Question Bank

Year: 2016- 2017 **Semester:** First

Subject Dept: CS **Subject Name:** Data Mining

……

Q1) What is data warehouse?

ANS.

"A data warehouse is a **subject-oriented**, **integrated**, **time-variant**, and **nonvolatile** collection of data in support of management's decision-making process." by William H. Inmon

Q2) What is the benefits of data warehouse?

1) It is a direct reflection of the business rules of the enterprise.
2) It is the collection point for strategic information.
3) It is the historical store of strategic information.
4) It is the source of information later delivered to data marts.
5) It is the source of stable data regardless of how the business processes may change.

Q3) What is the difference between OLTP and OLAP?

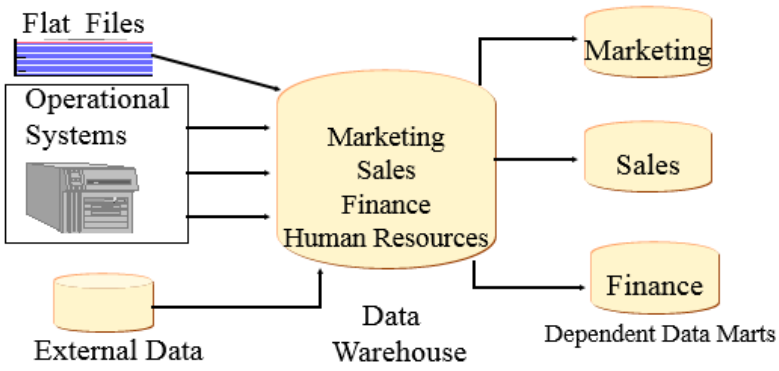
	OLTP	OLAP
users	Writer, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
access	read/write index/hash on primary key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB

Q4) Briefly state different between data ware house & data mart?

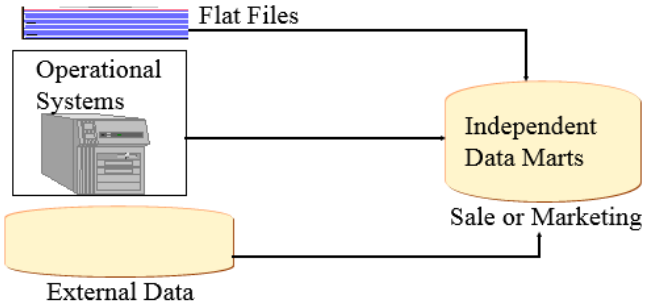
Property	Data Warehouse	Data Mart
Scope	Enterprise	Department
Subject	Multiple	Single-subject
Data Source	Many	Few
Size(typical)	100 GB to>n TB	<100 GB
Implementation time	Months to years	Months

Q5) What are the difference among dependent data mart, independent data mart and hybrid data mart?

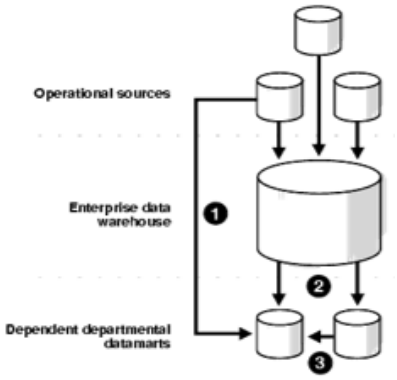
A **dependent data mart** is created with the use of a central data warehouse. This gives you the usual advantages of centralization. Figure below illustrates a dependent data mart.



An **independent data mart** is created without the use of a central data warehouse. This could be desirable for smaller groups within an organization. Figure below illustrates an independent data mart.



Hybrid data marts can draw data from operational systems or data warehouses. This could be useful for many situations, especially when you need a new group or product is added to the organization. This figure illustrates a hybrid data mart.



Q6) List some properties of data Marts.

1) A data mart is a smaller, more focused data warehouse. It reflects the business rules of a specific business unit.
2) The data mart does not need to cleanse its data because that was done when it went into the warehouse.
3) It is a set of tables for direct access by users.
4) It typically is not a source for traditional analysis.

Q7) List the Reasons for Creating a Data Mart

1) To give users more flexible access to the data they need to analyse most often.
2) To provide data in a form that matches the collective view of a group of users
3) To improve end-user response time.
4) Building a data mart is simpler compared with establishing a corporate data warehouse.
5) The cost of implementing data marts is far less than that required to establish a data warehouse.

Q8) What are the characteristics of data warehouse?

"A data warehouse is a **subject-oriented**, **integrated**, **time-variant**, and **nonvolatile** collection of data in support of management's decision-making process."

Data Warehouse—Subject-Oriented

- Organized around major subjects, such as customer, product, sales.
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

Data Warehouse—Integrated

1. **Constructed by integrating multiple, heterogeneous data sources**
   - relational databases, flat files, on-line transaction records
2. **Data cleaning and data integration techniques are applied.**
   - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
     - E.g., Hotel price: currency, tax, breakfast covered, etc.
   - When data is moved to the warehouse, it is converted.

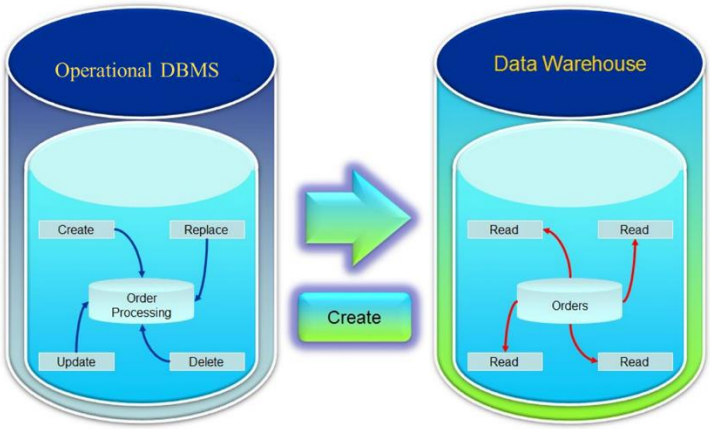
Data Warehouse—Time Variant

- **The time horizon for the data warehouse is significantly longer than that of operational systems.**
  - Operational database: current value data.
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- **Every key structure in the data warehouse**
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain "time element".

Data Warehouse—Non-Volatile

- A physically separate store of data transformed from the operational environment.
- Operational update of data does not occur in the data warehouse environment.
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only two operations in data accessing:
    - *initial loading of data* and *access of data*.

Q9) Differentiate between Data Warehouse versus Operational DBMS



Q10) multiple choice questions

1. Operational database is

A. A measure of the desired maximal complexity of data mining algorithms

B. A database containing volatile data used for the daily operation of an organization

C. Relational database management system

D. None of these

:

:

Q11) Give some alternative terms for data mining.

1. Knowledge mining
2. Knowledge extraction
3. Data/pattern analysis.
4. Data Archaeology
5. Information gathering,
6. Business intelligence

Q12) What is KDD.

KDD-Knowledge Discovery in Databases.

Q13) What are the steps involved in KDD process.

1. Data cleaning
2. Data Mining
3. Pattern Evaluation
4. Knowledge Presentation
5. Data Integration
6. Data Selection
7. Data Transformation

Q14) What are the benefits of data mining?

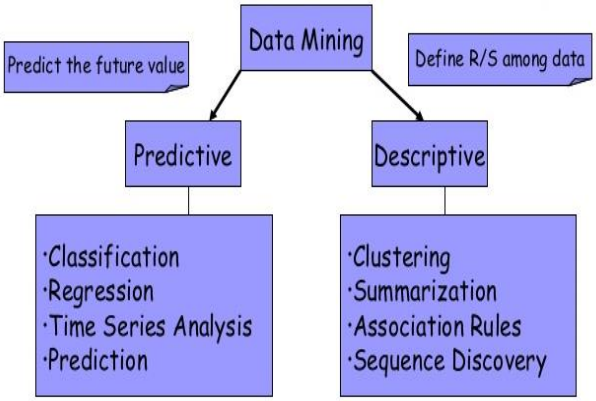
1) Data mining finds valuable information hidden in large volumes of data.
2) Data mining is the analysis of data and the use of software techniques for finding patterns in sets of data.
3) The computer is responsible for finding the patterns by identifying the underlying rules and features in the data.
4) The information hidden no-one has noticed them before.

Q15) Define prediction and description models

The goals of prediction and description are achieved by using the following primary data mining tasks:

1) Classification
2) Pattern Regression
3) Time serious analysis
4) Prediction
5) Clustering
6) Association rules

7) Summarization
8) Sequence discovery



Q16) Describe challenges to Separate Data Warehouse regarding performance and functions issues.

1. High performance for both systems
2. Different functions and different data

High performance for both systems

1) DBMS— tuned for OLTP: access methods, indexing, concurrency control, recovery
2) Warehouse—tuned for OLAP: complex OLAP queries, multidimensional view, and consolidation.

Different functions and different data:

1. **missing data**: Decision support requires historical data which operational DBs do not typically maintain
2. **data consolidation**: DS requires consolidation (aggregation, summarization) of data from heterogeneous sources
3. **data quality**: different sources typically use inconsistent data representations, codes and formats which have to be suitable.

Q17). Why Do We Need Data Warehouses?

1. **Unification of information resources**. Improved query performance "Separate research and decision support functions from the operational systems.
2. **The data stored in the warehouse is uploaded from the operational systems**. The data may pass through an operational data store for additional operations before it is used in the DW for reporting.

Q18). How is a Data Warehouse different from an Operational DBMS.

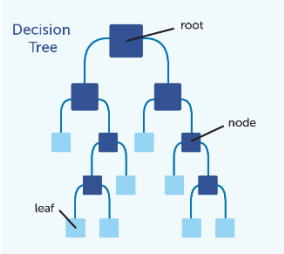
- **OLTP (on-line transaction processing)**
  1. Major task of traditional relational DBMS
  2. Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.
- **OLAP (on-line analytical processing)**
  1. Major task of data warehouse system
  2. Data analysis and decision making

Q19). Define an **operational system**

An **operational system** is a term used in data warehousing to refer to a **system** that is used to process the day-to-day transactions of an organization. These **systems** are designed in a manner that processing of day-to-day transactions is performed efficiently and the integrity of the transactional data is preserved.

Q20) What is Decision tree?

A decision tree is a flow chart like tree structures, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions. The top most in a tree is the root node.



Q21) What is meant by Pattern?

Pattern represents the knowledge.

Q22) What are the different Applications of Data Mining

1. E-commerce
2. Marketing and retail
3. Finance
4. Telecoms
5. Drug design
6. Process control
7. Space and earth sensing
8. Bioinformatics

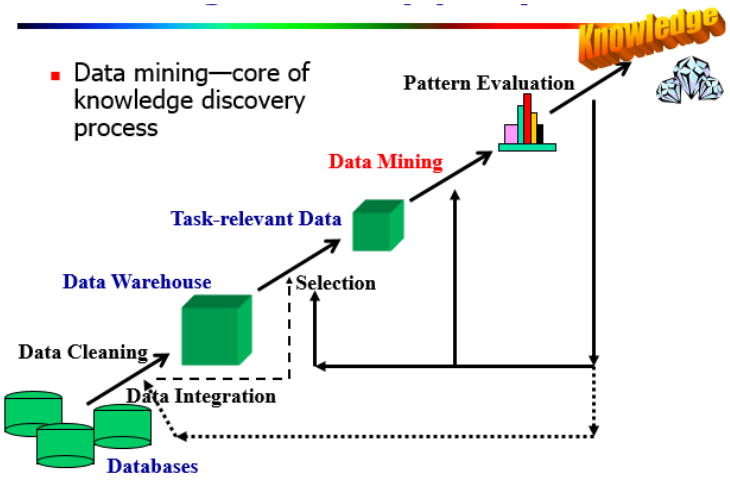
Q23) Explain the following Major Tasks in Data Processing

Data cleaning, Data integration and Data transformation.

1. Data cleaning
Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
2. Data integration
Integration of multiple databases, data cubes, or files
3. Data transformation

Normalization and aggregation

Q24) Explain the knowledge discovery phases.



Q25) Define the following terms:

1. Data cleaning

Data cleaning means removing the inconsistent data or noise and collecting necessary information

2. **Knowledge representation**

Knowledge representation techniques are used to present the mined knowledge to the user.

3. **Visualization**

Use of computer graphics to create visual images which aid in the understanding of complex, often massive representations of data.

4. **Visual Data Mining**

Visual Data Mining presents the data in some visual form, allowing users to mine and gain insight into the data, draw conclusions and directly interact with the data.

5. **Systems and Models**

System is a collection of interrelated objects and Model is a description of a system. Models are abstract, and conceptually simple.

6. **A decision tree**

It is a flow-chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions.

Decision tree is a predictive model. Each branch of the tree is a classification question and leaves of the tree are partition of the dataset with their classification.

7. **A data mart**

Is quite common, you may want to customize your warehouse's architecture for different groups within your organization. You can do this by adding data marts, which are systems designed for a particular line of business an example of data marts like: purchasing, sales, Inventories, Finance, and Human Resources.

Q26). What is Descriptive and predictive data mining?

- **Prediction** involves using some variables or fields in the database to predict unknown or future values of other variables of interest **Example :-**

1. Credit card fraud
2. Breast cancer early warning
3. Terrorist act

- **Description:** focuses on finding human-interpretable patterns describing the data **Example:-**

1. Segmenting marking area
2. Profiling student performances
3. Profiling GooglePlay/ AppleApps customer

Q27) **Write the preprocessing steps that may be applied to the data for prediction.**

a. Data Cleaning
b. Relevance Analysis
c. Data Transformation

Q28) **Why do you need data warehouse life cycle process? and what are the steps in the life cycle approach?**

Data warehouse life cycle approach is essential because it ensures that the project pieces are brought together in the right order and at the right time.

- Project Planning
- Business Requirements definition
- Data track: Dimensional modeling, Physical Design, Data Staging Design & Development
- Application track: End user Application Specification, End user Application Development
- Deployment
- Maintenance & Growth
- Project Management

Q29) **Explain OLAP and OLTP?**

OLTP (on-line transaction processing)

- Major task of traditional relational DBMS
- Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.

OLAP (on-line analytical processing)

- Major task of data warehouse system
- Data analysis and decision making

Q30) **Write the main steps of Data cleaning tasks**

1) Fill in missing values
2) Identify outliers and smooth out noisy data
3) Correct inconsistent data
4) Resolve redundancy caused by data integration

Q31) **Why Do We Need Data Warehouses?**

1. **Unification of information resources**. Improved query performance "Separate research and decision support functions from the operational systems.
2. **The data stored in the warehouse is uploaded from the operational systems**. The data may pass through an operational data store for additional operations before it is used in the DW for reporting.

Q32) **Name some of the data mining applications?**

1. E-commerce
2. Marketing and retail
3. Finance
4. Telecoms
5. Drug design
6. Process control
7. Space and earth sensing
8. Bioinformatics

Q33) **Why is data quality so important in a data warehouse environment?**

Data quality is important in a data warehouse environment to facilitate decision-making. In order to support decision-making, the stored data should provide information from a historical perspective and in a summarized manner.

Q34) **How can data visualization help in decision-making?**

Data visualization helps the analyst gain intuition about the data being observed. Visualization applications frequently assist the analyst in selecting display formats, viewer perspective and data representation schemas that faster deep intuitive understanding thus facilitating decision-making.

Q35) **Explain the different types of data repositories on which mining can be performed?**

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

- Relational Databases
- Data Warehouses
- Transactional Databases
- Advanced Databases
- Flat files
- World Wide Web

Q36) Explain the concept of Data Mining Classification Schemes

Different views lead to different classifications:

1. Data view: Kinds of data to be mined
2. Knowledge view: Kinds of knowledge to be discovered
3. Method view: Kinds of techniques utilized
4. Application view: Kinds of applications adapted

Q37) Define data warehouse. Draw the architecture of data warehouse and explain the main architecture.

"A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process." by William H. Inmon