# Data Warehousing and Mining – May 2005

**Time : 3 Hrs.]** **[Marks : 100**

**N.B. :** (1) Question No. **1** is **compulsory**.
(2) Attempt any **four** questions out of the remaining **six** questions.

**1.** Consider a university database that keeps track of student and their majors, transcripts and registration and the university's courses. Several sections of each course are offered and each section is related to the instructor who is teaching. It also keeps track of the sponsored research projects of faculty and graduate students of the academic departments of the particular college. The database also keeps track of research grants and contracts awarded to the university. A grant is related to one principle investigator and to all researchers it supports.

**For this problem answer the following :**

(a) Implement the data warehouse architecture. For this clearly state and implement the operational system, data warehouse, materialized view and ETL process. **[15]**

(b) Answer the following queries : **[5]**

(i) Print the name of the faculty member who is teaching any section of a course that is offered by the computer department, provided that the section is taken by at least one graduate student who is an research associate.

(ii) Define the user's view such that a course instance must have been taken by the student instance who may or may not be a graduate student provided that the course instance belongs to the computer department.

Assume the required data if any and state clearly. You can use any database of your choice for this implementation.

**2.** (a) Explain a three tier data warehousing architecture with suitable block diagram. **[5]**

(b) The Mumbai University wants you to help to design a star schema to record grades for course completed by students. There are four dimensional tables namely, **Course_section, Professor, Student**, Period with attributes as follows :

- **Course_section.** Attributes : Course_ID, Section_Number, Course_Name, Units, Room_id, Room_Capacity. During a given semester the college offers an average of 500 course sections.
- **Professor.** Attributes : Prof_ID, Prof_Name, Title, Department_ID, Department_Name.
- **Student.** Attributes : Student_ID, Student_Name, Major. Each course section has an average of 60 students.
- **Period.** Attributes : Semester_ID, Year. The database will contain data for 30 months periods. The only fact that is to be recorded in the fact table is Course_Grade.

**Answer the following questions :** **[15]**

(a) Design the star schema for this problem.

(b) Estimate the number of rows in the fact table, using the assumptions stated above and also Estimate the total size of the fact table (in bytes), assuming that each field has an average of 5 bytes.

(c) Can you convert this star schema to a snowflake schema ? Justify your answer and design a snowflake schema if it is possible.

**3.** (a) Explain data mining as a step in the process of knowledge discovery. Give the architecture of a typical data mining system. **[10]**

(b) Mess personnel would like to identify four groups of food items from a larger group of seven food items so that if the soldiers select at least one item from each of the group they will obtain a certain fat and protein content. The following is the table that gives the fat and protein content in the food items.

| Food item # | Protein content, P | Fat content, F |
|---|---|---|
| Food item # 1 | 1.1 | 60 |
| Food item # 2 | 8.2 | 20 |
| Food item # 3 | 4.2 | 35 |
| Food item # 4 | 1.5 | 21 |
| Food item # 5 | 7.6 | 15 |
| Food item # 6 | 2 | 55 |
| Food item # 7 | 3.9 | 39 |

Apply minimum spanning tree clustering algorithm to solve this problem. **[10]**

**4.** (a) What is classification. What are the issues in classification. Apply statistical based algorithm to obtain the actual probabilities of each event to classify the new tuple as a tall. Use the following data : **[10]**

| Person ID | Name | Gender | Height | Class |
|---|---|---|---|---|
| 1 | Kristina | Female | 1.6 m | Short |
| 2 | Jim | Male | 2m | Tall |
| 3 | Maggie | Female | 1.9m | Medium |
| 4 | Martha | Female | 2.1. | Tall |
| 5 | Stephanie | Female | 1.7m | Short |
| 6 | Bob | Male | 1.85m | Medium |
| 7 | Kathy | Female | 1.6m | Short |
| 8 | Dave | Male | 1.7m | Short |
| 9 | Worth | Male | 2.2m | Tall |

(b) Consider the following data for a number of examples of weather, for several days, with a classification 'Play Tennis'. **[10]**

(i) Construct the decision tree using classification algorithm to decide on which day you can Play Tennis.

(ii) Express the decision tree that you have constructed as an expression or if–then–else sentences :

| Day | Outlook | Temperature | Humidity | Wind | Play Tennis |
|---|---|---|---|---|---|
| D1 | Sunny | Hot | High | Weak | No |
| D2 | Sunny | Hot | High | Strong | No |
| D3 | Overeast | Hot | High | Weak | Yes |
| D4 | Rain | Mild | High | Weak | Yes |
| D5 | Rain | Cool | Normal | Weak | Yes |
| D6 | Rain | Cool | Normal | Strong | No |
| D7 | Overeast | Cool | Normal | Strong | Yes |
| D8 | Sunny | Mild | High | Weak | No |
| D9 | Sunny | Cool | Normal | Weak | Yes |
| D10 | Rain | Mild | Normal | Weak | Yes |
| D11 | Sunny | Mild | Normal | Strong | Yes |
| D12 | Overeast | Mild | High | Strong | Yes |
| D13 | Overeast | Hot | Normal | Weak | Yes |
| D14 | Rain | Mild | High | Strong | No |

**5.** (a) What is Association Rule Mining. Give the Aprioi Algorithm. Apply association rule mining to find the all–frequent item sets from the following table : **[10]**

| Trans–id | Items |
|---|---|
| 100 | 1,3,4,6 |
| 200 | 2,3,5,7 |
| 300 | 1,2,3,5,8 |
| 400 | 2,5,9,10 |
| 500 | 1,4 |

(b) What is Web Mining ? Explain web content mining with reference to crawlers, harvest system, virtual web view and personalization. **[10]**

**6.** (a) What is Web Structure Mining ? Give the page rank and HITS Algorithm used in search engine. **[10]**

(b) What is visualization ? How can you mine descriptive statistical measures in large databases ? **[10]**

**7.** Write detailed notes on the following (any Two) **[20]**

(i) Dataware house project planning and management.

(ii) Temporal Association rules

(iii) Comparison between OLTP and OLAP

(iv) Spatial clustering algorithms.

❍❍❍

## Data Warehousing and Mining – November 2005

**Time : 3 Hrs.]** **[Marks : 100**

**N.B. :** (1) Question No. **1** is **compulsory**.

(2) Solve any **four** questions out of remaining **six** questions.

(3) All questions carry equal marks.

(4) **Figures** to the **right** indicate **full** marks.

**1.** (a) State & explain key issues to be considered while planning for a data warehouse for Information Technology department. **[20]**

(b) Explain general trends in data warehouse design.

**2.** (a) Explain related to data warehouse design **[20]**

(i) Deployment of data warehouse

(ii) Maintenance & growth of data warehouse

(b) List and explain activities of ETL process.

**3.** (a) Explain data mining as a step in KDD. Give the architecture of typical data mining system. **[20]**

(b) define data mining ? What is regression ? Explain association rules.

**4.** (a) What is web mining ? Explain content mining with respect to personalization, crawlers and virtual web view. **[20]**

(b) Explain spatial clustering algorithms.

**5.** (a) Differentiate between star schema, snowflake schema & fact constellation. **[20]**

(b) What is click stream mining ? Explain page rank algorithm used in search engines.

**6.** (a) Explain K–mean clustering algorithm. Suppose the data for clustering – **[20]**
{1, 3, 5, 15, 23, 11, 25} Consider K = 2, cluster the given data using above algorithm.

(b) What is visualization ? How can you mine descriptive statistical measures in large database ?

**7.** Write short notes on the following (any four) **[20]**

(i) Data warehouse project planning & management

(ii) Temporal mining

(iii) Trends in data mining

(iv) Comparison between OLAP and OLTP

(v) DMQL

❍❍❍

## Data Warehousing and Mining – May 2006

**Time : 3 Hrs.]** **[Marks : 100**

**N.B. :** (1) Question No. **1** is **compulsory**.
(2) Attempt any **four** out of remaining **six** questions.
(3) Figures to the **right** indicate **marks** for **each** question.
(4) Illustrate answers with **neat sketches** whenever **required**.
(5) Answers to the sub-questions of an individual question should be written together and one below the other.

**1.** (a) Define Data Warehouse with features ? Explain the architecture of data warehouse with suitable block diagram ? **[10]**
(b) What are techniques and application of data mining ? **[10]**

**2.** (a) Why metadata is important ? How to provide metadata ? **[10]**
(b) Explain the different methods of visualization of suitable example ? **[10]**

**3.** (a) Explain ETL of data warehousing in detail ? **[10]**
(b) Explain the architecture of data mining ? Differentiate between OLAP v/s OLTP ? **[10]**

**4.** (a) What are the types of OALP Servers ? Explain the different operation OLAP with suitable example ? **[10]**
(b) What is Association Rule ? Describe the suitable tech. using FP trees of mining frequent pattern ? **[10]**

**5.** (a) Explain the different clustering algorithm applicable to data mining ? **[10]**
(b) All Electronics company have sales department. Sales consider four dimensions namely time, item, branch and location. The schema contain a central fact tables sales with two measures dollars sold and unit sold. **[10]**
(i) Defined star schema and Snowflake schema for above case using DMQL ?
(ii) Design star schema and Snowflake schema for same.

**6.** (a) Explain the different algorithm for spatial mining ? **[10]**
(b) Explain the general trend in data warehousing ? **[10]**

**7.** Write short note on **[20]**
(a) Web mining
(b) Project Planning and management of data warehouse
(c) Different Classification algorithm in Data mining.

❍❍❍

## Data Warehousing and Mining – November 2006

**Time : 3 Hrs.]** **[Marks : 100**

**N.B.:** (1) Question No.**1** is **compulsory**.
(2) Solve any **four** questions out of remaining **six** questions.

**1.** "Dictionary of the living World" is an ideal type of Multimedia presentation using at the maximum text, sound, images and video. You are appointed as a consultant to implement this application in distributed environment. Assume required data if any and specify clearly. **[20]**
(a) How can you manage multimedia object servers?
(b) Give the design for managing distributed objects.

**2.** (a) Describe the algorithm for CCITT group 4 2D compression. **[10]**
(b) Describe the algorithm for the MPEG. Clearly state and explain the mathematical treatment and building blocks used in detail. **[10]**

**3.** (a) Explain JPEG DIB file format for still and motion images. **[10]**
(b) Explain MIDI file format with reference to its chunks and communication protocol. **[10]**

**4.** Product advertisement using Multimedia : The most impressive advertisement of products use mostly animation, sound and video combined together. A well designed multimedia presentation based basically on animation, video and sound attracts client's attention and through the use of sound data can pass easily a message to the user. Almost all of this kind of presentations follows a slide–show. The user basically watches the application that is being presented in a predetermined way. You are appointed as a consultant to automate the computerised system to sell the various electronic goods and softwares of multimedia such as maya, photoshop etc. Assume required data and specify clearly.
Answer the following questions : **[20]**
(a) Give the Multimedia database schema design.
(b) Design the Multimedia Authoring System.

**5.** Continuous education program : One of the most important application using both technologies, networks and multimedia, is distance learning. Computers offer the chance for new educational procedures, which in combination to networks reach levels beyond imagination. Anyone could be educated by the greatest teachers of the world. Computers cannot replace the teacher, but they can bring him closer to the student. Assume required data if any and specify clearly. You are appointed as a consultant to implement this application. **[20]**
(a) Design the performance requirements if this application is to be used in distributed environment.
(b) Give the Workflow design.
(c) Model the various objects and design special Multimedia User Interface.

**6.** (a) Consider an RTP session consisting of four users, all of which are sending and receiving RTP packets into the same multicast address. Each user sends the video at 100 kbps. **[10]**
(i) RTCP will limit its traffic to what rate ?
(ii) A particular receiver will be allocate how much RTCP bandwidth ?
(iii) A particular sender will be allocated how much RTCP bandwidth ?
(b) Explain hypermedia messaging will suitable example. **[10]**

**7.** Write detailed note on (Any two) : **[20]**
(a) Video images, animation and full motion video
(b) Image Scanners
(c) Storage and Retrieval Technology

❍ ❍ ❍

## Data Warehousing and Mining – May 2007

**Time: 3Hrs.]** **[Marks : 100**

**N.B.** (1) Question No. **1** is **compulsory**.
(2) Attempt any **four** out of remaining.
(3) Figures to the **right** indicate full marks.

**1.** (a) A bank wants to develop a data warehouse for effective decision-making about their loan schemes. The bank provides loans to customers for various purposes like House Building Loan, Car Loan, Educational Loan, Personal Loan, etc. .The whole country is categorized into a number of regions, namely, North, South, East and West. Each region consists of a set of states. Loan is disbursed to customers at interest rates that change from time to time. Also, at any given point of time, the different types of loans have different rates. The data warehouse should record an entry for each disbursement of loan to customer. With respect to the above business scenario.
(1) Design an information package diagram. Clearly explain all aspects of the diagram. **[5]**
(2) Draw a star schema for the data warehouse clearly identifying the Fact table(s), Dimension table(s), their attributes and measures. **[5]**

(b) Consider the following transaction database :

| TID | Items |
|---|---|
| 01 | A, B, C, D |
| 02 | A, B, C, D, E, G |
| 03 | A, C, G, H, K |
| 04 | B, C, D, E, K |
| 05 | D, E, F, H, L |
| 06 | A, B, C, D, L |
| 07 | A, D, F, L |
| 08 | B, I, E, K, L |
| 09 | A, B, D, E, K |
| 10 | C, D, H, I, K |
| 11 | A, E, F, H, L |
| 12 | B, C, D, F |
| 13 | A, B, C, D |
| 14 | A, D, H, K |
| 15 | B, C, D, E, H, L |

Apply the **Apriori** algorithm with minimum support of 30% and minimum confidence of 75%, and find all the association rules in the data set. **[10]**

**2.** Define the following by giving examples **[20]**
(a) Factless Fact Tables
(b) Snowflake Schema
(c) Outliers in data mining
(d) Supervised learning in data mining
(e) Family of stars

**3.** (a) Consider a data warehouse for a hospital, where there are three dimensions: (1) Doctor, (2) Patient, and (3) Time, and two measures: (1) Count and (2) Charge, where charge is the fee that the doctor charges a patient for a visit. Using the above example describe the following OLAP operations
(1) Slice (2) Dice (3) Rollup (4) Drill Down (5) Pivot **[10]**
(b) Describe the different clustering algorithms. Discuss the advantages and disadvantages of each. **[10]**

**4.** (a) Consider an online travel agency that helps customers to plan and schedule their holidays. The agency maintains all past history in a data warehouse. Describe the different classes of users who could access this data warehouse and design the information delivery framework for this data warehouse. **[8]**
(b) What is web mining? Illustrate the working of the HITS algorithm by using an Example query on a search engine (example – search for "web mining applications") **[12]**

**5.** (a)

| Transaction | Income | Credit | Decision |
|---|---|---|---|
| 1 | Very High | Excellent | AUTHORIZE |
| 2 | High | Good | AUTHORIZE |
| 3 | Medium | Excellent | AUTHORIZE |
| 4 | High | Good | AUTHORIZE |
| 5 | Very High | Good | AUTHORIZE |
| 6 | Medium | Excellent | AUTHORIZE |
| 7 | High | Bad | REQUEST ID |
| 8 | Medium | Bad | REQUEST ID |
| 9 | High | Bad | REJECT |
| 10 | Low | Bad | CALL POLICE |

Using the above table illustrate any one classification technique. Further indicate how we can classify a new transaction, with (Income = Medium and Credit = Good). **[10]**

(b) Describe the ETL cycle in a data warehouse. **[10]**

**6.** (a) What are concept hierarchies ? Explain with an example. Describe the concept hierarchy using DMQL. **[8]**

(b) What is the importance of metadata in a data warehouse? What are the different types of metadata stored in a data warehouse? Illustrate with a simple Customer Sales data warehouse. **[12]**

**7.** (a) With a neat diagram describe the KDD process. **[8]**

(b) Discuss the importance of visualization in a data warehouse and in data mining. **[12]**

❍ ❍ ❍

## Data Warehousing and Mining – November 2007

**Time: 3Hrs.]** **[Marks : 100**

**N.B.:** (1) Question No.**1** is **compulsory**.
(2) Attempt any **four** out of remaining **six** questions.
(3) Figures to the **right** indicate **full** marks.
(4) Illustrate answers with neat sketches wherever **required**.

**1.** (a) Define Data warehouse with features ? Explain the architecture of data warehouse with suitable block diagram. **[10]**

(b) Draw KDD process diagram ? Explain in detail ? **[10]**

**2.** (a) What are the types of OLAP servers ? Explain the different operation of OLAP with suitable example ? **[10]**

(b) Explain the different methods of visualization with suitable example ? **[10]**

**3.** (a) Differentiate between star schema, showflake schema and fact constellation. **[10]**

(b) Define classification ? Explain any two algorithms with suitable example ? **[10]**

**4.** (a) Name any five types of activities that are part of ETL process ? Which of these are time consuming ? Explain any three ? **[10]**

(b) What is Association Rule mining ? Give the Apriori Algorithm. Apply association Rule to find all frequent item sets from following table ? **[10]**

| Trans-ID | Items |
|---|---|
| 100 | 1, 3, 4, 6 |
| 200 | 2, 3, 5, 7 |
| 300 | 1, 2, 3, 5, 8 |
| 400 | 2, 5, 9, 10 |
| 500 | 1, 4 |

(Let min-support = 60% and min-confidence = 80)

**5.** (a) Why metadata is important ? How to provide metadata ? **[10]**

(b) What are techniques and application of data mining ? **[10]**

**6.** (a) What is MDDB ? What types of business requirements determine use of MDDB in Data warehouse ? **[10]**

(b) Explain K-mean clustering algorithm ? Suppose the data for clustering is {1, 3, 5, 15, 23, 11, 25}. Consider k = 2, cluster the given data using above algorithm. **[10]**

**7.** Write short notes on the following (any four) : **[20]**

(a) Data warehouse project planning and management
(b) Temporal mining
(c) Comparison between OLAP arid OLTP
(d) Web structure mining
(e) Trends in data mining

❍ ❍ ❍

## Data Warehousing and Mining – May 2008

**Time : 3 Hrs.]** **[Marks : 100**

**N.B.** (1) Question No. **1** is **compulsory**.
(2) Attempt any **four** questions out of the remaining six questions.
(3) All Question carry equal marks.
(4) Illustrate answers with neat sketches whenever required.

**1.** (a) How are top-down and bottom up approaches for building data warehouse differ ? Discuss the merits and limitation of each approach? **[10]**
(b) Explain Data mining as a step in KDD. Give the architecture of typical DM system. **[10]**

**2.** (a) Give information package for recording information requirement for "Hotel Occupancy" considering dimensions like time, Hotel etc. Design star schema from information package. **[10]**
(b) What is Association rule? Describe the suitable technique using FP trees of mining frequent pattern? **[10]**

**3.** (a) Describe features of web enabled Data warehouse? Why is data security a major concern for Web-enabled data warehouse? **[10]**
(b) Define classification. Explain K-nearest neighbor classification algorithm. **[10]**

**4.** (a) What are the types of OLAP servers? Explain the different operations of OLAP with suitable example? **[10]**
(b) List and describe five primitives for specifying data mining task. **[10]**

**5.** (a) Explain ETL of data warehousing in detail. **[10]**
(b) What is clustering? Explain k – means clustering algorithm. Suppose the data for clustering – {2, 4, 10, 12, 3, 20, 11, 25}
Consider k = 2, cluster the given data using above algorithm. **[10]**

**6.** (a) State key issues to be considered while planning for data warehouse. Explain any four of them. **[10]**
(b) What is visualization? How can you mine descriptive statistical measures in large databases? **[10]**

**7.** Write short notes on the following (any two) **[20]**
(a) Comparison between Data warehouse and Data Mart
(b) Application and Trends in Data mining
(c) Temporal mining
(d) DMQL

❍ ❍ ❍

## Data Warehousing and Mining – November 2008

**Time : 3 Hrs.]** **[Marks : 100**

(2) Attempt any **four** questions out of the remaining **six** questions.
(3) **All** question carry **equal** marks.

**1.** (a) Define Data Warehouse with feature. Explain the architecture warehouse with suitable block diagram. **[10]**
(b) Explain Data mining as a step in KDD. Give the architecture of typical DM system. **[10]**

**2.** (a) Why metadata is important? How to provide metadata? **[10]**
(b) Explain regression and association rules in Data mining along with example. **[10]**

**3.** (a) What are the type of OLAP Server? Explain the different operation of OLAP with suitable example? **[10]**
(b) Explain the different methods of visualization with suitable example. **[10]**

**4.** (a) All Electronics company have sales department. Sales consider four dimensions namely time, item, branch, and location. The schema contains a central fact tables sales with two measures dollars-sold and unit sold.
(i) Define star schema and Snowflake schema for above case using DMQL.
(ii) Design star schema and Snowflake schema for same. **[10]**
(b) List and describe five primitives for specifying data mining task. **[10]**

**5.** (a) Explain ETL of data warehousing in detail. **[10]**
(b) What is clustering? Explain k-means clustering algorithm. Suppose the data for clustering – {2, 4, 10, 12, 3, 20, 11, 25}.
Consider k = 2, cluster the given data using above algorithm. **[10]**

**6.** (a) State key issue to be considered while planning for data warehouse. Explain any four of them. **[10]**
(b) Explain the general trend in data warehousing. **[10]**

**7.** Write short note on (any two): **[20]**
(a) Web mining
(b) Comparison between OLAP and OLTP.
(c) DMQL
(d) Spatial clustering algorithm.

❍ ❍ ❍