Quantum - Data Warehousing & Data Mining (KOE-093)

Data Warehouse and Data Mining (Dr. A.P.J. Abdul Kalam Technical University)

CONTENTS

KOE 093 : Data Warehousing & Data Mining

1 UNIT

Data Warehousing

CONTENTS

PART-1

Overview, Definition.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.1. What do you mean by data warehouse ? Discuss its key features with suitable example.

Answer

1. A data warehouse (DW) is a collection of corporate information and data derived from operational systems and external data sources.
2. A data warehouse is designed to support business decisions by allowing data consolidation, analysis and reporting at different aggregate levels.

Key features of data warehouse are :

a. **Subject-oriented :** A data warehouse can be used to analyze a particular subject area. For example, sales, marketing, etc can be a particular subject.

b. **Integrated :** A data warehouse integrates data from multiple data sources. For example, application *A* and *B* stores information in different way, but in a data warehouse, all this information is stored in common format.

c. **Time-variant :** Historical data is kept in a data warehouse. For example, one can retrieve data from 3 months, 6 months, 12 months, or even older data from a data warehouse *i.e.*, a data warehouse can hold all addresses associated with a customer.

d. **Non-volatile :** Data warehouse is also non-volatile means the previous data is not erased when new data is entered in it. Data is read-only and periodically refreshed. For example, changes in transaction status.

e. **Data granularity :** Data granularity can be defined as the level of details of data. In OLTP, the data granularity is the number of units for each unique product. For example, user will look the sale of products across all the stores, which region have recorded the maximum sale, region has given maximum sale etc.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.2. Describe the components of data warehouse.

Answer

Following are the various components of data warehouse :

a. **Data warehouse database :** The database is implemented on the RDBMS technology. Due to some constraints, different approaches to database are used :
   1. RDBMS are deployed in parallel to allow scalability.
   2. New index structures are used to bypass relational table scans and improve speed.
   3. Multidimensional databases are used to overcome any limitations due to relational data model.

b. **ETL tools :** The functionality of sourcing, acquisition, cleanup and transformation tools also called as ETL tools includes :
   1. Removing unwanted data from operational databases.
   2. Converting to common data names and definitions.
   3. Establishing defaults for missing data.

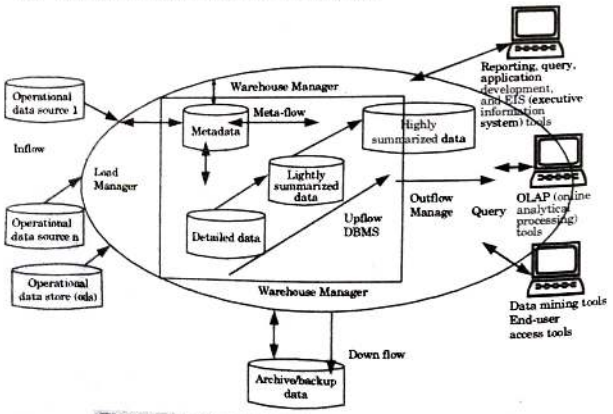


Fig. 1.3.1. Information flow of a data warehouse.

c. **Metadata :** Metadata is data about data that describes the data warehouse. It is used for building, maintaining, managing and using the data warehouse.

d. **Access tools :** Access tools are divided into four main categories :
   1. Query and reporting tools
   2. Application development tools
   3. Online analytical processing tools
   4. Data mining tools

e. **Data warehouse bus architecture :** Data warehouse bus determines the flow of data in our warehouse. The data flow in a data warehouse can be categorized as Inflow, Upflow, Downflow, Outflow and Meta flow.

PART-3

Building a Data Warehouse, Warehouse Database.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.3. Explain the concept of building a data warehouse.

Answer

Following steps should be adopted to build a successful data warehouse :

1. **Business considerations :**
   a. **Approach :** For data warehouse development, one of the two approaches is used :
      i. **Top-down approach :** In the top-down approach, data warehouse is built first. The data marts are then created from the data warehouse.
      ii. **Bottom-up approach :** In the bottom-up approach, data marts are created first and then data warehouse is built.
   b. **Organizational issues :** Most IS organizations have expertise in developing operational systems.
2. **Design considerations :** There are several points related to data warehouse design :
   a. **Data content :** The data warehouse system should not contain as much detail-level data as the operational system used to source this data in

b. **Metadata :** Metadata is data about data. It means it is a description and context of the data. It helps to organize, find and understand data.

c. **Data distribution :** It become necessary to know how the data should be divided across multiple servers and which users should get access to which type of data.

d. **Tools :** The tools provide the facilities for defining the transformation and cleanup rules, data movement, user query, reporting and data analysis.

e. **Performance considerations :** An ideal data warehouse system should support interactive query processing.

3. **Technical considerations :** A number of technical issues are to be considered when implementing and building a data warehouse system.
   a. **Hardware platform :** The data warehouse server has to be able to support large data volumes and complex queries.
   b. The database management system that supports the warehouse database.
   c. **Communication infrastructure :** A data warehouse user requires a large bandwidth to interact with the data warehouse and retrieve a large amount of data for analysis.
   d. The hardware platform and software to support the metadata repository.
   e. The systems management framework that enables centralized management and administration of the entire environment.
4. **Implementation consideration :** The implementation of data warehouse requires the integration of many products.
   a. **Access tools :** Ranking, statistical analysis, time series analysis, artificial intelligence, information mapping are some of the examples of access tools types.
   b. Data extraction, cleanup and transformation and migration.
   c. **Data placement strategies :** As a data warehouse grows, there should be a way to store the data in a storage media and distribute the data in the data warehouse across multiple servers.
   d. **Metadata :** Metadata is data about data. It means it is a description and context of the data. It helps to organize, find and understand data.
   e. **User sophisticated levels :** A certain degree of sophistication is required to effectively use the warehouse.

PART-4

Mapping the Data Warehouse to a Multiprocessor Architecture.

CONCEPT OUTLINE

- Mapping the relational database to the multiprocessing hardware architectures allows successful implementation of data warehouse.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.4. **Enumerate the steps involved in mapping the data warehouse to a multiprocessor architecture.**

AKTU 2016-17, Marks 10

OR

What is the architecture of data warehouse operations ?

AKTU 2013-14, Marks 05

Answer

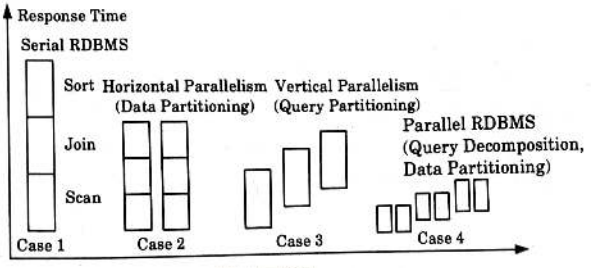
Steps involved in mapping the data warehouse to a multiprocessor architecture are :

1. **Relational database technology for data warehouse :**
   a. **Linear speed up :** The ability to increase the number of processor to reduce response time.
   b. **Linear scale up :** The ability to provide same performance on the same requests as the database size increases.

Types of parallelism :

i. **Horizontal parallelism :** In this, different server threads or processes handle multiple requests at the same time.

ii. **Vertical parallelism :** This form of parallelism decomposes the serial SQL query into lower level operations such as scan, join, sort etc.



iii. **Data partitioning :** Data partitioning is the key component for effective parallel execution of database operations. Partition can be done randomly or intelligently :

a. **Random partitioning :** Includes random data striping across multiple disks on a single server.

b. **Intelligent partitioning :** Assumes that DBMS knows where a specific record is located and does not waste time searching for it across all disks.

c. **Hash partitioning :** A hash algorithm is used to calculate the partition number based on the value of the partitioning key for each row.

d. **Key range partitioning :** Rows are placed and located in the partitions according to the value of the partitioning key.

e. **Schema partitioning :** An entire table is placed on one disk; another table is placed on different disk etc. This is useful for small reference tables.

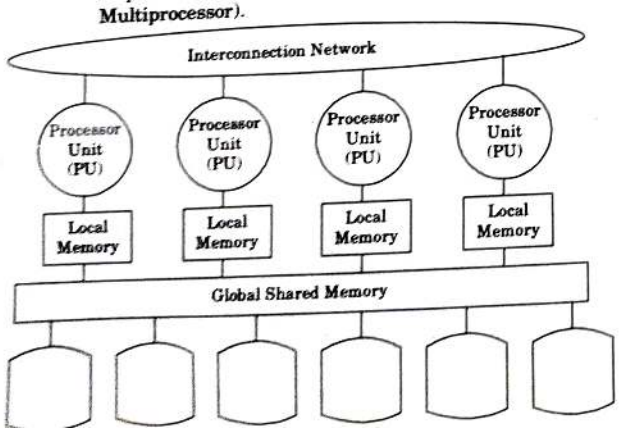
f. **User defined partitioning :** It allows a table to be partitioned on the basis of a user defined expression.

2. **Database architectures of parallel processing :** There are three DBMS software architecture styles for parallel processing :

a. **Shared memory or shared-everything architecture :** It has the following characteristics :

i. Multiple Processing Units (PU) share memory.

ii. It is simple to implement and provide a single system image, implementing an RDBMS on SMP (Symmetric Multiprocessor).



b. **Shared disk architecture :** Shared disk architecture implements a concept of shared ownership of the entire database between RDBMS servers, each of which is running on a node of a distributed memory system.

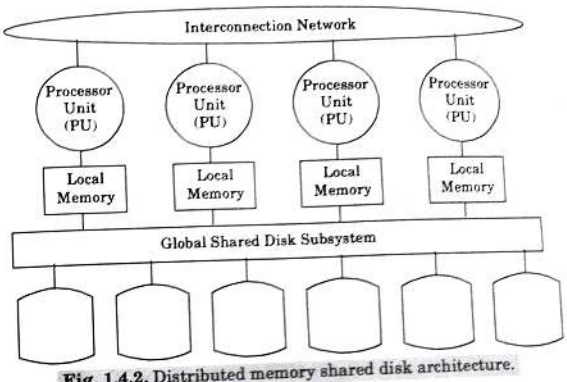


Fig. 1.4.2. Distributed memory shared disk architecture.

c. **Shared nothing architecture :** In shared architecture systems, only one CPU is connected to a given disk. If a table or database is located on that disk shared nothing systems are concerned with access to disks, not with access to memory.

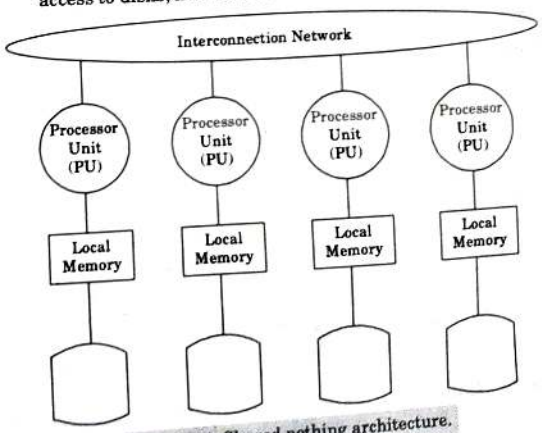


Fig. 1.4.4. Shared nothing architecture.

3. **Parallel DBMS features :**
   a. Parallel environment
   b. DBMS management tools
   c. Price / Performance
   d. Scope and techniques of parallel DBMS operations
   e. Optimized implementations
   f. Application transparency
4. **Alternative technologies :** For improving performance in data warehouse environment include following :
   a. Advanced database indexing products
   b. Multidimensional databases
   c. Specialized RDBMS
5. **Parallel DBMS Vendors :**
   a. **Oracle :** Support parallel database processing.
   b. **Informix :** It supports full parallelism.
   c. **IBM :** It is a parallel client/server database product-DB2-E (parallel edition).
   d. **SYSBASE :** It implemented its parallel DBMS functionality in a product called SYSBASE MPP (SYSBASE+NCR).

Que 1.5. **Define data warehouse. What strategies should be taken care while designing a warehouse ?** **AKTU 2017-18, Marks 10**

Answer

Data warehouse : Refer Q 1.1, Page 1–2D, Unit-1.

The strategies that should be taken care while designing a warehouse are :

1. **Educate yourself :** We must understand what users want because the purpose of a data warehouse system is to provide decision-makers the accurate, timely information they need to make the right choices.
2. **Determine business requirements :** To determine business requirements are should understand the following :
   a. Why the requestor needs a data warehouse.
   b. What are they trying to accomplish – saving time in collecting data, higher quality of data, supporting certain applications etc., we need to tie these business objectives to data sources.
   c. What business rules to follow and what users and/or applications to support.
3. **Make a timeline :** Break up business objectives mentioned above into two to three month incremental deliverables.
4. Choosing architecture, methodology and technology and building a team.

PART-5

Difference between Database System and Data Warehouse, Multidimensional Data Model.

CONCEPT OUTLINE

- A database system describes processing at operational sites whereas a data warehouse describes processing at warehouse.
- A multidimensional data model is used for the design of corporate data warehouses.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.6. **What is data warehouse ? How does it differ from a database ?** **AKTU 2013-14, Marks 05**

Answer

Data warehouse : Refer Q. 1.1, Page 1–2D, Unit-1.

Difference :

S. No.	Data warehouse	Database
1	It involves historical processing of information.	It involves day-to-day processing.
2.	It is used to analyze the business.	It is used to run the business.
3.	It focuses on information out.	It focuses on data in.
4.	It contains historical data.	It contains current data.

Que 1.7. **Differentiate between OLTP and OLAP with example.** **AKTU 2013-14, Marks 10**

OR

Differentiate between OLAP and OLTP. **AKTU 2017-18, Marks 10**

Answer

S. No.	Basis	OLAP	OLTP
1.	Abbreviation	It stands for 'Online Analytical Processing'.	It stands for 'Online Transaction Processing'.
2.	Use	It is used for Query Processing.	It is used for transaction Processing
3.	Data	It holds historical data. It stores only relevant data.	It holds current data. It stores all data.
4.	Type	It is analysis driven.	It is application driven.
5.	Source	The data comes from various OLTP sources	It is the original source of data.
6.	Purpose	To help with planning problem solving and decision support.	To control and run fundamental business tasks.
7.	Business	It reveals the multidimensional view of all types of business activities.	It reveals the ongoing business process.
8.	Speed	It is slow depending on the data.	It is very fast.
9.	Market	It is customer oriented.	It is market oriented.
10.	Database design	It is de-normalized with fewer tables and makes use of star or snow flake schemas.	It is highly normalized with many tables.
11.	View	It represents managerial view.	It represents clerical or operator view.
12.	Users	It has few concurrent users.	It has many concurrent users.

Que 1.8. **What is metadata and why is it important ? Discuss the multidimensional data.** **AKTU 2013-14, Marks 10**

Answer

Metadata :

1. Metadata is data about data. It means it is a description and context of the data. It helps to organize, find and understand data.
2. In data warehouse, metadata are the data that defines warehouse objects.
3. Metadata can be classified into two types : Technical metadata and Business metadata.

Importance of metadata :

1. Metadata drives data warehouse processes.
2. Metadata gives user the meaning of each data element.
3. Metadata establishes the context for data elements.

Multidimensional data :

1. Multidimensional data model stores data in the form of data cube. A data cube allows data to be viewed in multiple dimensions.
2. Data warehouses and Online Analytical Processing (OLAP) tools are based on a multidimensional data model.
3. Multidimensional data model provide both a mechanism to store data and a way for business analysis.
4. The two primary component of multidimensional data model are dimensions and facts. Dimensions are the entities with respect to which an organization wants to keep records and facts are the numerical measures.
5. There are three types of multidimensional data model :
   a. Star schema model
   b. Snow flake schema model
   c. Fact constellations

PART-6

Data Cubes.

CONCEPT OUTLINE

- A data cube is a multidimensional array of values which is used to view the data.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.9. **Describe data cubes with suitable example.**

Answer

a. A data cube is a method of storing data in a multidimensional form. It is a structure that enables OLAP to achieve multidimensional functionality.

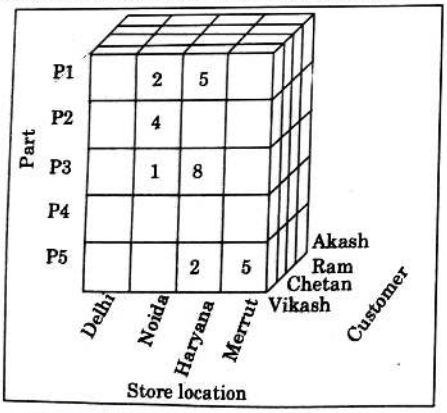
b. Data cubes are mainly categorized into two categories :
   i. Multidimensional data cube
   ii. Relational data cube

c. The data cube is used to represent data along some measure of interest.

d. Three important concepts associated with data cubes are :
   i. Slicing ii. Rotating
   iii. Dicing

Example :

1. We have a database that contains transaction information relating company sales of a part to a customer at a store location.
2. The data cube formed from this database is a three dimensional representation, with each cell (p, c, s) of the cube representing a combination of values from part, customer and store-location.
3. The content of each cell is the count of the number of times that specific combination of values occurs together in the database.
4. Cells that appear blank in fact have a value of zero. The cube can then be used to retrieve information within the database.



PART-7

Stars, Snow Flakes, Fact Constellations.

CONCEPT OUTLINE

- A multidimensional model can exists in the following three forms :
  i. Star schema
  ii. Snow flake schema
  iii. Fact constellations

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.10. **Explain with diagram the star, snow flake and fact constellation schemas for multidimensional databases.**

AKTU 2014-15, Marks 10

OR

Give the difference between the star and fact constellation multidimensional data model. **AKTU 2015-16, Marks 10**

OR

"A data warehouse can be modeled by either a star schema or a snow flake schema". With relevant examples discuss the two types of schema. **AKTU 2016-17, Marks 10**

OR

What are different database schemas shows with an example ?

AKTU 2017-18, Marks 05

Answer

Star schema :

1. The simplest data warehouse schema is star schema because its structure resembles a star.
2. Star schema consists of data in the form of facts and dimensions. The fact table present in the center of star and points of the star are the dimension tables.
3. In star schema fact, table contains a large amount of data, with no redundancy. Each dimension table is joined with the fact table using a primary or foreign key.
4. The main characteristics of star schema are that it is easy to understand and small number of tables can join.
5. The advantage of star schema is that it provides highly optimized performance for typical star users.

Example : Let us consider the "Employment" data warehouse. We have three dimension tables and one fact table. The star schema is shown in Fig. 1.10.1.

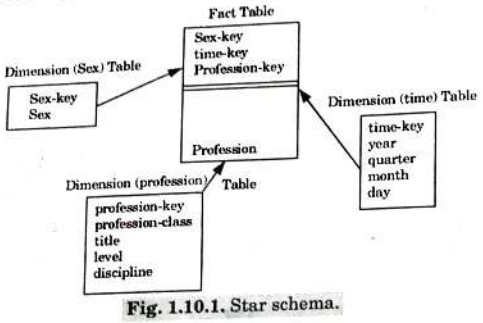


Fig. 1.10.1. Star schema.

Snow flake schema :

1. The snow flake schema is different from star schema because dimension tables of the snow flake are normalized.
2. The snow flake schema is represented by centralized fact table which is connected to multiple dimension table.
3. The major difference between the snow flake and star schema models is that the dimension tables of the snow flake model are normalized to reduce redundancies.

Example : Snow flake schema for a company XYZ electronics. Dimension table is normalized resulting in two tables.

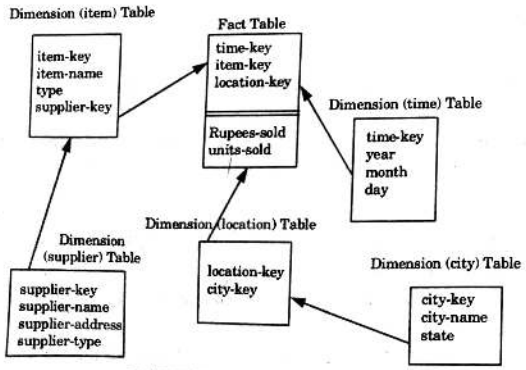


Fig. 1.10.2. Snow flake schema.

Fact constellations :

1. A fact constellation can have multiple fact tables that share many dimension tables.
2. This type of schema can be viewed as a collection of stars, snow flake and hence is called a galaxy schema or a fact constellation.
3. The main disadvantage of fact constellation schemas is its more complicated design.

Example : Let us assume that Deccan Electronics would like to have another fact table for supply and delivery. It may contain five dimensions, or keys : time, item, delivery-agent, origin, destination along with the numeric measure as the number of units supplied and the cost of delivery. It can be seen that both fact tables can share the same item-dimension table as well as time-dimension table. A fact constellation schema is shown in Fig. 1.10.3.

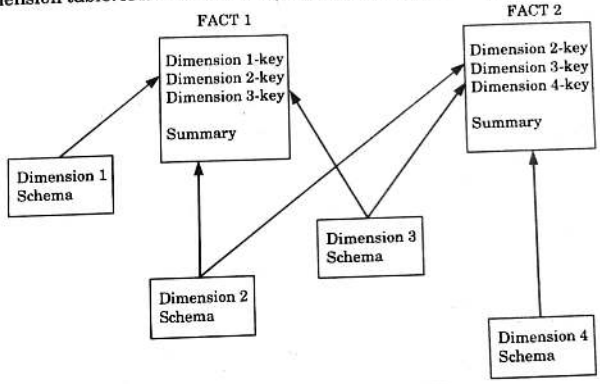


Fig. 1.10.3. Fact constellation.

Difference :

S. No.	Star schema	Fact constellation
1	In star schema, each dimension is represented by only one table.	In fact constellation, each dimension is represented by multiple fact tables.
2.	It is simple to understand and easily designed.	It is more complex and hard to design.
3.	It does not use normalization.	It uses normalization.
4.	It saves the space due to single fact table.	It does not save space due to multiple fact table.

Que 1.11. **Suppose that a data warehouse for a University consists of the following four dimensions : student, course, semester and instructor, and two measures such as count and avg_grade.**

When at the lowest conceptual level (for example, for a given student, course, semester and instructor combination) the avg_grade measure stores the actual course grade of the student. At higher conceptual levels, avg grade stores the average grade for the given combination.

i. **Draw a snow flake schema diagram for the data warehouse.**

ii. **Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations (for example, roll-up from semester to year) should one perform in order to list the average grade of CS courses for each student of the University.** **AKTU 2016-17, Marks 15**

Answer

i.

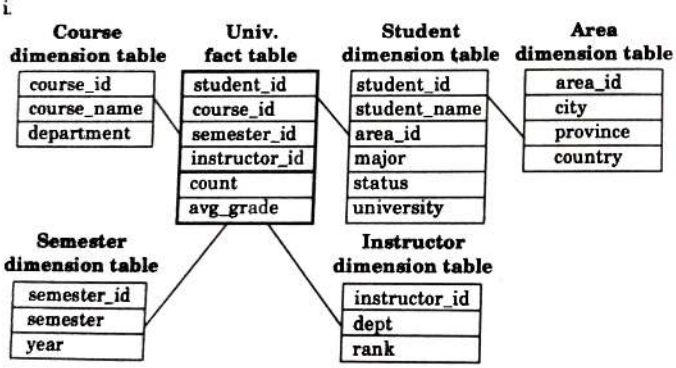


Fig. 1.11.1.

ii. Starting with the base cuboid [student, course, semester, instructor] :

1. Roll-up on course from (course_key) to major.
2. Roll-up on student from (student_key) to University.
3. Dice on course, student with department = "CS" and University = "Big University".
4. Drill-down on student from University to student name.

PART-8

Concept Hierarchy.

CONCEPT OUTLINE

- Concept hierarchy is a directed acyclic graph of concepts, where each of the concepts is identified by a unique name.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

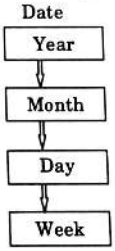
Que 1.12. **Describe concept hierarchy with example.**

AKTU 2013-14, Marks 05

AKTU 2017-18, Marks 2.5

Answer

1. Concept hierarchy represents the relationship between data elements in such a way that they can relate to each other as one above another, one below another.
2. Example of concept hierarchy is Date hierarchy which forms a relationship as Year → Month → Day → Week etc.



3. There are three main types of hierarchies in data warehouse design :
   a. Balanced hierarchy
   b. Unbalanced hierarchy
   c. Ragged hierarchy
4. Concept hierarchy reduces the data by collecting and replacing low level concepts by higher level concepts.

Que 1.13. **What do you mean by granularity ? What is partitioning ?** **AKTU 2013-14, Marks 05**

a. Determining the dimensions that are to be included.

b. Determining the location to place the hierarchy of each dimension of information.

Partitioning : Refer Q 1.4, Page 1–6D, Unit-1.

CONTENTS

PART-1

Warehousing Strategy, Warehouse / Management and Support Processes.

CONCEPT OUTLINE

- A warehouse strategy involves many important decisions such as the investment and operation costs that make up the logistics overhead.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 2.1. List the elements of warehouse strategy.

Answer

The data warehouse strategy should include the following elements :

1. **Preliminary data warehouse rollout plan :** All the user requirements cannot be met in one data warehouse project as the project would necessarily be large.
2. **Preliminary data warehouse architecture :** It defines the overall and initial architecture of each rollout.
3. **Shortlisted data warehouse environment and tools :** Create a shortlist for the tools and environment that appear to meet warehousing needs.

Que 2.2. Explain warehouse management and support processes.

Answer

Warehouse management and support processes are designed to address the aspects of the planning and managing DW project, subject to successful implementation and extension of software.

Steps in warehouse management and support processes :

1. **Define issue tracking and resolution process :** It includes following guidelines : Issue description, urgency, raised by, assigned to, date opened, date closed, resolved by and resolution description.
2. **Perform capacity planning :** It can be done is following forms :

   a. **Space required :** Space requirements are determined by schema design, backup and recovery strategy, indexing strategy, aggregation, metadata etc.
   b. **Machine processing power :** It chooses a configuration that is scalable and meets the processing requirements.
   c. **Network bandwidth :** It verify all assumptions about the network bandwidth before proceeding with each rollout.
3. **Define warehouse purging rules :** It define the mechanism for archiving or removing older data from the data warehouse and check for any legal, regulatory or auditing requirements.
4. **Define security management :** It keeps the data warehouse secure to prevent the loss of information either due to disaster or due to an unauthorized user. Various steps involved in security management are :
   i. Determine and evaluate IT assets
   ii. Analyze risk
   iii. Define security practices
   iv. Implement practices
   v. Monitor violations and take corresponding actions
   vi. Re-evaluate IT assets and risk
5. **Define backup and recovery strategy :**
   i. **Data to be backed up :** Identify the data that must be backed up on a regular basis. This gives us an indication of the regular backup size.
   ii. **Batch window of the warehouse :** It determines the maximum allowable down time for the warehouse.
   iii. **Maximum acceptable time for recovery :** It determines the maximum acceptable time for the warehouse data and metadata to be restored.
   iv. **Acceptable costs for backup and recovery :** Different backup mechanisms imply different backup costs.
6. **Set up collection of warehouse usage statistics :** Warehouse usage statistics are collected to provide the data warehouse designer with inputs for further refining the data warehouse design and to track the general usage and acceptance of warehouse.

PART-2

Warehouse Planning and Implementation.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 2.3. **Write a short note on data warehouse planning.**

Answer

The data warehouse planning describes the activities related to planning on rollout of the data warehouse. Different approaches for data warehouse planning are :

1. **Assemble and orient the team :**
   i. Identify the employees and brief them about the project.
   ii. Distribute copies of DW strategy.
   iii. Set up teams and specify roles.
   iv. Give training if required.
   v. Set up milestones and check points.
2. **Conduct decisional requirements analysis :** It means gain a thorough understanding of the information needs of decision-makers.
3. **Conduct decisional source system audit :** Survey current source of data for data warehouse.
4. **Design logical and physical warehouse schema :** It includes two schema design techniques :
   i. **Normalization :** Normalize scaled attribute data so as to fall within a small specified range, such as 0.0 to 1.0.
   ii. **Dimensional modeling :** This technique produces denormalized, star schema designs consisting of fact and dimension tables. A variation of the dimensional star schema also exists (*i.e.*, snowflake schema).
5. **Produce source-to-target field mapping :** The source-to-target field mapping documents how fields in the operational systems are transformed into the data warehouse fields.
6. **Select development and production environment and tools :** It finalizes the computing environment and tool set for rollout based on the results of development and production environment.
7. **Create prototype for this rollout :** It creates a prototype of the data warehouse using the final tools and production environment.
8. **Create implementation plan for this rollout :** It drafts an implementation plan for the rollout.

Que 2.4. **Explain all steps and guidelines for data warehouse implementation.** **AKTU 2014-15, Marks 10**

Answer

Steps for data warehouse implementation :

1. **Requirements analysis and capacity planning :** The first step in data warehousing involves defining enterprise needs, defining architecture, carrying out capacity planning and selecting the hardware and software tools.
2. **Hardware integration :** Once the hardware and software have been selected, they need to be put together by integrating the servers, the storage devices and the client software tools.
3. **Modeling :** Modeling is a major step that involves designing the warehouse schema and views. This may involve using the modeling tool if the data warehouse is complex.
4. **Physical modeling :** This involves designing the physical data warehouse organization, data placement, data partitioning, deciding on access methods and indexing.
5. **Sources :** The data for the data warehouse is likely to come from a number of data sources. This step involves identifying and connecting the sources using gateways, ODBC drives or other wrappers.
6. **ETL :** The data from the source systems will need to go through an ETL process. The step of designing and implementing the ETL process may involve identifying a suitable ETL tool vendor and purchasing and implementing the tool.
7. **Populate the data warehouse :** Once the ETL tools have been agreed upon, testing the tools will be required, perhaps using a staging area.
8. **User applications :** For the data warehouse to be useful there must be end-user applications. This step involves designing and implementing applications required by the end users.
9. **Roll-out the warehouse and applications :** Once the data warehouse has been populated and the end-user applications are tested, the warehouse system and the applications may be rolled out for the user community to use.

Guidelines for data warehouse implementation :

1. **Build incrementally :** Data warehouses must be built incrementally. It is recommended that a data part may first be built with one particular project in mind and then data warehouse can be implemented in an iterative manner allowing all data parts to extract information from the data warehouse.
2. **Need a champion :** A data warehouse project must have a champion who is willing to carry out considerable research into expected costs and benefits of the project.
3. **Senior management support :** A data warehouse project must be fully supported by the senior management. Give the resource intensive nature of such projects and the time they take to implement, a warehouse project calls for a sustained commitment from senior management.
4. **Ensure quality :** Only data that has been cleaned should be loaded in the data warehouse.

5. **Corporate strategy :** A data warehouse project must fit with corporate strategy and business objectives.
6. **Business plan :** The financial costs (hardware, software, and peopleware), expected benefits and a project plan (including an ETL plan) for a data warehouse project must be clearly outlined and understood by all stakeholders.
7. **Training :** A data warehouse project must not overlook data warehouse training requirements.
8. **Adaptability :** The project should build in adaptability so that changes can be made to the data warehouse when required. Like any system, a data warehouse will need to change, as needs of an enterprise change.
9. The project must be managed by both IT and business professionals in the enterprise.

PART-3

Hardware and Operating Systems for Data Warehousing.

Questions-Answers

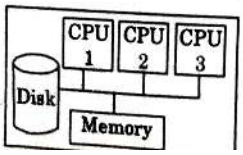
Long Answer Type and Medium Answer Type Questions

Que 2.5. Explain hardware and operating systems used in data warehouse.

Answer

Hardware and operating system refers to the server platforms and operating system that serve as the computing environment of the data warehouse. Hardware and operating system used in data warehouse are :

1. **Parallel hardware technology :**
   a. **Symmetric multiprocessing :** These system consists of pair of about 64 processor that share a common memory and operating system. Since, system resources are shared hence, they can be easily managed and make use of high speed interconnections, high bandwidth and low latency is also achieved.



   b. **Massively parallel processor systems :** It uses a large number of processors which communicate using some message interface. Each processor has its own CPU, memory and disk subsystem.

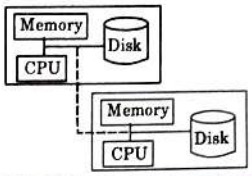


Fig. 2.5.2. MPP architecture.

2. **Clustered system :** These systems are configured with multiported array so that nodes which have direct disk access enjoy same disk I/O rates as standalone SMP systems. Nodes which not have direct disk access must use the high-speed cluster interconnect mechanism.

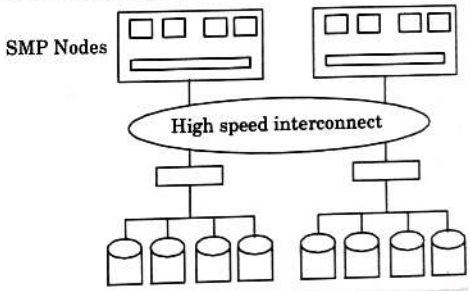


Fig. 2.5.3. Cluster of four SMP systems.

Que 2.6. Mention criteria for hardware selection.

Answer

The following selections are recommended for hardware selection :

1. Scalability
2. Financial stability
3. Price/Performance
4. Delivery lead time
5. Reference sites
6. Availability of support

PART-4

Client/Server Computing Model and Data Warehousing, ... and Cluster Systems.

CONCEPT OUTLINE

- *Client/server is a program client which requests a service or resource from the server.*
- *Parallel processing is a method of dividing the program into multiple fragments to speed up the execution of programs.*

Questions-Answers

Long Answer Type and Medium Answer Type Questions

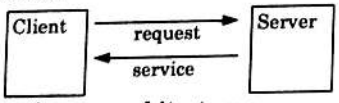
Que 2.7. Explain client/server architecture.

Answer

1. Client/server architecture is a network architecture in which each computer on the network is either a client or a server.
2. Client/server architecture works when the client sends a request to the server over the network connection, which is then processed and delivered to the client.

Components of client/server architecture :

1. **Client :** It is a computer which processes the request service from the server.
2. **Server :** Any computer can provide services to the client.
3. **Communication middleware :** A computer through which client and server communicate.



Advantages of client/server architecture :

1. Scalability
2. Resource sharing
3. Data integrity
4. Communication cost is reduced.
5. Ease of effort and maintenance.

Disadvantages of client/server architecture :

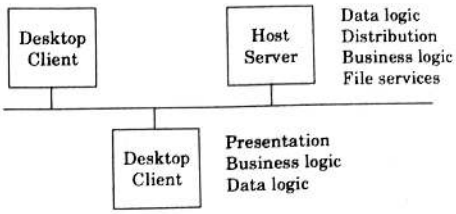
1. Single point of failure
2. Costly to maintain DB server

Que 2.8. What are the types of client/server architecture ?

Answer

Types of client/server architecture are :

1. **Two-tier architecture :** A two-tier architecture is where a client talks directly to a server. It is typically used in small environments.



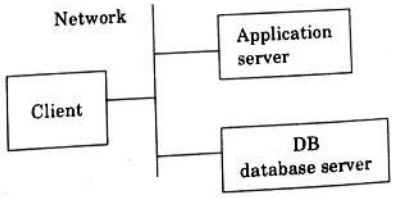
Advantages of two-tier architecture :

1. Interoperability
2. Portability
3. Integration
4. Transparency
5. Security

Disadvantages of two-tier architecture :

1. Network traffic is handled less efficiently.
2. The client and server are tightly coupled.

2. **Three-tier architecture :** In the three-tier architecture, a middleware is used between the client environment and the database management server environment. It is used in large environment.



Advantages of three-tier architecture :

1. Improve performance
2. Improve flexibility

Disadvantages of three-tier architecture :

1. The development environment is more difficult to use.

Que 2.9. Describe distributed memory architecture. What are its types ?

Answer

In distributed memory architecture all processors in the system are directly connected to own memory and caches. Any processor cannot directly access another processor's memory.

Two types of distributed memory architecture are :

1. **Shared nothing architecture :**
   a. Shared nothing architecture is used in distributing computing in which each node have their own memory, storage and independent input/output interfaces.
   b. Each node do not shares any resources with other nodes and communicate with each other by passing messages.

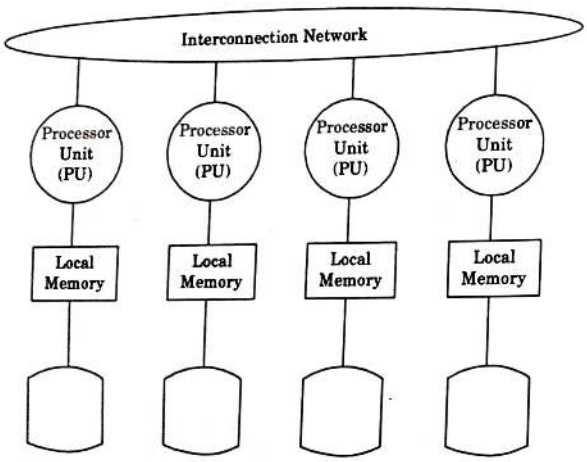


Fig. 2.9.1. Distributed memory shared nothing architecture.

2. **Shared disk architecture :**
   a. A shared disk architecture is a distributed computing architecture in which all disks are accessible from all cluster nodes.

   b. Multiple processors can access all disks directly via intercommunication network and every processor has local memory.

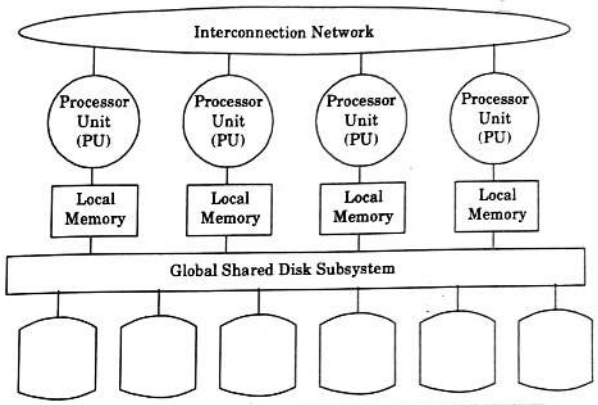


Fig. 2.9.2. Distributed memory shared disk architecture.

Que 2.10. Write a short note on cluster system.

Answer

1. In a cluster system, every processor unit (PU) executes a copy of operating system, and the inter-PU communications are performed over an open-systems-based interconnection.
2. Cluster system is designed for high availability by providing shared access to disks.
3. Cluster system describes many characteristics of MPP system, including a very high-speed scalable interconnection mechanism and support for hundreds of PUs.

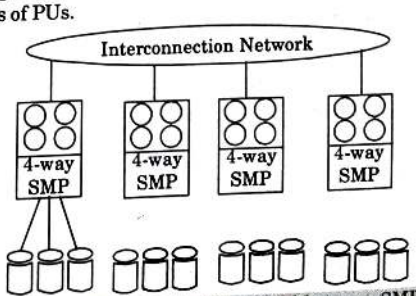


Fig. 2.10.1. Distributed memory cluster of four-way SMP nodes.

PART-5

Distributed DBMS Implementations, Warehousing Software and Warehouse Schema Design.

CONCEPT OUTLINE

- Schema is a logical description of the entire database.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 2.11. Explain warehousing software tools in detail.

Answer

Warehousing software tools are :

1. **Connectivity tools :** These tools provide transparent access to source system in heterogeneous environment.

   For example :

   i. **IBM :** Data joiner

   ii. **Oracle :** Transparent gateway

   iii. **SAS :** SAS/connect

   iv. **Sybase :** Enterprise connect

2. **Extraction tools :** There are two primary methods to use extraction tools *i.e.*, bulk extraction and change-based replication.

   For example :

   i. **Apertus carleton :** Passport

   ii. **Platinum :** InfoPump

3. **Transformation tools :** These tools has following features :

   i. Field splitting and consolidation

   ii. Standardization

   iii. De-duplication

   For example :

   i. **Apertus carleton :** Enterprise/Integrator

   ii. **Data mirror :** Transformation server

4. **Data quality tools :** It helps to identify and correct data error at source systems.

   For example :

   i. **Data flux :** Data quality workbench

   ii. **Prism :** Quality manager

   iii. **Pine cone systems :** Content tracker

5. **Data loaders :** It transforms data into data warehouse.

6. **Data access and retrieval tools :** These tools are classified into two categories :

   i. **OLAP tools :** These allow users to make ad hoc queries or generate queries against warehouse database.

   ii. **Reporting tools :** These allow users to produce scanned and sophisticated reports based on warehouse data.

7. **Data modeling tools :** These tools allow users to prepare and maintain an information model of both source and target database.

   For example :

   i. Cayenne software, Terrain

   ii. Relational matters, Syntagma designer

   iii. Sybase, PowerDesigner WarehouseArchitect

8. **Warehouse management tools :** These tools assist warehouse admin in the day-to-day management and administration of the warehouse.

   For example :

   i. Pine cone systems, usage tracker, refreshment tracker.

   ii. Red brick systems, enterprise control and coordination.

Que 2.12. Discuss various warehouse schema design techniques.

Answer

Various warehouse schema design techniques are :

1. OLTP systems use normalized data structures.
2. **Dimensional modeling for decisional systems :** It provides a number of techniques for denormalizing database to create schema.
3. **Star schema :** Refer Q.1.10, Page 1–14D, Unit-1.
4. **Dimensional hierarchies :** Each dimension will have hierarchies that imply grouping and structure.
5. **Granularity of the fact table :** The first step in designing a fact table is to determine the granularity of the fact table. By granularity, we mean the lowest level of information that will be stored in the fact table. This constitutes two steps :

   a. Determine which dimensions will be included.

   b. Determine where along the hierarchy of each dimension the information will be kept.

6. **Aggregates or summaries :** Aggregates are the summarization of fact related data for the purpose of improved performance. Aggregates are to be considered for use when the number of detailed records to be processed is large and/or the processing of the customer queries begins to impact the performance.
7. **Dimensional attributes :** The attribute values are used to establish the context of the facts.
8. **Multiple star schemas :** A data warehouse will have multiple star schemas, *i.e.*, many fact tables.

3
UNIT

Data Mining

CONTENTS

PART-1

Overview, Motivation, Definition and Functionalities.

CONCEPT OUTLINE

- Data mining is a process used by organizations to turn raw data into useful information.
- Functionalities of data mining :

1. Characterization
2. Discrimination
3. Classification
4. Outlier analysis
5. Evolution analysis

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 3.1. **Explain data, information and knowledge.**

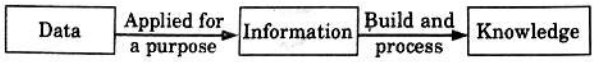
AKTU 2014-15, Marks 05

Answer

Data : Data are raw facts and figures that can be processed or stored by a computer. For example, text, numbers, symbols, etc.

Information : Information is data that has been processed into a form that gives it meaning. For example, analysis of retail of sale data can provide information on which products are selling.

Knowledge : Knowledge is the understanding of rules needed to interpret information. For example, information on retail market sales can be analyzed with promotional efforts to yield knowledge of customer behaviour.



Que 3.2. **What is data mining ? Define the major issues in data mining.**

AKTU 2014-15, Marks 05

OR

Describe challenges to data mining regarding data mining methodology and user interaction issues.

Answer

Data mining : Data mining is defined as a process used to extract usable data from a larger set of any raw data.

Key features of data mining :

1. Automatic pattern predictions based on trend and behaviour analysis.
2. Prediction based on likely outcomes.
3. Creation of decision oriented information.
4. Focus on large data sets and databases for analysis.
5. Clustering based on groups of facts not previously known.

Major issues in data mining :

1. **Mining methodology and user interaction issues :**
   - a. **Mining different kinds of knowledge in databases :** Different users may be interested in different kinds of knowledge.
   - b. **Interactive mining of knowledge at multiple levels of abstraction :** It allows users to focus the search for patterns from different angles.
   - c. **Incorporation of background knowledge :** Background knowledge is used to guide discovery process and to express the discovered patterns.
   - d. **Data mining query languages and adhoc data mining :** Data mining query language should be integrated with data warehouse query language.
   - e. **Presentation and visualization of data mining results :** Once the patterns are discovered it needs to be expressed in high level languages.
   - f. **Handling noisy or incomplete data :** The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities.
   - g. **Pattern evaluation :** The patterns discovered should be interesting because they represent common knowledge.
2. **Performance issues :**
   - a. **Efficiency and scalability of data mining algorithms :** To extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.
   - b. **Parallel, distributed, and incremental mining algorithms :** The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms.
3. **Diverse data types issues :**

b. Mining information from heterogeneous databases and global information systems.

Que 3.3. Explain the data mining / knowledge extraction process in detail ? **AKTU 2017-18, Marks 10**

Answer

Knowledge Discovery in Databases (KDD) refers to the process of discovering useful knowledge from data.

Steps involved in the knowledge discovery process are :

1. **Data cleaning :**
   a. Data cleaning is defined as removal of noisy and irrelevant data from collection.
   b. It includes :
      i. Cleaning in case of missing values.
      ii. Cleaning noisy data, where noise is a random or variance error.
      iii. Cleaning with data discrepancy detection and data transformation tools.
2. **Data integration :**
   a. Data integration is defined as heterogeneous data from multiple sources combined in a common source (Data Warehouse).
   b. It includes :
      i. Data integration using data migration tools.
      ii. Data integration using data synchronization tools.
      iii. Data integration using ETL (Extract-Load-Transformation) process.
3. **Data selection :**
   a. Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.
   b. It includes :
      i. Data selection using neural network.
      ii. Data selection using decision trees.
      iii. Data selection using Naive Bayes.
      iv. Data selection using clustering, regression, etc.
4. **Data transformation :**
   a. In this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
   b. Data transformation is a two step process :
      i. **Data mapping :** Assigning elements from source base to destination to capture transformations.

      ii. **Code generation :** Creation of the actual transformation program.
5. **Data mining :**
   a. Data mining is defined as clever techniques that are applied to extract patterns potentially useful.
   b. It includes :
      i. Transforms task relevant data into patterns.
      ii. Decides purpose of model using classification or characterization.
6. **Pattern evaluation :** Pattern evaluation is defined as an identifying strictly increasing patterns representing knowledge based on given measures.
7. **Knowledge representation :** Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.

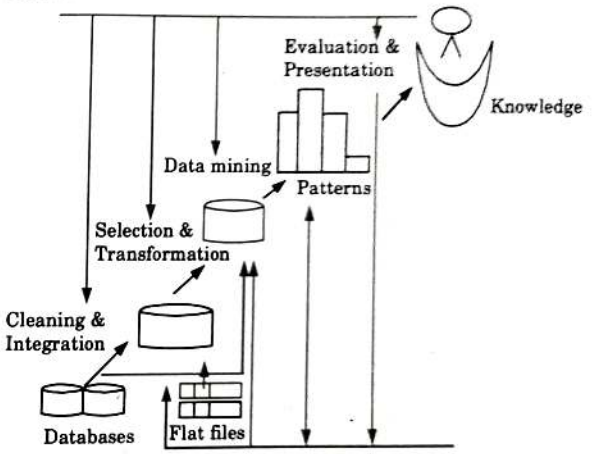


Fig. 3.3.1.

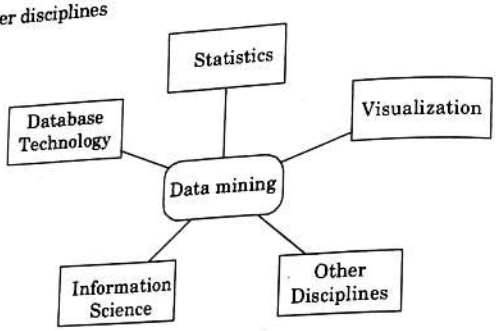
Que 3.4. How data mining systems are classified ? Describe each classification with example. **AKTU 2016-17, Marks 10**

Answer

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

1. Database technology
2. Statistics
3. Machine learning

4. Information science
5. Visualization
6. Other disciplines



Data mining system can also be classified as :

a. **Classification based on the databases mined :** Database system can be classified according to different criteria such as data models, types of data, etc. For example, if we classify a database according to the data model, then we may have a relational, transactional, object-relational, or data warehouse mining system.

b. **Classification based on the kind of knowledge mined :** It means the data mining system is classified on the basis of functionalities such as characterization, discrimination, association analysis, classification, prediction, outlier analysis, evolution analysis. A comprehensive data mining system usually provides multiple integrated data mining functionalities.

c. **Classification based on the techniques utilized :** We can classify a data mining system according to the kind of techniques used in user autonomous systems, interactive exploratory systems, query-driven systems or the methods of analysis employed such as machine learning, statistics, visualization, pattern recognition, neural networks.

d. **Classification based on the applications adapted :** We can classify a data mining system according to the applications adapted. The applications are as follows : finance, telecommunications, DNA, stock markets, E-mail.

Que 3.5. Explain data mining functionalities.

Answer

Following are the data mining functionalities :

1. **Data characterization :** It is a summarization of the general characteristics or features of a target class of data.

2. **Data discrimination :** It refers to the mapping or classification of a class with some predefined group or class.
3. **Association analysis :** It analyses the set of items that frequently appear together in a transactional dataset.
4. **Classification :** In classification, data are grouped into predefined classes.
5. **Prediction :** It refers to predict some unavailable data values rather than class labels.
6. **Cluster analysis :** Classification and prediction analyze class labeled data objects, where as clustering analyzes data objects without consulting a known class label.
7. **Outlier analysis :** Outliers are data elements that cannot be grouped in a given class or cluster.
8. **Evolution analysis :** Evolution analysis refers to the description and model regularities or trends for objects whose behaviour changes over time.

Que 3.6. Describe the difference between the following approaches for the integration of data mining system with database or data warehouse systems : no coupling, loose coupling and semi tight coupling. **AKTU 2015-16, Marks 7.5**

Answer

If a data mining system is not integrated with a database or a data warehouse system, then there will be no system to communicate with. This scheme is known as the non-coupling scheme.

Various integration schemes are as follows :

a. **No coupling :** In this scheme, the data mining system does not utilize any of the database or data warehouse functions. It fetches the data from a particular source and processes that data using some data mining algorithms.

b. **Loose coupling :** In this scheme, the data mining system may use some of the functions of database and data warehouse system. It fetches the data from the data respiratory and performs data mining on that data.

c. **Semi-tight coupling :** In this scheme, the data mining system is linked with a database or a data warehouse system and efficient implementations of a few data mining primitives can be provided in the database.

d. **Tight coupling :** In this scheme, the data mining system is smoothly integrated into the database or data warehouse system. The data mining subsystem is treated as one functional component of an information system.

PART-2

Data Processing, Form of Data Pre-Processing.

CONCEPT OUTLINE

- Data processing is the conversion of data into usable and desired form.
- Forms of data processing are :

1. Data cleaning
2. Data integration
3. Data transformation
4. Data reduction

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 3.7. **What are the different forms of data processing ?**

AKTU 2014-15, Marks 05

OR

Explain the data cleaning, data integration and transformation in brief.

AKTU 2014-15, Marks 05

Answer

Different forms of data processing are :

1. **Data cleaning :** Data cleaning is a process to remove the noisy data, clean the data by filling in the missing values and correct the inconsistencies in data.
2. **Data integration :** Data integration is a technique that combines the data from multiple heterogeneous data sources into a coherent data store. Data integration may involve inconsistent data and therefore needs data cleaning.
3. **Data transformation :** In this step, data is transformed or consolidated

   c. **Generalization :** In generalization low-level data are replaced with high-level data by using concept hierarchies climbing.
   d. **Normalization :** Normalization scaled attribute data so as to fall within a small specified range, such as 0.0 to 1.0. It is of two types:
      i. **Min-max normalization :** It is a technique that helps to normalize data. It will scale the data between 0 and 1.
      ii. **z-score normalization :** Transform the data by converting the values to a common scale with an average of zero and a standard deviation of one.
   e. **Attribute/feature construction :** New attributes constructed from the given ones.
4. **Data reduction :** Data reduction is used to obtain reduced representation of data in small values by maintaining the integrity of original data.

Que 3.8. **Data consolidation is data modeling activity. This statement is true or not ? Justify.**

AKTU 2013-14, Marks 05

Answer

1. The statement is true as data consolidation means transforming data into the forms that are appropriate for mining by performing certain operations.
2. The normal data which we obtain from different data sources is not in suitable form to be stored in data warehouses or for performing data mining operations. So, data is modeled for further activities after performing data consolidation.
3. **Data consolidation involve the following operations :**
   Refer Q. 3.7, Page 3–8D, Unit-3.

PART-3

Data Cleaning : Missing Values, Noisy Data (Binning, Clustering Regression, Computer and Human Inspection), Inconsistent Data.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

How to handle noisy data ?

Answer

Noise is a random error or variance in a measured variable.

Following are the data smoothing techniques :

1. **Binning :** It is a technique in which first of all we sort the data and then partition the data into equal frequency bins. For example,

   Price = 4, 8, 15, 21, 21, 24, 25, 28, 34

   a. **Partition into (equal-frequency) bins :**

      Bin a: 4, 8, 15, Bin b: 21, 21, 24, Bin c: 25, 28, 34

   b. **Smoothing by bin means :** In smoothing by bin, each value in a bin is replaced by the mean value of the bin.

      Bin a: 9, 9, 9, Bin b: 22, 22, 22, Bin c: 29, 29, 29

   c. **Smoothing by bin boundaries :** In smoothing by bin boundaries, each bin value is replaced by the closest boundary value.

      Bin a: 4, 4, 15, Bin b: 21, 21, 24, Bin c: 25, 25, 34

2. **Regression :**
   a. Data can be smoothed by fitting the data into a regression functions. Linear regression and multiple linear regression are type of regression.
   b. A regression task begins with a dataset in which the target values are known.
   c. For example, a regression model could be used to predict the value of a house based on location, number of rooms, lot size, and other factors.
3. **Clustering :**
   a. Outliers may be detected by clustering, where similar values are organized into groups, or clusters.

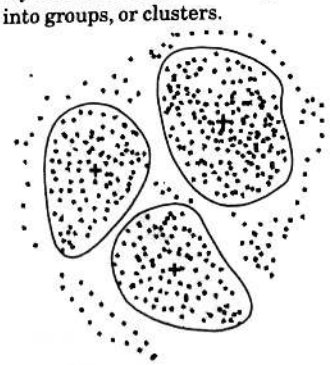


Fig. 3.9.1. Clustering.

   b. Values that fall outside of the set of clusters may be considered outliers.
   c. For example, clustering analysis can be used in area such as market research, pattern recognition, data analysis, and image processing.
4. **Combined computer and human inspection :** The outliers can also be identified with the help of computer and human inspection. The outliers patterns can be informative or garbage. Humans can sort out the garbage patterns.

Que 3.10. Elaborate the different strategies for data cleaning.

AKTU 2017-18, Marks 10

Answer

Data is cleaned through processors such as data migration, data scrubbing, and data auditing :

1. **Data migration :**
   a. During data migration, transformation rules are specified (for example, replacing sex by gender) to clean the data.
   b. Transcription errors, incomplete information, and lack of standard formats are also addressed during data migration.
2. **Data scrubbing :**
   a. It involves detecting and removing errors and inconsistencies from data in order to improve the quality of data.
   b. Data scrubbing involves a complex cleaning and mapping process that is the most labor intensive part of building a data warehouse.
   c. During the cleaning process, desired information is filtered out and its quality is maintained for the target system.
3. **Data auditing :**
   a. Data auditing tools make it possible to discover rules and relationships or to signal violation of stated rules by scanning data.
   b. It enhances the systems reliability and makes it possible to prevent, detect, and eliminate data errors, irregularities, and fraud.

Que 3.11. List the ways to handle the missing values. What do you mean by inconsistent data ?

Answer

Ways to handle missing values are :

1. **Ignore the tuple :** This is usually done when class label is missing.

How to handle noisy data ?

Answer

Noise is a random error or variance in a measured variable.

Following are the data smoothing techniques :

1. **Binning :** It is a technique in which first of all we sort the data and then partition the data into equal frequency bins. For example,

   Price = 4, 8, 15, 21, 21, 24, 25, 28, 34

   a. **Partition into (equal-frequency) bins :**

      Bin a: 4, 8, 15, Bin b: 21, 21, 24, Bin c: 25, 28, 34

   b. **Smoothing by bin means :** In smoothing by bin, each value in a bin is replaced by the mean value of the bin.

      Bin a: 9, 9, 9, Bin b: 22, 22, 22, Bin c: 29, 29, 29

   c. **Smoothing by bin boundaries :** In smoothing by bin boundaries, each bin value is replaced by the closest boundary value.

      Bin a: 4, 4, 15, Bin b: 21, 21, 24, Bin c: 25, 25, 34

2. **Regression :**
   a. Data can be smoothed by fitting the data into a regression functions. Linear regression and multiple linear regression are type of regression.
   b. A regression task begins with a dataset in which the target values are known.
   c. For example, a regression model could be used to predict the value of a house based on location, number of rooms, lot size, and other factors.

3. **Clustering :**
   a. Outliers may be detected by clustering, where similar values are organized into groups, or clusters.

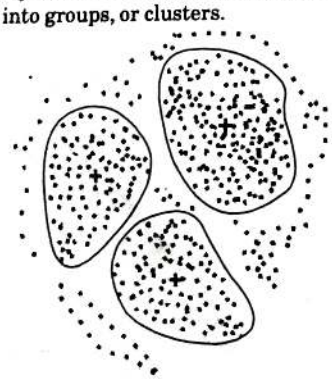


Fig. 3.9.1. Clustering.

   b. Values that fall outside of the set of clusters may be considered outliers.
   c. For example, clustering analysis can be used in area such as market research, pattern recognition, data analysis, and image processing.

4. **Combined computer and human inspection :** The outliers can also be identified with the help of computer and human inspection. The outliers patterns can be informative or garbage. Humans can sort out the garbage patterns.

Que 3.10. Elaborate the different strategies for data cleaning.

AKTU 2017-18, Marks 10

Answer

Data is cleaned through processors such as data migration, data scrubbing, and data auditing :

1. **Data migration :**
   a. During data migration, transformation rules are specified (for example, replacing sex by gender) to clean the data.
   b. Transcription errors, incomplete information, and lack of standard formats are also addressed during data migration.

2. **Data scrubbing :**
   a. It involves detecting and removing errors and inconsistencies from data in order to improve the quality of data.
   b. Data scrubbing involves a complex cleaning and mapping process that is the most labor intensive part of building a data warehouse.
   c. During the cleaning process, desired information is filtered out and its quality is maintained for the target system.

3. **Data auditing :**
   a. Data auditing tools make it possible to discover rules and relationships or to signal violation of stated rules by scanning data.
   b. It enhances the systems reliability and makes it possible to prevent, detect, and eliminate data errors, irregularities, and fraud.

Que 3.11. List the ways to handle the missing values. What do you mean by inconsistent data ?

Answer

Ways to handle missing values are :

1. **Ignore the tuple :** This is usually done when class label is missing.

2. **Fill in the missing value manually :** This approach is time consuming and may not be feasible with many missing values.
3. **Use a global constant to fill in the missing value :** Replace all the missing attribute values by the same constant.
4. Use the attribute mean to fill in the missing value.
5. **Use the most probable value to fill in the missing value:** This may be determined with regression or decision tree induction.
6. Use the attribute mean for all samples belonging to the same class as the given tuple.

Inconsistent data : Data inconsistency occur when similar data is kept in different formats in two different files, or when matching of data must be done between files. The inconsistency can be recorded in some transactions during data entry or arising from integrating data from different database.

Que 3.12. Explain Chi-square test method. Show using Chi-square test that gender and preferred reading are independent or not from given table. (Given are the observed counts).

	Male	Female	Total
Fiction	**250**	**200**	**450**
Non-Fiction	**50**	**1000**	**1050**
Total	**300**	**1200**	**1500**

AKTU 2015-16, Marks 15

Answer

1. A correlation relationship between two categorical (discrete) attributes, A and B, can be discovered by a χ^2 (Chi-square) test.
2. The χ^2 value (also known as the Pearson χ^2 statistics) is computed as :

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where o_{ij} is the observed frequency (*i.e.*, actual count) of the joint event (A_i, B_j) and e_{ij} is the expected frequency of (A_i, B_j), which can be computed as

$$e_{ij} = \frac{\text{count}(A = a_i) \times \text{count}(B = b_j)}{N}$$

where,

Numerical :

	Male	Female	Total
Fiction	250	200	450
Non-Fiction	50	1000	1050
Total	300	1200	1500

1. Suppose that a group of 1,500 people was surveyed. The gender of each person was noted. Each person was polled as to whether their preferred type of reading material was fiction or non-fiction. Thus, we have two attributes, gender and preferred reading.
2. The observed frequency (or count) of each possible joint event is summarized in the contingency table as shown, where the numbers in parentheses are the expected frequencies.

	Male	Female	Total
Fiction	250 (90)	200 (360)	450
Non-Fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

3. The expected frequency for the cell (male, fiction) is

$$e_{11} = \frac{\text{count(male)} \times \text{count(fiction)}}{N} = \frac{300 \times 450}{1500} = 90$$

and so on.

4. Using equation for χ^2 computation, we get

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$

$$= 284.44 + 121.90 + 71.41 + 30.48 = 507.93$$

5. For this 2×2 table, the degrees of freedom are $(2-1)(2-1) = 1$. For 1 degree of freedom, the χ^2 value needed to reject the hypothesis at the 0.001 significance level is 10.828. Since our computed value is above this, we can reject the hypothesis that gender and preferred reading are independent and conclude that the two attributes are (strongly) correlated for the given group of people.

Answer

Methods for attribute subset selection are :

1. **Stepwise forward selection :** In this method, the best of the original attributes is determine and added to the reduced set.

   For example : Initial attribute set : {A1, A2, A3, A4, A5}

   Initial reduced set : { } = {A1} = {A1, A4}

   Reduced attribute set : {A2, A3, A5}

2. **Stepwise backward elimination :** It removes the worst attribute remaining in the set.

   For example : Initial attribute set : {A1, A2, A3, A4, A5}

   {A1, A3, A4, A5} = {A1, A4, A5}

   Reduced attribute set : {A1, A5}

3. **Combination of forward selection and backward elimination :** This procedure selects the best attribute and removes the worst from remaining attributes.

   For example :

   Initial attribute set : {A1, A2, A3, A4, A5}

   Reduced attribute set in stepwise forward selection : {A2, A3, A5}

   Reduced attribute set in stepwise backward elimination : {A1, A5}

   Reduced attribute set : {A1, A2, A3, A5}

4. **Decision tree induction :** It constructs a flowchart where the best attribute is chosen to partition the data into individual classes.

   For example : Initial attribute set : $\{A_1, A_2, A_3, A_4, A_5, A_6\}$

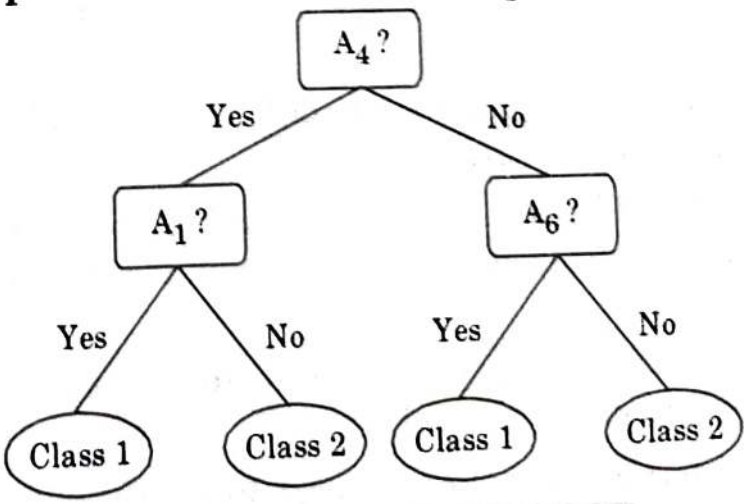


Fig. 3.14.1. Decision tree.

Reduced attribute set : $\{A_1, A_4, A_6\}$

Que 3.15. Write a short note on dimensionality reduction.

Answer

1. Data encoding or transformation are applied so as to obtain a reduced or compressed representation of the original data.
2. There are two components of dimensionality reduction :
   a. **Feature selection :** Feature selection is a process of removing features that are not relevant or are redundant.
   b. **Feature extraction :** Feature extraction is a process of transformation of raw data into features suitable for modeling.
3. The various methods used for dimensionality reduction include :
   a. **Wavelet transform :** It is a linear signal processing technique which transforms the data vector into numerically different vector of wavelet coefficients.
   b. **Principal Component Analysis (PCA) :** In this, the data in a higher dimensional space is mapped to data in a lower dimension space. It involves the following steps :
      i. Construct the covariance matrix of the data.
      ii. Compute the eigen vectors of this matrix.
      iii. Eigen vectors corresponding to the largest eigen values are used to reconstruct a large fraction of variance of the original data.

Que 3.16. Discuss numerosity reduction in detail.

Answer

In numerosity reduction, data volume can be reduced by choosing alternative forms of data representation. The various methods used for numerosity reduction include :

a. **Regression and log-linear model :** These models are used to approximate the given data.
b. **Histograms :** Histograms uses binning to approximate data distributions. It divide data into buckets and store average sum for each bucket.
c. **Clustering :** Partition data set into clusters based on similarity and store cluster representation only.
d. **Sampling :** It allows a large data set to be represented by a much smaller random sample of the data.

Que 3.17. Distinguish between dimensionality reduction and numerosity reduction. **AKTU 2014-15, Marks 05**

Answer

S. No.	Dimensionality reduction	Numerosity reduction
1.	In dimensionality reduction, data encoding or transformations are applied to obtain a reduced or compressed representation of original data.	In numerosity reduction, data volume is reduced by choosing alternating, smaller forms of data representation.
2.	Methods for dimensionality reduction are : a. Wavelet transforms b. Principal Component Analysis (PCA)	Methods for numerosity reduction are : a. Regression and log-linear model (parametric) b. Histograms, clustering, sampling (non-parametric).
3.	It can be used for removing irrelevant and redundant attributes.	It is merely a representation technique of original data to smaller form.
4.	In this method, some data can be lost which is irrelevant.	In this method, there is no loss of data.

Que 3.18. Write a short note on concept hierarchy generation for numeric data.

Answer

Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts with high-level concepts.

Concept hierarchy generation for numerical data methods :

1. **Binning :**
   a. Binning is a top-down splitting technique based on a specified number of bins.
   b. Binning is an unsupervised discretization technique.
2. **Histograms analysis :**
   a. Histograms partition the values for an attribute into disjoint ranges called buckets.
   b. Histograms analysis is an unsupervised discretization technique.
3. **Cluster analysis :** It is used to partition the data into clusters or groups.

Answer

1. Categorical data are discrete data.
2. Categorical attributes have finite number of distinct values, with no ordering among the values.
3. There are several methods for generation of concept hierarchies for categorical data :

   a. **Specification of a partial ordering of attributes explicitly at the schema level by experts :** Concept hierarchies for categorical attributes or dimensions typically involve a group of attributes. A user or an expert can easily define concept hierarchy by specifying a partial or total ordering of the attributes at a schema level.

   b. **Specification of a portion of a hierarchy by explicit data grouping :** In a large database, it is unrealistic to define an entire concept hierarchy by explicit value enumeration. However, it is realistic to specify explicit groupings for a small portion of the intermediate level data.

   c. **Specification of a set of attributes but not their partial ordering :** A user may specify a set of attributes forming a concept hierarchy, but omit to specify their partial ordering. The system can then try to automatically generate the attribute ordering so as to construct a meaningful concept hierarchy.

   d. **Specification of only of partial set of attributes :** To handle partially specified hierarchies, it is important to embed data semantics in the database schema so that attributes with tight semantic connections can be pinned together.

Que 3.20. What do you mean by data mining ? Differentiate between data mining technique and data mining strategy.

AKTU 2013-14, Marks 05

Answer

Data mining : Refe - Q. 3.2, Page 3–2D,Unit-3.

CONTENTS

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.1. Define the terms data generalization and analytical characterization with example. AKTU 2013-14, Marks 05

Answer

Data generalization :

1. Data generalization summarizes data by replacing relatively low level values with higher level concepts.
2. Data generalization approaches include : Data cube approach and attribute oriented induction approach.
3. Data generalization is a form of descriptive data mining.
4. For example, let us consider the database of XYZ electronics, instead of examining individual customer transactions, sales manager may prefer to view the generalized data to higher levels, such as summarized by customers groups according to regions, income, etc.

Analytical characterization :

1. Analytical characterization performs attribute and dimension relevance analysis in order to filter out irrelevant or weakly attributes.
2. It is performed to overcome the various limitations of class characterization.
3. For example, employee birth_date, birth_month, birth_year are not relevant to the employee's salary but experience is highly relevant to the salary of employee.

Que 4.2. **Explain data cube approach and attribute oriented approach.** **AKTU 2014-15, Marks 05**

OR

Discuss basic approaches of data generalization.

Answer

There are two basic approaches of data generalization :

1. **Data cube approach :**
   a. It is also known as OLAP approach.
   b. In this approach, computation and results are stored in the data cube.
   c. It is an efficient approach as it is helpful to make the past selling graph.
   d. It uses roll-up and drill-down operations on a data cube.
2. **Attribute oriented induction :**
   a. It is an online data analysis, query oriented and generalization based approach.
   b. In this approach, we perform generalization on the basis of different values of each attributes within the relevant data set. After that, same tuples are merged and their respective counts are accumulated in order to perform aggregation.
   c. Attribute oriented induction approach used two methods :
      i. Attribute removal
      ii. Attribute generalization

PART-2

Mining Class Comparisons.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.3. **Why class comparisons needed in data mining ? Discuss the steps of class comparisons.**

Answer

In many applications, users may not be interested in having a single class description but they need to compare two or more classes that distinguish

target class to its contrasting classes. For example, the three classes : person, address and item are not comparable.

Steps of class comparisons are :

1. Data collection
2. Dimension relevance analysis
3. Synchronous generalization
4. Presentation of the derived comparison

Que 4.4. **What is the role of statistics in data mining ?**

AKTU 2014-15, Marks 05

Answer

1. Statistics is a component of data mining that provides the tools and analytics techniques for dealing with large amounts of data.
2. It is the science of learning from data and includes everything from collecting and organizing to analyzing and presenting data. Statistics focuses on probabilistic models, specifically inference, using data.
3. Statistics is used in data mining for computing skills required to manage the data and its analysis and in automation of data analysis.
4. Main areas where statistical approach used in data mining are :
   i. Visualization
   ii. Size of data
   iii. Sampling
   iv. Data analysis

Que 4.5. **Explain various measures of central tendency.**

Answer

Measures of central tendency are :

1. **Mean :**
   a. It is a center of the data set.
   b. Let data set X are in values as $x_1, x_2, \ldots\ldots x_n$.

      Mean of data set X is : $X' = \frac{1}{n}\sum_{i=1}^{n} x_i$

2. **Median :**
   a. It is the middle value of the ordered set if the number of values n is an odd number or it is the average of middle two values if n is an even number.
   b. Median $= L_1 + \left(\frac{n/2(\sum f)l}{f_{medium}}\right)C$

3. **Mode :**
   a. It is a most frequently occur value from a large data set.
   b. Mode = 3 Median – 2 Mean.
4. **Midrange :** It is the average of the largest and smallest value of data set.

PART-3

Statistical Measures in Large Databases, Statistical-Based Algorithms, Distance-Based Algorithms.

CONCEPT OUTLINE

- Two descriptive statistics are used in statistical measures :
  1. Measuring the central tendency
  2. Measuring the dispersion of data
- Distance-based algorithms are :
  1. Simple approach
  2. *k*-nearest neighbours

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.6. Discuss various measures of dispersion of data.

OR

What are the properties of standard deviation and give its formula ?

AKTU 2014-15, Marks 05

Answer

Measures of dispersion of data are :

1. **Range :** The range of the data set is the difference between highest and lowest value.

$$\text{Range} = H - L$$

where H is the highest and L is the lowest value in the data set.

2. **Quartiles :** The first quartile is denoted by $Q1$, is the 25th percentile. The third quartile is denoted by $Q3$, is the 75th percentile. The distance between the 1st and 3rd quartiles is the simple measure of distribution which gives the range covered by the middle half of the data. This distance is called as Interquartile Range (IQR), defined as :

$$\text{IQR} = Q3 - Q1$$

3. **Outliers :** Outliers are the values higher/lower than 1.5*IQR.

4. **Boxplot :** Boxplots are a popular way of visualizing a distribution. A boxplot incorporates the five number summary as :
   a. Typically, the ends of the box are the quartiles, so that the box length is the Interquartile Range (IQR).
   b. The median is marked by a line within the box.
   c. Two lines (called whiskers) outside the box extend to the smallest (Minimum) and largest (Maximum) observations.
5. **Standard deviation and variance :** The standard deviation of a data set gives a measure of how each value in a data set varies from the mean.

   The standard deviation of a set of n observations, $x_1, x_2, \ldots x_n$, is given by :

$$\sigma = \sqrt{\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{2}\left[\sum x_i^2 - \frac{1}{n}(\sum x_i)^2\right]}$$

The basic properties of the standard deviation are :

a. Σ measures spread about the mean and should be used only when the mean is chosen as the measure of center.

b. $\sigma = 0$ only when there is no spread, that is, when all observations have the same value. Otherwise $\sigma > 0$, the variance is the mean of the squared deviations about the by σ^2. The variance of n observations, $x_1, x_2 \ldots x_n$, is given by :

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\left[\sum x_i^2 - \frac{1}{n}(\sum x_i)^2\right]$$

Que 4.7. Draw a box-and-whisker plot for the following data set : 126, 132, 138, 140, 141, 141, 142, 143, 144, 144, 144, 145, 146, 147, 148, 148, 149, 149, 150, 150, 150, 154, 155, 158, 158.

Also find the outliers.

AKTU 2015-16, Marks 10

Answer

Given : 126, 132, 138, 140, 141, 141, 142, 143, 144, 144, 144, 145, 146, 147, 148, 148, 149, 149, 150, 150, 150, 154, 155, 158, 158

Since there are 25 data points, the median Q_2 will be = 146

The first half has twelve values, so the median is the average of the middle two :

$$Q_1 = \frac{(141 + 142)}{2} = 141.5$$

The median of the second half is :

$$Q_3 = \frac{(150 + 150)}{2} = 150$$

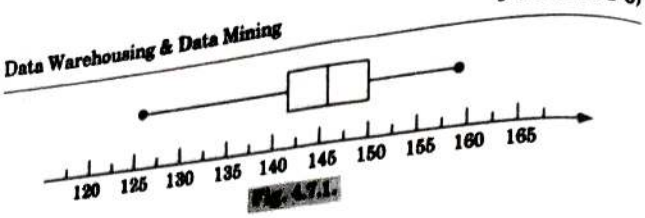


Fig. 4.7.1.

$Q_1 = 141.5$

$Q_3 = 150$

Interquartile range = 150 – 141.5 = 8.5

An outlier is any data point that is more than 1.5 times the *IQR* from either end of the box.

i.e., $8.5 \times 1.5 = 12.75$

At upper end, outlier is any data point more than

150 + 12.75 = 162.75

There are no data points larger than 162.75 so there are no outliers at the upper end.

At the lower end an outlier is any data point less than

141.5 – 12.75 = 128.75

Data point 126 is less than 128.75 therefore it is an outlier.

Que 4.8. **Given the following set of values {1, 3, 9, 15, 20), determine the Jack knife estimate for both the mean and standard deviation of the mean.**

AKTU 2013-14, Marks 05

Answer

Given : $n = 5$,

$x_i = \{1, 3, 9, 15, 20\}$

$x_1 = 1, \quad x_2 = 3, \quad x_3 = 9, \quad x_4 = 15, \quad x_5 = 20$

Mean, $\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$

$= \frac{1+3+9+15+20}{5} = 9.6$

By ignoring x_1, $\theta_1 = \frac{x_2 + x_3 + x_4 + x_5}{4}$

$= \frac{3+9+15+20}{4} = 11.75$

By ignoring x_2, $\theta_2 = \frac{x_1 + x_3 + x_4 + x_5}{4}$

$= \frac{1+9+15+20}{4} = 11.25$

By ignoring x_3, $\theta_3 = \frac{x_1 + x_2 + x_4 + x_5}{4}$

$= \frac{1+3+15+20}{4} = 9.75$

By ignoring x_4, $\theta_4 = \frac{x_1 + x_2 + x_3 + x_5}{4}$

$= \frac{1+3+9+20}{4} = 8.25$

By ignoring x_5, $\theta_5 = \frac{x_1 + x_2 + x_3 + x_4}{4}$

$= \frac{1+3+9+15}{4} = 7$

$\hat{\theta} = \frac{\theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5}{5}$

$= \frac{11.75+11.25+9.75+8.25+7}{5} = 9.6$

Jack knife estimate for mean is given by :

$= n(\bar{x}) - (n-1)\hat{\theta}$

$= 5(9.6) - 4(9.6)$

$= 9.6$

Standard deviation :

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{N}}$$

$$\sigma = \sqrt{\frac{1}{5}[1-9.6)^2 + (3-9.6)^2 + (9-9.6)^2 + (15-9.6)^2 + (20-9.6)^2]}$$

$$\sigma = \sqrt{\frac{1}{5} \times 255.2} = \sqrt{51.04} = 7.144$$

By ignoring x_1 :

$$\sigma_1 = \sqrt{\frac{1}{4}[(3-9.6)^2 + (9-9.6)^2 + (15-9.6)^2 + (20-9.6)^2]}$$

$$\sigma_1 = \sqrt{\frac{1}{4} \times 181.24} = \sqrt{45.31} = 6.73$$

By ignoring x_2 :

$$\sigma_2 = \sqrt{\frac{1}{4}[(1-9.6)^2 + (9-9.6)^2 + (15-9.6)^2 + (20-9.6)^2]}$$

$$\sigma_2 = \sqrt{\frac{1}{4} \times 211.64} = \sqrt{52.91} = 7.27$$

By ignoring x_3 :

$$\sigma_3 = \sqrt{\frac{1}{4}[(1-9.6)^2 + (3-9.6)^2 + (15-9.6)^2 + (20-9.6)^2]}$$

$$\sigma_3 = \sqrt{\frac{1}{4} \times 254.84} = \sqrt{63.71} = 7.98$$

By ignoring x_4 :

$$\sigma_4 = \sqrt{\frac{1}{4}[(1-9.6)^2 + (3-9.6)^2 + (9-9.6)^2 + (20-9.6)^2]}$$

$$\sigma_4 = \sqrt{\frac{1}{4} \times 226.04} = \sqrt{56.51} = 7.51$$

By ignoring x_5 :

$$\sigma_5 = \sqrt{\frac{1}{4}[(1-9.6)^2 + (3-9.6)^2 + (9-9.6)^2 + (15-9.6)^2]}$$

$$\sigma_5 = \sqrt{\frac{1}{4} \times 147.04} = \sqrt{36.76} = 6.06$$

$$\hat{\sigma} = \frac{\sigma_1 + \sigma_2 + \sigma_3 + \sigma_4 + \sigma_5}{5}$$

$$\hat{\sigma} = \frac{6.73 + 7.27 + 7.98 + 7.51 + 6.06}{5}$$

$$\hat{\sigma} = 7.11$$

Jack knife estimate for standard deviation is given by :

$$= n(\sigma) - (n - 1)\hat{\sigma}$$
$$= 5(7.144) - (5 - 1)(7.11)$$
$$= 35.72 - 28.44 = 7.28$$

Que 4.9. **Write short notes on :**

i. **Quartiles**
ii. **Histograms**
iii. **Scatter plots**

AKTU 2014-15, Marks 05

OR

Explain the various graphs for statistical class description.

Answer

Different types of graphs are :

1. **Histogram :** In this, we partition the data distribution of an attribute into disjoint sets but the width of each subset should be uniform. Each subset is drawn by a rectangle whose height is equal to the count of the subset.
2. **Scatter plots :** This graphical method is used for determining the existence of any relationship, pattern between two numerical attributes. In this method, every pair of value considered as a pair of coordinates in an algebraic sense and plotted as points in the plane.

3. **LOESS curve :** LOESS is locally estimated scatterplot smoothing. It adds smooth curve to existing scatterplot to provide better perception of the pattern of dependence.
4. **Quartile plots :** A quartile plot is a simple and effective way to have a first look at a univariate data distribution. First, it displays all of the data for the given attribute. Second, it plots quartile information. The mechanism used in this step is slightly different from the percentile computation.
5. **Q-Q (Quartile-Quartile) plot :** A quartile-quartile plot graphs the quartiles of one univariate distribution against the corresponding quartiles of another. It is a powerful visualization tool that allows the user to view whether there is a shift in going from one distribution to another.

Que 4.10. **Write a short note on Bayesian classification.**

AKTU 2013-14, Marks 05

Answer

1. Bayesian classifiers are the statistical classifiers.
2. Bayesian classifiers can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.
3. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.
4. Bayesian classification is based on Bayesian theorem.

Bayesian theorem : The purpose of Bayesian theorem is to predict the class label for a given tuple. Let X be a data tuple. In Bayesian terms, X is considered "evidence." Let H be some hypothesis, such as that the data tuple X belongs to a specified class C. There are two types of probabilities :

1. Posterior Probability $[P(H/X)]$
2. Prior Probability $[P(H)]$

where X is data tuple and H is some hypothesis. According to Bayes theorem,

$$P(H/X) = P(X/H)P(H) / P(X)$$

Que 4.11. **Write a short note on Naïve Bayes classifiers.**

Answer

1. A Naive Bayes classifier uses probability theory to classify data. Naive Bayes is also known as simple Bayes or independence Bayes.
2. Naive Bayes is a kind of classifier which uses the Bayes theorem.
3. It predicts membership probabilities for each class such as the probability that given record or data point belongs to a particular class.

4. The class with the highest probability is considered as the most likely class. This is also known as Maximum A Posteriori (MAP).
5. Naive Bayes classifier assumes that all the features are unrelated to each other.

For example : A fruit may be considered to be an apple if it is red, round, and about 5 in diameter. Even if these features depend on each other or upon the existence of the other features, a Naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

Que 4.12. **Classify the tuple X = (Colour = 'RED', Type = 'SUV' Origin = 'DOMESTIC') using Naive Bayesian classification. Training data is given in the following table where class label is (STOLEN).**

Colour	Type	Origin	Stolen
Red	Sports	Domestic	Yes
Red	Sports	Domestic	No
Red	Sports	Domestic	Yes
Red	Sports	Domestic	No
Red	Sports	Imported	Yes
Yellow	SUV	Imported	No
Yellow	SUV	Imported	Yes
Yellow	SUV	Domestic	No
Yellow	SUV	Imported	No
Red	Sports	Domestic	Yes

AKTU 2015-16, Marks 15

Answer

Colour			Type			Origin		
	Yes	No		Yes	No		Yes	No
Red	4	2	Sports	4	2	Domestic	3	3
Yellow	1	3	SUV	1	3	Imported	2	2

Stolen	
Yes	No
5	5

Colour			Type			Origin		
	Yes	No		Yes	No		Yes	No
Red	4/5	2/5	Sports	4/5	2/5	Domestic	3/5	3/5
Yellow	1/5	3/5	SUV	1/5	3/5	Imported	2/5	2/5

Stolen	
Yes	No
1/2	1/2

Likelihood of yes = $\frac{4}{5} \times \frac{1}{5} \times \frac{3}{5} \times \frac{1}{2} = \frac{6}{125} = 0.048$

Likelihood of no = $\frac{2}{5} \times \frac{3}{5} \times \frac{3}{5} \times \frac{1}{2} = \frac{9}{125} = 0.072$

Therefore the prediction is no.

Que 4.13. **What is Laplacian correction in Bayesian classifier ? Compute the class of the four following tuple by using Bayesian classification for given database in table. X = (age = senior, credit rating = fair, income= medium, student= no).**

Table 4.13.1.

Age	Income	Student	Credit rating	Class : buys computer
youth	high	No	Fair	No
youth	high	No	Excellent	No
middle aged	high	No	Fair	Yes
senior	medium	No	Fair	Yes
senior	low	Yes	Fair	Yes
senior	low	Yes	Excellent	No
middle aged	low	Yes	Excellent	Yes
youth	medium	No	Fair	No
youth	low	Yes	Fair	Yes
senior	medium	Yes	Fair	Yes
youth	medium	Yes	Excellent	Yes
middle aged	medium	No	Excellent	Yes
middle aged	high	Yes	Fair	Yes
senior	medium	No	Excellent	No

AKTU 2017-18, Marks 10

Answer

1. Laplacian correction is a technique used for avoiding zero probability values.
2. When training set is large enough that adding one to each count will make negligible difference in estimated probability (avoiding zero probability value) we use Laplacian correction.
3. If we have q counts to which we each add one, then we must remember to add q to the corresponding denominator used in the probability calculation.

Numerical :

X = (age = senior, income = medium, student = no, credit rating = fair)

We need to maximize $P(X|C_i)P(C_i)$, for $i = 1, 2$. $P(C_i)$, the priori probability of each class, can be estimated based on the training tuples :

P(buys_computer = yes) = 9/14 = 0.643

P(buys_computer = no) = 5/14 = 0.357

To compute $P(X|C_i)$, for i = 1, 2, we compute the following conditional probabilities:

P(age=senior | buys_computer=yes)= 3/9= 0.333

P(age=senior | buys_computer=no)=2/5=0.400

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

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

P(student=no | buys_computer=yes)=3/9=0.333

P(student=no | buys_computer=no)=4/5=0.800

P(credit_rating=fair | buys_computer=yes)=6/9=0.667

P(credit_rating=fair | buys_computer=no)=2/5=0.400

Using the above probabilities :

$P(X|$buys_computer=yes) = P(age=senior | buys_computer=yes) × P(income=medium | buys_computer=yes) × P(student=no | buys_computer=yes) × P(credit_rating=fair | buys_computer=yes) = 0.033

$P(X|$buys_computer=no)= P(age=senior | buys_computer=no) ×P(income=medium | buys_computer=no)× P(student=no | buys_computer=no) ×P(credit_rating=fair | buys_computer=no) = 0.051

Compute $P(X|C_i)P(C_i)$ for each class :

$P(X|$buys_computer=yes) × P(buys_computer=yes)=0.033 ×0.643=0.021

$P(X|$buys_computer=no) × P(buys_computer=no)=0.051 ×0.357=0.018

The Bayesian Classifier predicts buys_computer=yes for tuple X.

Que 4.14. Explain distance-based algorithms in detail.

Answer

1. Distance-based algorithms are non-parametric methods that can be used for classification.
2. These algorithms classify objects by the dissimilarity between them as measured by distance functions.
3. There are two types of distance-based algorithm :
   a. **Simple approach :** It assumes that each class is represented by its center or centroid. The new item is placed in the class with the largest similarity value.
   b. ***k*-nearest neighbour :** The KNN scheme requires not only training set but also the desired classification for each item. When a classification is to be made for a new item, its distance to each item in the training set must be determined. Only the k closest entries in the training set are considered. The new item is then placed in the class that contains the most items for this set of k closest items.

Algorithm :
Input :
T // Training data
K // Number of neighbours
t // Input tuple to classify
Output :
c // class to which t is assigned
KNN algorithm :
// Algorithm to classify tuple using KNN
$N = \phi$;
// Find set of neighbours, N, for t
for each $d \in T$ do
if $|N| \leq K$, then
$N = N \cup \{d\}$;
else
if $\exists\, x \in N$ such that
sim $(t, u) \leq$ sim (t, d), then
begin
$N = N - \{u\}$;
$N = N \cup \{d\}$;
end
// Find class for classification
c = class to which the most $u \in N$ are classified.

PART-4

Decision Tree-Based Algorithms.

CONCEPT OUTLINE

- Decision tree-based algorithms are :
  1. ID3
  2. C4.5
  3. CART
  4. Scalable DT Techniques

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.15. How are decision trees useful in data mining? Explain.

AKTU 2013-14, Marks 10

Answer

1. Decision trees are used for processing a large amount of data and thus find use in data mining.
2. The decision tree approach is most useful in classification problems. With this technique, a tree is constructed to model the classification process.
3. Decision tree can handle high dimensional data and easily understand by humans.
4. The learning and classification steps of decision tree induction are simple and fast.
5. Decision tree induction algorithm has been useful in many application areas like medicine, manufacturing and production, financial analysis, astronomy, etc.

Que 4.16. Write the algorithm of decision tree induction. What are the methods that can be used for selecting the splitting criteria?

AKTU 2015-16, Marks 10

OR

Write a short note on gain ratio.

AKTU 2017-18, Marks 2.5

Answer

1. A decision tree is a structure that includes a root node, branches, and leaf nodes.
2. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.
3. Once the tree is built, it is applied to each tuple in the database and results in a classification for tuple. There are two basic steps in the technique: building the tree and applying the tree to the database.

Decision tree induction algorithm :

1. **Create a node N;**
2. **If tuples in D are all of the same class, C then**
3. **Return N as a leaf node labeled with the class C;**
4. **If attribute-list is empty then**
5. **Return N as a leaf node labeled with the majority class in D; // majority voting.**
6. **Apply Attribute_selection _method(D, attribute_list) to find the "best" splitting-criterion;**
7. **Label node N with splitting criterion;**
8. **If splitting-attribute is discrete-valued and multiway splits allowed then // not restricted to binary trees**
9. **Attribute_list ← attribute list – splitting-attribute; // remove splitting_attribute**
10. **For each outcome j of splitting-criterion**

    // partition the tuples and grow subtrees for each partition
11. **Let D_j be the set of data tuples in D satisfying outcome j; II a partition**
12. **If D_j is empty then**
13. **Attach a leaf labeled with the majority class in D to node N;**
14. **Else attach the node returned by generate decision_tree (D_j, attribute _list) to node N;**

    endfor
15. **Return N;**

Methods used for selecting the splitting-criterion:

1. **Information gain :** Information gain is used as an attribute selection measure. In this, we pick the attribute that has the highest information gain.

$$\text{Gain}(D, A) = \text{Entropy}(D) - \sum_{j=1}^{v} \frac{|D_j|}{|D|} \text{Entropy}(D_j)$$

2. **Gain ratio :** It is the modification of the information gain that reduces its bias towards multi-valued attributes. Gain ratio takes number and size of branches into account when choosing an attribute. It corrects the information gain by taking the intrinsic information of a split into account.

$$GR(S, A) = \frac{\text{Gain}(S, A)}{\text{IntI}(S, A)}$$

3. **Gini index :** The gini index is used in CART. It considers a binary split for each attribute. When considering a binary split, we compute a weighted sum of the impurity of each resulting partition.

$$\text{Gini}(S) = 1 - \sum p_i^2$$

Que 4.17. **Compute the decision rules by deriving a decision tree classifier and information gain as selection measure for the given database in table. 4.13.1.**
Given : Gain (age) = 0.246, Gain (student) = 0.151 and Gain (credit rating) = 0.048

AKTU 2017-18, Marks 10

Answer

Given : Gain (age) = 0.246, Gain (student) = 0.151 and Gain (credit rating) = 0.048

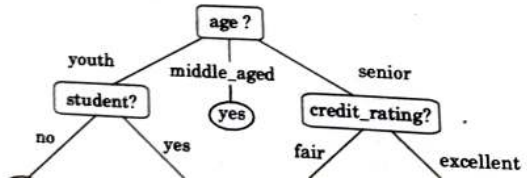
Age has the highest gain, therefore it is used as the decision attribute in the root node. Since, age has three possible values, the root node has three branches :

Gain (youth, student) = 0.970
Gain (youth, income) = 0.570
Gain (youth, credit_rating) = 0.019

The above calculations show that the attribute student shows the highest gain. Therefore, it should be used as the next decision node for the branch youth. This process is repeated until all data are classified perfectly or no attribute is left for the child nodes.

The corresponding rules are :

1. If age = youth and student = no then buys_computer = no
2. If age = youth and student = yes then buys_computer = yes
3. If age = middle-age then buys_computer = yes
4. If age = senior and credit_rating = excellent then buys_computer = yes
5. If age = senior and credit_rating = fair then buys_computer = no



the decision tree induction algorithm with appropriate examples. Discuss the disadvantages of this approach ? What is over fitting, and how can it be prevented for decision trees ?

AKTU 2016-17, Marks 10

Answer

Issues to be consider when employing a decision tree :

1. Choosing splitting attributes
2. Ordering of splitting attributes
3. Splits
4. Tree structure
5. Stopping criteria
6. Training data
7. Pruning

Decision tree induction algorithm with example : Refer Q. 4.16, Page 4–15D and Q. 4.17, Page 4–17D; Unit-4.

Disadvantages of decision tree :

1. They do not easily handle continuous data.
2. Handling missing data is difficult.
3. Difficult to use when we have smooth boundaries.

Overfitting : Since the decision tree is constructed from the training data, overfitting may occur. This can be overcome via tree pruning.

PART-5

Clustering : Introduction, Similarity and Distance Measures.

CONCEPT OUTLINE

- Clustering is the grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities.
- Clustering is of two types :
  1. Hierarchical clustering 2. Partitional clustering

iv. [illegible]

v. Model-based method

Que 4.20. **Explain the different data types used in cluster analysis.**

AKTU 2014-15, Marks 05

OR

Describe the types of data that often occur in cluster analysis and briefly explain how to preprocess that data for clustering.

AKTU 2015-16, Marks 10

Answer

Data types used in cluster analysis are :

1. **Interval scaled variables :** Interval scaled variables are continuous measurements of roughly linear scale. Typical examples include weight and height, latitude and longitude coordinates (for example, when clustering houses), and weather temperature.
2. **Binary variables :** A binary variables has only two states : 0 and 1, where 0 means that the variable is absent, and 1 means that it is present. Given the variable smoker describing a patient, for instance, 1 indicates that the patient smokes, while 0 indicates that the patient does not.
3. **Categorical variables :** A categorical variable is a generation of the binary variable in that it can take on more than two states. For example, map colour is a categorical variable that may have, say, five states : red, yellow, green, pink and blue.
4. **Ordinal variables :** Ordinal variables are very useful for registering subjective assessments of qualities that cannot be measured objectively. For example, professional ranks are often enumerated in a sequential order, such as assistant, associate end full for professors.
5. **Ratio scaled variables :** A ratio scaled variable makes a positive measurement on a non-linear scale, such as an exponential scale, approximately following the formula :

CONCEPT OUTLINE

- Types of hierarchical clustering :
  1. CURE
  2. Chameleon

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.21. **What is hierarchical method for clustering ? Explain BIRCH method.**

AKTU 2015-16, Marks 10

AKTU 2017-18, Marks 10

Answer

Hierarchical method for clustering :

1. Hierarchical clustering creates hierarchy of clusters on the data set.
2. This hierarchical tree shows levels of clustering with each level having a larger number of smaller clusters. CURE and Chameleon are the examples of hierarchical clustering.
3. **There are two approaches of hierarchical algorithm :**
   a. **Agglomerative clustering :** Agglomerative hierarchical clustering is a bottom-up clustering method where clusters have sub-clusters. Each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
   b. **Divisive clustering :** It starts with a single cluster containing all objects, and then successively splits resulting clusters until only clusters of individual objects remain. This is a "top-down" approach

in which all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) :

1. It is a scalable clustering method designed for very large data sets. In this, only one scan of data is necessary.
2. It is based on the notation of CF (Clustering Feature) tree. CF tree is a height balanced tree that stores the clustering features for a hierarchical clustering.
3. Cluster of data points is represented by a triple of numbers (N, LS, SS) where N = Number of items in the sub cluster, LS = Linear sum of the points, SS = Sum of the squared of the points.

BIRCH clustering algorithm :

a. **Phase 1 :** Build the CF Tree. Load the data into memory by building a cluster-feature tree. Optionally, condense this initial CF tree into a smaller CF.

b. **Phase 2 :** Global Clustering. Apply an existing clustering algorithm on the leaves of the CF tree. Optionally, refine these clusters.

4. BIRCH is sometimes referred to as two-step clustering, because of the two phases.

Que 4.32. Write a short note on hierarchical and non-hierarchical clustering. **AKTU 2013-14, Marks 05**

Answer

1. Non-hierarchical or partitional clustering is faster than hierarchical clustering.
2. Hierarchical clustering requires only a similarity measure, while partitional clustering requires stronger assumptions such as number of clusters and the initial centers.
3. Hierarchical clustering does not require any input parameters, while partitional clustering algorithms require the number of clusters to start running.
4. Hierarchical clustering returns a much more meaningful and subjective division of clusters but partitional clustering results in exactly k clusters.
5. Hierarchical clustering algorithms are more suitable for categorical data

Answer

1. **CURE :** (Clustering Using Representatives) is an agglomerative algorithm.

a. CURE is an efficient clustering algorithm for databases, which is more robust to outliers compared with other clustering methods, and identifies clusters having non-spherical shapes and wide variances in size.

b. One objective for the CURE (Clustering Using Representative) clustering algorithm is to handle outliers well. It has both a hierarchical component and a portioning component.

c. **Algorithm :**

Input :

$D = \{t_1, t_2, ..., t_n\}$ // Set of elements

k //Desired number of clusters

Output :

Q //Heap containing one entry for each cluster

CURE algorithm :

T = build(D);

Q = heapify(D); // Initially build heap with one entry per item;

repeat
u = min(Q) ;
delete (Q, u. close) ;
w = merge(u, v) ;
delete(T, u);
delete(T, v);
insert(T, w);
for each $x \in Q$ do
x . close = find closest cluster to x;
if x is closest to w, then
w. close = x;
insert (Q, w);
until number of nodes in Q is k;

2. **Chameleon :**

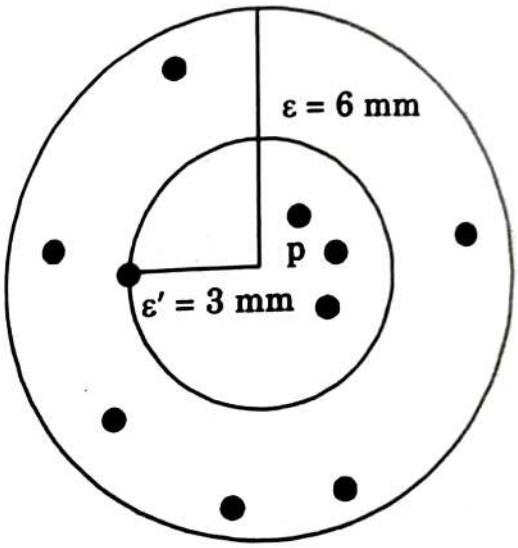
a. Chameleon is clustering using dynamic modeling.

b. Chameleon is a hierarchical clustering that uses dynamic modeling between pairs of clusters.

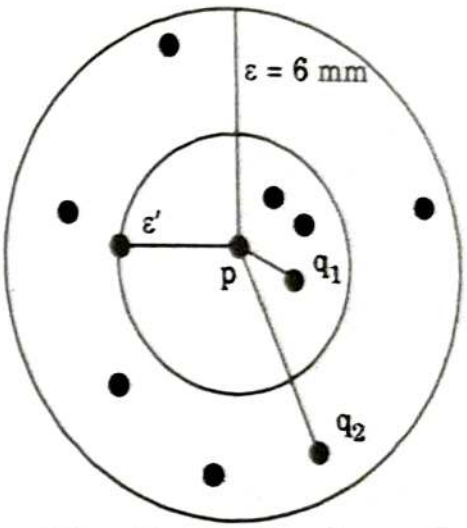
Two values need to be stored for each object, core-distance and reachability-distance.

1. **Core-distance :** The core-distance of an object p is the smallest ε value that makes {p} a core object. If p is not a core object, the core-distance of p is undefined.

2. **Reachability-distance :** The reachability-distance of an object q with respect to another object p is the greater value of the core-distance of p and the Euclidean distance between p and q. If p is not a core object, the reachability-distance between p and q is undefined.



Core-distance of p



Reachability-distance (p, q_1) = ε' = 3 mm
Reachability-distance (p, q_2) = d(p, q_2)

Fig. 4.24.1.

Que 4.25. Write the *k*-mean algorithm. Suppose that the data mining task is to cluster points (with (*x*, *y*) representing location) into three clusters, where the points are : A1 (2,10), A2 (2, 5), A3 (8, 4) B1 (5, 8), B2 (7, 5), B3 (6, 4), C1 (1, 2), C2 (4, 9)

The distance function is Euclidian distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the *k*-means algorithm to show only the three cluster centers after the first round of execution.

AKTU 2017-18, Marks 10

Answer

***k*-mean algorithm :** Refer Q. 4.24, Page 4–23D, Unit-4.

Numerical : The Euclidean distances between the given points are in the following matrix :

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

a. seed1 = A1 = (2, 10), seed2 = A4 = (5, 8), seed3 = A7 = (1, 2)

epoch1 – start :

A1 :

d(A1, seed1) = 0 as A1 is seed1

d(A1, seed2) = $\sqrt{13} > 0$

d(A1, seed3) = $\sqrt{65} > 0$

→ A1 ∈ cluster 1

A2 :

d(A2, seed1) = $\sqrt{25} = 5$

d(A2, seed2) = $\sqrt{18} = 4.24$

d(A2, seed3) = $\sqrt{10} = 3.16$ ← smaller

→ A2 ∈ cluster 3

A3 :

d(A3, seed1) = $\sqrt{36} = 6$

d(A3, seed2) = $\sqrt{25} = 5$ ← smaller

d(A3, seed3) = $\sqrt{53} = 7.28$

→ A3 ∈ cluster 2

B1 :

d(B1, seed1) = $\sqrt{13}$

d(B1, seed2) = 0 as B1 is seed2

d(B1, seed3) = $\sqrt{52} > 0$

→ B1 ∈ cluster 2

B2 :

d(B2, seed1) = $\sqrt{50} = 7.07$

d(B2, seed2) = $\sqrt{13} = 3.60$ ← smaller

d(B2, seed3) = $\sqrt{45} = 6.70$

→ B2 ∈ cluster 2

B3 :

d(B3, seed1) = $\sqrt{52} = 7.21$

d(B3, seed2) = $\sqrt{17} = 4.12$ ← smaller

d(B3, seed3) = $\sqrt{29} = 5.380$

→ B3 ∈ cluster 2

C1 :

d(C1, seed1) = $\sqrt{65} > 0$

d(C1, seed2) = $\sqrt{52} > 0$

d(C1, seed3) = 0 as A7 is seed3

→ C1 ∈ cluster 3

C2 :

d(C2, seed1) = $\sqrt{5}$

d(C2, seed2) = $\sqrt{2}$ ← smaller

d(C2, seed3) = $\sqrt{58}$

→ C2 ∈ cluster 2

end of epoch1

new clusters : 1 : {A1}, {A3, A4, A5, A6, A8} 3: {A2, A7}

b. Centers of the new clusters :

C1 = (2, 10), C2 = ((8 + 5 + 7 + 6 + 4)/5 = (6, 6), C3 = ((2 + 1)/2, (5 + 2)/2) = (1.5, 3.5)

Que 4.26. **Consider five points {X_1, X_2, X_3, X_4, X_5} with the following coordinates as a two dimensional sample for clustering : X_1 = (0, 2.25); X_2 = (0,0.25); X_3 = (1.25,0); X_4 = (4.5,0); X_5 = (4.5, 2.5); Illustrate the *k*-means partitioning algorithm (clustering algorithm) using the above data set.**

AKTU 2016-17, Marks 15

Answer

Given : $X_1 = (0, 2.25)$, $X_2 = (0, 0.25)$, $X_3 = (1.25, 0)$, $X_4 = (4.5, 0)$, $X_5 = (4.5, 2.5)$

Since, coordinates is a two-dimensional sample for clustering.

Initially, clusters are formed from random distribution of samples :

$$C_1 = \{X_1, X_2, X_4\} \text{ and } C_2 = \{X_3, X_5\}$$

Since, $$\text{centroid } (M_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} X_{ik}$$

So, the centroid for these two clusters are :

$$M_1 = \{(0 + 0 + 4.5)/3, (2.25 + 0.25 + 0)/3\} = \{1.5, 0.83\}$$

$$M_2 = \{(1.25 + 4.5)/2, (0 + 2.5)/2\} = \{2.875, 1.25\}$$

The error within-cluster variation

$$e_k^2 = \sum_{i=1}^{n_k} (X_{ik} - M_k)^2$$

Within-cluster variation after initial random distribution of sample, are

$$e_1^2 = [(0 - 1.5)^2 + (2.25 - 0.83)^2] + [(0 - 1.5)^2 + (0.25 - 0.83)^2] + [(4.5 - 1.5)^2 + (0 - 0.83)^2]$$

$$= 16.5417$$

$$e_2^2 = [(1.25 - 2.875)^2 + (0 - 1.25)^2] + [(4.5 - 2.875)^2 + (2.5 - 1.25)^2]$$

$$= 8.40625$$

The total square error is

$$E^2 = e_1^2 + e_2^2 = 16.5417 + 8.40625 = 24.94795$$

Depending on a minimum distance from centroids μ_1 and μ_2, the new redistribution of samples inside clusters will be

$$\left.\begin{aligned} d(M_1, X_1) &= (1.5^2 + 1.42^2)^{1/2} = 2.066 \\ d(M_2, X_1) &= (2.875^2 + 1^2)^{1/2} = 3.044 \end{aligned}\right\} \quad X_1 \in C_2$$

$$\left.\begin{aligned} d(M_1, X_2) &= (1.5^2 + 0.58^2)^{1/2} = 1.608 \\ d(M_2, X_2) &= (2.875^2 + 1^2)^{1/2} = 3.044 \end{aligned}\right\} \quad X_2 \in C_1$$

$$\left.\begin{aligned} d(M_1, X_3) &= (1.5^2 + 0.83^2)^{1/2} = 1.714 \\ d(M_2, X_3) &= (2.875^2 + 1.25^2)^{1/2} = 3.135 \end{aligned}\right\} \quad X_3 \in C_1$$

$$\left.\begin{aligned} d(M_1, X_4) &= (1.5^2 + 0.83^2)^{1/2} = 1.714 \\ d(M_2, X_4) &= (2.875^2 + 1.25^2)^{1/2} = 3.135 \end{aligned}\right\} \quad X_4 \in C_1$$

$$\left.\begin{aligned} d(M_1, X_5) &= (1.5^2 + 1.67^2)^{1/2} = 2.245 \\ d(M_2, X_5) &= (2.875^2 + 1.25^2)^{1/2} = 3.135 \end{aligned}\right\} \quad X_5 \in C_2$$

Above calculation is based on Euclidean distance formula,

$$d(X_i, X_j) = \sum_{k=1}^{m} (X_{ik} - X_{jk})^{1/2}$$

New clusters $C_1 = \{X_2, X_3, X_4\}$ and $C_2 = \{X_1, X_5\}$ have new centroids.

$$M_1 = \{1.679, 3.105\}$$

$$M_2 = \{2.156, 3.089\}$$

The corresponding within-cluster variations

$$e_1^2 = [(1.608 - 1.679)^2 + (3.044 - 3.105)^2] + [(1.714 - 1.679)^2 + (3.135 - 3.105)^2] + [(1.714 - 1.679)^2 + (3.135 - 3.105)^2] = 0.0129$$

$$e_2^2 = [(2.066 - 2.156)^2 + (3.044 - 3.089)^2] + [(2.245 - 2.156)^2 + (3.135 - 3.089)^2] = 0.0201$$

So, the total square error is

$$E^2 = e_1^2 + e_2^2 = 0.033$$

PART-7

Density-Based Methods : DBSCAN, OPTICS, Grid-Based Methods: STING, CLIQUE Model-Based Method : Statistical Approach.

CONCEPT OUTLINE

- Density-based methods are of two types :
  1. DBSCAN 2. OPTICS
- Grid-based methods are of two types :
  1. STING 2. CLIQUE

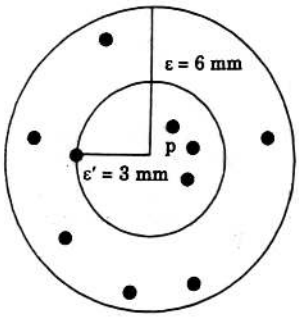
DBSCAN algorithm :

$k = 0$; // Initially there are no clusters.
for $i = 1$ to n do
if t_i is not in a cluster, then
$X = \{t_j \mid t_j$ is density-reachable from $t_i\}$;
if X is a valid cluster, then
$k = k + 1$;
$K_k = X$;

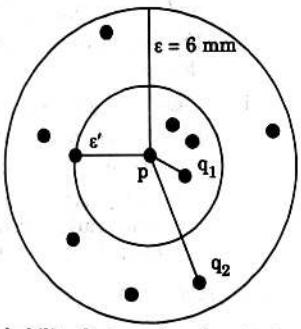
The expected time complexity of DBSCAN is $O(n \log n)$.

b. OPTICS :

i. OPTICS is a variation of DBSCAN which was designed to surmount issues occurs in DBSCAN.
ii. OPTICS does not explicitly produce a data set clustering.
iii. It instead gives us cluster ordering such that objects which are in a denser cluster are closer in a list.
iv. OPTICS stores two additional attributes *i.e.*, Core-distance and reachability distances, which are used to derive the ordering such that clusters with higher density will be finished first.



Core-distance of p



Reachability-distance (p, q_1) = ε' = 3 mm
Reachability-distance (p, q_2) = d(p, q_2)

Fig. 4.27.1.

Que 4.28. Discuss in detail various grid-based methods.

OR

Explain STING in detail. **AKTU 2017-18, Marks 05**

Answer

1. In this, the objects together form a grid.
2. The object space is quantized into finite number of cells that form a grid structure.
3. Grid-based methods are of two types :

a. STING (Statistical Information Grid) :

i. STING works with numerical attributes.
ii. Information's such as mean, maximum and minimum are pre-computed and stored in rectangular cells.
iii. Parameters at the higher level cells are drawn from the parameters of the bottom level cells.
iv. For each cell, there are attribute independent parameters and attribute dependent parameters.

Algorithm :

Input :

T // Tree
q // Query

Output :

R // Regions of relevant cells

STING algorithm :

$i = 1$
repeat
for each node in level i do
determine if this cell is relevant to q and mark as such;
$i = i + 1$
until all layers in the tree have been visited;
identify neighbouring cells of relevant cells to create regions of cells;

b. CLIQUE :

i. CLIQUE (Clustering in QUEst) is a bottom-up subspace clustering algorithm that constructs static grids.
ii. It uses apriori approach to reduce the search space.
iii. CLIQUE is a density and grid based and find out the clusters by taking density threshold and number of grids as input parameters.

iv. Steps in CLIQUE are :

1. The dimension space is partitioned into no overlapping units called cells.
2. Identify the dense and sparse cells.
3. Use the dense cells to assemble the clusters.
4. Starting with an arbitrary dense cell, we find the maximal region of all connected dense cells in all dimensions.
5. Repeat step 4 until all cells are covered.

Answer

Association rule mining : Refer Q 4.29, Page 4–31D, Unit-4.

Apriori algorithm :

a. The Apriori algorithm is used in association rule mining which uses the property of large itemset *i.e.*, any subset of a large itemset must be large.

b. Apriori uses bottom-up approach, where frequent subsets are extended one item at a time.

c. The entire algorithm can be divided into two steps :

Step 1 : Apply minimum support to find all the frequent sets with k items in a database.

Step 2 : Use the self-join rule to find the frequent sets with $k + 1$ items with the help of frequent k-itemsets. Repeat this process from $k = 1$ to the point when we are unable to apply the self-join rule.

Algorithm :

Input :

I // Itemsets

D // Database of transactions

s // Support

Output :

L // Large itemsets

Apriori algorithm :

$k = 0$; // k is used as the scan number.

$L = \phi$;

$C_1 = I$; // Initial candidates are set to be the items.

repeat

$k = k + 1$;

$L_k = \phi$;

for each $I_i \in C_k$ do

$c_i = 0$; // Initial counts for each itemset are 0.

for each $t_j \in D$ do

for each $I_j \in C_k$ do

if $I_i \in t_j$ then

$c_i = c_i + 1$;

for each $I_i \in C_k$ do

if $c_i \geq (s \times | D |)$ do

$L_k = L_k \cup I_i$;

$L = L \cup L_k$;

C_{k+1} = Apriori-gen(L_k)
until $C_{k+1} = \phi$;

Que 4.31. **Find frequent patterns under the association rules by using Apriori algorithm for the following transactional database :**

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

Let minimum support = 60 % and minimum confidence = 80 %

AKTU 2017-18, Marks 10

Answer

Using Apriori, we successively generate the sets C_k of candidate k-itemsets, and then verify these for minsup, obtaining the sets L_k of frequent k-itemsets. In the database D, this leads to :

C_1 =

Item	Count
A	A
C	2
D	1
E	4
I	1
K	5
M	3
N	2
O	3
U	1
Y	3

L_1 =

Item	Count
E	4
K	5
M	3
O	3
Y	3

C_2 =

Item	Count
EK	4
EM	2
EO	3
EY	2
KM	3
KO	3
KY	3
MO	1
MY	2
OY	2

L_2 =

Item	Count
EK	4
EO	3
KM	3
KO	3
KY	3

$C_3 = L_3$ =

Item	Count
EKO	3

Association rules :

$\{KO\} \Rightarrow \{E\}$
$\{EO\} \Rightarrow \{K\}$

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.32. **Explain parallel and distributed algorithms.**

Answer

Parallel and distributed algorithms : Parallel or distributed and algorithms strive to parallelize either the data, known as data parallelism, or the candidates, referred to as task parallelism.

Data parallelism :

1. One data parallelism algorithm is the Count Distribution Algorithm (CDA).
2. The database is divided into p partitions, one for each processor.
3. Each processor counts the candidates for its data and then broadcasts its counts to all other processors.
4. Each processor then determines the global counts.
5. These counts are used to determine the large itemsets and to generate the candidates for the next scan.

Task parallelism :

1. The Data Distribution Algorithm (DDA) demonstrates task parallelism.
2. Here the candidates as well as the database are partitioned amount the processors.
3. Each processor in parallel counts the candidates given to it using its local database partition.
4. Then each processor broadcasts its database partition to all other processors.
5. Each processor then uses this to obtain a global count for its data and broadcasts this count to all other processors.
6. Each processor then can determine globally large itemsets and generate

CONCEPT OUTLINE

- A neural network usually involves a large number of processors operating in parallel and arranged in tiers.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.33. What do you mean by neural network ?

Answer

1. An artificial neural network, often called a neural network, is a mathematical model based on biological neural networks.
2. A neural network consists of an interconnected group of artificial neurons, and its information system.
3. Neural networks are used to model complex relationships between inputs and outputs or to find patterns in data.
4. Neural network method is used for classification, clustering, feature mining, prediction and pattern recognition.
5. The neural network model can be broadly divided into the following three types :
   a. Feed-forward network
   b. Feedback network
   c. Self-organization networks
6. Using neural networks as a tool, data warehousing firms are harvesting information from datasets in the process known as data mining.

Que 4.34. Explain multilayer feed-back neural network. Differentiate between feed-forward and feedback system.

AKTU 2014-15, Marks 10

3. Feedback loops, has a profound impact on learning capability of the network and on its performance.

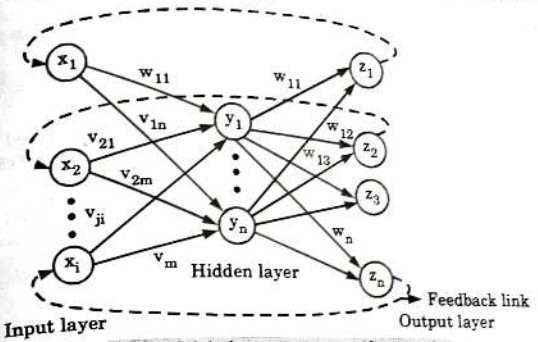


Fig. 4.34.1. A recurrent neural network.

Difference between feed-forward and feedback system :

S. No.	Feed-forward system	Feedback system
1.	Feed-forward system allows signals to travel one way only from input to output.	Feedback system allows signal travelling in both directions by introducing loops in the network.
2.	There is no feedback.	There is a feedback.
3.	Feed-forward tends to be straight forward network that associate input with output.	Computations derived from earlier input are feedback into the network.
4.	Feed-forward are static *i.e.*, in feed-forward, state does not changes.	Feedback are dynamic *i.e.*, in feedback, state changes continuously until equilibrium is reached.

...atic algorithm.

Answer

1. The genetic algorithm is derived from natural evolution. In genetic algorithm, first of all, the initial population is created.
2. This initial population consists of randomly generated rules. We can represent each rule by a string of bits.
3. For example, in a given training set, the samples are described by two boolean attributes such as A1 and A2. And this given training set contains two classes such as C1 and C2.
4. We can encode the rule *IF A1 AND NOT A2 THEN C2* into a bit string *100*. In this bit representation, the two leftmost bits represent the attribute A1 and A2, respectively.

Advantages of genetic algorithm :

1. Does not require any derivative information.
2. Faster and more efficient as compared to the traditional methods.
3. Has very good parallel capabilities.

Disadvantage of genetic algorithm :

1. Genetic algorithms are not suited for all problems, especially problems which are simple and for which derivative information is available.
2. Fitness value is calculated repeatedly which might be computationally expensive for some problems.

5 UNIT Data Visualization and Overall Perspective

CONTENTS

PART-1

Aggregation, Historical Information.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.1. What do you mean by the term aggregation ?

Answer

1. Data aggregation is a process in which information is gathered and expressed in a summary form, for purposes such as statistical analysis.
2. The purpose is to get more information about particular groups based on specific variables such as age, profession, or income.
3. Data aggregation may be performed manually or through specialized software.

PART-2

Query Facility, OLAP Function and Tools, OLAP Servers, ROLAP, MOLAP, HOLAP

CONCEPT OUTLINE

- OLAP is an acronym for Online Analytical Processing and it performs multidimensional analysis of business data.
- Various OLAP servers are :
  1. ROLAP
  2. MOLAP
  3. HOLAP

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.2. Explain OLAP in detail.

Answer

1. **OLAP** (Online Analytical Processing) is computer processing that enables a user to easily and selectively extract and view data from different points of view.
2. **OLAP** allows users to analyze database information from multiple database systems at one time.
3. **OLAP** data is stored in multidimensional databases. OLAP processing is often used for data mining.
4. There are the following key features of OLAP :
   i. Multidimensional views of data
   ii. Support for complex calculations
   iii. Time intelligence
5. Applications of OLAP are :
   i. OLE DB for OLAP
   ii. Marketing and sales analysis
   iii. Consumer goods industries
   iv. Financial services industry (insurance, banks etc.)
   v. Database marketing

Que 5.3. Explain different types of OLAP operations.

Answer

Different types of OLAP operations are :

1. **Roll-up :** Roll-up is also known as "consolidation" or "aggregation". In the roll-up process at least one or more dimensions need to be removed. The roll-up operation can be performed in two ways :
   i. Reducing dimensions
   ii. Climbing up concept hierarchy
2. **Drill-down :** In drill-down data is fragmented into smaller parts. It is the opposite of the roll-up process. It can be done via moving down the concept hierarchy and increasing a dimension.
3. **Slice :** One dimension is selected, and a new sub-cube is created.
4. **Dice :** This operation is similar to a slice. In Dice, we select two or more dimensions that result in the creation of a sub-cube.
5. **Pivot :** In pivot, we rotate the data axes to provide a substitute presentation of data.

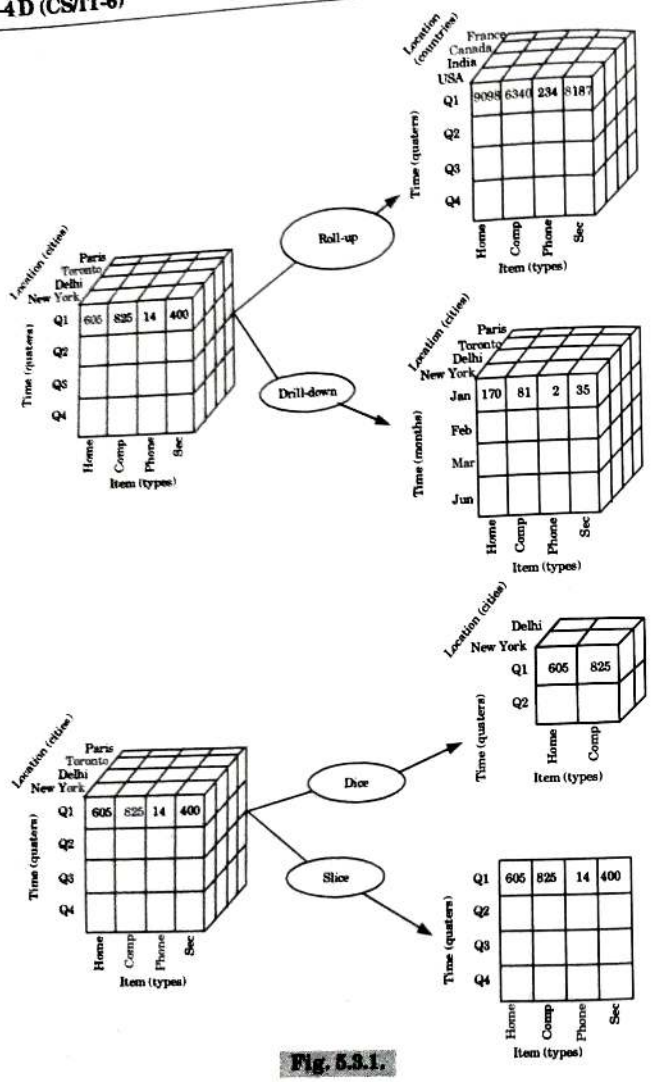


Fig. 5.3.1.

Que 5.4. Explain the various types of OLAP servers. What are the steps for efficient processing of OLAP queries ?

AKTU 2015-16, Marks 10

OR

Diagrammatically illustrate and discuss the architecture of MOLAP and ROLAP.

AKTU 2016-17, Marks 10

OR

Explain how query performance can be improved by cascading the operations.

AKTU 2015-16, Marks 10

Answer

Types of OLAP servers are :

1. **Relational OLAP :** ROLAP servers are placed between relational back-end server and client front-end tools. To store and manage warehouse data, ROLAP uses relational or extended-relational DBMS.

ROLAP architecture : ROLAP includes the following components :

1. Database server
2. ROLAP server
3. Front-end tool

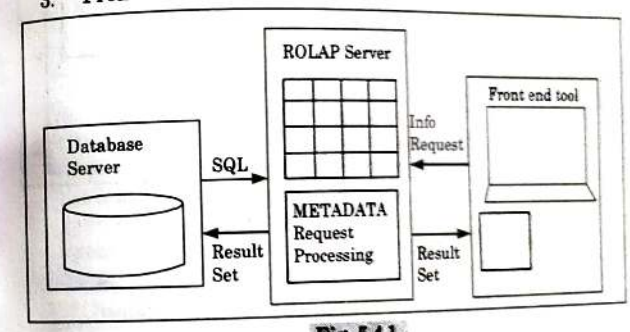


Fig. 5.4.1.

Advantages of ROLAP :

1. It can handle large amounts of data.
2. It can leverage functionalities inherent in the relational database.
3. ROLAP servers can be easily used with existing RDBMS.
4. Data can be stored efficiently, since no zero facts can be stored.
5. ROLAP tools do not use pre-calculated data cubes.
6. DSS server of micro-strategy adopts the ROLAP approach.

Disadvantages of ROLAP :

1. Performance can be slow
2. Limited by SQL functionalities
3. Hard to maintain aggregate tables

2. Multidimensional OLAP :

i. MOLAP stores in optimized multidimensional array storage, rather than in a relational database.

ii. With multidimensional data stores, the storage utilization may be low if the data set is sparse.

iii. Therefore, many MOLAP server use two levels of data storage representation to handle dense and sparse data sets.

MOLAP architecture : MOLAP includes the following components :

1. Database server
2. MOLAP server
3. Front-end tool

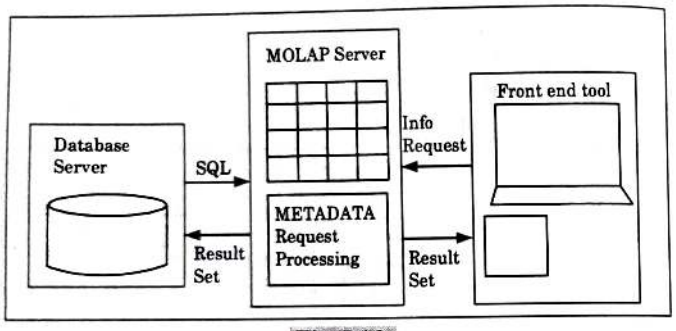


Fig. 5.4.2.

Advantages of MOLAP :

1. It is optimal for slice and dice operations.
2. Performance is better than ROLAP when data is dense.
3. It can perform complex calculations.
4. MOLAP allows fastest indexing to the pre-computed summarized data.
5. Helps the users connected to a network who need to analyze larger, less-defined data.
6. Easier to use, therefore MOLAP is suitable for inexperienced users.

Disadvantages of MOLAP :

1. Difficult to change dimension without re-aggregation.
2. MOLAP can handle limited amount of data.
3. Some MOLAP methodologies introduce data redundancy.
4. Requires additional investment.

3. Hybrid OLAP :

i. Hybrid OLAP is a combination of both ROLAP and MOLAP.

ii. It offers higher scalability of ROLAP and faster computation of MOLAP.

iii. HOLAP server allows storing large data volumes of detailed information.

iv. The aggregations are stored separately in MOLAP store.

Advantages of HOLAP :

1. HOLAP provides advantages of both MOLAP and ROLAP.
2. It provides fast access at all levels of aggregation.

Disadvantages of HOLAP : HOLAP architecture is very complex because it support both MOLAP and ROLAP servers.

Steps for efficient processing of OLAP queries :

Processing of OLAP queries :

To speed up the query processing in data cubes, the cuboids are materialized and OLAP index structures are constructed with following procedure :

1. **Determining which operation should be performed on the available cuboids :**
   a. This involves transformation of operations specified in the query into the corresponding SQL and/or OLAP operators.
   b. These operations include roll-up, drill-down, projection, selection, etc.
   c. For example, slicing and dicing operation on data cube can be transformed into selection and/or projection operations on materialized cuboids.
2. **Determining on which materialized cuboids(s) the relevant operations should be applied :** In this, all of the materialized cuboids are identified which may be useful for answering the query, pruning the relationships among the cuboids, estimating the cost of using the remaining materialized cuboids and selecting the cuboids with the least cost.

Que 5.5. **Define and describe the basic similarities and differences among ROLAP, MOLAP and HOLAP.** AKTU 2014-15, Marks 10 AKTU 2015-16, Marks 7.5

OR

Compare MOLAP vs HOLAP. AKTU 2013-14, Marks 05

OR

Write a short note on ROLAP vs MOLAP. AKTU 2017-18, Marks 2.5

Answer

Similarities between ROLAP, MOLAP and HOLAP : These three OLAP servers are used to implement data warehouses, and they are related to the logical model used to represent data.

Differences between ROLAP, MOLAP and HOLAP :

S. No.	Basis	ROLAP	MOLAP	HOLAP
1.	Storage location for detail data	Relational database	Multidimensional database	Relational database
2.	Storage location for summary aggregations	Relational database	Multidimensional database	Multidimensional database
3.	Storage space requirement	Large	Medium	Small
4.	Query-response time	Slow	Fast	Medium
5.	Processing time	Slow	Fast	Fast
6.	Latency	Low	High	Medium

Que 5.6. **Give E.F. Codd's 12 guidelines for OLAP.**

AKTU 2013-14, Marks 10

Answer

Dr. E.F. Codd the father of the relational model, created a list of rules to deal with the OLAP systems.

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

9. **Unrestricted cross-dimensional operations :** The OLAP tool should be able to perform operations across the dimensions of the cube.
10. **Intuitive data manipulation :** Data manipulation inherent in the consolidation path, such as drilling down or zooming out, should be accomplished via direct action on the analytical model's cells, and not require use of a menu or multiple trips across the user interface.
11. **Flexible reporting :** It is the ability of the tool to present the rows and column in a manner suitable to be analyzed.
12. **Unlimited dimensions and aggregation levels :** This depends on the kind of business, where multiple dimensions and defining hierarchies can be made.

PART-3

Data Mining Interface, Security, Backup and Recovery.

CONCEPT OUTLINE

- Data Mining Interface (DMI) is a web-based, interactive, dynamic report building module.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.7. **Write a short note on data mining interface.**

AKTU 2013-14, Marks 05

OR

Describe data mining interface in details. **AKTU 2014-15, Marks 05**

Answer

Data mining interface provides the medium that allows users to communicate with data mining processes. It is difficult to use data mining query languages. A graphical user interface (GUI) can be used to communicate with data mining systems.

A data mining interface may consist of following functional components :

i. **Data collection and data mining query composition :** It allows user to specify task relevant data sets and to compose data mining queries.

ii. **Presentation of discovered patterns :** It allows the display of discovered patterns in various forms like tables, graphs, charts, and other visualization techniques.

iii. **Hierarchy specification and manipulation :** It allows to do the specification of concept hierarchy, either manually or automatically.

iv. **Manipulation of data mining primitives :** It allows the dynamic adjustment of data mining operations like selection, display, and modification of concept hierarchies.

v. **Interactive multilevel mining :** It allows the roll-up or drill-down operations on discovered patterns.

The design of data mining interface should also consider the different classes of users. Users of data mining system can be classified into two categories : business analysts and business executives.

Que 5.8. **Write short note on backup and recovery.**

AKTU 2013-14, Marks 05

OR

Explain different backup and recovery models in data warehousing.

AKTU 2014-15, Marks 10

Answer

1. Backup and recovery refers to the process of backing up data in case of a loss and setting up systems that allow data recovery due to data loss.
2. A data warehouse is a complex system and it contains a huge volume of data.
3. Therefore, it is important to backup all the data so that it becomes available for recovery in future as per requirement.
4. Some of the backup terminologies are :
   a. **Complete backup :** It backup the entire database at the same time.
   b. **Partial backup :** Partial backup is very useful because various parts of the database are backed up in a round-robin fashion on a day-to-day basis.
   c. **Cold backup :** Cold backup is taken when the database is completely shut down.
   d. **Hot backup :** Hot backup is taken when the database engine is up and running.
   e. **Online backup :** It is quite similar to hot backup.

Following are different backup and recovery models :

1. **Full recovery model :** It provides the most flexibility for recovering database to an earlier point in time.

2. **Bulk-logged recovery model :** Bulk-logged recovery provides higher performance than lower log space consumption for certain large scale operations.
3. **Simple recovery model :** Simple recovery provides the highest performance and lowest log space consumption but with the significant exposure to data loss in the event of a system failure.

Que 5.9. **How data backup and data recovery is managed in data warehouse ?**

AKTU 2017-18, Marks 10

Answer

1. Managing the recovery of a large data warehouse is a difficult task and traditional OLTP backup and recovery strategies may not meet the needs of a data warehouse.
2. We should plan a backup strategy as part of our system design and consider what to backup and how frequently to backup.
3. The most important variables in our backup design are the amount of resources we have to perform a backup or recovery and the recovery time objective.
   a. NOLOGGING operations must be taken into account when planning a backup and recovery strategy. Traditional recovery, restoring a backup and applying the changes from the archive log, does not apply for NOLOGGING operations.
   b. Never make a backup when a NOLOGGING operation is taking place.
   c. Plan for one of the following or a combination of the following strategies :
      i. **The ETL strategy :** Recover a backup that does not contain non-recoverable transactions and replay the ETL that has taken place between the backup and the failure.
      ii. **The incremental backup strategy :** Perform a backup immediately after a non-recoverable transaction has taken place.

Strategies and best practices for backup and recovery : The following best practices can help us to implement our warehouse's backup and recovery strategy :

1. Use ARCHIVELOG mode
2. Use RMAN mode
3. Use read-only tablespaces
4. Plan for NOLOGGING operations
5. Not all tablespaces are equally important

PART-4

Tuning Data Warehouse and Testing Data Warehouse.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.10. Explain tuning in data warehouse.

Answer

1. Tuning in data warehouses are the processes of selecting adequate optimization techniques in order to make queries and updates run faster.
2. A data warehouse is usually accessed by complex queries for key business operations.
3. Therefore it becomes more difficult to tune a data warehouse system. The tuning of data warehouse can be done to improve the performance.
4. **Difficulties in data warehouse tuning are :**
   a. Data warehouse is dynamic; it never remains constant.
   b. It is very difficult to predict what query the user is going to post in the future.
   c. Business requirements change with time.
   d. Users and their profiles keep changing.

Que 5.11. Write a short note on testing data warehouse.

AKTU 2013-14, 2014-15; Marks 05

Answer

Testing is very important for data warehouse systems to make them work correctly and efficiently. There are three basic levels of testing performed on a data warehouse :

1. **Unit testing :** In unit testing, each component is separately tested. Each module, *i.e.*, procedure, program, SQL Script, Unix shell is tested. This test is performed by the developer.

2. **Integration testing :** In integration testing, the various modules of the application are brought together and then tested against the number of inputs. It is performed to test whether the various components do well after integration.
3. **System testing :** In system testing, the whole data warehouse application is tested together. The purpose of system testing is to check whether the entire system works correctly together or not. System testing is performed by the testing team.

Challenges of data warehouse testing are :

1. Data selection from multiple source and analysis that follows pose great challenge.
2. Volume and complexity of the data.
3. Redundant data in a data warehouse.
4. Inconsistent and inaccurate reports.

ETL testing is performed in five stages :

1. Identifying data sources and requirements
2. Data acquisition
3. Implement business logics and dimensional modeling
4. Build and populate data
5. Build reports

PART-5

Warehousing Applications and Recent Trends : Types of Warehousing Applications.

CONCEPT OUTLINE

- Applications of data warehouse are :
  i. Airline
  ii. Banking
  iii. Healthcare
  iv. Public sector

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.12. **What are the applications of data warehousing ?**

AKTU 2016-17, Marks 05

Answer

Applications of data warehousing are :

1. **Airline :** In the Airline system, it is used for operation purpose like crew assignment, analysis of route profitability, frequent flyer program promotions, etc.
2. **Banking :** It is widely used in the banking sector to manage the resources available on desk effectively. Few banks also used for the market research, performance analysis of the product and operations.
3. **Healthcare :** Healthcare sector also used data warehouse to strategize and predict outcomes, generate patient's treatment reports, share data with tie-in insurance companies, medical aid services, etc.
4. **Public sector :** In the public sector, data warehouse is used for intelligence gathering. It helps government agencies to maintain and analyze tax records, health policy records, for every individual.
5. **Investment and insurance sector :** In this sector, the warehouses are primarily used to analyze data patterns, customer trends, and to track market movements.
6. **Retain chain :** In retail chains, data warehouse is widely used for distribution and marketing. It also helps to track items, customer buying pattern, promotions and also used for determining pricing policy.
7. **Telecommunication :** A data warehouse is used in this sector for product promotions, sales decisions and to make distribution decisions.
8. **Hospitality industry :** This industry utilizes warehouse services to design as well as estimate their advertising and promotion campaigns where they want to target clients based on their feedback and travel patterns.

PART-6

Web Mining Spatial Mining and Temporal Mining.

CONCEPT OUTLINE

- Web mining is of three types :
  - i. Web content mining
  - ii. Web usage mining
  - iii. Web structure mining
- Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases.
- Temporal data mining refers to the extraction of implicit, non-trivial, and potentially useful abstract information from large collections of temporal data.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.13. **What is web mining ? Differentiate between web content mining, web structure mining and web usage mining.**

AKTU 2016-17, Marks 10

Answer

Web mining :

1. Web mining is an application of data mining techniques to find information patterns from the web data.
2. Web mining helps to improve the power of web search engine by identifying the web pages and classifying the web documents.
3. There are three types of web mining :
   a. **Web content mining :** Web content mining can be used for mining of useful data, information and knowledge from web page content. Web content mining performs scanning and mining of the text, images and groups of web pages according to the content of the input (query), by displaying the list in search engines.
   b. **Web usage mining :** Web usage mining is used for mining the web log records (access information of web pages) and helps to discover the user access patterns of web pages.
   c. **Web structure mining :** The web structure mining can be used to discover the link structure of hyperlink. The purpose of structure mining is to produce the structural summary of website and similar web pages.

View of data	a. Unstructured b. Structured	a. Semi-structured b. Website as DB	a. Link structure	a. Interactivity b. Server logs b. Browser logs	
Main data	a. Text documents b. Hypertext documents	a. Hypertext documents	a. Link structure	a. Server logs b. Browser logs	
Representation	a. Bag of words, *n*-gram terms b. Phrases, concepts or ontology c. Relational	a. Edge labeled graph b. Relational	a. Graph	a. Relational table b. Graph	
Method	a. Machine learning b. Statistical (including NLP)	a. Proprietary algorithms b. Association rules	a. Proprietary algorithms	a. Machine learning b. Statistical c. Association rules	
Applications categories	a. Categorization b. Clustering c. Finding extract rules d. Finding patterns in text	a. Finding frequent sub structures b. Web site schema discovery	a. Categorization b. Clustering	a. Site construction b. Adaptation and management	

Que 5.14. **Write a short note on spatial and temporal data mining.**

Answer

Spatial mining :

1. Spatial data mining is the application of data mining to spatial models.
2. In spatial data mining, analysts use geographical or spatial information to produce business intelligence or other results.
3. Challenges involved in spatial data mining include identifying patterns or finding objects that are relevant to the research project.

Temporal mining :

1. Temporal data mining is a single step in the process of knowledge discovery in temporal databases that enumerates structures over the temporal data.
2. Temporal data mining is concerned with the analysis of temporal data and for finding temporal patterns and regularities in sets of temporal data tasks of temporal data mining are :
   a. Data characterization and comparison
   b. Cluster analysis
   c. Classification
   d. Association rules
   e. Prediction and trend analysis
   f. Pattern analysis

Que 5.15. **Compare and contrast spatial, temporal mining with relevant examples.**

AKTU 2016-17, Marks 15

Answer

S. No.	Spatial mining	Temporal mining
1.	Spatial mining is the extraction of knowledge/ spatial relationships and interesting measures that are not explicitly stored in spatial database.	Temporal mining is the extraction of knowledge about occurrence of an event or values whether they follow cyclic, random, seasonal variations etc.
2.	It deals with spatial (location, geo-referenced) data.	It deals with implicit or explicit temporal content, from large quantities of data.

3.	It includes finding characteristic rules, discriminant rules, association rules and evaluation rules etc.	It aims at mining new and unknown knowledge, which takes into account the temporal aspects of the data.
4.	**For example :** Determining hotspots, unusual locations.	**For example :** An association rule which looks like – "Any person who buys a car also buys steering lock". By temporal aspect, this rule would be "Any person who buys a car also buys a steering lock after that".

Data Warehousing (2 Marks Questions)

1.1. Briefly explain important approaches to build the data warehouse. AKTU 2015-16, Marks 02

Ans. Two approaches to build a data warehouse are :

1. **Top-down approach :** In the top-down approach, data warehouse is built first. The data marts are then created from the data warehouse.
2. **Bottom-up approach :** In the bottom-up approach, data marts are created first and then data warehouse is built.

1.2. Why data warehouse is maintained separately from database ? AKTU 2015-16, Marks 02

Ans.

1. An operational database is constructed for well-known tasks and workloads such as searching particular records, indexing, etc. In contract, data warehouse queries are often complex and they present a general form of data.
2. An operational database query allows to read and modify operations, while an OLAP query needs only read only access of stored data.

1.3. How is the data warehouse different from a database ? AKTU 2016-17, Marks 02

Ans.

S. No.	Data warehouse	Database
1	It involves historical processing of information.	It involves day-to-day processing.
2.	It is used to analyze the business.	It is used to run the business.
3.	It focuses on information out.	It focuses on data in.
4.	It contains historical data.	It contains current data.

Ans. Benefits of data warehouse are :
1. Locating the right information.
2. Discovery of information.
3. Presentation of information.

1.9. What are the different components of data warehouse ?

Ans. Different components of data warehouse are :
1. Data warehouse database
2. ETL tools
3. Metadata
4. Access tools

1.10. Define the term data mart.

Ans. A data mart is a subject-oriented database that meets the demands of a specific group of users.

1.11. Mention the types of metadata.

Ans. Types of metadata are :
1. Descriptive metadata
2. Structural metadata
3. Administrative metadata

1.12. What is the difference between star and snowflake schema ?

Ans.

S. No.	Basis for Comparison	Star schema	Snowflake schema
1.	Structure of schema	Contains fact and dimension tables.	Contains sub-dimension tables including fact and dimension tables.
2.	Use of normalization	Does not use normalization.	Uses normalization and denormalization.
3.	Ease of use	Simple to understand and easily designed.	Hard to understand and design.
4.	Data model	Top-down	Bottom-up

1.13. List the different types of data warehouse.

Ans. Different types of data warehouses are :
1. Enterprise data warehouse
2. Operational data store
3. Data mart

1.14. Define parallelism.

Ans. Parallelism is the process to provide speed-up and scale up by

1.15. List the types of data mart.

Ans. Types of data mart are :

1. Independent data mart
2. Dependent data mart

Data Warehouse and Process Technology (2 Marks Questions)

2.1. Define workload matrix.

Ans. The workload matrix is a matrix that is created as the intersection of the tables in the data warehouse and the processes that will run in the data warehouse.

2.2. Describe key areas of security management.

Ans. Key areas of security management are :

1. Asset classification practices
2. Risk assessment and acceptance
3. Asset ownership
4. Security audits
5. Asset handling responsibilities

2.3. What are the causes of data errors ?

Ans. Causes of data errors are :

1. Missing values
2. Duplicates
3. Multiple hierarchies
4. Lack of referential integrity
5. Fields to be split up

2.4. What do you understand by fact and dimension tables ?

Ans. Fact tables are used to record actual facts or measures in the business, while dimension tables stores fields that describe the facts.

2.5. Explain the advantages of dimensional modeling.

Ans: Advantages of dimensional modeling are :

1. Dimensional modeling is simple.
2. Dimensional modeling promotes data quality.
3. Dimensional modeling makes use of relational database technology.

2.6. List the various functions of server.

Ans: Various functions of server are :

1. File sharing
2. Printer sharing
3. Database access
4. Communication

2.7. Discuss the factors while selecting the backup mechanism.

Ans: Factors while selecting the backup mechanism are :

1. Archive format
2. Automatic backup devices
3. Parallel data streams
4. Offsite backups
5. Incremental backups

2.8. Give some approaches to partitioning DB records.

Ans: Approaches to partitioning DB records are :

1. Range partitioning
2. Round-robin partitioning
3. Hash partitioning

2.9. Mention some features of transformation tools.

Ans: Features of transformation tools are :

1. Field splitting and consolidation
2. Standardization
3. De-duplication

3 UNIT

Data Mining (2 Marks Questions)

3.1. Define KDD. Identify the phases in KDD process.

AKTU 2015-16, Marks 02

Ans. Knowledge Discovery in Databases (KDD) refers to the process of discovering useful knowledge from data.

Phases in KDD process are :

1. Data cleaning
2. Data integration
3. Data selection
4. Data transformation
5. Data mining
6. Pattern evaluation
7. Knowledge presentation

3.2. Write short notes on linear regression.

AKTU 2015-16, Marks 02

Ans. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observe the data.

3.3. Why data cleaning routines are needed ?

AKTU 2016-17, Marks 02

Ans. Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

3.4. Sketch a neat diagram of architecture of a typical data mining system.

AKTU 2015-16, Marks 02

Ans.

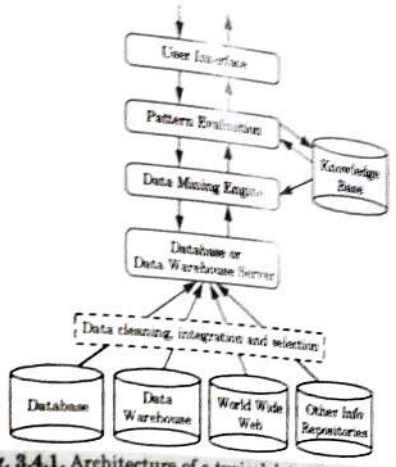


Fig. 3.4.1. Architecture of a typical data mining system.

3.5. Draw the diagram for key steps of data mining.

AKTU 2017-18, Marks 02

Ans.

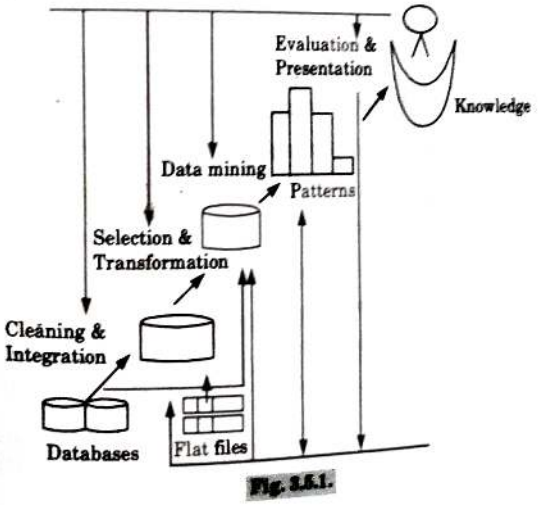


Fig. 3.5.1.

3.6. What is Chi-square test ? **AKTU 2017-18, Marks 02**

Ans. Chi-square test is a correlation relationship between two categorical (discrete) attributes.

3.7. Discuss the various factors which fueled the growth of data mining.

Ans. **The growth in the field of data mining has been fueled by a variety of factors :**

i. The explosive growth in data collection by businesses.

ii. The storing of the data in data warehouses for easy and reliable access by the entire enterprise.

iii. The availability of increased access to data from internet.

iv. Growth in computing power and storage capacity.

3.8. What is the motivation behind data mining ?

Ans. The data are added on daily basis to the databases. This makes the databases very large. Due to wide availability of huge amounts of data, there is a need for turning such data into useful information and knowledge. Hence the need of extracting important information from enormous amount of data motivated the mining of large amounts of data.

3.9. What are the components of data mining ?

Ans. **Components of data mining are :**

i. Databases

ii. Data warehouse server

iii. Knowledge base

iv. Data mining engine

v. Pattern evaluation module

vi. User interface

3.10. What is z-score normalization ?

Ans. In z-score normalization, the mean of the transformed set of data points is reduced to zero. For this, the mean and standard deviation of the initial set of data values are required. The transformation formula is :

$$v' = \frac{v - \overline{A}}{\sigma_A}$$

where, $\overline{A}$ and σ_A are the mean and standard deviation of the initial data values.

3.11. Define the term information gain.

Ans. Information gain is an attribute selection measure used in decision tree induction. The information gain is based on the decrease in entropy after a dataset is split on an attribute.

3.12. Define gain ratio.

Ans. Gain ratio is a ratio of information gain to the intrinsic information. It is a modification of information gain that reduces its bias by using a split information.

3.13. List the applications of data mining.

Ans. **Applications of data mining are :**

1. Market analysis and management.
2. Corporate analysis and risk management.
3. Fraud detection.

3.14. What are the disadvantages of data mining ?

Ans. **Disadvantages of data mining are :**

1. There are chances of companies may sell useful information of their customers to other companies for money.

2. Many data mining analytics software is difficult to operate and requires advance training to work on.
3. Different data mining tools work in different manners due to different algorithms employed in their design. Therefore, the selection of correct data mining tool is a very difficult task.

☺☺☺

4 UNIT

Classification and Clustering (2 Marks Questions)

4.1. What are attribute selection measures ? What is the drawback of information gain ? AKTU 2017-18, Marks 02

Ans: Attribute selection measure is mainly used to select the splitting criterion that best separates the given data partition. The popular attribute selection measures are information gain and gain ratio.

Drawback of information gain are :

1. Fragmentation
2. Problem occurs when information gain is applied to attributes that can take on a large number of distinct values.
3. Information gain is biased towards choosing attributes with a large number of values.

4.2. Write some of the facts of the association rule mining.

AKTU 2015-16, Marks 02

Ans: Association rule mining is a procedure which is meant to find frequent patterns, correlations, or associations from data sets found in various kinds of databases.

4.3. Briefly explain the concept of frequent item sets and closed item sets. AKTU 2015-16, Marks 02

Ans: **Frequent item set :** A set of items that appears in many baskets is said to be "frequent." To be formal, we assume there is a number s, called the support threshold. If I is a set of items, the support for I is the number of baskets for which I is a subset.

Closed item set : An item set is closed in a data set if there exists no superset that has the same support count as this original item set.

4.4. Describe the market basket analysis.

AKTU 2015-16, Marks 02

Ans. Market basket analysis is a modelling technique based upon the theory that if we buy a certain group of items, we are more (or less) likely to buy another group of items. It is used to determine what items are frequently bought together or placed in the same basket by customers.

4.5. Name main features of genetic algorithm.

AKTU 2017-18, Marks 02

OR

Name and describe the main features of genetic algorithm.

AKTU 2015-16, Marks 02

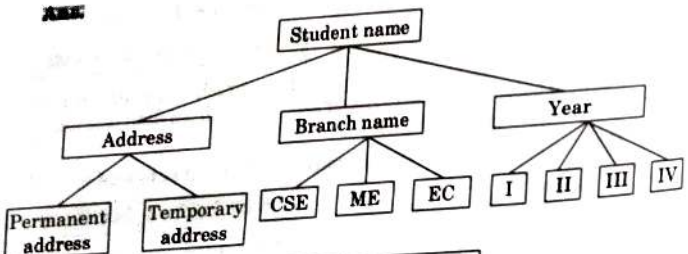
Ans. The main features of genetic algorithm are :

1. Encoding
2. Generalized uniform population
3. Fitness function
4. Uniform operator

4.6. Represent a decision tree for a student record database.

AKTU 2016-17, Marks 02

Ans.



Student name	
Address	
Branch name	
Year	

Fig. 4.6.1.

4.7. Define the term support and confidence.

AKTU 2017-18, Marks 02

OR

Give the definition of the terms 'frequent itemset', 'support' and 'confidence'.

AKTU 2016-17, Marks 02

Ans. Frequent item set : Refer Q. 4.3, Page SQ–13D, Unit-4, Two Marks Question.

Support and confidence :

Support (s) :

1. Support is an indication of how frequently the item set appears in the dataset.
2. Fraction of transactions that contain both X and Y.

$$s = \frac{\sigma(\{X, Y\})}{\text{\#of transactions}} = 0.38$$

Confidence (c) :

1. Confidence measures how often items in Y appear in transactions that contain X.

$$c = \frac{\sigma(\{X, Y\})}{\sigma(\{X\})} = 0.75$$

4.8. What are hierarchical methods for clustering ?

AKTU 2017-18, Marks 02

Ans. Hierarchical methods for clustering are :

1. Agglomerative method
2. Divisive method

4.9. Write the statement for Apriori algorithm.

AKTU 2017-18, Marks 02

Ans. The Apriori algorithm is used in association rule mining which uses the property of large item set "Any subset of large item set must be large".

4.10. What are the drawbacks of k-mean algorithm ?

AKTU 2017-18, Marks 02

Ans. Drawbacks of k-mean algorithm :

predict k-value.

2. With global cluster, it did not work well.
3. Different initial partitions can result in different final clusters.

4.11. Discuss data discrimination.

Ans: Data discrimination is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes. It can be implemented using the attribute-oriented induction or data cube approaches.

4.12. Define ID3 decision tree algorithm.

Ans: ID3 is a decision tree algorithm in which attributes are chosen in any order for the CLS algorithm. This can result in large decision trees if the ordering is not optimal.

ID3 = CLS + efficient ordering of attributes.

4.13. Write a short note on tree pruning.

Ans: The decision tree can contain many branches which will reflect anomalies in the training data due to noise or outliers. The method addresses this problem of overfitting the data. Such methods typically use statistical measure to remove the least reliable branches resulting in faster classification.

4.14. Name the two approaches of tree pruning.

Ans: **Two approaches of tree pruning are :**

i. Pre-pruning approach
ii. Post-pruning approach

4.15. Differentiate between classification and clustering.

AKTU 2017-18, Marks 02

Ans:

4.16. What are clustering requirements ?

Ans: **Requirements of clustering are :**

1. Scalability
2. Ability to deal with different kinds of attributes
3. Discovery of clusters with attribute shape
4. High dimensionality
5. Ability to deal with noisy data
6. Interpretability

5 UNIT

Data Visualization and Overall Perspective (2 Marks Questions)

5.1. Compare roll-up, drill-down, slice and dice operations.

AKTU 2015-16, Marks 02

AKTU 2017-18, Marks 02

Ans.

1. **Roll-up :** Roll-up is also known as "consolidation" or "aggregation". In the roll-up process at least one or more dimensions need to be removed. The Roll-up operation can be performed in two ways :
   i. Reducing dimensions
   ii. Climbing up concept hierarchy.
2. **Drill-down :** In drill-down data is fragmented into smaller parts. It is the opposite of the roll-up process. It can be done via moving down the concept hierarchy and increasing a dimension.
3. **Slice :** Here, one dimension is selected, and a new sub-cube is created.
4. **Dice :** This operation is similar to a slice. The difference in dice is we select two or more dimensions that result in the creation of a sub-cube.

5.2. Classify OLAP tools. **AKTU 2016-17, Marks 02**

Ans. The OLAP tools enable a user to easily and selectively extract and view data from different points of view. There are two types of OLAP tools :

i. MOLAP (Multidimensional OLAP)
ii. ROLAP (Relational OLAP)

5.3. Bring out any two points with respect to spatial mining.

AKTU 2016-17, Marks 02

Ans. Two points with respect to spatial mining are :

1. Spatial mining is the application of data mining to spatial models.
2. It plays an important role in the army's strategic, tactical and operational planning.

5.4. Give the characteristics of OLAP.

Ans. Characteristics of OLAP are :

i. Calculations and modeling applied across dimensions, through hierarchies or across members.
ii. Trend analysis over sequential time periods.
iii. Slicing subsets for on-screen viewing.
iv. Drill-down to deeper levels of consolidation.
v. Reach through to underlying detail data.

5.5. Write a short note on MQE.

Ans. MQE (Managed Query Environment) has been adopted by the industry to describe a query and reporting package that allows IT control over user access to data and application facilities in accordance with each user's level of expertise and business needs.

5.6. Give the functional components of data mining interfacing.

Ans. Functional components of data mining interface are :

i. Data collection and data mining query composition.
ii. Presentation of discovered patterns.
iii. Hierarchy specification and manipulation.
iv. Manipulation of data mining primitives.
v. Interactive multilevel mining.

5.7. What are the types of security in data warehouse at different levels ?

Ans. Following are the types of security at different levels :

i. Application development
ii. Load manager
iii. Warehouse manager
iv. Query manager

5.8. Define recovery. What are the types of recovery model ?

Ans. Recovery is the process of rebuilding a database after some part of a database has been lost. The recovery model of a new database is inherited from the model database, when new database is created

Types of recovery model are :

i. Full recovery model
ii. Bulk-logged recovery model
iii. Simple recovery model

5.9. What are the steps required for tuning a data warehouse ?

Ans. Steps required for tuning a data warehouse are :

i. Tune the business rules.
ii. Tune the data design.
iii. Tune the application design.
iv. Tune the logical structure of DB.
v. Tune DB operation.
vi. Tune the access paths.
vii. Tune memory allocation.

5.10. What are the types of warehousing applications ?

Ans. Types of warehousing application are :

i. Financial services
ii. Banking services
iii. Consumer goods
iv. Retail sectors
v. Controlled manufacturing

5.11. Give activities and issues of web usage mining.

Ans. Activities of web usage mining :

i. Preprocessing activities center around reformatting the web log data before processing.
ii. Pattern discovery activities form major portion of mining activities.
iii. Pattern analysis is the process of looking at and interpreting the results of activities that are discovered.

Issues in web usage mining :

i. Identification of exact user is not possible.
ii. It is difficult to uncover the sequence of pages that a user actually visits.
iii. There are many privacy, security and legal issues yet to be solve.

5.12. Give key features of OLAP.

Ans. Key features of OLAP are :

1. Multi-dimensional views of data
2. Support for complex calculations
3. Time intelligence

5.13. Write down the benefits of OLAP.

Ans. Benefits of OLAP are :

1. OLAP has consistency of information and calculations.
2. It allows a manager to pull down data from an OLAP database in broad or specific terms.
3. OLAP creates a single platform for all the information and business needs, planning, budgeting, forecasting, reporting and analysis.