Paper solution (2024)

404. Data mining & Data warehousing

Page No. 7
Date: / /

[70 marks]

1 [A] Attempt the followings [06 marks]

1] What is data warehouse? List out its characteristics.

⇒ A data warehouse is a subject-oriented integrated time-variant and non-volatile collection of data in support of management decision making process.

Four type's of characteristics.

1] subject-oriented
2] integrated
3] time-variant
4] non-volatile.

Define: Data Mart.

A data mart is a data storage system that contains information specific to an organization's business unit it contains a small and selected part of the data that the company stores in a larger storage system.

List out three possible approach of Data Warehouse.

1] Single tier architecture: A single-layer structure aimed at keeping data space minimal. This structure is rarely used in real life.

2] Two-tier :- If you want to store large amounts of data in a small amount of space then you should consider using a data warehouse.

3] three-tier :- The top, middle and Bottom tiers of this Architecture of Data warehouse are collectively referred to as the top tiers.

Q-1[b] Attempt Any two: [12 mark]

1] List and Explain all the characteristics of Data Warehouse.

=> there are four type of characteristics of Data Warehouse.

1] Subject oriented
2] Integrated
3] Nonvolatile
4] Time-variant

1] Subject oriented :- Decision makers of a business can analyze data easily by constricting to a particular subject area of the data warehouse. This makes understanding and analysis of the data concise and straight forward by excluding unwanted information on some subject that is not needed for decision-making. This means that the ongoing operations of an organization are not taken into consideration.

2] Integrated

Data warehouse consist of data from different variable sources integrated under one platform. this Data obtained is extracted and transformed maintaining uniformity without depending on the source it was obtained from. This is the integrated feature standards

are established which are universally acceptable for the data present in the warehouse.

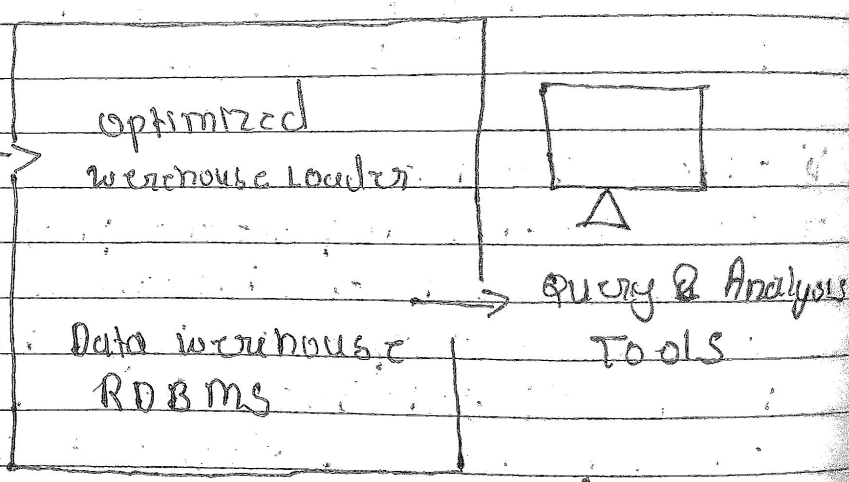
3) Non-volatile
=> Data is updated by uploading data in the data warehouse to protect data from momentary changes. This means that once data is fed there can be no alterations or changes. The inability to be executed is the non-volatile character of the data warehouse environment. Data is read-only and allows only two functions to be performed Access and Loding.

4) Time-variant
=> One of the important properties of the data warehouse is the historical perspective it holds. It keeps the huge volume of data from all database stored in accordance of a temporal element and an extensive time horizon. The inability to change the element of time is an essential aspect of time variance. The record key is used to display time variance.

2] Explain Basic principal of Data warehouse modelling.

⇒ Data warehouse modeling is the process of designing the schemas of the detail and summarized information of the data warehouse the goal of data warehouse modeling is to develop a schema describing the reality or at leat a part of the fact which the data warehouse is needed to support.

diagram



Data warehouse model

1] Dimensional modeling
2] Fact tables
3] Dimension tables
4] Granularity
5] Data integration
6] Historical Data.

1] Dimension Modeling: organizing data into dimensions Dimensions represent the perspective by which data is analyzed while facts are the numerical data being analyzed.

2] Fact tables: Fact tables contain business facts or measures while dimension tables contain descriptive attributes related to the facts Fact table are typically large and contain numerical data, while dimension tables are smaller and contain textual or categorical data.

3] Granularity:- Determining the level of detail at which data is stored in the warehouse Granularity affects the level of analysis and performance of queries.

4] Data integration:- a method of processing data from multiple heterogeneous sources of data and combining them coherently to create a unified view of the information.

5] Historical Data:- Historical data is not just a record of the past, it is a powerful tool for shaping the future of SMBs.

OR

3) Write a short note on temporal modelling.

=> Temporal modeling in data mining and data warehousing involves incorporating time related aspects into the data analysis process.

=> Temporal modeling in data mining and data warehousing is essential.

=> Temporal data mining can be defined as "process of knowledge discovery in temporal databases that enumerates structures (temporal patterns or models) over the temporal data and any algorithm that enumerates temporal patterns from or fits models to temporal data is a temporal data mining algorithm" (Lin etal 2002). The aim of temporal data mining is to discover temporal patterns, unexpected trends, or other hidden relations in the larger sequential data. which is composed of a sequence of nominal symbols from the alphabet known as a temporal sequence and a sequence of continous real-valued elements known as a time series by using a combination of techniques from machine learning.

=> Temporal modeling is a technique used to analyze data that varies over time.

=> It involves capturing patterns, trends, and dependencies in time.

=> It involves capturing patterns, trends
=> methods range from simple moving averages to complex algorithms like recurrent neural network (RNNs) or Hidden Markov Models (HMMs).

=> Effective temporal modeling considers seasonality, trend, cyclic patterns, and irregularity in the data.

=> Its applications span various fields including statistics, machine learning, and computer science.

=> Temporal models enable accurate forecasting and decision-making by understanding underlying temporal dynamics.

=> This approach is essential for extracting meaningful insights, making predictions, and understanding underlying processes in dynamic systems.

Q-2 [A] Attempt the following [05 mark]

1] Define: Data cube

⇒ In computer programming contcat a data cube is a multi-dimensional cerray of values. Typically the term data cube is applied in contcats where these cerrays are massively lerger than the hosting computer's main memory examples include multi-terabyte.

2] Full form of ROLAP

⇒ Relational online Analytical processing
(OLAP = online Analytical processing)
(MOLAP = Multidimensional online Analytical processing)
(HOLAP = Hybrid onlie Analytical processing).
(OLTP = Online Transaction processing.)

3] What is the use of OLAP?

⇒ online analytical processing (OLAP) is a database analysis technology that involves querying, extracting and studying summarized data.

4] What is metadata?

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

5] What is Star model?

⇒ Ans ⇒

7. A star schema is a multi-dimensional data model used to organize data in a database so that it is easy to understand and analyze.

Q-2 [b] Attempt any two: [12 marks]

i) Give the difference between OLAP vs OLTP.

	OLAP	OLTP
1	OLAP help you analyze large volumes of data to support decision-making	OLTP helps you manage and process real time transactions.
2	OLAP uses historical Data source aggregated data from multipele sources	OLTP uses real-time transactional data from a single source.
3	OLAP uses multidimensional (cubes) or relational database	OLTP uses relational database
4	OLAP uses star schema snowflake schema or other analytical models.	OLTP uses normalized or denormalized models.

5	OLAP has large storage requirements think terabytes (TB) and petabytes (PB)	OLTP has comparatively smaller storage requirements think gigabytes (GB)
6	OLAP has longer response times typically in seconds or minutes.	OLTP has shorter response times typically in milliseconds
7	OLAP is good for analyzing trends predicting customer behavior and identifying profitability.	OLTP is good for processing payments customer data management and order processing

online analytical processing (OLAP) and online transaction processing (OLTP) are data processing systems that help you store and analyze business data.

2] Give the difference between ROLAP v.s MOLAP.

⇒ ROLAP is Relational online analytical processing where data are stored in the form of tables, columns and rows MOLAP is a multidimensional online analytical processing where data is stored in multidimensional formatted database called Data cubes.

ROLAP	MOLAP
1 ROLAP Relational online Analytical processing	MOLAP is a Multidimensional online Analytical processing.
2 ROLAP stores a large amount of data	It stores limited data and keeps summarized data in multidimensional database.
3] ROLAP stores and retrive data from the main data warehouse	MOLAP stores and retrives data from multiple dimensional database
4] It stores data in the form of a relational database	It stores data in the form of array-based multidimensional data cubes.

5	the speed of response is slow	The speed of response is fast
6	It uses complex SQL queries to retrieve the data from the data warehouse.	It creates data cubes to retrieve the data from multiple dimensions To handle related scattered matrix technology is used.
7	It dynamically creates a multidimensional view of data.	It already stores data in the form of multidimensional array in multidemensional databases.
8	Latency Low.	Latency High.
9	OBMS facility strong	OBMS facility week

OR

3] What is metadata? Explain source of metadata.

⇒ Metadata is defined as the information that describes and explains data. It provides context with details such as the source type owner, and relationship to other data sets. So it can help you understand the relevance of a particular data set and guide you on how to use it.

Source of Metadata

1] Database Schema:

In data warehousing, the database schema defines the structure of the data warehouse including tables, columns, keys, constraints and relationships.

- Metadata extracted from the database schema provides information about the organization and semantics of the data facilitating data exploration and analysis.

2] Data Dictionary:

⇒ A data dictionary contains detailed descriptions of the data elements or attributes used in the database or data warehouse.

⇒ It includes metadata such as data type length format, constraints and definitions enabling users to understand the meaning and

characteristics of each data element.

4) ETL (Extract, transform, load)

⇒ Metadata generated associated with data mining models includes information about model parameters, algorithms, performance metrics and evaluation criteria.

⇒ Understanding the metadata of data mining models helps interpret the result, assess model quality and guide decision-making based on mining.

5) Data profiling and Analysis

⇒ Data profiling tools generate metadata about the statistical properties, distributions, patterns and anomalies present in the data.

6) User Documentation and Annotations

⇒ User documentation, annotations and annotations provide additional context and insights into the data, such as business rules, semantics and user-defined metadata tags.

⇒ These annotations enrich the metadata repository making it more comprehensive and user-friendly for analysts, data scientists and other stakeholders.

Q3 [A] Attempt the following [06 marks]

1] Full form of KDD
⇒ Knowledge discovery in Database (KDD)

2] List out Application of Data mining.
⇒ 1] Finacial Analysis
2] Telecommunication industry
3] Intrusion Detection
4] Retail Industry
5] Higher education
6] Energy Industry
7] Spatial data mining
8] Biological Data Analysis

3] Define the term classification.
⇒ Classification is a data mining function that assigns item in a collection to target categories or classes.

4] DWM :- Division of waste management

5 MDDM :- Multi-dimensional Data Model.

Q-3 [b] Attempt any two [12 marks]

1] Explain KDD vs Data mining.

=> KDD refers to a process of identifying valid, novel, potentially useful and ultimately understandable patterns and relationship in data.

=> Data mining refers to process of extracting useful and valuable information or patterns from large data sets.

=> It transforms task relevant data into patterns and decides purpose of model using classification or characterization.

KDD	Data Mining
1] KDD stand for knowlyge discovery In Database KDD refers to a process of identifying valid noval potentially usefull and ultimately understandable patterns and relationship in Data.	Data mining refers to a process of extracting useful and valuable information or patterns from large data sets.

	KDD	Data mining
2]	To find useful knowledge from data	To extract useful information from data.
3]	Data cleaning data integration data selection data transformation data mining pattern evaluation and knowledge representation and visualization	Association rules classification clustering regression decision trees neural networks and dimensionality reduction
4]	Structured information such as rules and models that can be used to make decision or predictions	patterns associations or insights that can be used to improve decision making or understanding.
5]	Focus is on the discovery of useful knowledge rather than simply finding patterns in data.	Data mining focus is on the discovery of patterns or relationships in data.

2) Write a short note on Automatic cluster Detection.

=> cluster analysis also known as clustering is a method of data mining that groups similar data points together. The goal of cluster analysis is to divide a dataset into groups such that the data points within each group are more similar to each other than to data points in other groups. this process is often used for exploratory data analysis and can help identify patterns or relationship within the data and it may not be immediately obvious.

=> cluster Analysis is the process to find similar groups of object in order to form cluster. It is an unsupervised a group or data point would comprise together form a cluster in which all the objects would belong to the same group.

=> the given data is divide into different group by combining similar objects into a group this group is nothing but a cluster a cluster is nothing but a collection of similar data which is grouped together.

⇒ the main idea of cluster analysis is that it would arrange all data points by forming clusters like cars cluster which contains all the cars, bikes clusters which contains all the bikes etc.

⇒ simply it is the partitioning of similar objects which are applied to unlabeled data.

OR

31 Give the difference between Discovery VS Verification Mode.

☆ Discovery Mode:

- Objective: Discovery mode in data mining and data warehousing focuses on exploring and uncovering new patterns trends and insights within the data.
- Exploratory Analysis
In discovery mode analysts and data scientists conduct exploratory analysis to identify hidden patterns, associations correlations and anomalies in the data.
- Hypothesis Generation
Analysts formulate hypotheses and theories based on observed pattern and relationship which are then tested and validated in subsequent steges.
- Iterative process
- Discovery mode is an iterative process where analysts iteratively explore, analyze and refine their understanding of the data often leading to new insights and discoveries.

☆ Varification Mode.

- objective :- Varification mode is a data mining and data warehousing aims to validate and confirm previously discovered patterns or models

- Confirmation of Findings
In verification mode analysts focus on verifying the accuracy, reliability and generalization of discovered patterns and insights using independent datasets or validation techniques.
- Model Evaluation
Analysts evaluate the performance of data mining models predictive algorithms or analytical hypotheses using holdout samples cross validation or other validation techniques.
- ~~Bias~~ Reduction validation
verification mode often employs various validation metrics such as accuracy precision recall, F1-score and ROC curves to assess the performance of predictive models and analytical
- Bias Reduction
verification mode helps mitigate the risk of overfitting and bias by independently validating the predictive power and generalization capability of data mining models.

Key Differences:
1) purpose
2) Analytical Approach
3) Tools and Techniques
4) Iteration

Q-4 [A] Attempt the following [05 marks]

1] List out trends in mining.

1 ⇒ Association rule learning
2 clustering
3 multimedia data mining
4 ubiquitous data mining
5 Automated data mining
6 classification
7 Anomaly detection
8 Data generalization
9 Data structuring
10 Text mining

2] What is web structure mining.

⇒ Web structure mining is a tool that can recognize the relationship between web pages linked by data or direct link connection.

3] Full form of MOLAP

⇒ Multidimensional online Analytical processing

4] Define web content mining.

⇒ Web content mining is the mining scanning and extraction of text videos, graphs and pictures from web documents.

5] What is spatial data mining?

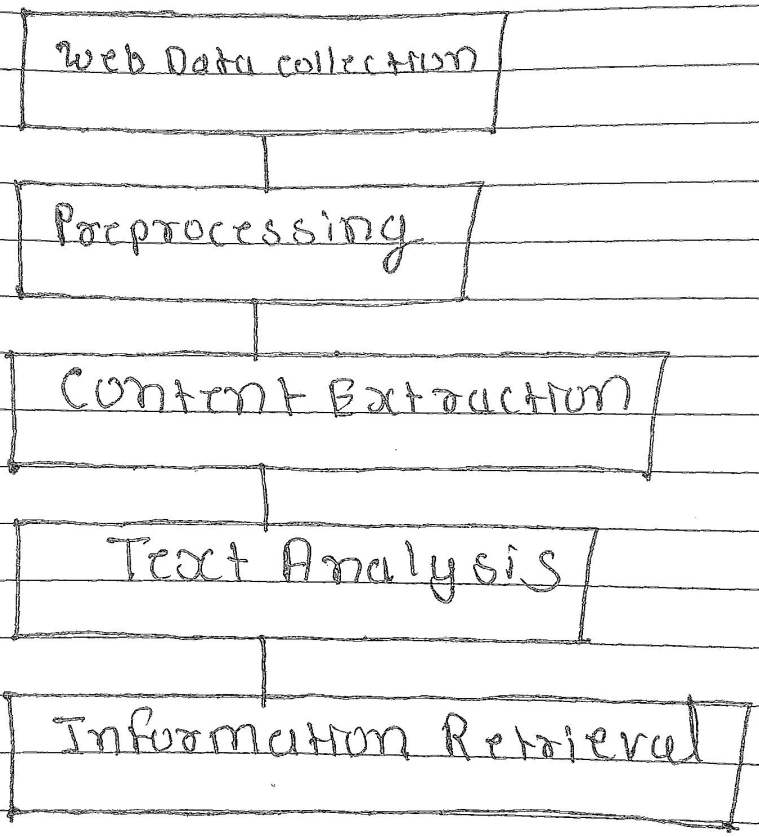
⇒ Spatial Data mining is the process of discovering interesting and useful patterns spatial relationships, which were otherwise stored in spatial database.

Q-4 [b] Attempt any two: [12 marks]

1] Write a short note on web content mining.

=>

Diagram



1] Data collection :- Web content mining involves collecting data from various sources on the internet including web pages, documents blogs social media platforms and online forums. This data may be structured semi-structured, or unstructured.

2. preprocessing :- Before analysis web data undergoes preprocessing to clean and transform it into a suitable format for mining this includes tasks as HTML parsing text extraction noise removal and data normalization.

3 Content Extraction :- content extraction identify and extract relevant information from web documents. This way involve text extraction, entity recognition, sentiment analysis and categorization of web content based on predefined categories or topics.

4 Information Retrieval :

⇒ Information retrieval methods are employed to retrieve specific information or documents from large collections of web data. Techniques such as keyword-based search, indexing, ranking algorithms and relevance feedback help users find relevant information.

5 Text Mining

Text mining techniques are applied to analyze the textual content of web documents. This include tasks such as text summarization, document analysis and named entity recognition to extract valuable insights from text data.

6] Link Analysis

⇒ Link analysis method examine the hyperlink structure of the web to uncover patterns relationships and connectivity between web pages. This includes tasks such as web page ranking link prediction community detection and identifying authoritative sources based on link analysis algorithms.

7] challenges

- Data volume and diversity: Handling large volumes of heterogeneous web data.
- Data Quality: Dealing with noise redundancy and inconsistencies in web content.
- privacy and ethical concerns: Ensuring the ethical use of web data and protecting user privacy.
- Scalability: Scaling mining algorithms to handle the dynamic and ever-growing nature of the web.

2] Explain Data Generalization and Summarization in detail.

Data Generalization

Data generalization involves transforming detailed data into a more abstract form by replacing specific values with more general ones. The goal is to reduce the level of detail while preserving the essential characteristics of the data. This process is particularly useful for protecting sensitive information, reducing storage requirements, and simplifying data analysis.

- Aggregation: Aggregating data involves combining multiple individual data points into summary values.
- Categorization: categorizing data involves grouping similar data items into predefined categories or ranges. For instance grouping ages into age ranges (e.g. 20-30, 31-40 etc.).
- Attribute Hierarchy: Creating hierarchies for attributes allow for the organization of data at different level of granularity. For example a time hierarchy could include years, months, weeks, and days.

★ Data Summarization

Data summarization involves condensing large volumes of data into more manageable and understandable representations while maintaining the essence of the information. Summarization techniques help in identifying significant patterns trends and outliers within the data.

- Statistical summaries :- Calculating summary statistics such as mean, median, mode, standard deviation and percentiles provides insights into the central tendency variability and distribution of the data.
- Sampling : sampling ~~is not~~ involves selecting a representative subset of data from the entire data set for analysis. Sampling techniques such as random sampling stratified sampling and cluster sampling help in reducing the computational burden while preserving the characteristics of the original data.
- Data Cube Aggregation
- Clustering

OR

3 Explain web structure and usage mining in detail.

=> Web structure

Web structure mining focuses on analyzing the structure of the world wide web, particularly the link structure between web pages. It involves extracting valuable information from the link topology of the web including hyperlinks web page relationship and connectivity patterns.

- Link Analysis :- Link analysis techniques examine the hyperlink structure of the web to uncover patterns, relationship and connectivity between web pages. This includes method such as pagerank, HITS and web link Analysis.

- Graph Theory :- Graph theory concepts are applied to model the web a graph where nodes represent web pages and edges represent hyperlinks between pages. Graph algorithms such as shortest path algorithm, clustering algorithms and community detection algorithms are used to analyze web structure.

Application

1) search Engine optimization
2) web crawling and indexing

3| Web Navigation and Recommendation System

Usage Mining

- Web usage mining involves analyzing user interaction data captured from web servers such as web server logs clickstream data and user session. The goal is to extract useful knowledge about behaviour preferences and navigation patterns on the web.

Techniques

- Sessionization: sessionization techniques group user interaction into session based on time intervals of user activities. this helps in analyzing user behaviour within individual sessions and understanding navigation patterns.

- Pattern Discovery: pattern discovery techniques identify frequent sequences paths, or patterns of user interaction on the web. this includes methods such as sequence mining, association rule mining and clustering of user sessions.

- Predictive Modeling :- predictive modeling techniques use historical user interaction data to predict future user behavior, such as click through rates, conversion rates and user preferences.

Application
1) personalization and Recommendation
2) website optimization
3) E-commerce and marketing