Data Warehousing and Data Mining (MB331IT): April 2008

Section A : Basic Concepts (30 Marks)

- **This section consists of questions with serial number 1 - 30.**
- **Answer all questions.**
- **Each question carries one mark.**
- **Maximum time for answering Section A is 30 Minutes.**

1. Which of the following is not produced in the technical blueprint stage of data warehouse delivery process?
(a) Overall system architecture
(b) Server and data mart architecture
(c) Essential components of database design
(d) Backup and recovery strategy
(e) Detailed design of database.

2. Which of the following is/are the function(s) of warehouse manger in a data warehouse?
I. Transforming and managing the data.
II. Backs up and archiving the data warehouse.
III. Directing and managing queries.
(a) Only (I) above
(b) Only (II) above
(c) Both (I) and (II) above
(d) Both (II) and (III) above
(e) All (I), (II) and (III) above.

3. The encrypted information stored in a database is an example of
(a) Shallow knowledge
(b) Multi-dimensional knowledge
(c) Hidden knowledge
(d) Deep knowledge
(e) Representation knowledge.

4. Which of the following statements is/are true about hardware architectures used in Data warehouse solutions?
I. A Symmetric Multi-Processing (SMP) machine is a set of loosely coupled CPUs, each of which has its own
memory and disk.
II. A Massively Parallel Processing (MPP) machine is a set of tightly coupled CPUs, that share memory and
disk.
III. A Non Uniform Memory Architecture (NUMA) machine is basically a tightly coupled cluster of Symmetric Multi-Processing (SMP) nodes.
(a) Only (III) above
(b) Both (I) and (II) above
(c) Both (I) and (III) above
(d) Both (II) and (III) above
(e) All (I), (II) and (III) above.

5. NEXT inc., a reputed BPO company is maintaining the database of the customers‘ office & home phone
numbers. But it is having a problem of wasting space when the customers are using either of them. So which of
the following processes can help the company in eliminating the redundant data?
(a) Structuring
(b) Randomizing
(c) Analyzing
(d) Normalizing
(e) Actualizing.

6. An accounts database has a table with invoices and each invoice is associated with a particular supplier. Supplier
details (such as address and name) are kept in a separate table; each supplier is given a 'supplier number' to
identify them. Each invoice record has an attribute containing the supplier number for that invoice. Identify the
1

primary key, foreign keys in the tables.
(a) Supplier number in the supplier table is foreign key and InvoiceNumber in the Invoices table is the primary key
(b) Supplier number in the supplier table is primary key and InvoiceNumber in the Invoices table is the foreign key
(c) Supplier number in the supplier table is primary key, InvoiceNumber in the Invoices table is the primary
key and Supplier number in Invoices table is the foreign key
(d) Supplier number in the supplier table is foreign key and InvoiceNumber, Supplier number in Invoices
table are primary keys
(e) Supplier name in the supplier table is the primary key, InvoiceNumber in the Invoices table is the primary key and Supplier number in Invoices table is the foreign key.

7. The ADSM backup software package was produced by
(a) HP
(b) Sequent
(c) IBM
(d) Epoch
(e) Legato.

8. A perceptron consists of a simple three-layered network, with output units called
(a) Photo-receptors
(b) Associators
(c) Responders
(d) Acceptors
(e) Rejectors.

9. In an organization, the relation between projects and employees is

(a) One-to-one relationship
(b) One-to-many relationship
(c) Many-to-one relationship
(d) Many-to-many relationship
(e) Symmetric relationship.

10. Collection of interesting and useful patterns in a database is called
(a) Prediction
(b) Knowledge
(c) Enrichment
(d) Heuristics
(e) Information.

11. Which of the following is the correct order of empirical cycle of scientific research?
(a) Analysis, Theory, Prediction and Observation
(b) Prediction, Analysis, Theory and Observation
(c) Observation, Analysis, Theory and Prediction
(d) Observation, Prediction, Analysis and Theory
(e) Prediction, Observation, Analysis and Theory.

12. The main organizational justification for implementing a data warehouse is to provide
(a) Cheaper ways of handling transactions
(b) Decision support
(c) Large scale transaction processing
(d) Storing large volumes of data
(e) Access to the data.

13. Which of the following tasks is performed by the copy management tool in a load manager?
(a) Fast load
(b) Simple transformation
(c) Complex checking
(d) Job control
(e) Query profile analysis.

14. Which of the following backups is a synonym for hot backup?
2

(a) Complete backup
(b) Partial backup
(c) Cold backup
(d) Online backup
(e) Both (a) and (d) above.

15. A salesman_master table contains attributes like salesman no, salesman_name , address 1, address 2, city,
pincode, state, sal_amt, tgt_to_get, ytd_sales, remarks. Write a SQL statement to find the names of the salesmen
who have a salary equal to Rs.3000.
(a) Select count(salesman_ name) from salesman_master where sal_amt=3000

(b) Select * from salesman_master where sal_amt=3000
(c) Select count(sal_amt) from salesman_master where sal_amt=3000
(d) Select salesman_ name from salesman_master where sal_amt=3000
(e) Select salesman_ name, state from salesman_master where sal_amt=3000.

16. Alexandria backup software package was produced by
(a) HP
(b) Sequent
(c) IBM
(d) Epoch Systems
(e) Legato.

17. What is Artificial intelligence?
(a) Putting your intelligence into computer
(b) Programming with your own intelligence
(c) Making a machine intelligent
(d) Playing a game
(e) Putting more memory into computer.

18. Which of the following Redundant Array of Inexpensive Disk (RAID) levels are commercially viable and widely
used?
(a) 0, 1 and 2
(b) 0, 1 and 4
(c) 0, 3 and 5
(d) 0, 4 and 5
(e) 0, 1 and 5.

19. In star schema, the surrounding reference tables around the central factual dimension table are called as
(a) Dimensional tables
(b) Fact tables
(c) Relational tables
(d) Temporary tables
(e) Meta data tables.

20. Which of the following managers is not a part of system managers in a data warehouse?
(a) Configuration manager
(b) Schedule manager
(c) Event manager
(d) Database manager
(e) Load manager.

21. In which of the following Redundant Array of Inexpensive Disks (RAID) levels bitwise striping of data with
parity disk is maintained?
(a) 1
(b) 2
(c) 3

(d) 4
(e) 5.

22. Which of the following type of knowledge is the information that can be easily retrieved from databases using
query tools?
(a) Shallow knowledge
(b) Multi-dimensional knowledge
3

(b) Multi-dimensional knowledge
(c) Hidden knowledge
(d) Deep knowledge
(e) Tacit knowledge.

23. A petabyte equals to
(a) 1024 terabytes
(b) 1024 gigabytes
(c) 1024 megabytes
(d) 1024 kilobytes
(e) 1024 bytes.

24. The non-trivial extraction of implicit, previously unknown and potentially useful knowledge from data is known
as
(a) Data selection
(b) Data mirroring
(c) Data cleaning
(d) Knowledge discovery in databases
(e) Data design.

25. The theory in which information content of the message is related to the probability that a certain message will
occur is
(a) Shannon‘s communication theory
(b) Kolmogorov complexity theory
(c) Rissanen theory
(d) Freud‘s theory
(e) Kohonen theory.

26. An approach to a problem that is not guaranteed to work but performs well in most cases is
(a) Heuristics
(b) Enumeration
(c) Falsification
(d) Naive prediction
(e) Enrichment.

27. Which of the following is not a stage in Knowledge Discovery Process?
(a) Data selection

(b) Cleaning
(c) Enrichment
(d) Reporting
(e) Data encapsulation.

28. Which of the following statements is/are true about ”Online Analytical Processing (OLAP)‘?
I. OLAP tools do not learn.
II. OLAP creates no new knowledge.
III. OLAP is more powerful than Data mining.
IV. OLAP cannot search for new solutions.
(a) Only (I) above
(b) Only (III) above
(c) Both (I) and (II) above
(d) Both (II) and (III) above
(e) (I), (II) and (IV) above.

29. In Freud‘s theory of psychodynamics, the human brain was described as a
(a) Decision tree
(b) Neural network
(c) Learning
(d) Knowledge
(e) Visualization technique.

30. —An individual learns how to carry out a certain task by making a transition from a situation in which the task
cannot be carried out to a situation in which the same task can be carried out under the same circumstances“. The
4

given definition is referred to as
(a) Learning
(b) Knowledge
(c) Machine learning
(d) Learning algorithm
(e) Meta learning.
Section B : Caselets (50 Marks)
• This section consists of questions with serial number 1 œ 6.
• Answer all questions.
• Marks are indicated against each question.
• Detailed explanations should form part of your answer.
• Do not spend more than 110 - 120 minutes on Section B.
Caselet 1
Read the caselet carefully and answer the following questions:
1. Why KPC insurance company need tools to manage Data warehouse? Explain. (5 marks)

2. If you are the manager of the KPC insurance company, what are the issues you will

consider before buying the ETL tool? (6 marks)
KPC insurance is implementing new data warehousing and reporting technology

which has brought a cultural transformation by installing new software‘s by the company and the way they are operated.
The key business reasons for bringing in new data warehousing and reporting technology:
From the business perspective there are two main things which make information even more important for an insurance company than others with regard to its price for a product. It means a good business model and good understanding of trends. The other thing is the nature of insurance business which is cyclical. The market will attract additional capital which drives prices down leading to losses-so there is always a period when industry is making money and period when it is not making money.
Are information systems not up to mark:
Because information is vital, KPC always wanted decent MIS in place. It had single scorecard system across the business-a PowerBuilder application that gives profit and loss across the whole business. Also a mainframe type reporting system apart from local databases and excels spreadsheets to look at the performance of particular products and territories and work with particular brokers.
What new system was adopted?
The Insurance company had a significantly large database, which was consolidated from various systems using an ETL (Extract, Transform, Load) tool. This tool would mine data from a database and then transform and store the mined data in a data warehouse. KPC met with a consultant before buying the ETL tool.
How does new system operate?
They had single data ware house but with many subject areas with in it. The first thing rolled out in July 2005 was profitability module which enables to look at results and in particular future trends. Then in October 2005 a broker module which allows to see profit and loss for each broker was the key. Then bought in claims management module. In may 2006 profitability management module looking at sales processes, new business. Then brokerage module was extended looking at
5

budget and tracking performance against this.
Having installed new reporting systems, how did managers actually use them?
What was done upfront was to articulate that KPC were in business of cultural transformation - it sounded pretentious. But we wanted to develop a culture where reliance on good MIS to what KPC is as a company.
END OF
CASELET
1
Caselet 2
Read the caselet carefully and answer the following questions:
3. Critically analyze the objectives of Disaster Recovery Plan (DRP). (10 marks)

4. What are the requirements Core mind might have been considered for Disaster

Recovery Plan (DRP) (Discuss in the context of Data warehousing environment)? (12 marks)
Disasters take place every day, from small fires to tornadoes to acts of terrorism. The attacks of September 11, the blackout that struck the East Coast and Michigan in USA, should have taught managers to be proactive to ensure operations can continue

during a crisis.
Many software companies may experience problems as a result of a fire, natural calamity or massive power outage. If software companies and all of the resources it has for day-to-day operations are no longer available, it would wreak havoc.
Core mind, World‘s one of the largest department store retailer with more than 500 stores, generated a tremendous amount of data spread across a number of operational systems. Core mind is implementing a Teradata Warehouse with the Teradata Database. One sudden day it was affected by the cyclone that hit Kolkata, but it was able to restore facilities within 60 hours because it had a well-defined Disaster Recovery Plan (DRP) and procedures.
Core mind is putting together a disaster recovery plan to ensure that its large global customers continue to get round-the-clock support, even if the subcontinent goes to war. It wants to set up disaster recovery sites in Singapore and Canada. The plan is to move employees to these sites and resume operations in the advent of an emergency.
END OF
CASELET
2
Caselet 3
Read the caselet carefully and answer the following questions:
5. Assume that you are a Tester at Allina, explain how you will test Backup Recovery. (12 marks)

6. What other factors (Other than mentioned in the caselet) Allina might have been

considered for successful implementation of data warehousing solution? Explain. (5 marks)
Allina Health System Implements Data Warehouse
Minneapolis-based Allina Health System is a non-profit healthcare system serving one million people living in Minnesota. They vertically integrated healthcare system includes 13,0000 physicians and 22,000 employees who own and manage 19 hospitals, 57 clinics and seven nursing homes networked across the country. Allina provides people with a life time of healthcare options and full continuum of care-from prevention and wellness services such as health screening and immunizations to high-quality and technologically advanced inpatient and outpatient services.
6

Allina is pressured from all sides to reduce costs, holding the premiums while providing its patients with best treatment possible. It must also integrate information systems and business practices of multiple organizations it has acquired through various mergers in the recent past. To meet these challenges, management has decided to develop and market a data warehouse strategy across the country. The goal is to enable Allina to pull information together and integrate it in ways never done before; for example extracting cost information from hospital and the health plans, comparing best practices in treatments and matching cost of level of service.
Meeting this was a data warehouse challenge. The data modeling effort had to ensure that each data mart of data warehouse could be tied together logically and physically. The key elements of the success of the project first a multitier database was developed so that so that both summary and detailed information is available.

Second, large implementation team divided in to management groups in-charge of architecture, data modeling and databases, implement new subsets of data warehouse across subject areas and data marts. Third, organizations used pilot projects to build data marts for administration reporting for hospital and large scale data warehouse for patient histories for all hospitals. Apart from these key elements one more element has contributed to the success of this project i.e., thorough test of the project. The team involved in testing the project tested all aspects pertaining to this project, for example, operational environmental aspects, database, backup recovery etc.

END OF

CASELET 3

END OF SECTION B

Section C : Applied Theory (20 Marks)

• This section consists of questions with serial number 7 - 8.
• Answer all questions.
• Marks are indicated against each question.
• Do not spend more than 25 - 30 minutes on Section C.

7. Discuss the Data mining primitives in SQL. (10 marks)

8. Discuss the different forms of Neural networks. (10 marks)