Based Cluster Analysis – Outlier Analysis – Data Mining Applications.

OUTCOMES:

After completing this course, the student will be able to:

- Apply data mining techniques and methods to large data sets.
- Use data mining tools.
- Compare and contrast the various classifiers.

TOTAL: 45 PERIODS

TEXT BOOKS:

1. Alex Berson and Stephen J.Smith, “Data Warehousing, Data Mining and OLAP”, Tata McGraw –

Hill Edition, Thirteenth Reprint 2008.
2. Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Third Edition, Elsevier, 2012.

REFERENCES:

1. Pang-Ning Tan, Michael Steinbach and Vipin Kumar, "Introduction to Data Mining", Person Education, 2007.
2. K.P. Soman, Shyam Diwakar and V. Aja, "Insight into Data Mining Theory and Practice", Eastern Economy Edition, Prentice Hall of India, 2006.
3. G. K. Gupta, "Introduction to Data Mining with Case Studies", Eastern Economy Edition, Prentice Hall of India, 2006.
4. Daniel T.Larose, "Data Mining Methods and Models", Wiley-Interscience, 2006.

TABLE OF CONTENTS

Aim and Objective of the Subject

AIM

- To learn the architecture and functionality of data warehouse
- To study various data mining algorithms
- To understand the application of data mining in different fields.

OBJECTIVES:

The student should be made to:

- Be familiar with the concepts of data warehouse and data mining,
- Be acquainted with the tools and techniques used for Knowledge Discovery in
- Databases

OUTCOMES

Upon completion of the course, the student should be able to:

- Apply data mining techniques and methods to large data sets.
- Use data mining tools.
- Compare and contrast the various classifiers.

Department of Information Technology

DETAILED LESSON PLAN

Name of the Subject& Code: IT6702-DATA WAREHOUSING AND DATA MINING

Name of the Faculty: Dr.A.Anitha

TEXT BOOKS:

T1. Alex Berson and Stephen J.Smith, "Data Warehousing, Data Mining and OLAP", Tata McGraw –Hill Edition, Thirteenth Reprint 2008.

T2. Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Third Edition,Elsevier, 2012.

REFERENCES:

1. Pang-Ning Tan, Michael Steinbach and Vipin Kumar, "Introduction to Data Mining",Person Education, 2007.
2. K.P. Soman, Shyam Diwakar and V. Aja, "Insight into Data Mining Theory and Practice", EasternEconomy Edition, Prentice Hall of India, 2006.
3. G. K. Gupta, "Introduction to Data Mining with Case Studies", Eastern Economy Edition, Prentice Hall of India, 2006.
4. Daniel T.Larose, "Data Mining Methods and Models", Wiley-Interscience, 2006.

Instruction Schedule

S. No	Week No	Topics	No of Hrs	Book No	Page No.
UNIT – I DATA WAREHOUSING			**(8)**		
1	I	Data warehousing Components	2	T1	113-127
2		Building a Data warehouse	1	T1	129-149
3		Mapping the Data Warehouse to a Multiprocessor Architecture	1	T1	151-167
4	II	DBMS Schemas for Decision Support	1	T1	169-185
5		Data Extraction, Cleanup, and Transformation Tools	2	T1	187-203
6		Metadata	1	T1	205-219
Remarks:					

S. No	Week No	Topics	No of Hrs	Book No	Page No.
UNIT II - BUSINESS ANALYSIS (9)					
7	III	Reporting and Query tools and Applications	1	T1	223-224
8		Tool Categories	1	T1	224-225
9		The Need for Applications	1	T1	226-227
10		Cognos Impromptu	1	T1	228-232
11		Online Analytical Processing (OLAP) – Need	1	T1	247
12	IV	Multidimensional Data Model	1	T1	248-250
13		OLAP Guidelines	1	T1	250-251
14		Multidimensional versus Multirelational OLAP- Categories of Tools	1	T1	251-256
15		OLAP Tools and the Internet	1	T1	262-265
Remarks:					
UNIT III- DATA MINING (9)					
16	V	Introduction	1	T2	1-9
17		Data – Types of Data	1	T2	9-21
18		Data Mining Functionalities	1	T2	21-27
19		Interestingness of Patterns	1	T2	27-29
20	VI	Classification of Data Mining Systems	1	T2	29-31
21		Data Mining Task Primitives	1	T2	31-34
22		Integration of a Data Mining System with a Data Warehouse	1	T2	34-36
23		Issues	1	T2	36-39
24		Data Preprocessing	1	T2	47-97
Remarks:					

S. No	Week No	Topics	No of Hrs	Book No	Page No.
UNIT IV ASSOCIATION RULE MINING AND CLASSIFICATION (11)					
25	VII	Mining Frequent Patterns, Associations and Correlations – Mining Methods	1	T2	227-250
26		Mining various Kinds of Association Rules – Correlation Analysis	1	T2	250-259
27		Constraint Based Association Mining	1	T2	259-265
28		Classification and Prediction - Basic Concepts	1	T2	265-271
29		Decision Tree Induction	1	T2	291-310
30		Bayesian Classification	1	T2	310-318
31	VIII	Rule Based Classification	1	T2	318-327
32		Classification by Back propagation	1	T2	327-337
33		Support Vector Machines	1	T2	337-344
34		Associative Classification – Lazy Learners	1	T2	344-351
35		Other Classification Methods – Prediction	1	T2	351-359
Remarks:					
UNIT V – CLUSTERING AND TRENDS IN DATA MINING (11)					
36	IX	Cluster Analysis	1	T2	383-386
37		Types of Data	1	T2	386-398
38		Categorization of Major Clustering Methods , k means partitioning	1	T2	398-408
39		Hierarchical Methods	1	T2	408-418
40		Density-Based Methods	1	T2	418-424
41	X	Grid Based Methods	1	T2	424-429
42		Model-Based Clustering Methods	1	T2	429-434
43		Clustering High Dimensional Data	1	T2	434-444

S. No	Week No	Topics	No of Hrs	Book No	Page No.
44		Constraint – Based Cluster	1	T2	444-451
45		Outlier Analysis	1	T2	451-460
46		Data Mining Applications	1	T2	649-684
Remarks:					

Total Hours: 48

DEPARTMENT OF INFORMATION TECHNOLOGY

MINIMUM STUDY MATERIAL

IT6702 – DATA WAREHOUSING AND DATA MINING

UNIT I

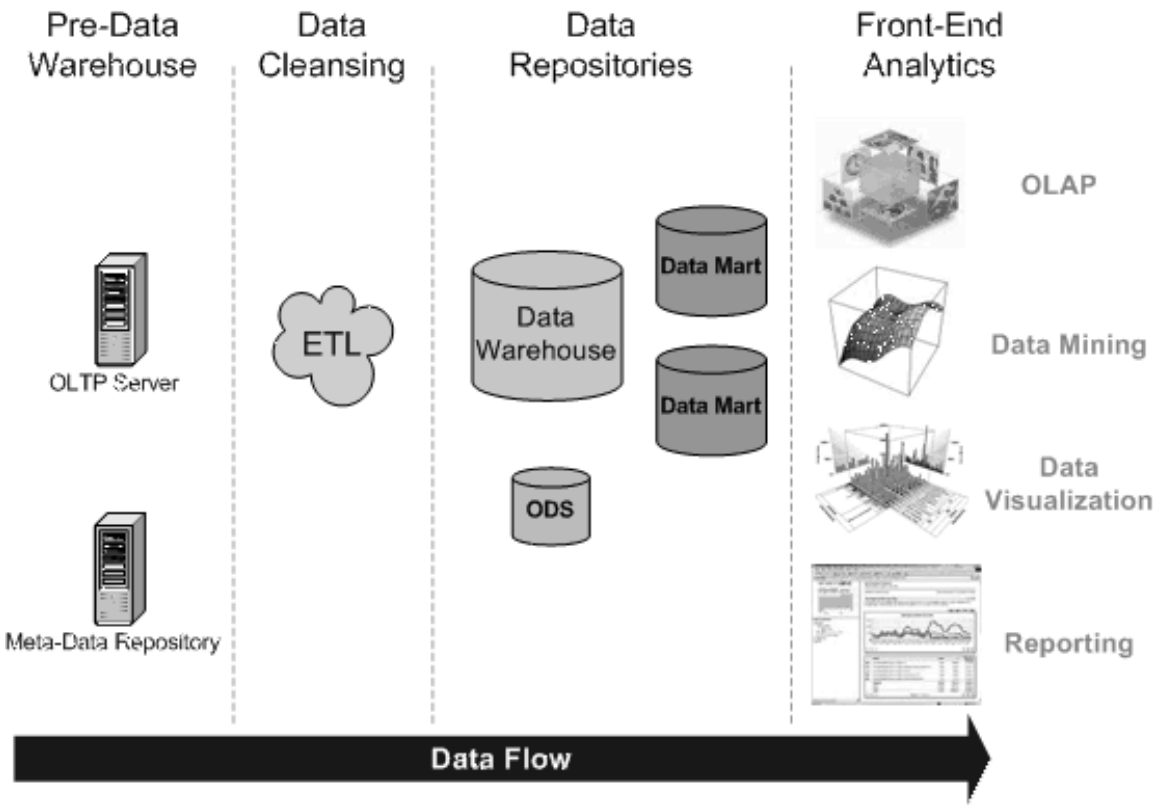
DATAWAREHOUSING

Data warehousing Components–Building a Datawarehouse–- Mapping the Data Warehouse to a Multiprocessor Architecture– for Decision Support–DataExtraction,Cleanup, and transformation Tools–Metadata

PART A

1) List the characteristics of Data warehouse

- Subject Oriented
- Integrated
- Nonvolatile
- Time Variant
- Some data is de-normalized for simplification and to improve performance
- Large amounts of historical data are used
- Queries often retrieve large amounts of data
- Both planned and ad hoc queries are common
- The data load is controlled



2) State why data partitioning is key requirement for effective parallel execution of DB operations NOV'15

Data partitioning is a key requirement for effective parallel execution of data base operations. It spreads data from data base tables across multiple disks so that I/O operations such as read and write can be performed in parallel

- Random partitioning
- Intelligent partitioning

3) What is metadata? May'15, Dec'14, Dec'13, May'11

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

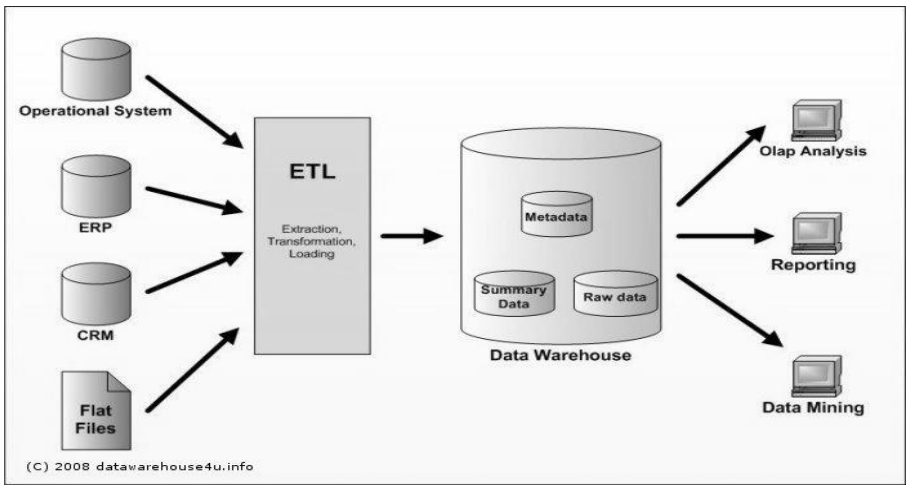
For example, the index of a book serves as a metadata for the contents in the book. Metadata is the summarized data that leads us to the detailed data.

In terms of data warehouse, we can define metadata as following:

- Metadata is a roadmap to data warehouse.
- Metadata in data warehouse defines the warehouse objects.

- Metadata acts as a directory. This directory helps the decision support system to locate the contents of a data warehouse.

4) **What is Data warehouse? May'15**



A data warehouse is a relational database that is designed for query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data, but it can include data from other sources.

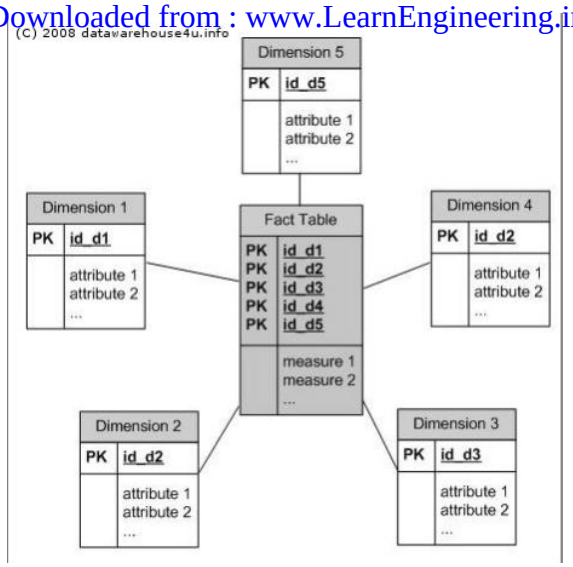
In addition to a relational database, a data warehouse environment includes an extraction, transportation, transformation, and loading (ETL) solution, an online analytical processing (OLAP) engine, client analysis tools, and other applications that manage the process of gathering data and delivering it to business users.

5) What is Star schema? Dec'14

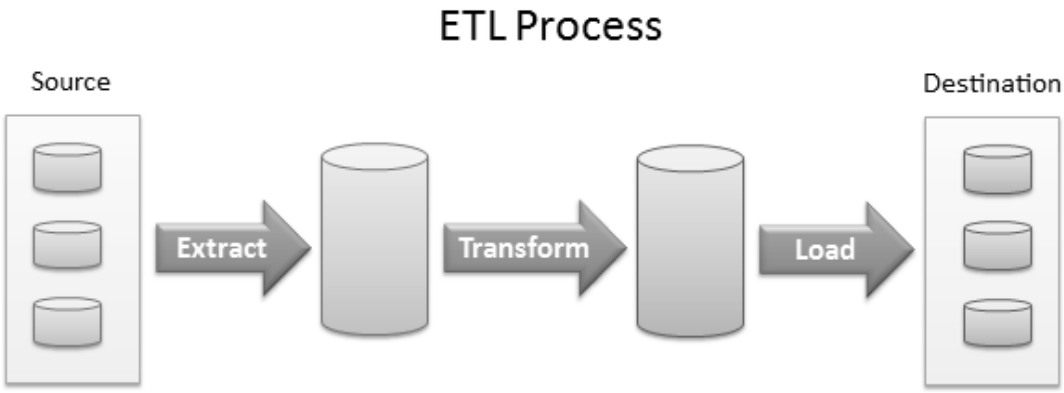
The star schema architecture is the simplest data warehouse schema. It is called a star schema because the diagram resembles a star, with points radiating from a center.

The center of the star consists of fact table and the points of the star are the dimension tables. Usually the fact tables in a star schema are in third normal form(3NF) whereas dimensional tables are de-normalized.

Despite the fact that the star schema is the simplest architecture, it is most commonly used nowadays and is recommended by Oracle.

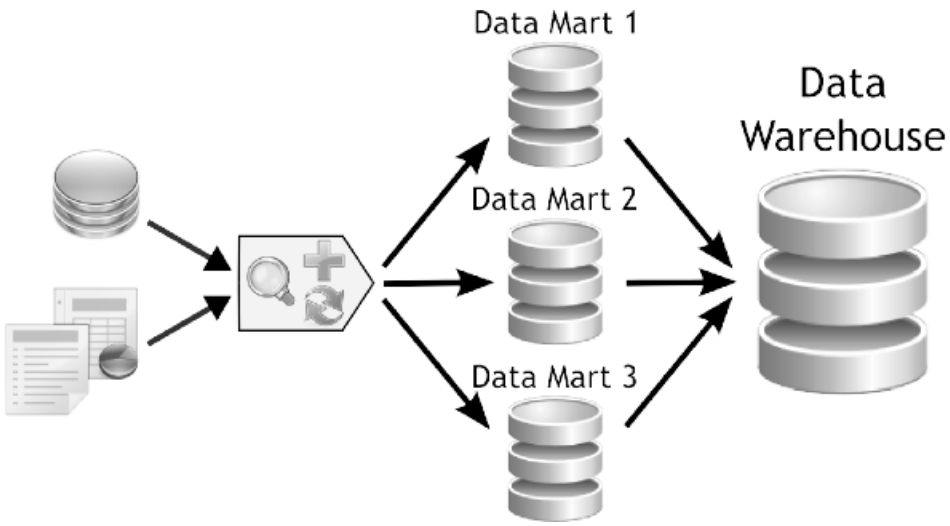


6) What is ETL process? Give its significance. Dec'13



The process of extracting data from source systems and bringing it into the data warehouse is commonly called **ETL**, which stands for extraction, transformation, and loading

7) What is Data Mart? June'13, Dec'11



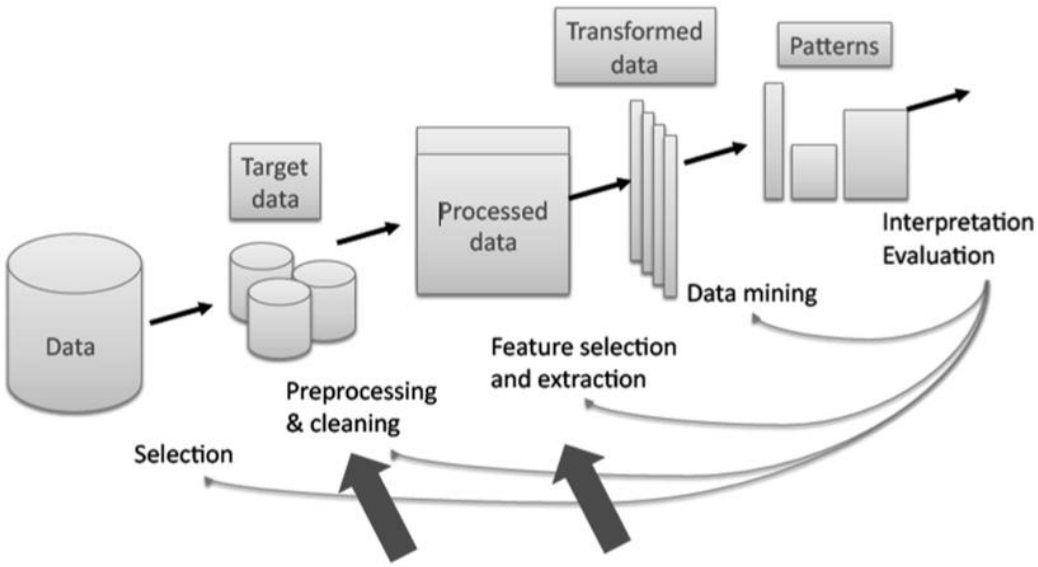
A **data mart** is the access layer of the data warehouse environment that is used to get data out to the users. The data mart is a subset of the data warehouse that is usually oriented to a specific business line or team. Data marts are small slices of the data warehouse.

8) What are the issues in data integration?

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

9) Define Data transformation. May'11,May '17

In this preprocessing step, the data are transformed or consolidated so that the resulting mining process may be more efficient, and the patterns found may be easier to understand.



10) How is a data warehouse different from a database? How are they similar? (Nov/Dec 2007, Nov/Dec 2010,Apr/May 2017)

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.

A relational databases is a collection of tables, each of which is assigned a unique name. Each table consists of a set of attributes(columns or fields) and usually stores a large set of tuples(records or rows).

Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values. Both are used to store and manipulate the data.

PART B

1) Explain star schema and snow flake schema with example and discuss their performance problems May'15, Dec'13, May'11/ Explain about multidimensional Schema with example Dec'15, Dec'14,Dec '16

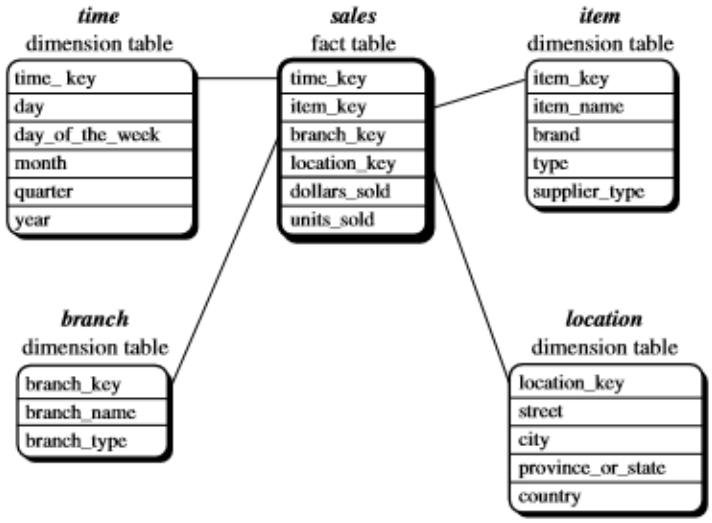
The entity-relationship data model is commonly used in the design of relational databases, where a database schema consists of a set of entities and the relationships between them. Such a data model is appropriate for on-line transaction processing.

A data warehouse, however, requires a concise, subject-oriented schema that facilitates on-line data analysis. The most popular data model for a data warehouse is a multidimensional model. Such a model can exist in the form of a **star schema, a snowflake schema**

Star schema: The most common modeling paradigm is the star schema, in which the data warehouse contains

- A large central table (fact table) containing the bulk of the data, with no redundancy
- A set of smaller attendant tables (dimension tables), one for each dimension.

The schema graph resembles a starburst, with the dimension tables displayed in a radial pattern around the central fact table.



Star schema of a data warehouse for sales

Sales are considered along four dimensions, namely, time, item, branch, and location. The schema contains a central fact table for sales that contains keys to each of the four dimensions, along with two measures: dollars sold and units sold. To minimize the size of the fact table, dimension identifiers (such as time key and item key) are system-generated identifiers.

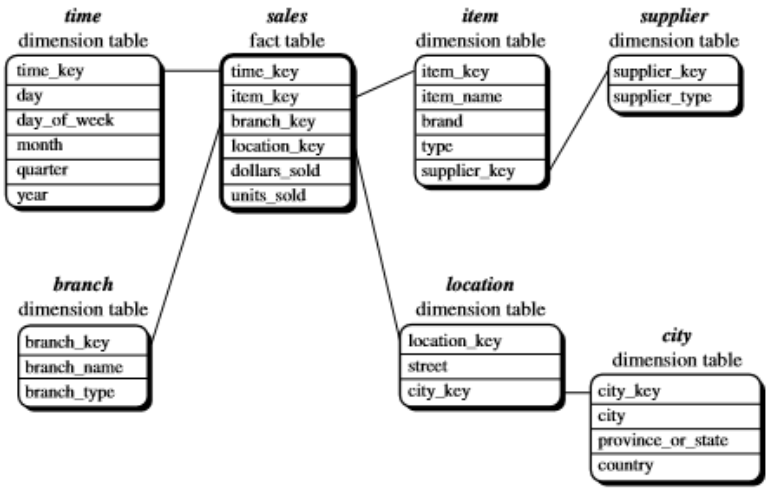
Notice that in the star schema, each dimension is represented by only one table, and each table contains a set of attributes. For example, the location dimension table contains the attribute set {Location key, street, city, province or state, country}.This constraint may introduce some redundancy.

Snowflake schema

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

The major difference between the snowflake and star schema models is that the dimension tables of the snowflake model may be kept in normalized form to reduce redundancies. Such a table is easy to maintain and saves storage space.

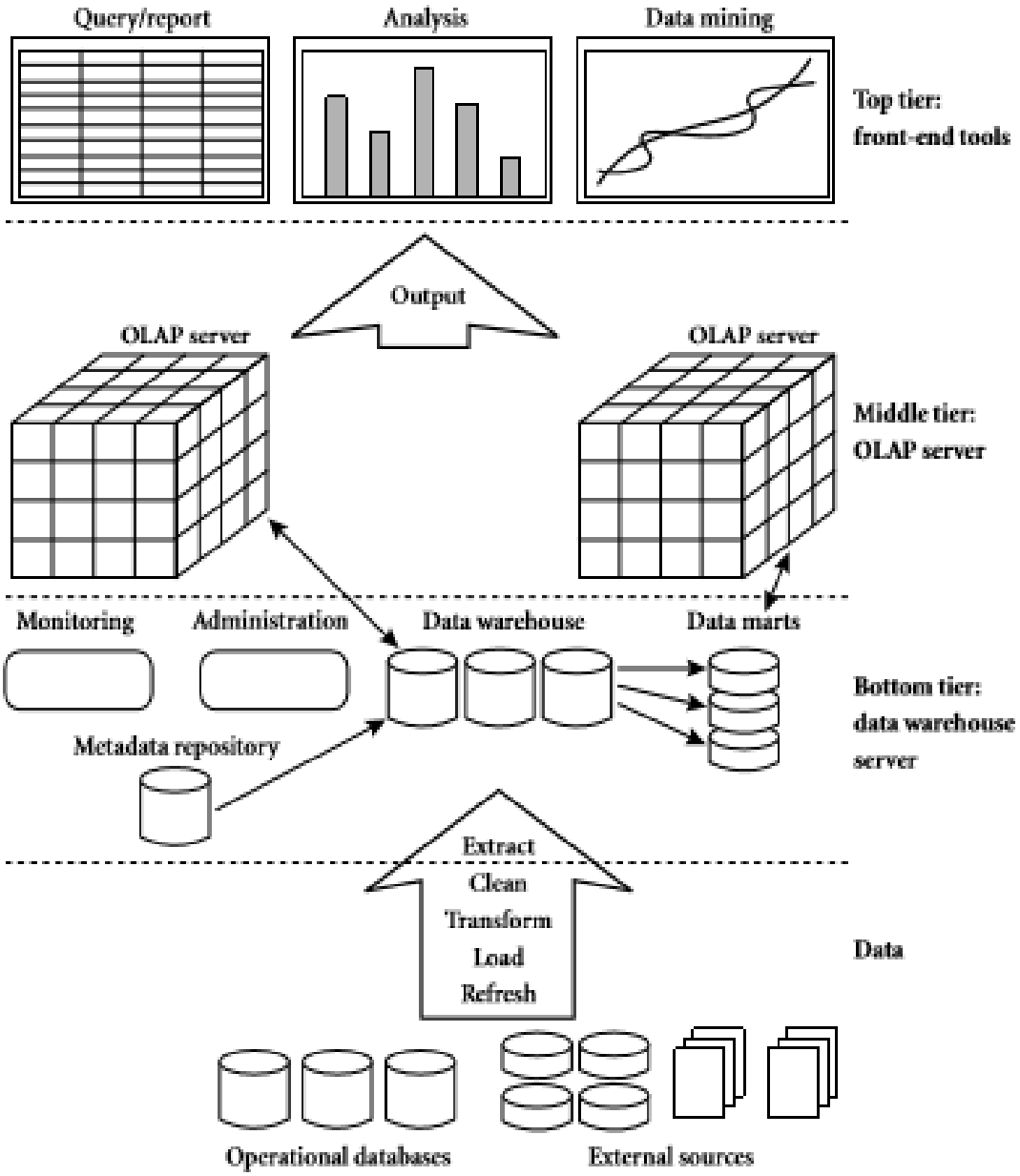
However, this saving of space is negligible in comparison to the typical magnitude of the fact table. Furthermore, the snowflake structure can reduce the effectiveness of browsing, since more joins will be needed to execute a query. Consequently, the system performance may be adversely impacted. Hence, although the snowflake schema reduces redundancy, it is not as popular as the star schema in data warehouse design.



Snowflake schema of a data warehouse for sales

A snowflake schema for AllElectronics sales is given in Figure "Snowflake schema for a data warehouse for sales". Here, the sales fact table is identical to that of the star schema in the previous figure, the main difference between the two schemas is in the definition of dimension tables. The single dimension table for item in the star schema is normalized in the snowflake schema, resulting in new item and supplier tables. For example, the item dimension table now contains the attributes item key, item name, brand, type, and supplier key, where supplier key is linked to the supplier dimension table, containing supplier key and supplier type information. Similarly, the single dimension table for location in the star schema can be normalized into two new tables: location and city. The city key in the new location table links to the city dimension. Notice that further normalization can be performed on province or state and country

2) **Explain the three tier architecture of Data warehouse with diagrammatic illustration. Dec'14, Jun'14, Dec'13, May'11**



Three-tier data warehousing architecture

Data warehouses often adopt a three-tier architecture

The bottom tier is a warehouse database server that is almost always a relational database system. Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources (such as customer profile informationprovidedbyexternalconsultants).Thesetoolsandutilitiesperformdata extraction, cleaning, and transformation (e.g., to merge similar data from different sources into a unified format), as well as load and refresh functions to update the data warehouse (The data are extracted using application program interfaces known as gateways.

A gateway is supported by the underlying DBMS and allows client programs to generate SQL code to be executed at a server. Examples of gateways include ODBC (Open Database Connection) and OLEDB (Open Linking and Embedding for Databases) by Microsoft and JDBC (Java Database Connection). This tier also contains a metadata repository, which stores information about the data warehouse and its contents.

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

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

From the architecture point of view, there are three data warehouse models:

- enterprise warehouse,
- data mart,
- Virtual warehouse.

3) Explain the mapping methodology from Data warehouse to multiprocessor architecture. May'15, Dec'14, Jun'14, Dec'11,May '17

Relational Data base Technology for data warehouse

The size of a data warehouse rapidly approaches the point where the search of a data warehouse rapidly approaches the point where the search for better performance and scalability becomes a real necessity.

The search is pursuing two goals

- Speed Up: the ability to execute the same request on the same amount of data in less time
- Scale-Up: The ability to obtain the same performance on the same request as the data base size increases.

Types of Parallelism

Parallel execution of tasks within the SQL statements can be done in either of two ways.

Horizontal parallelism: Which means that the data base is partitioned across multiple disks and the parallel processing occurs in the specific tasks, that is performed concurrently on different processors against different sets of data

Vertical Parallelism: which occurs among different tasks all components query operations are executed in parallel in a pipelined fashion. In other words an output from one task becomes an input into another task as soon as records become available

Data Partitioning

Data partitioning is a key requirement for effective parallel execution of data base operations. It spreads data from data base tables across multiple disks so that I/O operations such as read and write can be performed in parallel.

Random partitioning includes random data striping across multiple disks on single servers. In round robin partitioning, each new record id placed on the new disk assigned to the data base.

Intelligent partitioning assumes that DBMS knows where a specific record id located and does not waste time searching for it across all disks. This partitioning

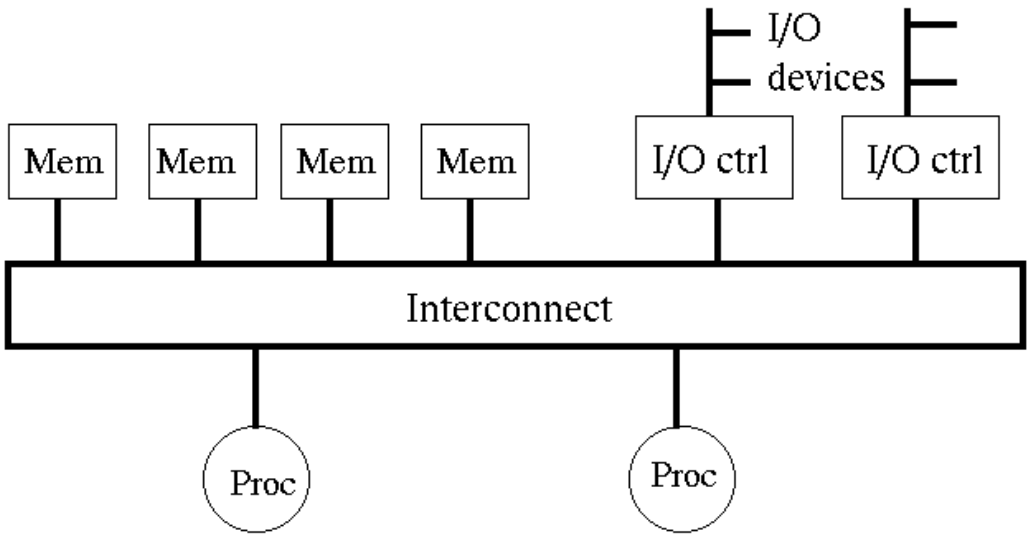
allows a DBMS to fully exploit parallel architectures and also enables higher availability.

Intelligent partitioning includes

- **Hash Partitioning :** Hash algorithm is used to calculate the partition number
- **Key range partitioning :** Partitions are based on the partition key
- **Schema partitioning :**Each table is placed in each disk, Useful for small references
- **User-defined partitioning:** Tables are partitioned based on user defined expressions

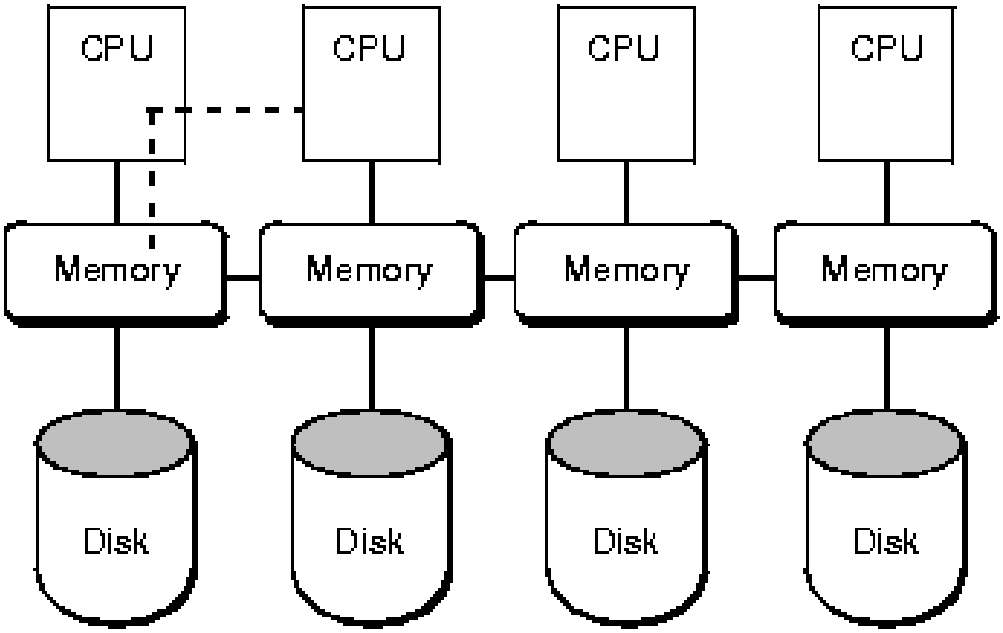
Database Architecture for parallel Processing

Shared-Memory Architecture



Also called as shared-everything style .Traditional approach to implement an RDBMS on SMP hardware. Simple to implement. The key point of this approach is that a single RDBMS server can potentially utilize all processors, access all memory, and access the entire database, thus providing the user with a consistent single system image

Shared-disk Architecture

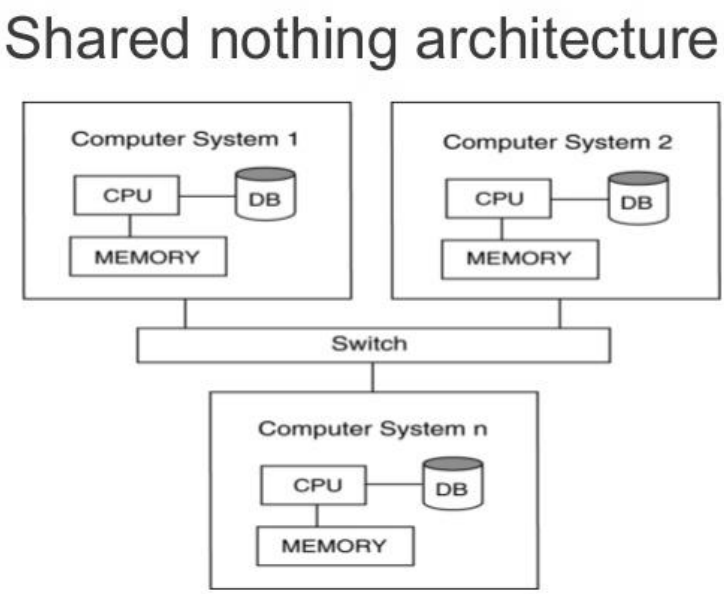


It implements the concept of shared ownership of the entire data base between RDBMS servers, each of which is running on a node of distributed memory system. Each RDBMS server can read, write, update and delete records from the same shared data base, which would require the system to implement a form of distributed lock manager (DLM).

Pining: In worst case scenario, if all nodes are reading and updating same data, the RDBMS and its DLM will have to spend a lot of resources synchronizing multiple buffer pool. This problem is called as pining Data skew: Un even distribution of data

Shared-disk architectures can reduce performance bottle-necks resulting from data skew

Shared-Nothing Architecture



The data is partitioned across many disks, and DBMS is "partitioned" across multiple conservers, each of which resides on individual nodes of the parallel system and has an ownership of its own disk and thus, its own data base partition.

It offers non-linear scalability. These requirements includes

- Support for function shipping
- Parallel join strategies
- Support for data repartitioning
- Query compilation
- Support for data base transactions
- Support for the single system image of the data base environment.

Combined Architecture

Inter server parallelism of the distributed memory architecture means that each query is parallelized across multiple servers. While intraserver parallelism of the shared memory architecture means that a query is parallelized with in the server.

4) Explain the components of the data warehousing system May'15, Dec'11,Nov'16

The data warehouse architecture is based on a relational database management system server that functions as the central repository for informational data. Operational data and processing is completely separated from data warehouse processing.

This central information repository is surrounded by a number of key components designed to make the entire environment functional, manageable and accessible by both the operational systems that source data into the warehouse and by end-user query and analysis tools.

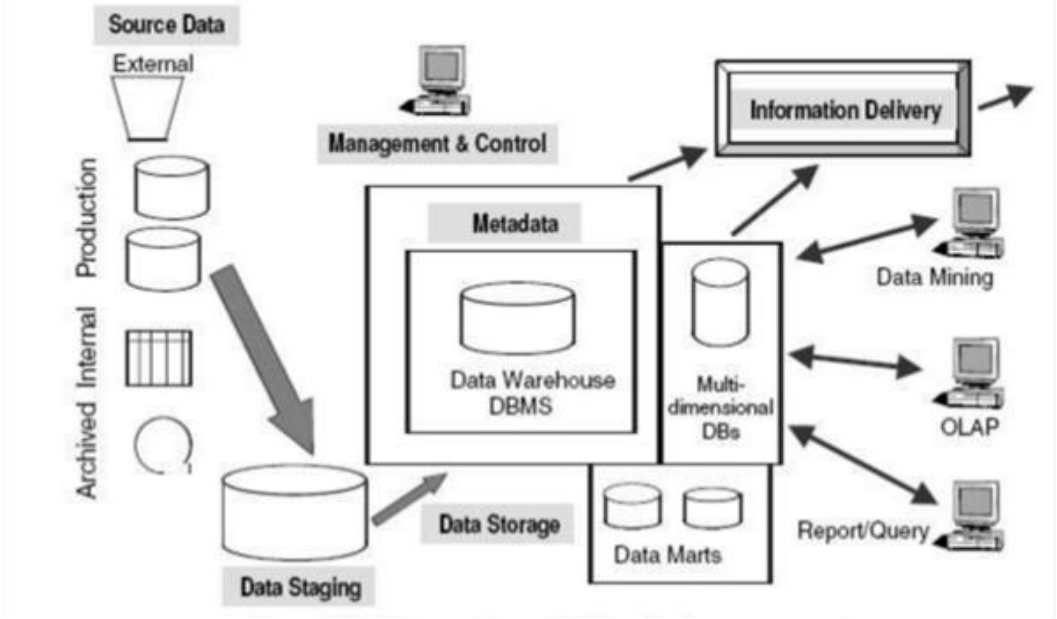
Typically, the source data for the warehouse is coming from the operational applications. As the data enters the warehouse, it is cleaned up and transformed into an integrated structure and format. The transformation process may involve conversion, summarization, filtering and condensation of data.

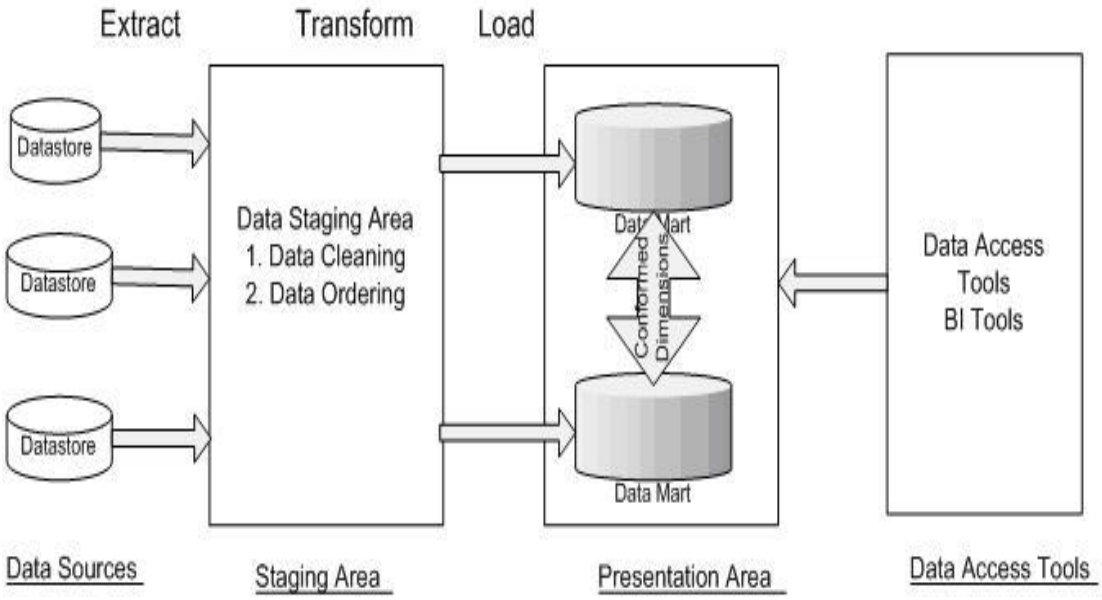
Because the data contains a historical component, the warehouse must be capable of holding and managing large volumes of data as well as different data structures for the same database over time.

Seven major components of data warehousing

1. Data Warehouse Database
2. Sourcing, Acquisition, Cleanup and Transformation Tools
3. Meta data
4. Access Tools
5. Data Marts
6. Data Warehouse Administration and Management
7. Information Delivery System

DW: BUILDING BLOCKS OR COMPONENTS





5) a.)Describe the various stages of building a data warehouse. Dec'14

Business Considerations:

Organizations embarking on data warehousing development can chose on of the two approaches

Top-down approach: Meaning that the organization has developed an enterprise data model, collected enterprise wide business requirement, and decided to build an enterprise data warehouse with subset data marts

Bottom-up approach: Implying that the business priorities resulted in developing individual data marts, which are then integrated into the enterprise data warehouse.

Organizational Issues The requirements and environments associated with the informational applications of a data warehouse are different. Therefore an organization will need to employ different development practices than the ones it uses for operational applications

Design Consideration In general, a data warehouse's design point is to consolidate data from multiple, often heterogeneous, sources into a query data base. The main factors include

- Heterogeneity of data sources, which affects data conversion, quality, time-lines
- Use of historical data, which implies that data may be" old"
- Tendency of database to grow very large

Data Content: Typically a data warehouse may contain detailed data, but the data is cleaned up and transformed to fit the warehouse model, and certain transactional attributes of the data are filtered out. The content and the structure of the data warehouses are reflected in its data model. The data model is a template for how information will be organized with in the integrated data warehouse framework.

Meta data: Defines the contents and location of the data in the warehouse, relationship between the operational databases and the data warehouse, and the business view of the warehouse data that are accessible by end-user tools. The warehouse design should prevent any direct access to the warehouse data if it does not use meta data definitions to gain the access.

Data distribution: As the data volumes continue to grow, the data base size may rapidly outgrow a single server. Therefore, it becomes necessary to know how the data should be divided across multiple servers. The data placement and distribution design should consider several options including data distribution by subject area, location, or time. **Tools:** Data warehouse designers have to be careful not to sacrifice the overall design to fit to a specific tool. Selected tools must be compatible with the given data warehousing environment each other.

Performance consideration: Rapid query processing is a highly desired feature that should be designed into the data warehouse.

Nine decisions in the design of a data warehouse:

i. Choosing the subject matter
ii. Deciding what a fact table represents
iii. Identifying and conforming the decisions
iv. Choosing the facts
v. Storing pre calculations in the fact table
vi. Rounding out the dimension table
vii. Choosing the duration of the data base
viii. The need to track slowly changing dimensions
ix. Deciding the query priorities and the query modes

Technical Considerations A number of technical issues are to be considered when designing and implementing a data warehouse environment .these issues includes. The hardware platform that would house the data warehouse. The data base management system that supports the warehouse data base. The communication infrastructure that connects the warehouse, data marts, operational systems, and end users. The hardware platform and software to support the meta data repository The systems management framework that enables the centralized management and administration of the entire environment.

Implementation Considerations A data warehouse cannot be simply bought and installed-its implementation requires the integration of many products within a data ware house.

1. Access tools

2. Data Extraction, clean up, Transformation, and migration
3. Data placement strategies
4. Meta data
5. User sophistication levels: Casual users, Power users, Experts

6) a.)What is Meta data? Classify Meta data and explain the same

Meta data is data about data that describes the data warehouse. It is used for building, maintaining, managing and using the data warehouse.

Meta data can be classified into:

Technical Meta data, which contains information about warehouse data for use by warehouse designers and administrators when carrying out warehouse development and management tasks.

Business Meta data, which contains information that gives users an easy-to-understand perspective of the information stored in the data warehouse.

Equally important, Meta data provides interactive access to users to help understand content and find data. One of the issues dealing with Meta data relates to the fact that many data extraction tool capabilities to gather Meta data remain fairly immature. Therefore, there is often the need to create a Meta data interface for users, which may involve some duplication of effort.

Meta data management is provided via a Meta data repository and accompanying software. Meta data repository management software, which typically runs on a workstation, can be used to map the source data to the target database; generate code for data transformations; integrate and transform the data; and control moving data to the warehouse.

As user's interactions with the data warehouse increase, their approaches to reviewing the results of their requests for information can be expected to evolve from relatively simple manual analysis for trends and exceptions to agent-driven initiation of the analysis based on user-defined thresholds.

The definition of these thresholds, configuration parameters for the software agents using them, and the information directory indicating where the appropriate sources for the information can be found are all stored in the Meta data repository as well.

b.) Explain the role played by sourcing, extraction, acquisition, cleanup & transformation tools in building a Data warehouse. Dec '16,May'17

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.

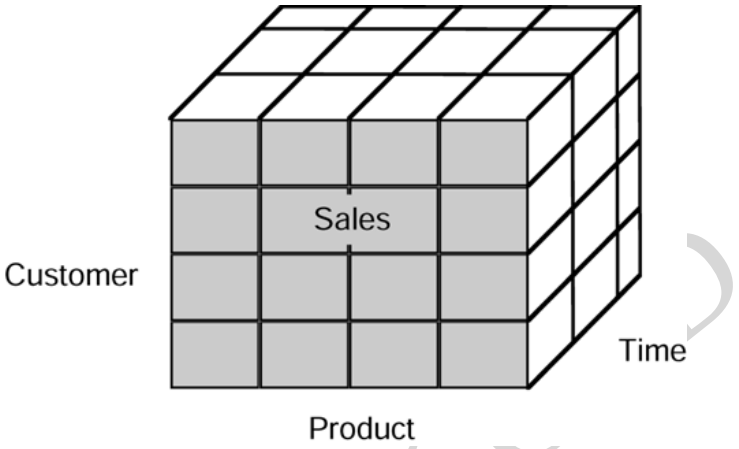
These tools can save a considerable amount of time and effort. However, significant shortcomings do exist. For example, many available tools are generally useful for simpler data extracts. Frequently, customized extract routines need to be developed for the more complicated data extraction procedures

UNIT II

BUSINESS ANALYSIS Reporting and Query –Tool Categories– The Need for Applications–Cognos Impromptu– (OLAP)–Need– Multidimensional Data Model– OLAP Guidelines–Multidimensional versus Multi- –Categories of Tools– OLAP Tools and theInternet.

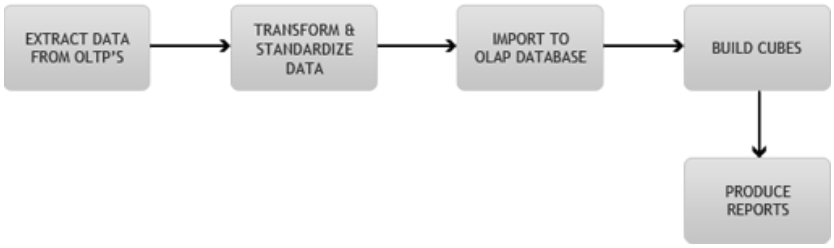
PART A

1) What is data cube? Dec'15, May'13



A data cube is a set of data that is usually constructed from a subset of a data warehouse and is organized and summarized into a multidimensional structure defined by a set of dimensions and measures.

2) Define OLAP. May'14



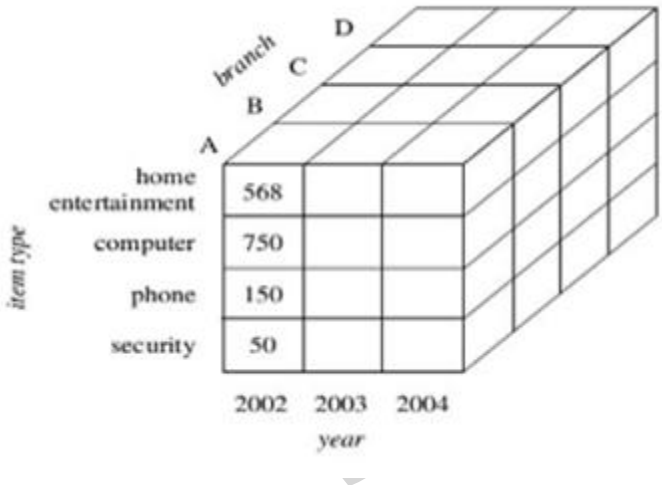
OLAP (Online Analytical Processing) is the technology behind many Business Intelligence (BI) applications. OLAP is a powerful technology for data discovery, including capabilities for limitless report viewing, complex analytical calculations, and predictive "what if" scenario (budget, forecast) planning.OLAP enables a user to easily and selectively extract and view data from different points of view

3) What are the advantages of dimensional modelling? Jun'14, Dec'15

- Understandability.
- Query performance
- Extensibility

4) What is apex cuboid? Dec'11, May'11

A cube at the highest level of abstraction is the apex cuboid. For the sales data, apex cuboid would give one total – the total sales for all the three years, for all item types and for all branches



- A data cube for the highest level of abstraction is the apex cuboid.
- A data cube for the lowest level of abstraction is the base cuboid.

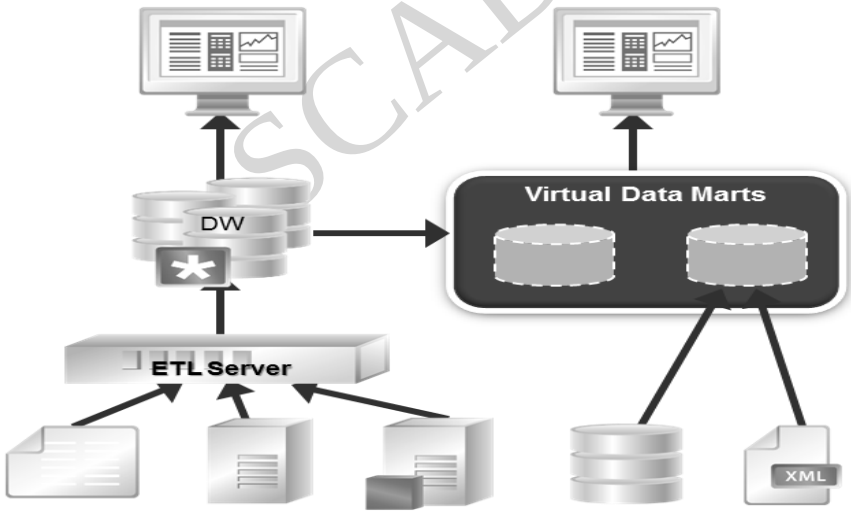
5) List different types of reporting tools. May'14, May'13

- Business Objects
- Info View
- Info Burst

6) Differentiate MOLAP & ROLAP. Dec'13

MOLAP	ROLAP
MOLAP (multidimensional OLAP) tools utilize a pre-calculated data set	ROLAP (relational OLAP) tools do not use pre-calculated data set
MOLAP tools feature very fast response, and the ability to quickly write back data into the data set	ROLAP tools feature the ability to ask any question (you are not limited to the contents of a cube) and the ability to drill down to the lowest level of detail in the database.
The most common examples of MOLAP tools are Hyperion (Arbor) Essbase and Oracle (IRI) Express	The most common examples of ROLAP tools are Micro Strategy and Sterling (Information Advantage).

7) What is virtual Warehouse? Dec'14



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

8) What is multi-dimensional data model and where is used? May'15, Apr/May 2017

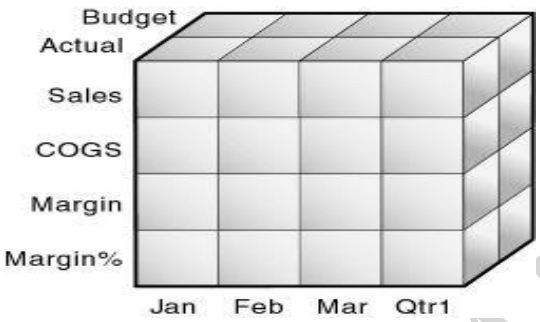
- OLAP database servers use multi-dimensional structures to store data and relationships between data.
- Multi-dimensional structures are best-visualized as cubes of data, and cubes within cubes of data. Each side of a cube is a dimension

9) Differentiate OLAP & OLTP. May'15 (M.E), Apr/May 2017,Dec 2016

	OLAP	**OLTP**
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
Purpose of data	To control and run fundamental business tasks	To help with planning, problem solving, and decision support

10) What is multidimensional DB? Dec'11



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.

11.) List OLAP guidelines. .(Nov/Dec 2016)

Multidimensional conceptual view

Transparency

Accessibility

Consistent reporting performance

Client/server architecture

Generic Dimensionality

Dynamic sparse matrix handling

Multi-user support

Unrestricted cross-dimensional operations

Intuitive data manipulation

Flexible reporting

Unlimited Dimensions and aggregation levels

12)Comment on OLAP tools Internet.(Nov/Dec 2016)

The mainly comprehensive premises in computing have been the internet and data warehousing thus the integration of these two giant technologies is a necessity. The advantages of using the Web for access are inevitable.(Reference 3) These advantages are:

- The internet provides connectivity between countries acting as a free resource.
- The web eases administrative tasks of managing scattered locations.
- The Web allows users to store and manage data and applications on servers that can be managed, maintained and updated centrally.

PART B

1) Highlight the features of Cognous Impromptu business analysis tool. Dec'15, May'15, Dec'13, May'13,Dec '16,May '17

Cognous 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.

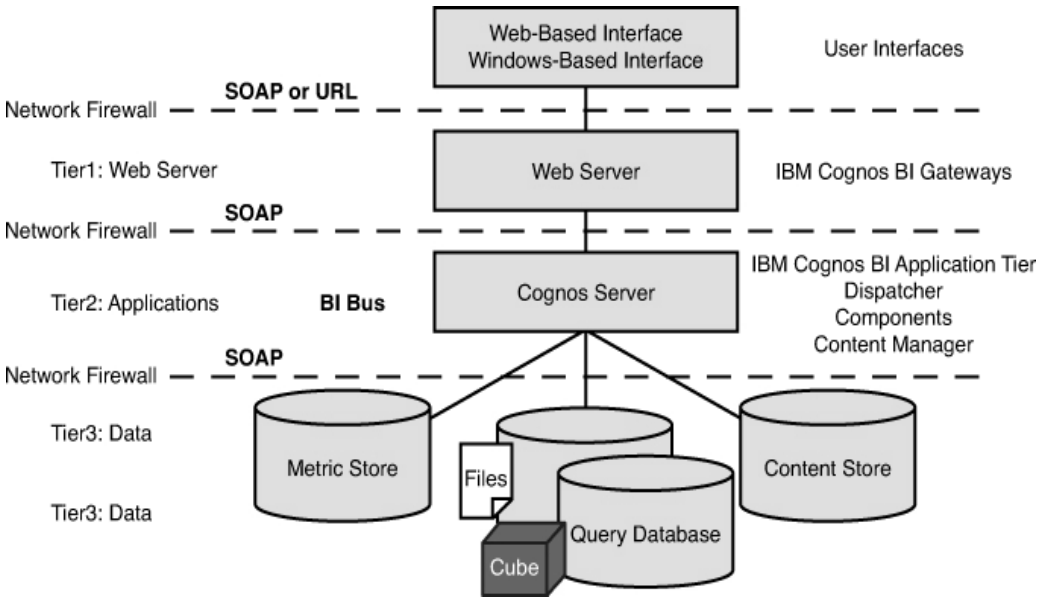
It is only capable of reading the data.

Impromptu's main features includes,

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

Catalogs Impromptu stores metadata in subject related folders. This metadata is what will be used to develop a query for a report. The metadata set is stored in a file called a catalog.

The catalog does not contain any data. It just contains information about connecting to the database and the fields that will be accessible for reports.



A catalog contains:

- Folders—meaningful groups of information representing columns from one or more tables
- Columns—individual data elements that can appear in one or more folders Calculations—expressions used to compute required values from existing data
- Conditions—used to filter information so that only a certain type of information is displayed
- Prompts—pre-defined selection criteria prompts that users can include in reports they create
- Other components, such as metadata, a logical database name, join information, and user classes

Catalog can be used to

- view, run, and print reports
- export reports to other applications
- disconnect from and connect to the database
- create reports
- change the contents of the catalog
- add user classes

2) List and explain typical OLAP operations for multidimensional data with suitable examples and diagrammatic illustrations. Dec'15, May'15, Dec'14, Dec'13, May'13

In the multidimensional 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.

A number of OLAP data cube Operations exist to materialize these different views, allowing interactive querying and analysis of the data at hand. Hence, OLAP provides a user-friendly environment for interactive data analysis.

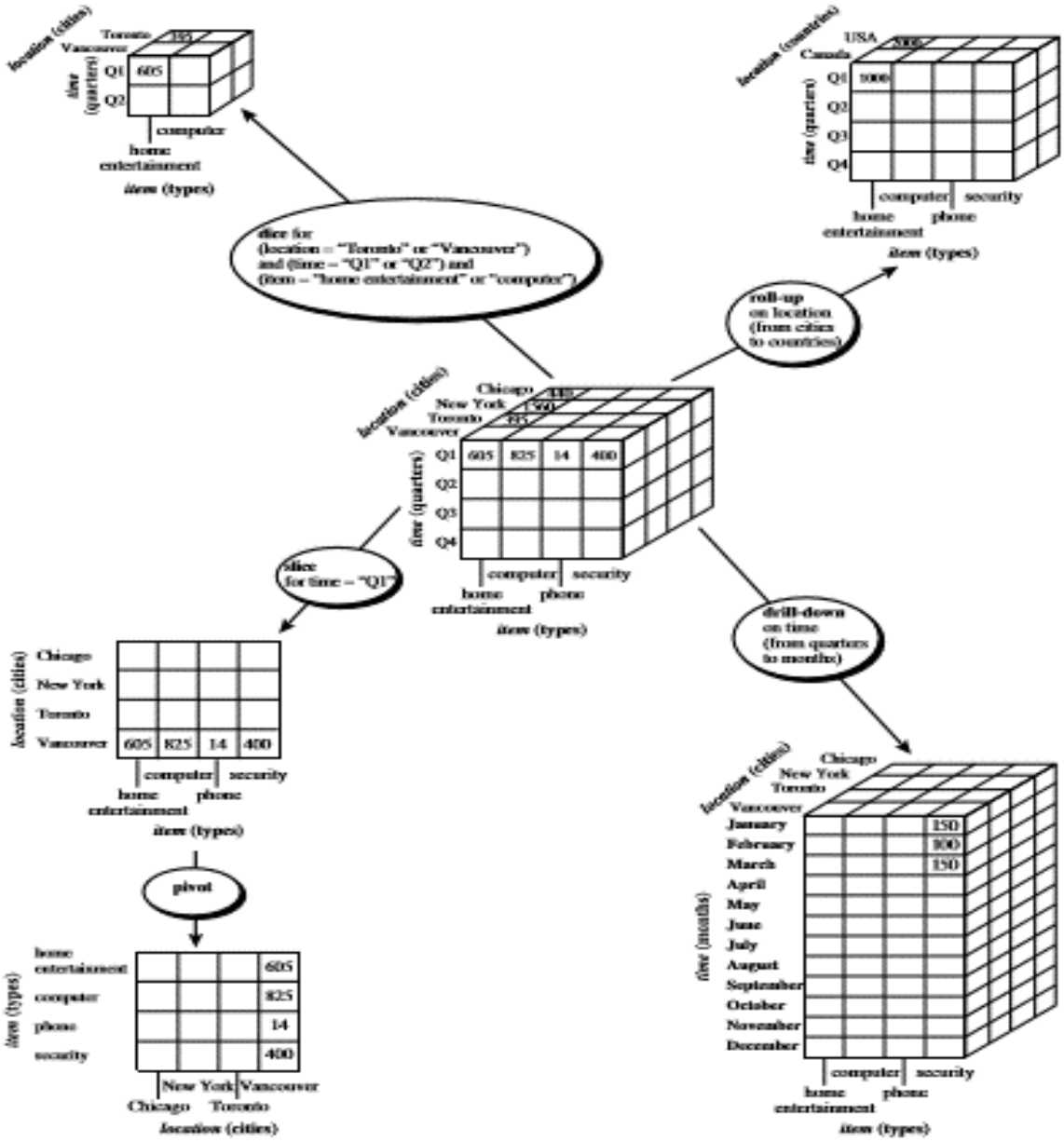
Roll-up: The roll-up operation (also called the drill-up operation by some vendors) performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction.

Drill-down: Drill-down is the reverse of roll-up. It navigates from less detailed data to moredetaileddata.

Drill-downcanberealizedbyeithersteppingdownaconcepthierarchy for a dimension or introducing additional dimensions.

Slice and dice: The slice operation performs a selection on one dimension of the given cube, resulting in a sub-cube.

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



OLAP OPERATIONS ON MULTI DIMENSIONAL DATA

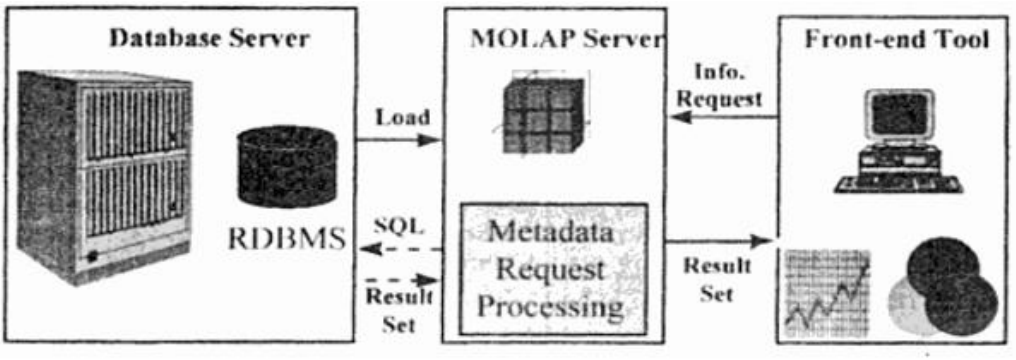
3. Write the difference between multi-dimensional OLAP and multi relational OLAP. May'14, May'13,May'11,May'17

S.NO	Multi relational OLAP	Multi-dimensional OLAP
1	Relational models can be very complex with hundreds of tables having long chains of relationship among them.	Multidimensional modeling are very simple. Each of the dimension table has a direct relationship with the fact table
2	Normal data modeling is quite flexible.	The Multidimensional modeling has a rigid structure
3	One of the goals of relational modeling is to confirm to the rules of normalization. In a normalized database each data value is stored only once.	Multidimensional modeling are radically de-normalized. The dimension tables have a high number of repeated values in their fields.
4	Standard relational models are optimized for On Line Transaction Processing. OLTP needs the ability to efficiently update data. This is provided in a normalized database that has each value stored only once.	Multidimensional modeling are optimized for On Line Analytical Processing. OLAP needs the ability to retrieve data efficiently. Efficient data retrieval requires a minimum number of joins. This is provided with the simple structure of relationship in a Multidimensional modeling, where each dimension table is only a single join away from the fact table.
6	Tables are units of relational data storage.	Cubes are units of multi-dimensional data storage.
7	Table fields of particular data type store the actual data.	Dimensions and measures stores actual data.

4) Explain different types of OLAP tools. May'14,May '17

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:

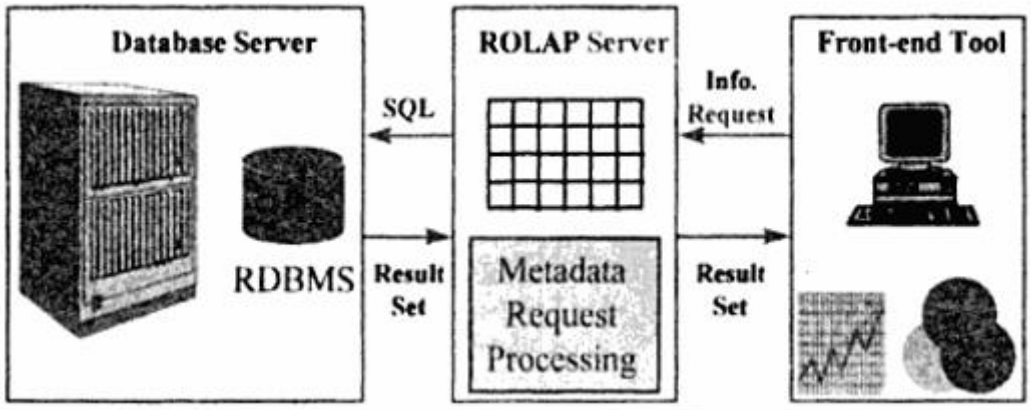
- **Excellent performance:** MOLAP cubes are built for fast data retrieval, and are optimal for slicing and dicing operations.
- **Can perform complex calculations:** All calculations have been pre-generated when the cube is created. Hence, complex calculations are not only doable, but they return quickly.

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.

5) Explain the data model suitable for Data warehouse with example. May'14

The three levels of data modeling,

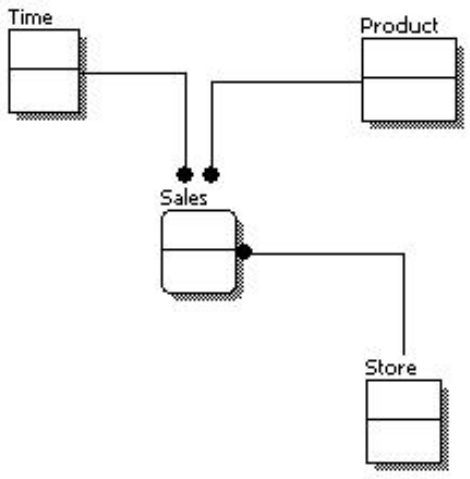
- Conceptual data model,
- Logical data model,
- Physical data model

Conceptual data model

A conceptual data model identifies the highest-level relationships between the different entities.Features of conceptual data model include:

- Includes the important entities and the relationships among them.
- No attribute is specified.
- No primary key is specified.

The figure below is an example of a conceptual data model.



From the figure above, we can see that the only information shown via the conceptual data model is the entities that describe the data and the relationships between those entities.

No other information is shown through the conceptual data model.

Logical Data Model

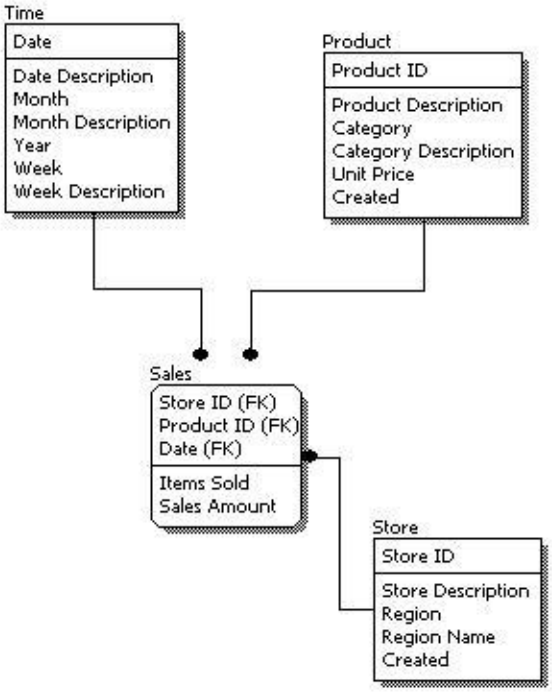
A logical data model describes the data in as much detail as possible, without regard to how they will be physical implemented in the database. Features of a logical data model include:

- Includes all entities and relationships among them.
- All attributes for each entity are specified.
- The primary key for each entity is specified.
- Foreign keys (keys identifying the relationship between different entities) are specified.
- Normalization occurs at this level.

The steps for designing the logical data model are as follows:

1. Specify primary keys for all entities.
2. Find the relationships between different entities.
3. Find all attributes for each entity.
4. Resolve many-to-many relationships.
5. Normalization.

The figure below is an example of a logical data model.



Comparing the logical data model shown above with the conceptual data model diagram, we see the main differences between the two:

- In a logical data model, primary keys are present, whereas in a conceptual data model, no primary key is present.
- In a logical data model, all attributes are specified within an entity. No attributes are specified in a conceptual data model.
- Relationships between entities are specified using primary keys and foreign keys in a logical data model. In a conceptual data model, the relationships are simply stated, not specified, so we simply know that two entities are related, but we do not specify what attributes are used for this relationship.

Physical Data Model

Physical data model represents how the model will be built in the database. A physical database model shows all table structures, including column name, column data type, column constraints, primary key, foreign key, and relationships between tables.

Features of a physical data model include:

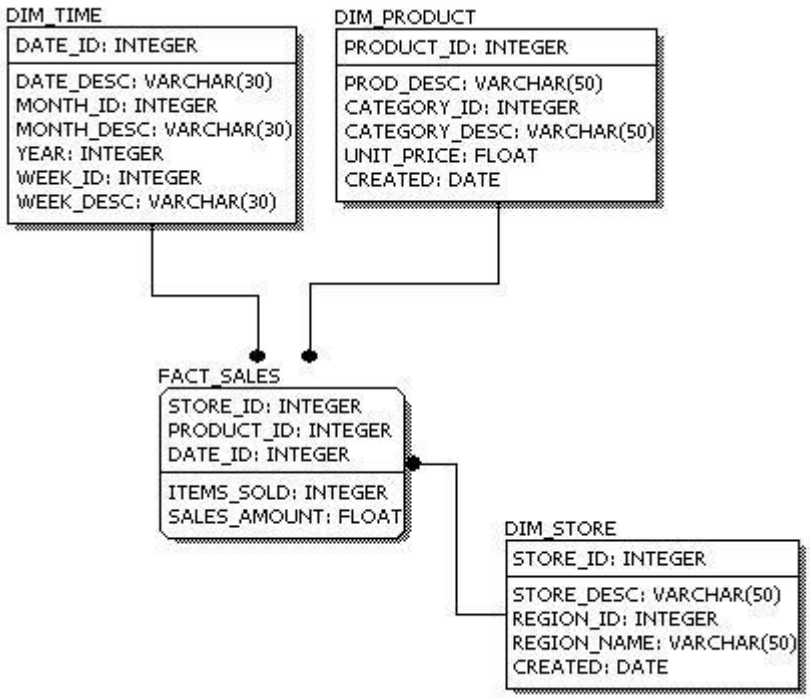
- Specification all tables and columns.
- Foreign keys are used to identify relationships between tables.
- De-normalization may occur based on user requirements.

- Physical considerations may cause the physical data model to be quite different from the logical data model.
- Physical data model will be different for different RDBMS. For example, data type for a column may be different between MySQL and SQL Server.

The steps for physical data model design are as follows:

1. Convert entities into tables.
2. Convert relationships into foreign keys.
3. Convert attributes into columns.
4. Modify the physical data model based on physical constraints / requirements.

The figure below is an example of a physical data model.



Comparing the physical data model shown above with the logical data model diagram, we see the main differences between the two:

- Entity names are now table names.
- Attributes are now column names.
- Data type for each column is specified. Data types can be different depending on the actual database being used.

The table below compares the different features:

Feature	Conceptual	Logical	Physical
Entity Names	✓	✓	
Entity Relationships	✓	✓	
Attributes		✓	
Primary Keys		✓	✓
Foreign Keys		✓	✓
Table Names			✓
Column Names			✓
Column Data Types			✓

We can see that the complexity increases from conceptual to logical to physical. This is why we always first start with the conceptual data model (so we understand at high level what are the different entities in our data and how they relate to one another), then move on to the logical data model (so we understand the details of our data without worrying about how they will actually implemented), and finally the physical data model (so we know exactly how to implement our data model in the database of choice). In a data warehousing project, sometimes the conceptual data model and the logical data model are considered as a single deliverable.

UNIT III

DATAMINING

Introduction–Data– Types of Data–Data Mining Functionalities– Interestingness of Patterns– Classification of Data Mining Systems– Data Mining Task Primitives– Integration of a Data Mining System with a Data Warehouse–Issues–Data Preprocessing.

PART A

1) Define pattern and pattern evaluation. Dec'15, May'13, Dec'11

- **Pattern** represents knowledge if it is easily understood by humans; valid on test data with some degree of certainty; and potentially useful.Measures of pattern interestingness, either objective or subjective, can be used to guide the discovery process.
- **Pattern evaluation** is to identify the truly interesting patterns representing knowledge based on some interestingness measures

2) List out the data mining functionalities. May'15, May'11

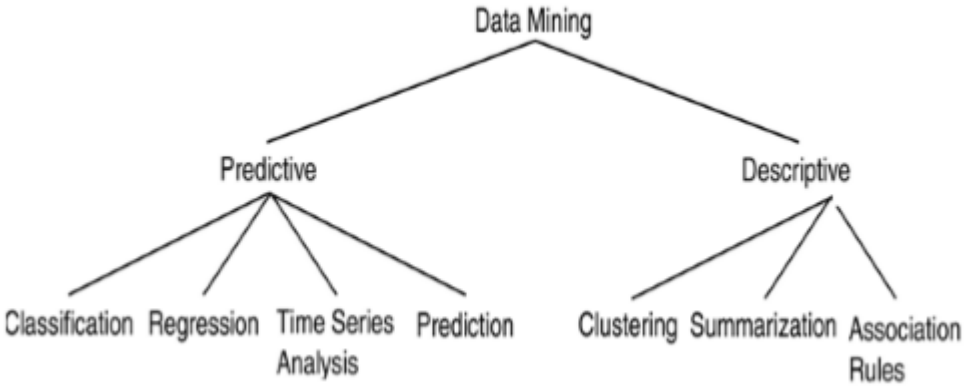
Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks.

In general, data mining tasks can be classified into two categories:

- Descriptive
- Predictive

Descriptive mining tasks characterize the general properties of the data in the database.

Predictive mining tasks perform inference on the current data in order to makepredictions.



3) Define Noisy data. Give example. May'15 (M.E)

- Noisy data is meaningless data or corrupt data. Any data that cannot be understood and interpreted correctly by machines.
- Any data that has been received, stored, or changed in such a manner that it cannot be read or used by the program that originally created it can be described as noisy.
- Noise is a random error or variance in a measured variable. Data smoothing tech is used for removing such noisy data.

4) What are the types of data? Dec'14

Data Type	**Supported Content Types**
Text	Cyclical, Discrete, Discretized, Key Sequence, Ordered, Sequence
Long	Continuous, Cyclical, Discrete, Discretized, Key, Key Sequence, Key Time, Ordered, Sequence, Time Classified
Boolean	Cyclical, Discrete, Ordered
Double	Continuous, Cyclical, Discrete, Discretized, Key, Key Sequence, Key Time, Ordered, Sequence, Time Classified
Date	Continuous, Cyclical, Discrete, Discretized, Key, Key Sequence, Key Time, Ordered

5) Define the term interestingness of pattern. Dec'14, May'13

A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm

6) Distinguish between data characterization and discrimination. Dec'13

Data characterization	Data discrimination
It is a summarization of the general characteristics or features of a target class of data.	Comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes

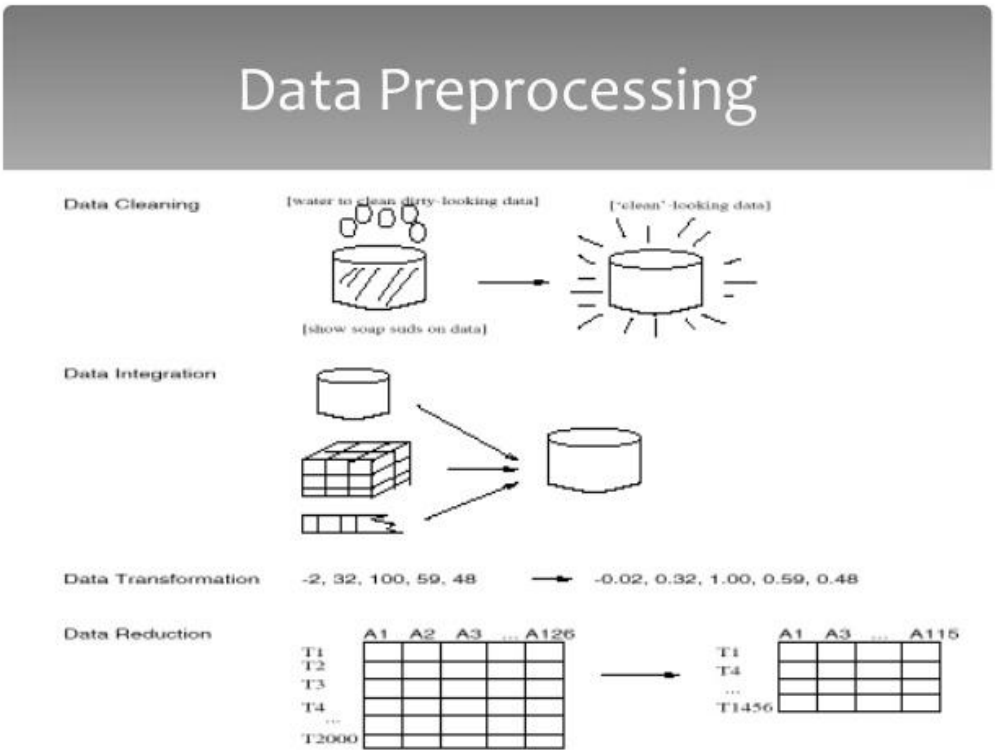
7) What is legacy DB? May'14

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

8) What is data pre-processing?

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.

Data preprocessing is used database-driven applications such as customer relationship management and rule-based applications (like neural networks).



9) What are the advantages and disadvantages of data warehousing?

Advantages	Disadvantages
Potential high returns on investment	Underestimation of resources of data loading
Competitive advantage	Hidden problems with source systems
Increased productivity of corporate decision-makers	Required data not captured
More cost-effective decision-making	Increased end-user demands
Better enterprise intelligence.	Data homogenization
	High demand for resources
	High maintenance
	Long-duration projects
	Complexity of integration

10) What are the classification of Data mining system?

- Classification of data mining systems according to the type of data sources mined
- Classification of data mining systems according to the database involved
- Classification of data mining systems according to the kind of knowledge discovered
- Classification of data mining systems according to mining techniques used

PART B

1) List and discuss about the primitives involved in a data mining task. Dec'15, May'15, Jun'14, Dec'13, May'13.,May '17

Primitives for specifying a data mining task

- Task-relevant data
- Database or data warehouse name
- Database tables or data warehouse cubes
- Conditions for data selection
- Relevant attributes or dimensions
- Data grouping criteria
- Knowledge type to be mined
- Characterization
- Discrimination
- Association/correlation
- Classification/prediction
- Clustering
- Background knowledge
- Concept hierarchies
- User beliefs about relationships in the data
- Pattern interestingness measures
- Simplicity Certainty (e.g., confidence)
- Utility (e.g., support) Novelty
- Visualization of discovered patterns
- Rules, tables, reports, charts, graphs, decision trees, and cubes
- Drill-down and roll-up

A data mining task can be specified in the form of a data mining query, which is input to the datamining system. A datamining query is defined in terms of data mining task primitives. These primitives allow the user to interactively communicate with the data mining system during discovery in order to direct the mining process, or examine the findings from different angles or depths.

A data mining query language can be designed to incorporate these primitives, allowing users to flexibly interact with datamining systems. Having a datamining query language provides a foundation on which user-friendly graphical interfaces can be built

This facilitates a data mining system's communication with other information systems and its integration with the overall information processing environment. Designing a comprehensive datamining language is challenging because datamining covers a wide spectrum of tasks, from data characterization to evolution analysis.

Each task has different requirements. The design of an effective data mining query language requires a deep understanding of the power, limitation, and underlying mechanisms of the various kinds of data mining tasks.

2) What is concept hierarchy and data discretization? Explain how they are useful for data mining. Dec'15 ,Dec '16Apr/May 2017

Concept hierarchy generation are powerful tools for data mining, in that they allow the mining of data at multiple levels of abstraction

Data discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values. Replacing numerous values of a continuous attribute by a small number of interval labels thereby reduces and simplifies the original data. This leads to a concise, easy-to-use, knowledge-level representation of mining results

Discretization techniques can be categorized based on how the discretization is performed, such as whether it uses class information or which direction it proceeds (i.e., top-down vs. bottom-up).

If the discretization process uses class information, then we say it is supervised discretization. Otherwise, it is unsupervised. If the process starts by first finding one or a few points (called split points or cut points) to split the entire attribute range, and then repeats this recursively on the resulting intervals, it is called top-down discretization or splitting. This contrasts with bottom-up discretization or merging, which starts by considering all of the continuous values as potential split-points,

removes some by merging neighborhood values to form intervals, and then recursively applies this process to the resulting intervals.

Discretization can be performed recursively on an attribute to provide a hierarchical or multiresolution partitioning of the attribute values, known as a concept hierarchy. Concept hierarchies are useful for mining at multiple levels of abstraction.

Concept Hierarchy example

A concept hierarchy for a given numerical attribute defines a discretization of the attribute. Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts (such as numerical values for the attribute age) with higher-level concepts(such as youth, middle-aged, or senior).Al though detail is lost by such data generalization, the generalized data may be more meaningful and easier to interpret.

This contributes to a consistent representation of datamining results among multiple mining tasks, which is a common requirement .In addition, mining on a reduced dataset requires fewer input/output operations and is more efficient than mining on a larger, generalized dataset. Because of these benefits, discretization techniques and concept hierarchies are typically applied before datamining as a preprocessing step, rather than during mining.

3) State and explain the various classification of data mining systems with example. Dec'14, Dec'13, Dec'11, May'11

There are many data mining systems available or being developed. Some are specialized systems dedicated to a given data source or are confined to limited data mining functionalities, other are more versatile and comprehensive.

The data mining system can be classified according to the following criteria:

- Database Technology Statistics
- Machine Learning
- Information Science
- Visualization

Other classification are the following:

Classification according to the type of data source mined: this classification categorizes data mining systems according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.

Classification according to the data model drawn on: this classification categorizes data mining systems based on the data model involved such as relational database, object oriented database, data warehouse, transactional, etc.

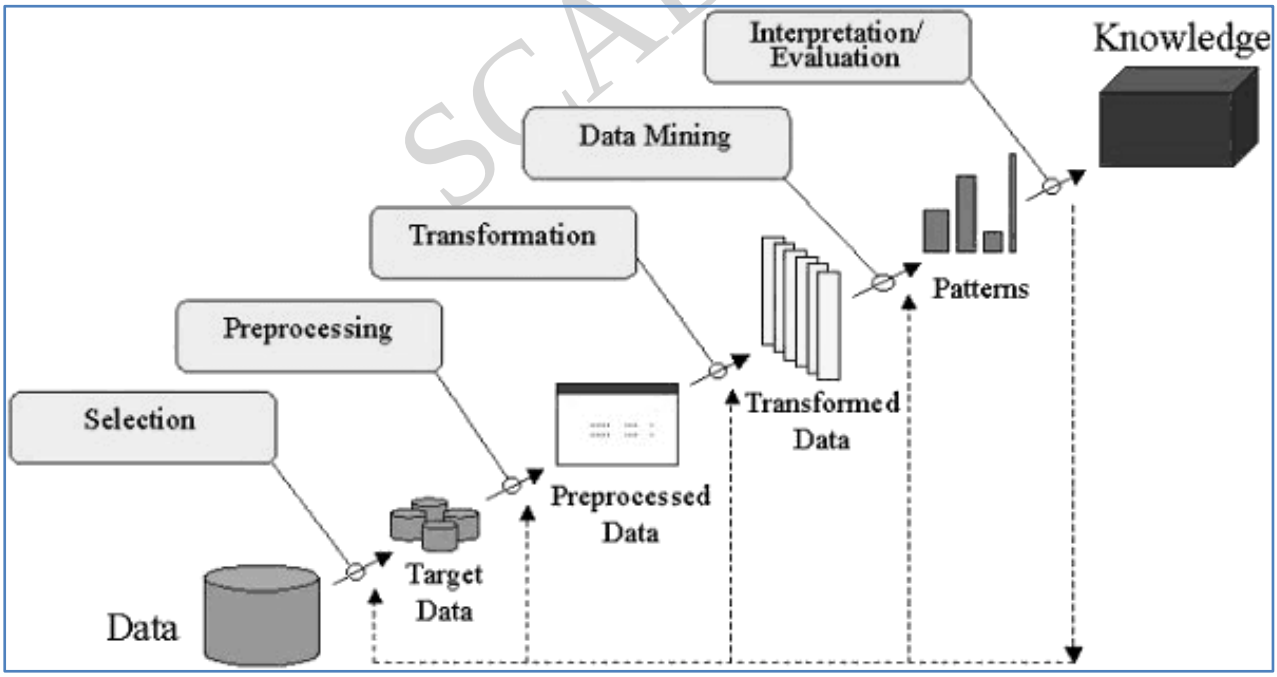
Classification according to the king of knowledge discovered: this classification categorizes data mining systems based on the kind of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several data mining functionalities together.

Classification according to mining techniques used: Data mining systems employ and provide different techniques. This classification categorizes data mining systems according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, etc. The classification can also take into account the degree of user interaction involved in the data mining process such as query-driven systems, interactive exploratory systems, or autonomous systems.

4) Explain with diagrammatic illustration the steps involved in the process of knowledge discovery from database. Dec'15, May'15, May13.

Knowledge discovery process consists of an iterative sequence of the following steps:

- data cleaning: to remove noise or irrelevant data
- data integration: where multiple data sources may be combined
- data selection: where data relevant to the analysis task are retrieved from the database
- data transformation: where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations
- data mining :an essential process where intelligent methods are applied in order to extract data patterns
- Pattern evaluation: to identify the truly interesting patterns representing knowledge based on some interestingness measures knowledge presentation: where visualization and knowledge representation techniques are used to present the mined knowledge to the user.



5) Explain various methods of data cleaning /Data Preprocessing in detail. May'14, Dec'13,May '17,Dec'16

Data cleaning:

Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. Various methods for handling this problem:

The various methods for handling the problem of missing values in data tuples include:

Ignoring the tuple: This is usually done when the class label is missing (assuming the mining task involves classification or description). This method is not very effective unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.

Manually filling in the missing value: In general, this approach is time-consuming and may not be a reasonable task for large data sets with many missing values, especially when the value to be filled in is not easily determined.

Using the attribute mean for quantitative (numeric) values or attribute mode for categorical (nominal) values, for all samples belonging to the same class as the given tuple:

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

Using the most probable value to fill in the missing value: This may be determined with regression, inference-based tools using Bayesian formalism, or decision tree induction.

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

Noisy data:

Noise is a random error or variance in a measured variable. Data smoothing tech is used for removing such noisy data.

Data smoothing techniques:

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

In this technique,

- The data for first sorted
- Then the sorted list partitioned into equi-depth of bins.
- Then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

Smoothing by bin means: Each value in the bin is replaced by the mean value of the bin.

Smoothing by bin medians: Each value in the bin is replaced by the bin median.

Smoothing by boundaries: The min and max values of a bin are identified as the bin boundaries. Each bin value is replaced by the closest boundary value.

Example: Binning Methods for Data Smoothing

Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

Partition into (equi-depth) bins(equi depth of 3 since each bin contains three values):

Bin 1: 4, 8, 9, 15

Bin 2: 21, 21, 24, 25

Bin 3: 26, 28, 29, 34

Smoothing by bin means:

Bin 1: 9, 9, 9, 9

Bin 2: 23, 23, 23, 23

Bin 3: 29, 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 4, 15

Bin 2: 21, 21, 25, 25

Bin 3: 26, 26, 26, 34

In smoothing by bin means, each value in a bin is replaced by the mean value of the bin. For example, the mean of the values 4, 8, and 15 in Bin 1 is 9.

Therefore, each original value in this bin is replaced by the value 9. Similarly, smoothing by bin medians can be employed, in which each bin value is replaced by the bin median. In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value.

Suppose that the data for analysis include 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.

(a) 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.

The following steps are required to smooth the above data using smoothing by bin means with a bin depth of 3.

Step 1: Sort the data. (This step is not required here as the data are 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: 14, 14, 14

Bin 2: 18, 18, 18

Bin 3: 21, 21, 21

Bin 4: 24, 24, 24

Bin 5: 26, 26, 26

Bin 6: 33, 33, 33

Bin 7: 35, 35, 35

Bin 8: 40, 40, 40

Bin 9: 56, 56, 56

Data Reduction

Why Data Reduction?

- A database of data warehouse may store terabytes of data
- Complex data analysis or mining will take long time to run on the complete data set

- Obtaining a reduced representation of the complete dataset
- Produces same result or almost same mining / analytical results as that of original.

1 Data cube Aggregation
2 Dimensionality reduction – remove unwanted attributes
3 Data Compression
4 Numerosity reduction – Fit data into mathematical models
5 Discretization and Concept Hierarchy Generation

1. Data Cube Aggregation:

- The lowest level of data cube is called as base cuboid.

- Single Level Aggregation - Select a particular entity or attribute and Aggregate based on that particular attribute.

  Eg. Aggregate along 'Year' in a Sales data.

- Multiple Level of Aggregation – Aggregates along multiple attributes – Further reduces the size of the data to analyze.
- When a query is posed by the user, use the appropriate level of Aggregation or data cube to solve the task
- Queries regarding aggregated information should be answered using the data cube whenever possible.

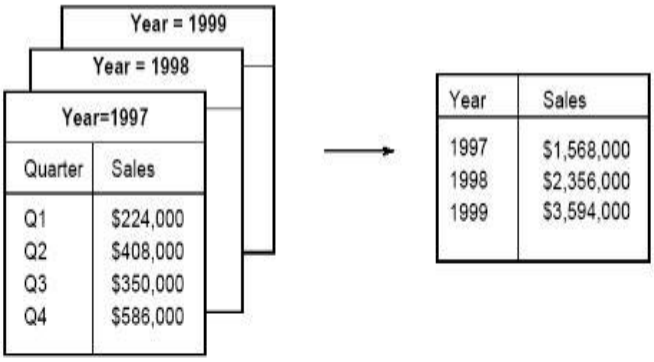


Figure 3.4: Sales data for a given branch of *AllElectronics* for the years 1997 to 1999. In the data on the left, the sales are shown per quarter. In the data on the right, the data are aggregated to provide the *annual_sales*.

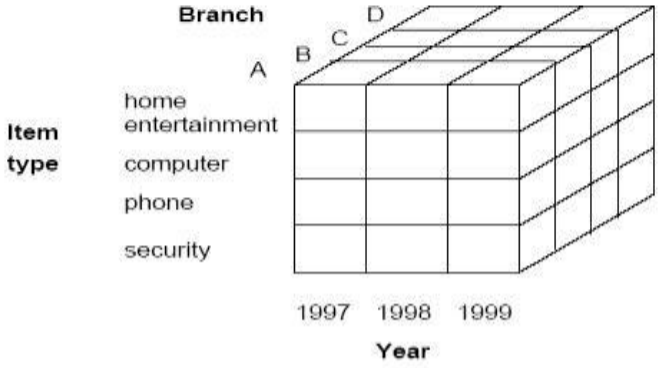


Figure 3.5: A data cube for sales at *AllElectronics*.

2. Attribute Subset Selection

Feature Selection: (attribute subset selection)

- The goal of attribute subset selection is to find the minimum set of Attributes such that the resulting probability distribution of data classes is as close as possible to the original distribution obtained using all Attributes.

- This will help to reduce the number of patterns produced and those patterns will be easy to understand

Heuristic Methods: (Due to exponential number of attribute choices)

- Step wise forward selection
- Step wise backward elimination
- Combining forward selection and backward elimination
- Decision Tree induction - Class 1 - A1, A5, A6; Class 2 - A2, A3, A4

6) Describe in detail data mining functionalities and the different kinds of patterns can be mined. (May '17)

Data Mining Functionalities

Data mining is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses or other information repositories.

Data Mining Functionalities – What kinds of Patterns can be mined?

Two Categories – (i) Descriptive (ii) Predictive

Descriptive – describes the general properties of the data in the database

Predictive – Makes Predictions from the data

Data Mining Functionalities should allow:

Mining of multiple kinds of patterns that accommodates different user expectations and applications.Discover patterns at different levels of granularity.Hints / Specifications / Queries to focus the search for interesting patterns.Each discovered pattern is measured for its "trustworthiness" based on data in the database.

1) Characterization:

- Concept description / Class description
- Data is associated with concept or class
- Eg. Classes of items for sale – (i) Computers (ii) Printers
- Eg. Concepts of customers – (i) Big Spenders (ii) Budget

Spenders o Class / Concept descriptions can be delivered via:

(1) Data Characterization

(2) Data Discrimination

(3) Data Characterization and Data Discrimination

Data Characterization

Summarization of general characteristics or features of a target class of data

Data specific to a class are collected by query

Types of data summarization:

- Summarization based on simple statistical measures
- Data summarization along a dimension – user controlled – OLAP rollup
- Attribute oriented induction – without user interaction.

Output of data characterization can be represented in different forms:

- Pie Charts, Bar Charts, Curves
- Multidimensional Data cubes, Multidimensional tables
- Generalized relations – in rule form – called as "Characteristics Rules"
- Eg. "Find Summarization of characteristics of customers who spend more than Rs. 50000 in shop S1 in a year"
- Result = "Customers are 40–50 years old, employed and have high credit rating"
- Users can drill down on any dimension – Eg. "Occupation of customers"

Data Discrimination

- Comparison of general features of target class data objects with the general features of objects from one or a set of contrasting classes.
- Target and Contrasting classes are specified by the users
- Data objects retrieved through database queries
- Eg. "Users wants to compare general features of S/W products whose sales increased by 10% in the last year with those whose sales decreased by 30% during the same period."

- Output of data discrimination is same as output of data characterization.
- Rule form is called as "Discriminate Rules".
- Eg. Compare two groups of customers.
  - Group1 – Shops frequently – at least 2 times a month
  - Group 2 – Shops rarely – less than 3 times a year
- Result = "80% of frequent shopping customers are between 20-40 years old & have university education." & "60% of infrequent shopping customers are seniors or youths with no university degree."
- Users can drill down on income level dimension for better discriminative features between the two classes of customers.

2) Mining Frequent Patterns, Associations and Correlation:

Frequent Patterns – Patterns that occur frequently in data.

- Many kinds of Frequent Patterns exists:
  - (1) Itemsets (2) Subsequences (3) Substructures
  - Frequent Itemsets: (Simple)
    - Set if items that frequently appear together in a database.
    - Eg. Bread & Jam
  - Frequent Subsequences: (Advanced)
    - Frequent sequential patterns
    - Eg. Purchase PC Purchase Digital Camera Memory Card
  - Frequent Structured Patterns: (Advanced)
    - Structural forms that occur frequently
    - Structural forms – Graphs, Trees, Lattices

- Result = Discovery of interesting associations and correlations within data.

- Eg. Association Analysis:
  - Example 1: "Find which items are frequently purchased together in the same transactions".

    Buys (X. "Computer") => Buys (X, "Software") [Support = 1%,
    - Confidence = 50%]

    X is a variable representing customers

    Confidence = % of chance that a customer buying computer buys a software

    Support = % of transactions in the whole database that showed computers and software's were purchased together.

    This association rule has a single repeated predicate "Buys"

    Such association rules are called "Single Dimensional Association Rules"
  - Example 2:

    Age (X, "20…29") ^ Income (X, "20K…29K") => Buys (X, "CD Player") [Support = 2%, Confidence = 60%]

    Association Rule = "2% of total customers in the database are between
    - 20-29 years of age and with income Rs.20000 to Rs.29000 and have purchased CD player." & "There is 60% probability that a customer in this age group and income group will purchase a CD player"

    This is an association between more than one predicate (ie. Age, Income and Buys)

    This is called as "Multidimensional Association Rule".

Association rules that do not satisfy minimum support threshold and minimum confidence threshold are discarded.

3) Classification and Prediction:

Classification:

Process of finding a model that describes data classes or concepts

Based on a set of training data

This model can be represented in different forms

- Classification Rules
- Decision Trees
- Mathematical Formulae
- Neural Networks

Decision Trees

- Flowchart like tree structure
- Each Node = Test on the attribute value
- Each Branch = Outcome of the test
- Tree Leaves = Classes or class distributions
- Decision trees can be converted into classification rules

Neural Networks

Collection of neuron-like processing units + weighted connections between the units.

Other methods of Classification

i. Naïve Bayesian Classification
ii. Support Vector Machines
iii. K-nearest neighbor Classification

Classification is used to predict missing or unavailable numeric data values => Prediction.

- Regression Analysis:- is a statistical methodology used for numeric prediction.

- Prediction also includes distribution trends based on the available data.
- Classification and Prediction may be used to be preceded by Relevance Analysis.
- Relevance Analysis:- attempts to identify attributes that do not contribute to the classification or prediction process which can be excluded.

Example – Classification and Prediction:

1) IF-THEN rules – Classification Model:

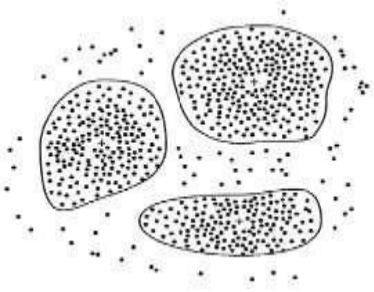
2) A Decision Tree – Classification Model:

3) A Neural Network Classification Model:

4) Cluster Analysis:

- Analyzes data objects without consulting a known class label.
- The objects are clustered or grouped based on the principles of "Maximizing the intra-class similarity" and "Minimizing the inter-class similarity".
- Objects within a cluster have high similarity compared to objects in other clusters.
- Each cluster formed is a class of objects.
- From this class of objects rules can be derived.
- Clustering allows "Taxonomy Formation" Hierarchy of classes that groups
  - similar events together.

- Eg. Customers with respect to customer locations in a city.



- 3 Data Clusters; Cluster center marked with a '+'

5) Outlier Analysis:

Data that do not comply with general behavior of data are called as Outliers.

- Most Data Mining methods discard outliers as noise or exceptions.
- Some applications like fraud detection, rare events can be interesting than regular ones.
- Analysis of such outliers is called as Outlier Analysis / Outlier Mining.
- Outliers detected using:
    i. Statistical Methods
    ii. Distance Measures
    iii. Deviation Based Methods
    iv. Difference in characteristics of an object in a group o Example – Outlier Analysis:
    v. Fraudulent usage of credit cards by detecting purchase of extremely large amount for a given credit card account compared to its general charges incurred.Same applies for Type of purchase, Place of purchase, Frequency of purchase.

6) Evolution Analysis:

Describes the trends of data whose behavior change over time.

This step includes:

i. Characterization & Discrimination
ii. Association & Correlation Analysis
iii. Classification & Prediction
iv. Clustering of time-related data
v. Time-series data analysis
vi. Sequence or periodicity pattern matching
vii. Similarity based data analysis

- Example – Evolution Analysis:
  i. Stock exchange data for past several years available.
  ii. You want to invest in TATA Steel Corp.

Data mining study / Evolution analysis on previous stock exchange data can help prediction of future trends in stock exchange prices.This will help in decision making in stock investment

UNIT IV

ASSOCIATION RULE MININGAND CLASSIFICATION

Mining Frequent Patterns, Associations and Correlations– Mining Methods–Mining various Kinds of Association Rules–Correlation Analysis–Constraint Based Association Mining– Classification and Prediction-Basic Concepts-Decision Tree Induction-Bayesian Classification–Rule Based Classification–Classificationby Backpropagation–SupportVectorMachines–Associative Classification– LazyLearners –OtherClassificationMethods–Prediction.

PART A

1) What is correlation analysis? Dec'15, Dec'11, May'11

Many of the attributes in the data may be redundant. Correlation analysis can be used to identify whether any two given attributes are statistically related

2) What is tree pruning? Dec'14, May'13

When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outlier. Tree pruning methods Identify and remove branches that reflect noise or outliers

Tree pruning Approaches: • Pre pruning • Post pruning

3) What are eager learners & lazy learners? Give example. Dec'15, Dec'14,May '17

Lazy learning (e.g., instance-based learning):

- Simply stores training data (or only minor processing) and waits until it is given a test tuple
- less time in training but more time in predicting

- Lazy method effectively uses a richer hypothesis space since it uses many local linear functions to form an implicit global approximation to the target function

Eager learning:

- Given a set of training tuples, constructs a classification model before receiving new data to classify
- Eager learning commit to a single hypothesis that covers the entire instance space

4) How prediction differs from classification data mining? May'14 May '17

- A classification problem could be seen as a predictor of classes
- Predicted values are usually continuous whereas classifications are discreet.
- Predictions are often (but not always) about the future whereas classifications are about the present.
- Classification is more concerned with the input than the output

Prediction	Classification
It is used to predict missing or unavailable numerical data values rather than class labels. Regression Analysis is generally used for prediction. Prediction can also be used for identification of distribution trends based on available data.	It predicts the class of objects whose class label is unknown. Its objective is to find a derived model that describes and distinguishes data classes or concepts. The Derived Model is based on the analysis set of training data i.e. the data object whose class label is well known.
• **Accuracy** − Accuracy of classifier refers to the ability of classifier. It predict the class label correctly and the accuracy of the 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 the ability to construct the classifier or predictor efficiently; given large amount of data.
- **Interpretability** − It refers to what extent the classifier or predictor understands.

5) What is decision tree? Mention two phases in decision tree induction.(Nov/Dec 2016)

- A decision tree is a graph that uses a branching method to illustrate every possible outcome of a decision.Decision tree software is used in data mining to simplify complex strategic challenges and evaluate the cost-effectiveness of research and business decisions.
- The two phases are Decision tree Construction and Tree Pruing.

6) Distinguish between classification and clustering. May'15

Clustering and classification can seem similar because both data mining algorithms divide the data set into subsets, but they are two different learning techniques, used in data mining for the purpose of getting reliable information from a collection of raw data.

	Clustering	**Classification**
Definition	Clustering is an unsupervised learning technique used to group similar instances on the basis of features.	Classification is a supervised learning technique used to assign predefined tags to instances on the basis of features.

Aim	The aim of clustering is, grouping a set of objects in order to find whether there is any relationship between them.	The aim of classification is to find which class a new object belongs to from the set of predefined classes.
Training Set	A training set is not used in clustering.	A training set is used to find similarities in classification.
Process	Statistical concepts are used, and datasets are split into subsets with similar features.	Classification uses the algorithms to categorize the new data according to the observations of the training set.
Labels	There are no labels in clustering.	There are labels for some points.

7) Define support vector machine. May'15, May'11

- A Support Vector Machine (SVM) is an algorithm for the classification of both linear and nonlinear data. It transforms the original data in a higher dimension, from where it can find a hyperplane for separation of the data using essential training tuples called support vectors and margins (defined by the support vectors).
- SVM uses a nonlinear mapping to transform the original training data into a higher dimension. Within this new dimension, it searches for the linear optimal

separating hyperplane (that is, a "decision boundary" separating the tuples of one class from another).

8) Define frequent item set. Dec'13

Frequent item sets play an essential role in many Data Mining tasks that try to find interesting patterns from databases, such as association rules, correlations, sequences, episodes, classifiers and clusters.

9) What is market basket analysis? May'13

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

10) State rule based classification with example. Dec'11.

Rule-based classifier makes use of a set of IF-THEN rules for classification

Let us consider a rule R1,

R1: IF age=youth AND student=yes

THEN buy_computer=yes

- The IF part of the rule is called rule antecedent or precondition.

- The THEN part of the rule is called rule consequent.
- The antecedent part the condition consist of one or more attribute tests and these tests are logically AND'ed.
- The consequent part consists of class prediction.

11)List the techniques to improve the efficiency of Apriori algorithm. (May/June 2010) (Nov/Dec 2016)

- Hash based technique
- Transaction Reduction
- Portioning Sampling
- Dynamic item counting

PART B

1) Apriori algorithm for discovering frequent item set. Dec'15, may'15, Dec'14, May'14, Dec'13, May'13, Dec'11, May'11

Discuss the single dimensional Boolean association rule mining for transaction database. May '17

The Apriori Algorithm: Finding Frequent Itemsets Using Candidate Generation

Aprioriis a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994 for miningfrequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses *prior knowledge* of frequent itemset properties

Apriori property

Apriori property is used to reduce the search space if an itemset*I* does not satisfy the minimum support threshold, *min sup*, then *I* is not frequent; thatis, *P*(*I*) <*min sup*. If an item *A* is added to the itemset*I*, then the resulting itemset (i.e., *I* [*A*) cannot occur more frequently than *I*. Therefore, *I* [*A*is not frequent either; thatis, *P*(*I* [*A*) <*min sup*.

Example of Apriori algorithm

Transaction ID	Items Bought
T1	{Mango, Onion, Nintendo, Key-chain, Eggs, Yo-yo}
T2	{Doll, Onion, Nintendo, Key-chain, Eggs, Yo-yo}
T3	{Mango, Apple, Key-chain, Eggs}
T4	{Mango, Umbrella, Corn, Key-chain, Yo-yo}
T5	{Corn, Onion, Onion, Key-chain, Ice-cream, Eggs}

Now, we follow a simple golden rule: we say an item/itemset is frequently bought if it is bought at least 60% of times. So for here it should be bought at least 3 times.

For simplicity

M = Mango

O = Onion

And so on……

So the table becomes

Original table:

Transaction ID	Items Bought
T1	{M, O, N, K, E, Y }
T2	{D, O, N, K, E, Y }
T3	{M, A, K, E}
T4	{M, U, C, K, Y }
T5	{C, O, O, K, I, E}

Step 1: Count the number of transactions in which each item occurs, Note 'O=Onion' is bought 4 times in total, but, it occurs in just 3 transactions.

Item	No of transactions
M	3
O	3
N	2
K	5
E	4
Y	3
D	1
A	1

U	1
C	2
I	1

Step 2: Now remember we said the item is said frequently bought if it is bought at least 3 times. So in this step we remove all the items that are bought less than 3 times from the above table and we are left with

Item	Number of transactions
M	3
O	3
K	5
E	4
Y	3

This is the single items that are bought frequently. Now let's say we want to find a pair of items that are bought frequently. We continue from the above table (Table in step 2)

Step 3: We start making pairs from the first item, like MO,MK,ME,MY and then we start with the second item like OK,OE,OY. We did not do OM because we already did MO when we were making pairs with M and buying a Mango and Onion together is same as buying Onion and Mango together. After making all the pairs we get,

Item pairs
MO
MK
ME
MY
OK
OE
OY

KE
KY
EY

Step 4: Now we count how many times each pair is bought together. For example M and O is just bought together in {M,O,N,K,E,Y}

While M and K is bought together 3 times in {M,O,N,K,E,Y}, {M,A,K,E} AND {M,U,C, K, Y}

After doing that for all the pairs we get

Item Pairs	Number of transactions
MO	1
MK	3
ME	2
MY	2
OK	3
OE	3
OY	2
KE	4
KY	3
EY	2

Step 5: Golden rule to the rescue. Remove all the item pairs with number of transactions less than three and we are left with

Item Pairs	Number of transactions
MK	3
OK	3
OE	3

KE	4
KY	3

These are the pairs of items frequently bought together.

Now let's say we want to find a set of three items that are brought together.

We use the above table (table in step 5) and make a set of 3 items.

Step 6: To make the set of three items we need one more rule (it's termed as self-join),

It simply means, from the Item pairs in the above table, we find two pairs with the same first Alphabet, so we get

- OK and OE, this gives OKE
- KE and KY, this gives KEY

Then we find how many times O,K,E are bought together in the original table and same for K,E,Y and we get the following table

Item Set	Number of transactions
OKE	3
KEY	2

While we are on this, suppose you have sets of 3 items say ABC, ABD, ACD, ACE, BCD and you want to generate item sets of 4 items you look for two sets having the same first two alphabets.

- ABC and ABD -> ABCD
- ACD and ACE -> ACDE

And so on … In general you have to look for sets having just the last alphabet/item different.

Step 7: So we again apply the golden rule, that is, the item set must be bought together at least 3 times which leaves us with just OKE, Since KEY are bought together just two times.

Thus the set of three items that are bought together most frequently are O,K,E.

2) Explain the working of Bayesian classification& rule based classification with an example. Dec'15, May'15, Dec'14, May,13,May'11

Bayesian Classification:

Predicts class membership probabilities = Probability that a given sample belongs to a particular class. This is based on Bayes Theorem

Exhibits high accuracy and speed when applied to large databases. One type of Bayesian Classification is This method has performance comparable to Decision Tree induction. This method is based on the assumption
given class is independent of the values.This assumption is called "Class Condit
Another type of Bayesian Classification

Bayesian Classifiers have minimum error rate when compared to all other classifiers

Bayes Theorem:

Let X be a data sample whose class label is unknown
Ex: Data Samples consists of fruits described by their colour and shape.
Say, X is red and round
Let H be some hypothesis such that the data sample X belongs to a class C
Ex: H is the hypothesis that X is an apple.

Then, we have to determine P(H/X) = Probability that the hypothesis H holds for the data sample X

P(H/X) is called as the Posterior Probability of H on X

Ex: Probability that X is an apple given that X is red and round.

P(X/H) = Posterior Probability of X on H

Ex: Probability that X is red and round given that X is an apple.

P(H) = Prior Probability of H

Ex: Probability that any given data sample is an apple regardless of its colour and shape.

P(X) = Prior Probability of X

Ex: Probability that X is red and round given that X is an apple.

Bayes Theorem is =>P(H/X) = P(X/H) P(H) / P(X)

Naïve Bayesian Classification:

1. Each data sample is represented by an n-dimensional vector X = [Where, we have n-attributes => A1, A2, … An].

Say there are m classes, C1, C2, … Cm; no class label.

Then Naïve Bayesian Classifier assigns X to the class Ci; Where Ci is the class having highest Posterior Probability Conditioned on X.

i.e. $P(C_i|X) > P(C_j|X)$ for $1 \leq j \leq m, j \neq i.$ = Maximum Posteriori Hypothesis

Here we calculate P(Ci/X) by $P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}.$ = Bayes Theorem

P(X) is constant for all classes. Hence only P(X/Ci) &P(Ci) need to be maximized.

Here P(Ci) = si/s; Where si = Samples in class Ci; s = Total number of samples in all classes.

To evaluate P(X/Ci) we use the Naïve assumption Independence".

Hence $$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i).$$

The probabilities P(x1/Ci), P(x2/Ci), … training samples;

Where P(xk/Ci) = sik/si; Where sik = No. Of samples in Class Ci that has value = xk for Ak;

si = No. Of samples in class Ci.

Evaluate P(X/Ci) P(Ci) for each class C

X is assigned to the class Ci for which P(X/Ci) P(Ci) is maximum.

Example- Predicting a class label using Naïve Bayesian Classification:

rid	age	income	student	credit_rating	Class: buys_computer
1	<30	high	no	fair	no
2	<30	high	no	excellent	no
3	30-40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	30-40	low	yes	excellent	yes
8	<30	medium	no	fair	no
9	<30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<30	medium	yes	excellent	yes
12	30-40	medium	no	excellent	yes
13	30-40	high	yes	fair	yes
14	>40	medium	no	excellent	no

Let C1 = Class buys_computer = yes; C2 = Class buys_computer = no

Let us try to classify an unknown sample:

- X = (age = "<30", income-rating= = medium,fair) stud

We need to maximize P(X/Ci) P(Ci) for i = 1,2; So, Compute P(C1) & P(C2)

- P(buys_computer = yes) = 9/14 = 0.643 P(buys_computer = no) = 5/14 = 0.357

Next, Compute P(X/C1) & P(X/C2)

P(age = "<30" / buys_computer = yes) =

P(income = medium / buys_computer = yes) = 4/9 = 0.444

P(student = yes / buys_computer = yes) = 6/9 = 0.667

P(credit-rating = fair / buys_computer = yes) = 6/9 = 0.667

P(age = "<30" / buys_computer = no) = 3

P(income = medium / buys_computer = no) = 2/5 = 0.400

P(student = yes / buys_computer = no) = 1/5 = 0.200

P(credit-rating = fair / buys_computer = no) = 2/5 = 0.400

Hence P(X/buys_computer = yes) = 0.222 * 0.444 * 0.667 * 0.667 = 0.044

P(X/buys_computer = no) = 0.600 * 0.400 * 0.200 * 0.400 = 0.019

Finally P(X/buys_computer = yes) P(buys_computer = yes) = 0.044 * 0.643 = 0.028

P(X/buys_computer = no) P(buys_computer = no) = 0.019 * 0.357 = 0.007

Hence Naïve Bayesian Classifier predict

3) Explain various attribute selection measure in classification. May'14

An attribute selection measure is a heuristic for selecting the splitting criterion that "best" separates a given data partition, D, of class-labeled training tuples into individual classes. If we were to split D into smaller partitions according to the outcomes of the splitting criterion, ideally each partition would be pure (i.e., all of the tuples that fall into a given partition would belong to the same class).

Conceptually, the "best" splitting criterion is the one that most closely results in such a scenario. Attribute selection measures are also known as splitting rules because

they determine how the tuples at a given node are to be split. The attribute selection measure provides a ranking for each attribute describing the given training tuples. The attribute having the best score for the measure6 is chosen as the splitting attribute for the given tuples.

If the splitting attribute is continuous-valued or if we are restricted to binary trees then, respectively, either a split point or a splitting subset must also be determined as part of the splitting criterion.

The tree node created for partition D is labeled with the splitting criterion, branches are grown for each outcome of the criterion, and the tuples are partitioned accordingly. This section describes three popular attribute selection measures—information gain, gain ratio, and Gini index.

The **notation used herein is as follows.**

Let D, the data partition, be a training set of class-labeled tuples. Suppose the class label attribute has m distinct values defining m distinct classes, Ci (for i = 1, :::, m). Let Ci,D be the set of tuples of class Ci in D. Let jDj and jCi,Dj denote the number of tuples in D and Ci, D, respectively.

Information gain

The expected information needed to classify a tuple in D is given by

$$Info(D) = \sum_{i=1}^{m} pi \log2(pi);$$

Where pi is the probability that an arbitrary tuple in D belongs to class Ci and is estimated by jCi,Dj/jDj. A log function to the base 2 is used, because the information is encoded in bits. Info(D) is just the average amount of information needed to identify the class label of a tuple in D. Note that, at this point, the information we have is based solely on the proportions of tuples of each class. Info(D) is also known as the entropy of D.

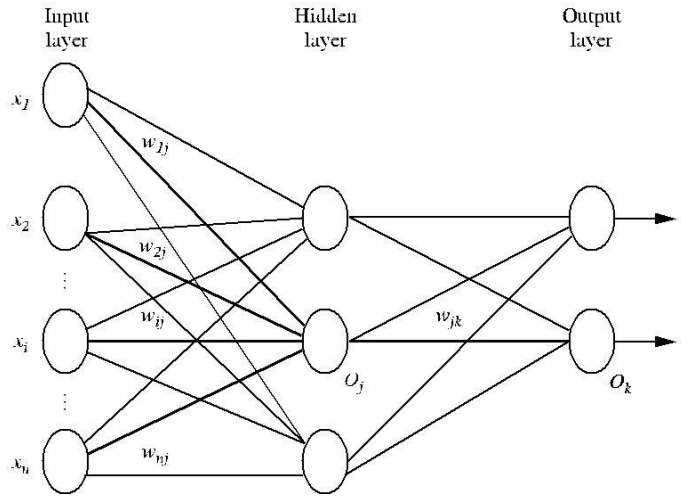
4) Explain as to how neural networks are used for classification of data. Dec'13

The backpropagation algorithm performs learning on a multilayer feed-forward neural network. It iteratively learns a set of weights for prediction of the class label of tuples. A multilayer feed-forward neural network consists of an input layer, one or more hidden layers, and an output layer. An example of a multilayer feed-forward network is shown in Figure 1.

Each layer is made up of units. The inputs to the network correspond to the attributes measured for each training tuple. The inputs are fed simultaneously into the units making up the input layer. These inputs pass through the input layer and are then weighted and fed simultaneously to a second layer of "neuronlike" units, known as a hidden layer. The outputs of the hidden layer units can be input to another hidden layer, and so on.

The number of hidden layers is arbitrary, although in practice, usually only one is used. The weighted outputs of the last hidden layer are input to units making up the output layer, which emits the network's prediction for given tuples.

The units in the input layer are called input units. The units in the hidden layers and output layer are sometimes referred to as neurodes, due to their symbolic biological basis, or as output units. The multilayer neural network shown in Figure 1.



of output units. Therefore, we say that it is a two-layer neural network. (The input layer is not counted because it serves only to pass the input values to the next layer.) Similarly, a network containing two hidden layers is called a three-layer neural network, and so on. The network is feed-forward in that none of the weights cycles back to an input unit or to an output unit of a previous layer. It is fully connected in that each unit provides input to each unit in the next forward layer.

Algorithm: Backpropagation. Neural network learning for classification or prediction, usingthe backpropagation algorithm.

Input:

D, a data set consisting of the training tuples and their associated target values;
l, the learning rate;
network, a multilayer feed-forward network.

Output: A trained neural network.

Method:

(1) Initialize all weights and biases in *network*;
(2) **while** terminating condition is not satisfiedf
(3) **for** each training tuple***X***in*D*f
(4) // Propagate the inputs forward:
(5) **for** each input layer unit*j*f
$O_j = I_j$; // output of an input unit is its actual input
(6) value
(7) **for** each hidden or output layer unit*j*f
(8) $I_j = \sum_i w_{ij} O_i + \theta_j$;//compute the net input of unit *j*

(9) j=i

(10) // Backpropagate the errors:

(11) **for** each unit*j*in the output layer

(12) $Err_j = O_j(1 - O_j)(T_j - O_j)$; // compute the error

(13) **for** each unit*j*in the hidden layers, from the last to the first hidden layer

(14) $Err_j = O_j(1 - O_j)\sum_k Err_k w_{jk}$; // compute the error with respect to the next higher layer, *k*

(15) **for** each weightw_{ij}in*network*f

$w_{ij} = (l)Err_j O_i$; // weight

(16) increment

$w_{ij} = w_{ij} + w_{ij}$;g// weight

(17) update

(18) **for** each bias$\theta_j$in*network*f

(19) $\theta_j = (l)Err_j$; // bias increment

(20) $\theta_j = \theta_j + \theta_j$; g // bias update

5) Explain how support vector machines (SVM) can be used for classification. May'15, Dec'11

Support Vector Machines, a promising new method for the classification of both linear and nonlinear data. In a nutshell, a support vector machine (or SVM) is an algorithm that works as follows. It uses a nonlinear mapping to trans-form the original training data into a higher dimension. Within this new dimension, it searches for the linear optimal separating hyperplane (that is, a "decision boundary" separating the tuples of one class from another).

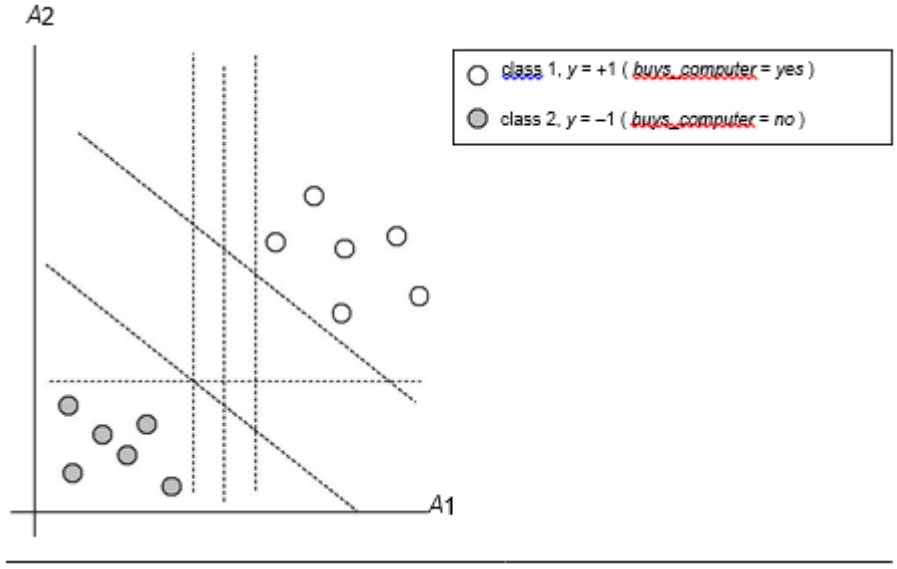
With an appropriate nonlinear map-ping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane. The SVM finds this

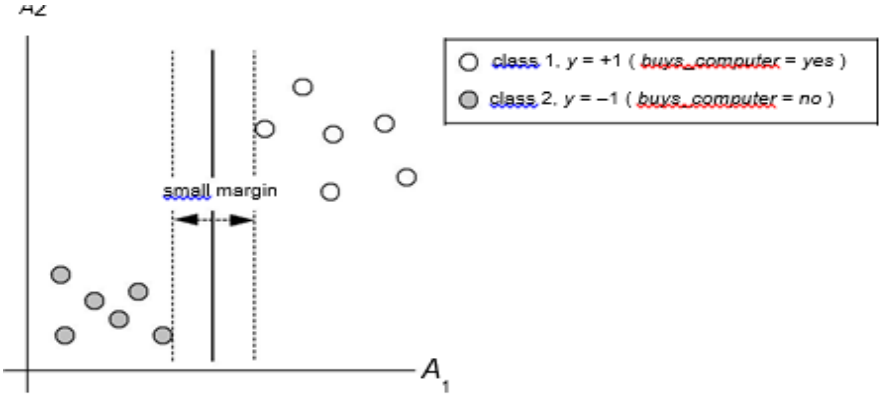
hyperplane using support vectors ("essential" training tuples) and margins (defined by the support vectors)

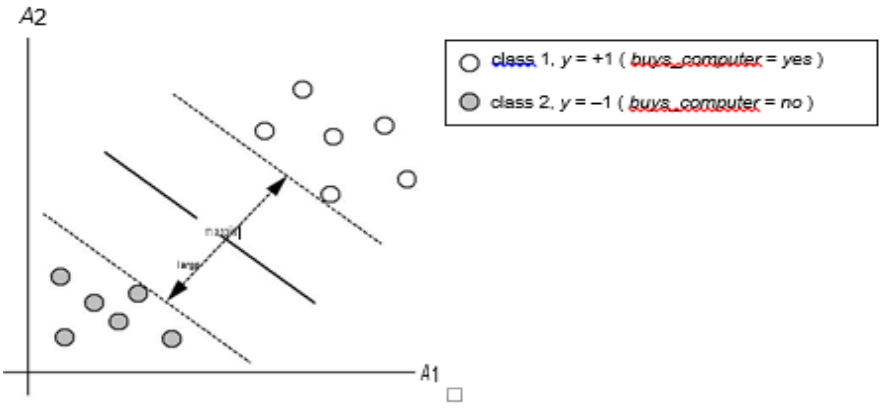
1) The Case When the Data Are Linearly Separable

An SVM approaches this problem by searching for the maximum marginal hyperplane. Consider the below Figures, which shows two possible separating hyperplanes and their associated margins. Before we get into the definition of margins. Both hyperplanes can correctly classify all of the given data tuples. Intuitively, however, we expect the hyperplane with the larger margin to be more accurate at classifying future data tuples than the hyperplane with the smaller margin.

This is why (during the learning or training phase), the SVM searches for the hyperplane with the largest margin, that is, the maximum marginal hyperplane (MMH). The associated margin gives the largest separation between classes. Getting to an informal definition of margin, we can say that the shortest distance from a hyperplane to one side of its margin is equal to the shortest distance from the hyperplane to the other side of its margin, where the ―sides‖ of the margin are parallel to the hyperplane. When dealing with the MMH, this distance is, in fact, the shortest distance from the MMH to the closest training tuple of either class.







2) The Case When the Data Are Linearly Inseparable

We learned about linear SVMs for classifying linearly separable data, but what if the data are not linearly separable no straight line cajn be found that would separate the classes. The linear SVMs we studied would not be able to find a feasible solution here. Now what?

The good news is that the approach described for linear SVMs can be extended to create nonlinear SVMs for the classification of linearly inseparable data (also called nonlinearly separable data, or nonlinear data, for short). Such SVMs are capable of finding nonlinear decision boundaries (i.e., nonlinear hypersurfaces) in input space.

"So," you may ask, "how can we extend the linear approach?" We obtain a nonlinear SVM by extending the approach for linear SVMs as follows. ***There are two main steps.***

In the first step, we transform the original input data into a higher dimensional space using a nonlinear mapping. Several common nonlinear mappings can be used in this step.

Once the data have been transformed into the new higher space, the second step searches for a linear separating hyperplane in the new space.

We again end up with a quadratic optimization problem that can be solved using the linear SVM formulation. The maximal marginal hyperplane found in the new space corresponds to a nonlinear separating hypersurface in the original space

6.Explain about constraint based association rule mining (May 2017)

Constraint-Based Association Mining

A data mining process may uncover thousands of rules from a given set of data, most of which end up being unrelated or uninteresting to the users. Often, users have a good sense of which directionof mining may lead to interesting patterns and the form of the

patterns or rules they would like to find. Thus, a good heuristic is to have the users specify such intuition or expectations as constraints to confine the search space. This strategy is known as constraint-based mining. The constraints can include the following:

- **Knowledge type constraints:** These specify the type of knowledge to be mined, such as association or correlation.
- **Data constraints:** These specify the set of task-relevant data.
- **Dimension/level constraints:** These specify the desired dimensions (or attributes) of the data, or levels of the concept hierarchies, to be used in mining.
- **Interestingness constraints:** These specify thresholds on statistical measures of rule interestingness, such as support, confidence, and correlation.
- **Rule constraints:** These specify the form of rules to be mined. Such constraints may be expressed as metarules (rule templates), as the maximum or minimum number of predicates that can occur in the rule antecedent or consequent, or as relationships among attributes, attribute values, and/or aggregates.

a. Metarule-Guided Mining of Association Rules

"How are metarules useful?" Metarules allow users to specify the syntactic form of rules that they are interested in mining. The rule forms can be used as constraints to help improve the efficiency of the mining process. Metarules may be based on the analyst's experience, expectations, or intuition regarding the data or may be automatically generated based on the database schema.

Metarule-guided mining:- Suppose that as a market analyst for AllElectronics, you have access to the data describing customers (such as customer age, address, and credit rating) as well as the list of customer transactions. You are interested in finding associations between customer traits and the items that customers buy. However, rather than finding all of the association rules reflecting these relationships, you are particularly interested only in determining which pairs of customer traits SCE Department of Information Technology promote the sale of office software.A metarule can be used to specify this information describing the form of rules you are interested in finding. An example of such a metarule is

$$P_1(X, Y) \wedge P_2(X, W) \Rightarrow buys(X, \text{“office software”}),$$

where P1 and P2 are predicate variables that are instantiated to attributes from the given database during the mining process, X is a variable representing a customer, and Y and W take on values of the attributes assigned to P1 and P2, respectively. Typically, a user will specify a list of attributes to be considered for instantiation with P1 and P2. Otherwise, a default set may be used.

b. Constraint Pushing: Mining Guided by Rule Constraints

Rule constraints specify expected set/subset relationships of the variables in the mined rules, constant initiation of variables, and aggregate functions. Users typically employ their knowledge of the application or data to specify rule constraints for the mining task. These rule constraints may be used together with, or as an alternative to, metarule-guided mining. In this section, we examine rule constraints as to how they can be used to make the mining process more efficient. Let’s study an example where rule constraints are used to mine hybrid-dimensional association rules.

Our association mining query is to “Find the sales of which cheap items (where the sum of the prices is less than $100) may promote the sales of which expensive items (where the minimum price is $500) of the same group for Chicago customers in 2004.” This can be expressed in the DMQL data mining query language as follows,

(1) mine associations as
(2) $lives\_in(C, \_, \text{“}Chicago\text{”}) \wedge sales^{+}(C, ?\{I\}, \{S\}) \Rightarrow sales^{+}(C, ?\{J\}, \{T\})$
(3) from sales
(4) where S.year = 2004 and T.year = 2004 and I.group = J.group
(5) group by C, I.group
(6) having sum(I.price) < 100 and min(J.price) ≥ 500
(7) with support threshold = 1%
(8) with confidence threshold = 50%

UNIT V

. CLUSTERING AND TRENDS IN DATAMINING

Cluster Analysis - Types of Data – Categorization of Major Clustering Methods – K-means– Partitioning Methods –Hierarchical Methods - Density-Based Methods –Grid Based Methods – Model- Based Clustering Methods–Clustering HighDimensional Data-Constraint– Based Cluster Analysis– OutlierAnalysis–Data Mining Applications

PART A

1) Define Outlier & its applications. Dec'15, May'15, Dec'14, Dec'13, May'13, Dec'11

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. Most data mining methods discard outliers as noise or exceptions. Applications – Credit card fraud detection, Crime detection, Medical diagonosis etc.,

These can be categorized into four approaches:

- Statistical approach,
- Distance based approach,
- Density-based local outlier approach,
- Deviation-based approach.

2) What are the applications of data mining? Dec'15, May'11,May 2017

- Financial Data Analysis
- Retail Industry
- Telecommunication Industry

- Biological Data Analysis
- Scientific Applications
- Intrusion Detection

3) Define Divisive hierarchical clustering. / Types of hierarchical clustering. Dec'14, May'13

Divisive Hierarchical clustering method works on the top-down approach. In this method all the objects are arranged within a big singular cluster and the large cluster is continuously divided into smaller clusters until each cluster has a single object

Types of hierarchical clustering

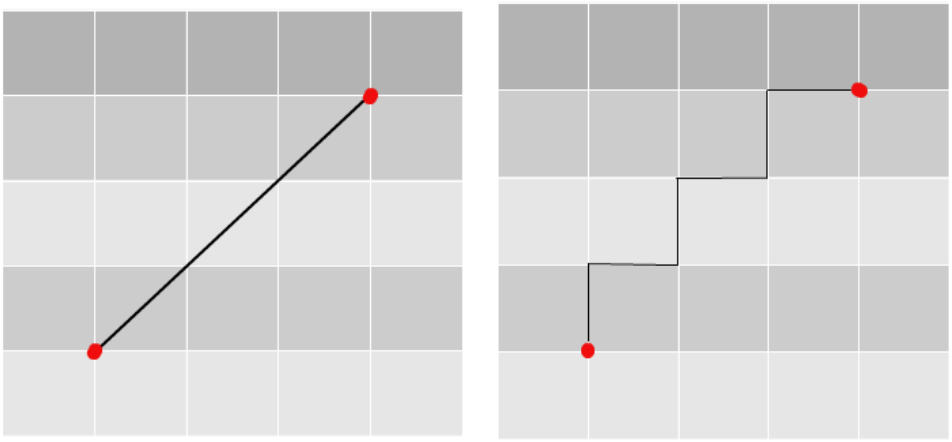
- **Agglomerative:** This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- **Divisive:** This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

4) Define Euclidean distance &Manhattan distance. May'15, May'11

Euclidean distance – Euclidean distance is a classical method helps compute distance between two objects A and B in Euclidean space (1- or 2- or n- dimension space). In Euclidean geometry, the distance between the points can be found by traveling along the line connecting the points. Inherently in the calculation you use the Pythagorean Theorem to compute the distance.

$$dist = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$$

Taxicab or Manhattan distance – Similar to Euclidean distance between point A and B but only difference is the distance is calculated by traversing the vertical and horizontal line in the grid base system. Example, Manhattan distance used to calculate distance between two points that are geographically separated by the building blocks in the city.The difference between these two distance calculations is
best seen visually. Figure illustrates the difference.



Euclidean distance

Taxicab or Manhattan distance

5) Define CLARANS.

CLARANS(Cluster Large Applications based on Randomized Search) to improve the quality of CLARA we go for CLARANS. It Draws sample with some randomness in each step of search.□

It overcome the problem of scalability that K-Medoids suffers from.

6) Define Density based method?

Density based method deals with arbitrary shaped clusters. In density-based method, clusters are formed on the basis of the region where the density of the objects is high.

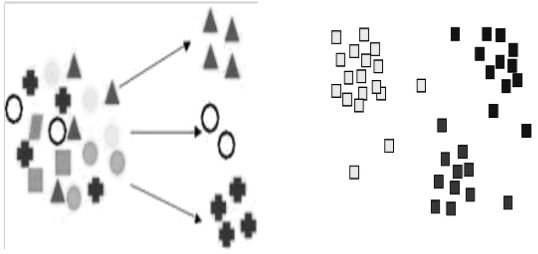
7) Define BIRCH,ROCK and CURE.

- **BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies):** Partitions objects hierarchically using tree structures and then refines the clusters using other clustering methods.it defines a clustering feature and an associated tree structure that summarizes a cluster. The tree is a height balanced tree that stores cluster information. BIRCH doesn't Produce spherical Cluster and may produce unintended cluster.
- **ROCK (RObust Clustering using links):** Merges clusters based on their interconnectivity. Great for categorical data. Ignores information about the looseness of two clusters while emphasizing interconnectivity.
- **CURE (Clustering Using Representatives):** Creates clusters by sampling the database and shrinks them toward the center of the cluster by a specified fraction. Obviously better in runtime but lacking in precision.

8) State the role of clustering. Dec'11 (Nov/Dec 2016)

Clustering is a process of grouping the physical or conceptual data object into clusters.

Examples of Clustering



9) Define STING. May'14

Statistical Information Grid is called as STING; it is a grid based multi resolution clustering method. In STING method, all the objects are contained into rectangular cells, these cells are kept into various levels of resolutions and these levels are arranged in a hierarchical structure.

10) What is wave cluster? May'14

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.

11) What is a DBSCAN?

Density Based Spatial Clustering of Application Noise is called as DBSCAN. DBSCAN is a density based clustering method that converts the high-density objects regions into clusters with arbitrary shapes and sizes. DBSCAN defines the cluster as a maximal set of density connected points.

12) Give an example of outlier analysis for the library management system. (Nov/Dec 2016)

In library management systems, among set of books accessed frequently, if one book is found to be accessed never, that is considered as an outlier. Likewise, a missing book can also be considered as an outlier.

13) Give the reason on why clustering is needed in data mining?(Nov/Dec 2016)

Clustering is needed to identify set of similar data objects. The objects are similar in terms of multiple dimensions. By considering only similar data objects for further mining improves the interestingness of retrieved knowledge and accuracy

PART B

1) Illustrate the (K-means) partition based clustering algorithm with a suitable example. Dec'15, May'15 (M.E), May'13, Dec'11, May'11

Themostwell-knownandcommonlyused partitioningmethodsarek-means,k-medoids, and their variations

Given a database of n objects or data tuples, a partitioning method constructs k partitions of the data, where each partition represents a cluster and k≤n. That is, itclassifies the data into k groups, which together satisfy the followingrequirements:

(1) Each groupmustcontain at least one object,

(2) Each object must belong to exactly one group

K means Clustering

Clustering is the process of partitioning a group of data points into a small number of clusters. For instance, the items in a supermarket are clustered in categories (butter, cheese and milk are grouped in dairy products). Of course this is a qualitative kind of partitioning. A quantitative approach would be to measure certain features of the products, say percentage of milk and others, and products with high percentage of milk would be grouped together.

In general, we have n data points xi,i=1...nthat have to be partitioned in k clusters. The goal is to assign a cluster to each data point. K-means is a clustering method that aims to find the positions μi,i=1...k of the clusters that minimize the distance from the data points to the cluster. K-means clustering solves

$$\mathbf{argmin_c \sum_{i=1}^{k} \sum_{x \in ci} d(x,\mu i) = argmin_c \sum_{i=1}^{k} \sum_{x \in ci} \|x-\mu i\|_2^2}$$

Where ci is the set of points that belong to cluster i. The K-means clustering uses the square of the Euclidean distance $d(x,\mu i)=\|x-\mu i\|22$. This problem is not trivial (in fact it is NP-hard), so the K-means algorithm only hopes to find the global minimum, possibly getting stuck in a different solution

K-means algorithm

The Lloyd's algorithm, mostly known as k-means algorithm, is used to solve the k-means clustering problem and works as follows. First, decide the number of clusters k. Then:

1. Initialize the center of the clusters	μi= some value ,i=1,...,k
2. Attribute the closest cluster to each data point	$ci=\{j:d(xj,\mu i)\leq d(xj,\mu l),l\neq i,j=1,...,n\}$
3. Set the position of each cluster to the mean of all data points belonging to that cluster	$\mu i=1\|ci\\|\sum j\in cixj,\forall i$
4. Repeat steps 2-3 until convergence	
Notation	|c|= number of elements in c

Simple illustration of a k-means algorithm

Consider the following data set consisting of the scores of two variables on each of seven individuals:

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

This data set is to be grouped into two clusters. As a first step in finding a sensible initial partition, let the A & B values of the two individuals furthest apart (using the Euclidean distance measure), define the initial cluster means, giving:

	Individual	Mean Vector (centroid)
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

The remaining individuals are now examined in sequence and allocated to the cluster to which they are closest, in terms of Euclidean distance to the cluster mean. The mean vector is recalculated each time a new member is added. This leads to the following series of steps:

	Cluster 1		Cluster 2	
Step	Individual	Mean Vector (centroid)	Individual	Mean Vector (centroid)
1	1	(1.0, 1.0)	4	(5.0, 7.0)
2	1, 2	(1.2, 1.5)	4	(5.0, 7.0)
3	1, 2, 3	(1.8, 2.3)	4	(5.0, 7.0)
4	1, 2, 3	(1.8, 2.3)	4, 5	(4.2, 6.0)
5	1, 2, 3	(1.8, 2.3)	4, 5, 6	(4.3, 5.7)
6	1, 2, 3	(1.8, 2.3)	4, 5, 6, 7	(4.1, 5.4)

Now the initial partition has changed, and the two clusters at this stage having the following characteristics:

	Individual	Mean Vector (centroid)
Cluster 1	1, 2, 3	(1.8, 2.3)
Cluster 2	4, 5, 6, 7	(4.1, 5.4)

But we cannot yet be sure that each individual has been assigned to the right cluster. So, we compare each individual's distance to its own cluster mean and to that of the opposite cluster.

And we find:

Individual	Distance to mean (centroid) of Cluster 1	Distance to mean (centroid) of Cluster 2
1	1.5	5.4
2	0.4	4.3
3	2.1	1.8
4	5.7	1.8
5	3.2	0.7
6	3.8	0.6
7	2.8	1.1

Only individual 3 is nearer to the mean of the opposite cluster (Cluster 2) than its own (Cluster 1). In other words, each individual's distance to its own cluster mean should be smaller that the distance to the other cluster's mean (which is not the case with individual 3). Thus, individual 3 is relocated to Cluster 2 resulting in the new partition:

	Individual	Mean Vector (centroid)
Cluster 1	1, 2	(1.3, 1.5)
Cluster 2	3, 4, 5, 6, 7	(3.9, 5.1)

The iterative relocation would now continue from this new partition until no more relocations occur. However, in this example each individual is now nearer its own cluster

mean than that of the other cluster and the iteration stops, choosing the latest partitioning as the final cluster solution.

Also, it is possible that the k-means algorithm won't find a final solution. In this case it would be a good idea to consider stopping the algorithm after a pre-chosen maximum of iterations.

2) What is hierarchical clustering and density based clustering? / Agglomerative hierarchical. Dec'15, May'15, Dec'14, Dec'13,May'17

Hierarchical Clustering Methods

- Hierarchical clustering (or hierarchic clustering) outputs a hierarchy, a structure that is more informative than the unstructured set of clusters returned by flat clustering.
- Hierarchical clustering does not require us to pre specify the number of clusters and most hierarchical algorithms that have been used in IR are deterministic. These advantages of hierarchical clustering come at the cost of lower efficiency

Hierarchical clustering methods can be further classified as

- Agglomerative hierarchical clustering
- Divisive hierarchical clustering:

Agglomerative hierarchical clustering:

- This bottom-up strategy starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters, until all of the objects are in a single cluster or until certain termination conditions are satisfied.
- Most hierarchical clustering methods belong to this category. They differ only in their definition of inter cluster similarity.

Divisive hierarchical clustering:

- This top-down strategy does the reverse of agglomerative hierarchical clustering by starting with all objects in one cluster.

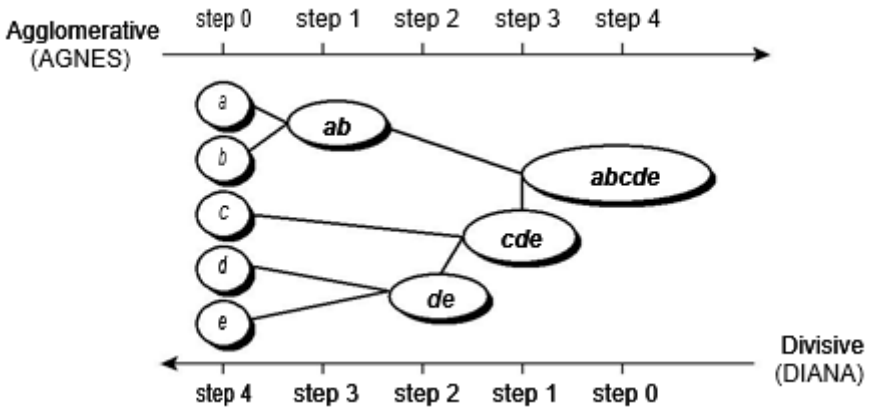
- It subdivides the cluster into smaller and smaller pieces, until each object forms a cluster on its own or until it satisfies certain termination conditions, such as a desired number of clusters is obtained or the diameter of each cluster is within a certain threshold

AGNES (AGglomerativeNESting), an agglomerative hierarchical clustering method, and **DIANA (DIvisiveANAlysis)**, a divisive hierarchical clustering method, to a data set of five objects, fa, b, c, d, eg.

Initially, AGNES places each object into a cluster of its own. The clusters are then merged step-by-step according to some criterion.

For example, clusters C1 and C2 may be merged if an object in C1 and an object in C2 form the minimum Euclidean distance between any two objects from different clusters. This is a single-linkage approach in that each cluster is represented by all of the objects in the cluster, and the similarity between two clusters is measured by the similarity of the closest pair of data points belonging to different clusters. The cluster merging process repeats until all of the objects are eventually merged to form one cluster.

In DIANA, all of the objects are used to form one initial cluster. The cluster is split according to some principle, such as the maximum Euclidean distance between the clos-est neighboring objects in the cluster. The cluster splitting process repeats until, eventually, each new cluster contains only a single object



Hierarchical clustering dendrogram - Figure 1

Agglomerative and divisive hierarchical clustering on data objectsf***a, b, c, d, e***g.

In either agglomerative or divisive hierarchical clustering, the user can specify the desired number of clusters as a termination condition.

A tree structure called a dendrogram is commonly used to represent the process of hierarchical clustering. It shows how objects are grouped together step by step. Figure 2 shows a dendrogram for the five objects presented in Figure 1, where $l = 0$ shows the five objects as singleton clusters at level 0. At $l = 1$, objects a and b are grouped together to form the first cluster, and they stay together at all subsequent levels. We can also use a vertical axis to show the similarity scale between clusters. For example, when the similarity of two groups of objects, fa, bg and fc, d, eg, is roughly 0.16, they are merged together to form a single cluster.

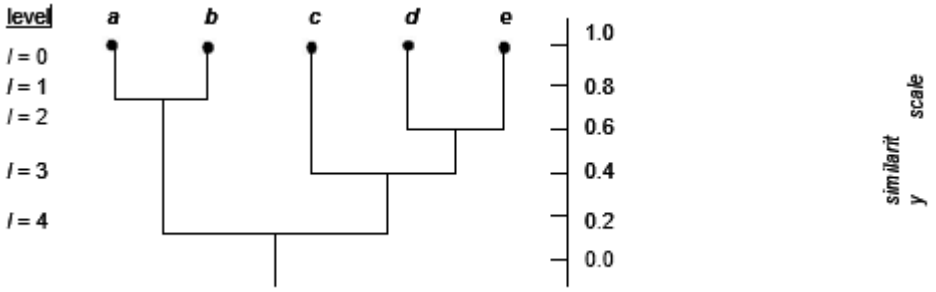
Minimum distance : $d_{min}(C_i,C_j) = min_{\mathbf{p}2Ci,\mathbf{p}}0_{2Cj}\mathrm{j}\mathbf{p} \;\; \mathbf{p}^0\mathrm{j}$

Maximum distance : $d_{max}(C_i,C_j) = max_{\mathbf{p}2Ci,\mathbf{p}}0_{2Cj}\mathrm{j}\mathbf{p} \;\; \mathbf{p}^0\mathrm{j}$

Mean distance

$$: d_{mean}(C_i,C_j) = \mathrm{j}\boldsymbol{m_i m_j}\mathrm{j}$$

When an algorithm uses the minimum distance, dmin(Ci, Cj), to measure the distance between clusters, it is sometimes called **a nearest-neighbor clustering algorithm.** More-over, if the clustering process is terminated when the distance between nearest clusters exceeds an arbitrary threshold, it is called a single-linkage algorithm. If we view the data points as nodes of a graph, with edges forming a path between the nodes in a cluster, then the merging of two clusters, Ci and Cj, corresponds to adding an edge between



Hierarchical clustering dendrogram -- Figure 2

the nearest pair of nodes in C_i and C_j. Because edges linking clusters always go between distinct clusters, the resulting graph will generate a tree. Thus, an agglomerative hierarchical clustering algorithm that uses the minimum distance measure is also called a **minimal spanning tree algorithm**.

When an algorithm uses the *maximum distance*, $d_{max}(C_i, C_j)$, to measure the distance between clusters, it is sometimes called a **farthest-neighbor clustering algorithm**. If the clustering process is terminated when the maximum distance between nearest clusters exceeds an arbitrary threshold, it is called a **complete-linkage algorithm**. By viewing data points as nodes of a graph, with edges linking nodes, we can think of each cluster as a *complete* subgraph, that is, with edges connecting all of the nodes in the

clusters. The distance between two clusters is determined by the most distant nodes in the two clusters. Farthest-neighbor algorithms tend to minimize the increase in diameter of the clusters at each iteration as little as possible. If the true clusters are rather compact and approximately equal in size, the method will produce high-quality clusters. Otherwise, the clusters produced can be meaningless.

The above minimum and maximum measures represent two extremes in measuring the distance between clusters. They tend to be overly sensitive to outliers or noisy data. The use of *mean* or *average distance* is a compromise between the minimum and maximum distances and overcomes the outlier sensitivity problem. Whereas the *meandistance* is the simplest to compute, the *average distance* is advantageous in that it canhandle category as well as numeric data.The computation of the mean vector for categorical data can be difficult or impossible to define.

Density-Based Methods

To discover clusters with arbitrary shape, density-based clustering methods have been developed. These typically regard clusters as dense regions of objects in the data space that are separated by regions of low density (representing noise). DBSCAN grows clusters according to a density-based connectivity analysis. OPTICS extends DBSCAN to produce a *cluster ordering* obtained from a wide range of parameter settings. DENCLUE clusters objects based on a set of density distribution functions.

DBSCAN: A Density-Based Clustering Method Based on Connected Regions with Sufficiently High Density

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm. The algorithm grows regions with sufficiently high density into clusters and discovers clusters of arbitrary shape in spatial databases with noise. It defines a cluster as a maximal set of *density-connected* points.

The basic ideas of density-based clustering involve a number of new definitions. We intuitively present these definitions, and then follow up with an example.

- The neighborhood within a radius ε of a given object is called the ε-**neighborhood** of the object.

- If the ε-neighborhood of an object contains at least a minimum number, *MinPts*, of objects, then the object is called a **core object**.

- Given a set of objects, *D*, we say that an object ***p*** is **directly density-reachable** from object ***q*** if ***p*** is within the ε-neighborhood of ***q***, and ***q*** is a core object.

- An object ***p*** is **density-reachable** from object ***q*** with respect to ε and *MinPts* in a set of objects, *D*, if there is a chain of objects $\boldsymbol{p_1}$, :::, $\boldsymbol{p_n}$, where $\boldsymbol{p_1} = \boldsymbol{q}$ and $\boldsymbol{p_n} = \boldsymbol{p}$ such that $\boldsymbol{p_{i+1}}$is directly density-reachable from $\boldsymbol{p_i}$with respect toεand*MinPts*, for1*i n*, $\boldsymbol{p_i}$2*D*.

- An object ***p*** is **density-connected** to object ***q*** with respect to ε and *MinPts* in a set of objects, *D*, if there is an object **o** 2 *D* such that both ***p*** and ***q*** are density-reachable from ***o*** with respect to ε and *MinPts*.

Density reachability is the transitive closure of direct density reachability, and this relationship is asymmetric. Only core objects are mutually density reachable. Density connectivity, however, is a symmetric relation.

Example Density-reachability and density connectivity. Consider Figure 3 for a givenεrepresented by the radius of the circles, and, say, let *MinPts* = 3. Based on the above definitions:

- Of the labeled points, ***m***, ***p***, ***o***, and ***r*** are core objects because each is in an ε-neighborhood containing at least three points.

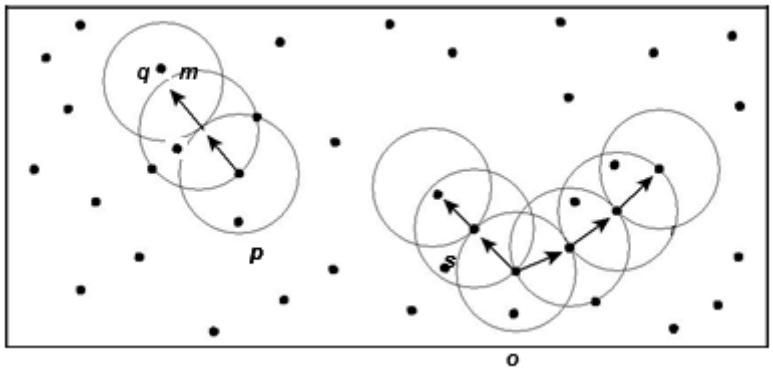
- ***q***is directly density-reachable from ***m***. ***m*** is directly density-reachable from ***p*** and viceversa.

- ***q***is (indirectly) density-reachable from ***p*** because ***q*** is directly density-reachable from ***m*** and ***m*** is directly density-reachable from ***p***. However, ***p*** is not density-reachablefrom ***q*** because ***q*** is not a core object. Similarly, ***r*** and ***s*** are density-reachable from ***o***, and ***o*** is density-reachable from ***r***.

- ***o***, ***r***, and ***s*** are all density-connected.

A **density-based cluster** 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*.

"How does DBSCAN find clusters?" DBSCAN searches for clusters by checking theε-neighborhood of each point in the database. If the ε-neighborhood of a point ***p*** contains more than *MinPts*, a new cluster with ***p*** as a core object is created. DBSCAN then iteratively collects directly density-reachable objects from these core objects, which may involve the merge of a few density-reachable clusters. The process terminates when no new point can be added to any cluster.



Density reachability and density connectivity in density-based clustering – Figure 3

If a spatial index is used, the computational complexity of DBSCAN is $O(n \log n)$, where n is the number of database objects. Otherwise, it is $O(n^2)$. With appropriate set-tings of the user-defined parameters ε and *MinPts*, the algorithm is effective at finding arbitrary-shaped clusters.

3) Write short notes on

a.) Outlier analysis. May'15, Dec'13, May'13

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

There exist data objects that 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. Many data mining algorithms try to minimize the influence of outliers or eliminate them all together. This, however, could result in the loss of important hidden information because one person's noise could be another person's signal. In other words, 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 task, referred to as outlier mining. It can be used in fraud detection,

For example, by detecting unusual usage of credit cards or telecommunication services. In addition, it is useful in customized marketing for identifying the spending behavior of customers with extremely low or extremely high incomes, or in medicalanalysis for finding unusual responses to various medical treatments.

Outlier mining can be described as follows:

Given a set of n data points or objectsand k, the expected number of outliers, find the top k objects that are considerablydissimilar, exceptional, or inconsistent with respect to the remaining data.

The outliermining problem can be viewed as two subproblems:

Define what data can be considered as inconsistent in a given data set, and Find an efficient method to mine the outliers so defined.

Types of outlier detection:

- Statistical Distribution-Based Outlier Detection
- Distance-Based Outlier Detection
- Density-Based Local Outlier Detection
- Deviation-Based Outlier Detection

4) Write short notes on Spatial data mining. May'15

Spatial Data Mining refers to the extraction of knowledge, spatial relationships or other interesting patterns not explicitly stored in spatial databases. A spatial database stores a large amount of space-related data, such as maps, preprocessed remote sensing or medical imaging data, and VLSI chip layout data. Statistical spatial data analysis has been a popular approach to analyzing spatial data and exploring geographic information. The term 'geostatistics' is often associated with continuous geographic space, whereas the term 'Spatial statistics' is often associated with discrete space.

Spatial Data Mining Applications:

- Geographic information systems

- Geo marketing
- Remote sensing
- Image database exploration
- Medical Imaging Navigation
- Traffic Control
- Environmental Studies

Spatial Data Cube Construction and Spatial OLAP:

Spatial data warehouse is a subject-oriented integrated, time-variant and non-volatile collection of both spatial and non-spatial data in support of spatial data mining and spatial data related decision-making process.

There are three types of dimensions in a Spatial Data Cube:

A non-spatial dimension contains only non-spatial data, each contains nonspatial data whose generalizations are non-spatial. A Spatial-to-nonspatial dimension is a dimension whose primitive-level data are spatial but whose generalization, starting at a certain high level, becomes non-spatial. A Spatial-to-Spatial dimension is a dimension whose primitive level and all of its high level generalized data are spatial. **Measures of Spatial Data Cube:**

A numerical measure contains only numeric data. A Spatial measure contains a collection of pointers to spatial objects.

Computation of Spatial Measures in Spatial Data Cube Construction:

Collect and store the corresponding spatial object pointers but do not perform precomputation of spatial measures in the spatial data cube.

Precompute and store a rough approximation of the spatial measures in the spatial data cube. Selectively pre-compute some spatial measures in the spatial data cube.

Mining Spatial Association and Co-Location Pattern:

Spatial Association rules can be mined in spatial databases. A Spatial association rule is of the form A -> B [s%, c%] where A & B are sets of spatial or non-spatial predicates. S% is the support of the rule; c% is the confidence of the rule

Spatial Association mining needs to evaluate multiple spatial relationships among a large number of spatial objects, the process could be quite costly. An interesting mining optimization called 'progressive refinement' can be adopted in spatial association analysis. The method first mines large data sets roughly using a fast algorithm and then improves the quality of mining in a pruned data set using a more expensive algorithm

Superset Coverage Property:

It should allow a false-positive test, which might include some data sets that do not belong to the answer sets, but it should not allow a 'false-negative test', which might exclude some potential answers.

For mining spatial associations related to the spatial predicate close to and collect the candidates that pass the minimum support threshold by applying certain rough spatial evaluation algorithms. Evaluating the relaxed spatial predicate, 'g close to', which is generalized close to covering a broader context that includes 'close to', 'touch' and intersect'

Spatial Clustering methods:

Spatial data clustering identifies clusters, or densely populated regions, according to some distance measurement in a large, multi dimensional data set.

Spatial Classification and Spatial Trend Analysis:

Spatial Classification analyzes spatial objects to derive classification schemes in relevance to certain spatial properties. Example: Classify regions in a province into rich Vs poor according to the average family income. Trend analysis detects changes with time, such as the changes of temporal patterns in time-series data. Spatial trend analysis replaces time with space and studies the trend of non-spatial or spatial data changing with space. Example: Observe the trend of changes of the climate or vegetation with the

increasing distance from an ocean. Regression and correlation analysis methods are often applied by utilization of spatial data structures and spatial access methods.

Mining Raster Databases:

Spatial database systems usually handle vector data that consists of points, lines, polygons (regions) and their compositions, such as networks or partitions. Huge amounts of space-related data are in digital raster forms such as satellite images, remote sensing data and computer tomography.

5) Write Short notes on Text mining. May'15

Text Databases and Information Retrieval:

Text Databases (Document Databases):

Large collections of documents from various sources, new articles, research papers, books, digital libraries, email messages and web pages, library databases etc.

Data stored is usually semi-structured

Traditional information retrieval techniques become inadequate for the increasingly vast amounts of text data.

Information Retrieval:

A field developed in parallel with database systems.

Information is organized into a large number of documents

Information retrieval problem:

locating relevant documents based on user input, such as keywords or example documents.

Information Retrieval:

Typical IR Systems:

Online Library Catalogs

Online document management systems Information Retrieval Vs Database Systems:

Some DB problems are not present in IR, eg., update, transaction management, complex objects.

Some IR problems are not addressed well in DBMS, eg., unstructured documents, approximate search using keywords and relevance.

Precision: the percentage of retrieved documents that are in fact relevant to the query.

Precision = |{Relevant} ^ {Retrieved}|

|{Retrieved}|

Recall: the percentage of documents that are relevant to the query and were in fact retrieved.

Recall = |{Relevant} ^ {Retrieved}|

|{Relevant}|

Keyword–Based retrieval:

A document is represented by a string, which can be identified by a set of keywords. Queries may use expressions of keywords.

Eg. Car amd Repair shop, tea, coffee, DBMS but not Oracle

Queries and retrieval should consider synonyms, eg. Repair and maintenance.

Major difficulties of the Model:

- Synonymy: A keyword T does not appear anywhere in the document, even though the document is closely related to T, eg. Data mining
- Polysemy: The same keyword may mean different things in different contexts eg. Mining.

Latent Semantic Indexing:

Basic Idea:

- Similar documents have similar word frequencies.
- Difficulty: the size of the term frequency matrix is very large.

- Use a singular value decomposition (SVD) techniques to reduce the size of the frequency table.
- Retain the K most significant rows of the frequency table. Method:
- Create a term frequency matrix, freq-matrix.
- SVD Construction: Compute the singular valued decomposition of the freq-matrix by splitting it into 3 matrices, U, S, V.

Vector Identification:

- For each document d, replace its original document vector by a new excluding the eliminated terms.

Index Creation:

- Store the set of all vectors, indexed by one of a number of techniques (such as TV-tree)

Other Text Retrieval Indexing Techniques: Inverted Index:

Maintains two hash or B +tree indexed tables. Document Table:

a set of documents records <doc_id, postings_list> Term-table: a set of term records, < term, postings_list>

Answer Query: Find all docs associated with one or a set of terms. Advantage: Easy to implement

Disadvantage: Do not handle well synonymy and polysely and posting lists could be too long (storage could be very large)

Signature File:

Associate a signature with each document

A signature is a representation of an ordered list of terms that describe the document.

Order is obtained by frequency analysis, stemming and stop lists

.

Types of Text Data Mining:

- Keyword –based association analysis.

- Automatic document classification

  Similarity detection
- Cluster documents by a common author
- Cluster documents containing information from a common source Link analysis: Unusual Correlation between entities.

Sequence Analysis: Predicting a recurring event.

Anomaly Detection: Find information that violates usual patterns.

Hypertext Analysis:

- Patterns in anchors / links
- Anchor text correlations with linked objects. Keyword based Association Analysis:

Keyword based Association analysis:

- Collect sets of keywords or terms that occur frequently together and then find the association or correlation relationships among them.
- First preprocess the text data by parsing, stemming, removing stop words etc. Then evoke association mining algorithms.

  Consider each document as a transaction

  View a set of keywords in the document as a set of items in the transaction. Term level Association Mining:

  No need for human effort in tagging documents

  The number of meaningless results and the execution time is greatly reduced.

Automatic Document Classification:

Motivation

Automatic Classification for the tremendous number of on-line text documents.

A Classification Problem:

Training set: Human experts generate a training data set. Classification: The computer system discovers the classification rules.

Application: The discovered rules can be applied to classify new / unknown documents.

Text Document Classification differs from the classification of relational data Document databases are not structured according to attribute-value pairs.

Association-Based Document Classification:

Extract keywords and terms by information retrieval and simple association analysis techniques.

Obtain concept hierarchies of keywords and terms using:

- Available term classes, such as Word Net
- Expert Knowledge
- Some keyword classification systems.
- Classify documents in the training set into class hierarchies

Apply term association mining method to discover sets of associated terms. Use the terms to maximally distinguish one class of documents from others. Derive a set of association rules associated with each document class

Order the classification rules based on their occurrence frequency and discriminative power.

Used the rules to classify new documents.

6) a.) Explain the types of data in cluster analysis in detail with example Dec'14 May'17 (16)

The different types of data used for cluster analysis are

- interval scaled,
- binary,
- nominal,
- ordinal

- ratio scaled data

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering.

A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression. Cluster analysis tools based on k-means, k-medoids, and several methods have also been built into many statistical analysis software packages or systems, such as S-Plus, SPSS, and SAS.

Applications:

Cluster analysis has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing. In business, clustering can help marketers discover distinct groups in their customer bases and characterize customer groups based on purchasing patterns. In biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionality, and gain insight into structures inherent in populations.

Clustering may also help in the identification of areas of similar land use in an earth observation database and in the identification of groups of houses in a city according to house type, value,and geographic location, as well as the identification of groups of automobile insurance policy holders with a high average claim cost. Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity.

Clustering can also be used for outlier detection,Applications of outlier detection include the detection of credit card fraud and the monitoring of criminal activities in electronic commerce

What is Cluster Analysis?

The process of grouping a set of physical objects into classes of similar objects is called clustering.

Cluster – collection of data objects

– Objects within a cluster are similar and objects in different clusters are dissimilar.

Cluster applications – pattern recognition, image processing and market research.

- helps marketers to discover the characterization of customer groups based on purchasing patterns
- Categorize genes in plant and animal taxonomies
- Identify groups of house in a city according to house type, value and geographical location
- Classify documents on WWW for information discovery

Clustering is a preprocessing step for other data mining steps like classification, characterization.

Clustering – Unsupervised learning – does not rely on predefined classes with class labels.

Typical requirements of clustering in data mining:

2. Scalability – Clustering algorithms should work for huge databases
3. Ability to deal with different types of attributes – Clustering algorithms should work not only for numeric data, but also for other data types.
4. Discovery of clusters with arbitrary shape – Clustering algorithms (based on distance measures) should work for clusters of any shape.

4. Minimal requirements for domain knowledge to determine input parameters – Clustering results are sensitive to input parameters to a clustering algorithm (example

– number of desired clusters). Determining the value of these parameters is difficult and requires some domain knowledge.

5. Ability to deal with noisy data – Outlier, missing, unknown and erroneous data detected by a clustering algorithm may lead to clusters of poor quality.

6. Insensitivity in the order of input records – Clustering algorithms should produce same results even if the order of input records is changed.
7. High dimensionality – Data in high dimensional space can be sparse and highly skewed, hence it is challenging for a clustering algorithm to cluster data objects in high dimensional space.
8. Constraint-based clustering – In Real world scenario, clusters are performed based on various constraints. It is a challenging task to find groups of data with good clustering behavior and satisfying various constraints.
9. Interpretability and usability – Clustering results should be interpretable, comprehensible and usable. So we should study how an application goal may influence the selection of clustering methods.

Types of data in Clustering Analysis

2. Data Matrix: (object-by-variable structure)

Represents n objects, (such as persons) with p variables (or attributes) (such as age, height, weight, gender, race and so on. The structure is in the form of relational table or n x p matrix as shown below:

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$ called as "two mode" matrix

3. Dissimilarity Matrix: (object-by-object structure)

This stores a collection of proximities (closeness or distance) that are available for all pairs of n objects. It is represented by an n-by-n table as shown below.

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$ called as "one mode" matrix

Where d (i, j) is the dissimilarity between the objects i and j; d (i, j) = d (j, i) and d (i, i) = 0

Many clustering algorithms use Dissimilarity Matrix. So data represented using Data Matrix are converted into Dissimilarity Matrix before applying such clustering algorithms.

Clustering of objects done based on their similarities or dissimilarities. Similarity coefficients or dissimilarity coefficients are derived from correlation coefficients.

b.) What are the social impacts of data mining? May'15 (8)

Data Mining can offer the individual many benefits by improving customer service and satisfaction, and lifestyle in general. However, it also has serious implications regarding one's right to privacy and data security.

Is Data Mining a Hype or a persistent, steadily growing business?

Data Mining has recently become very popular area for research, development and business as it becomes an essential tool for deriving knowledge from data to help business person in decision making process.

Several phases of Data Mining technology is as follows:

- Innovators
- Early Adopters

- Chasm
- Early Majority
- Late Majority
- Laggards

Is Data Mining Merely Managers Business or Everyone's Business?

Data Mining will surely help company executives a great deal in understanding the market and their business. However, one can expect that everyone will have needs and means of data mining as it is expected that more and more powerful, user friendly, diversified and affordable data mining systems or components are made available.

Data Mining can also have multiple personal uses such as:

Identifying patterns in medical applications to choose best companies based on customer service. To classify email messages etc.

Is Data Mining a threat to Privacy and Data Security? With more and more information accessible in electronic forms and available on the web and with increasingly powerful data mining tools being developed and put into use, there are increasing concern that data mining may pose a threat to our privacy and data security.

Data Privacy: In 1980, the organization for Economic co-operation and development (OECD) established as set of international guidelines, referred to as fair information practices. These guidelines aim to protect privacy and data accuracy.

They include the following principles:

- Purpose specification and use limitation.
- Openness
- Security Safeguards
- Individual Participation

Data Security: Many data security enhancing techniques have been developed to help protect data. Databases can employ a multilevel security model to classify and restrict data according to various security levels with users permitted access to only their authorized level.

Some of the data security techniques are:

Encryption Technique Intrusion Detection In secure multiparty computation in data obscuration

7. Describe the applications and trends in data mining in detail (Dec '16) (16)

A wide range of companies have deployed successful applications of data mining. While early adopters of this technology have tended to be in information-intensive industries such as financial services and direct mail marketing, the technology is applicable to any company looking to leverage a large data warehouse to better manage their customer relationships. Two critical factors for success with data mining are: a large, well-integrated data warehouse and a well-defined understanding of the business process within which data mining is to be applied (such as customer prospecting, retention, campaign management, and so on).

Some successful application areas include:

A pharmaceutical company can analyze its recent sales force activity and their results to improve targeting of high-value physicians and determine which marketing activities will have the greatest impact in the next few months. The data needs to include competitor market activity as well as information about the local health care systems. The results can be distributed to the sales force via a wide-area network that enables the representatives to review the recommendations from the perspective of the key attributes

in the decision process. The ongoing, dynamic analysis of the data warehouse allows best practices from throughout the organization to be applied in specific sales situations.

A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product. Using a small test mailing, the attributes of customers with an affinity for the product can be identified. Recent projects have indicated more than a 20-fold decrease in costs for targeted mailing campaigns over conventional approaches.

A diversified transportation company with a large direct sales force can apply data mining to identify the best prospects for its services. Using data mining to analyze its own customer experience, this company can build a unique segmentation identifying the attributes of high-value prospects. Applying this segmentation to a general business database such as those provided by Dun & Bradstreet can yield a prioritized list of prospects by region.

A large consumer package goods company can apply data mining to improve its sales process to retailers. Data from consumer panels, shipments, and competitor activity can be applied to understand the reasons for brand and store switching. Through this analysis, the manufacturer can select promotional strategies that best reach their target customer segments.

Each of these examples have a clear common ground. They leverage the knowledge about customers implicit in a data warehouse to reduce costs and improve the value of customer relationships. These organizations can now focus their efforts on the most important (profitable) customers and prospects, and design targeted marketing strategies to best reach them.There are a number of applications that data mining has. The first is called market segmentation. With market segmentation, you will be able to find behaviors that are common among your customers. You can look for patterns

among customers that seem to purchase the same products at the same time. Another application of data mining is called customer churn. Customer churn will allow you to estimate which customers are the most likely to stop purchasing your products or services and go to one of your competitors.

For example, by using data mining a retail store may be able to determine which products are stolen the most. By finding out which products are stolen the most, steps can be taken to protect those products and detect those who are stealing them. While direct mail marketing is an older technique that has been used for many years, companies who combine it with data mining can experience fantastic results. For example, you can use data mining to find out which customers will respond favorably to a direct mail marketing strategy. You can also use data mining to determine the effectiveness of interactive marketing. Some of your customers will be more likely to purchase your products online than offline, and you must identify them.

While many businesses use data mining to help increase their profits, many of them don't realize that it can be used to create new businesses and industries. One industry that can be created by data mining is the automatic prediction of both behaviors and trends. Imagine for a moment that you were the owner of a fashion company, and you were able to precisely predict the next big fashion trend based on the behavior and shopping patterns of your customers? It is easy to see that you could become very wealthy within a short period of time. You would have an advantage over your competitors. Instead of simply guessing what the next big trend will be, you will determine it based on statistics, patterns, and logic.

Another example of automatic prediction is to use data mining to look at your past marketing strategies. Which one worked the best? Why did it work the best? Who were the customers that responded most favorably to it? Data mining will allow you to answer these questions, and once you have the answers, you will be able to avoid making any

mistakes that you made in your previous marketing campaign. Data mining can allow you to become better at what you do. It is also a powerful tool for those who deal with finances. A financial institution such as a bank can predict the number of defaults that will occur among their customers within a given period of time, and they can also predict the amount of fraud that will occur as well.

Another potential application of data mining is the automatic recognition of patterns that were not previously known. Imagine if you had a tool that could automatically search your database to look for patterns which are hidden. If you had access to this technology, you would be able to find relationships that could allow you to make strategic decisions.

Data mining is becoming a pervasive technology in activities as diverse as using historical data to predict the success of a marketing campaign, looking for patterns in financial transactions to discover illegal activities or analyzing genome sequences. From this perspective, it was just a matter of time for the discipline to reach the important area of computer security.

Applications of Data Mining in Computer Security presents a collection of research efforts on the use of data mining in computer security.

Data mining has been loosely defined as the process of extracting information from large amounts of data. In the context of security, the information we are seeking is the knowledge of whether a security breach has been experienced, and if the answer is yes, who is the perpetrator. This information could be collected in the context of discovering intrusions that aim to breach the privacy of services, data in a computer system or alternatively, in the context of discovering evidence left in a computer system as part of criminal activity.

Applications of Data Mining in Computer Security concentrates heavily on the use of data mining in the area of intrusion detection. The reason for this is twofold. First, the volume of data dealing with both network and host activity is so large that it makes it an ideal candidate for using data mining techniques. Second, intrusion detection is an extremely critical activity. This book also addresses the application of data mining to computer forensics. This is a crucial area that seeks to address the needs of law enforcement in analyzing the digital evidence.

Applications of Data Mining in Computer Security is designed to meet the needs of a professional audience composed of researchers and practitioners in industry and graduate level students in computer science.

Industrial Connectivity of the subject

- Students will be familiar with OLAP tools used in many business
- WEKA tool is used in many companies for performing mining task
- New data mining techniques and its applications
  are used in many companies.

B.E/B.Tech Degree Examinations Nov/Dec 2016

Sixth/Seventh Semester

Information Technology

IT6702-DATA WAREHOUSING AND DATA MINING

(Common to Computer Science and Engineering)

(Regulations 2013)

Time: Three hours Maximam: 100 Marks

Answer all the questions

Part A – (10 x 2=20 marks)

1. What are the nine decisions in the design of data warehouse? [Pg No.26]
2. Define Star schema. [Pg No.10]
3. List OLAP guidelines.[Pg. No 33]
4. Comment on OLAP tools Internet. [Pg. No 34]
5. Give an example of outerlier analysis for the library management system.[Pg.No.89]
6. What are the different steps in Data Transformation?[Pg.No.12]
7. Elucidate two phase involved in decision tree induction?[Pg No 74]
8. List the methods to improve Apriori's efficiency.[Pg.No 77]
9. State the role of Cluster analysis [Pg.No.100]
10. Give the reason on why clustering is needed in data mining?[Pg.No 100]

Part B – (5 x16=80 marks)

11. (a) Explain seven components of Data warehouse architecture with neat diagram.[Pg.No 23] (16)

Or

(b) Discuss DBMS schemas for decision support. Describe performance problems with star schema [Pg.No 14] (16)

12. (a) Discuss different tool categories in data warehouse business analysis [Pg. No 28] (16)

Or

(b) (i) Summarize the major differences between OLTP and OLAP system [Pg.No 32] (8)

(ii) Describe about Cognus Impromptu. [Pg.No 35] (8)

13. (a)Explain different strategies of Data Reduction. [Pg.No 57] (16)

Or

(b) Describe Data discretization and concept hierarchy generation. State why concept hierarchies are useful in data mining. [Pg.No 53] (8+8)

14. (a) Find all frequent items sets for the given training set using Apriori and FP growth, respectively. Compare the efficiency of the two mining processes [Pg.No 78] (10+6)

TID	items_bought
T100	(M,O,N,K,E,Y)
T200	(D,O,N,K,E,Y)
T300	(M,A,K,E)
T400	(M,U,C,K,Y)
T500	(C,O,O,K,I,E)

Or

(b) Explain Naive Bayesian classifications with algorithm and sample example. [Pg.No 83] (16)

15. (a) Describe the applications and trends in data mining in detail. [Pg. No 127] (16)

Or

(b) Describe different partition methods in cluster analysis. [Pg. No 101] (16)

B.E/B.Tech Degree Examinations Apr/May 2017

Sixth/Seventh Semester

Information Technology

IT6702-DATA WAREHOUSING AND DATA MINING

(Common to Computer Science and Engineering)

(Regulations 2013)

Time: Three hours Maximam: 100 Marks

Answer all the questions

Part A – (10 x 2=20 marks)

1. How is a data warehouse different from database? How they are similar?[Pg.No 13]
2. What is data discretization?[Pg.No 55]
3. List the distinct feature of OLTP and OLAP.[Pg.No 32]
4. What is multidimensional data model? Give Example. [Pg.No 32]
5. Why we need data transformation? Mention the ways by which data can be transformed.[Pg.No 12]
6. List the five primitives for specification of a data mining task.[Pg.No 52]
7. How do you evaluate accuracy of a classifier?[Pg No 73]
8. What is lazy learner? Give an example. [Pg No 72]
9. What is meant by K-Nearest Neighbor Algorithm?[Pg.No 98]
10. List the some applications of data mining.[Pg.No 96]

Part B – (5 x16=80 marks)

11. (a) Explain mapping data warehouse with multi processor architecture with the concept of parallelism and data partitioning. [Pg.No19] (16)

Or

(b) Discuss Data Extraction, Clean up and transformation tools with meta data management. [Pg.No. 28] (16)

12. (a)Explain different categories of OLAP tools with diagram.[Pg. No 40] (16)

Or

(b)(i) Summarize multi dimensional data model[Pg. No 39] (8)

(ii) Discuss about Cognus Impromptu. [Pg.No 35] (8)

13. (a)Why do we need to preprocess data? What are the different forms of preprocessing? [Pg.No 57] (16)

Or

(b) Describe in detail data mining functionalities and the different kinds of patterns can be mined. [Pg. No 64] (16)

14. (a) Discuss the single dimensional Boolean association rule mining for transaction database.[Pg.No78] (16)

Or

(b) Discuss about constraint based association rule mining with examples and state how association mining to correlation analysis is dealt with. [Pg.No.93] (16)

15. (a) Describe different types of data in Cluster Analysis.
[Pg.No 121] (16)

Or

(b) Describe different Hierarchical methods in Cluster Analysis.
[Pg.No 106] (16)