Question Paper
Data Warehousing and Data Mining (MB3G1IT): January 2009

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.

<Answer>

1. Which of the following data warehouse process managers transform and manages the data?

(a) Query manager
(b) Database manager
(c) Event manager
(d) Warehouse manager
(e) Load manager.

<Answer>

2. Which of the following task is performed by the copy management tool in a load manager?

(a) Job control
(b) Complex checking
(c) Backup and archive
(d) Fast load
(e) Simple transformation.

<Answer>

3. Which of the following statements is/are **true** about Massively Parallel-Processing (MPP) machine?

I. MPP machines require the use of a distributed lock manager to maintain the integrity of the distributed resources across the system as a whole.
II. Design and management of a data warehouse on an MPP machine is considerably more difficult than on Symmetric Multi-Processing (SMP) system.
III. The amount of CPU power that can be made available in an MPP machine is independent.

(a) Only (I) above
(b) Only (II) above
(c) Only (III) above
(d) Both (I) and (II) above
(e) Both (II) and (III) above.

<Answer>

4. It is certain that large data sorts are going to be required within the data warehouse. If we can gauge the size of the largest transaction that will realistically be run, we can use this to size the temporary requirements. If not, the best we can do is tie it to the size of a partition. If the number of concurrent queries allowed are 4 and the size of the partition is 3GB, then we need to set the temporary space (T) to

(a) 7 GB
(b) 12 GB
(c) 24 GB
(d) 27 GB
(e) 36 GB.

<Answer>

5. Which of the following backup software packages is produced by 'HP'?

(a) OmnibackII
(b) ADSM
(c) Alexandria
(d) Epoch
(e) Networker.

6. Which of the following statements is/are **true** about various types of partitioning? <Answer>

I. Vertical partitioning can take two forms: normalization and row splitting.
II. Before using a vertical partitioning we need to be very sure that there will be no requirements to perform major join operations between the two partitions.
III. Horizontal hardware partitioning technique spreads the processing load by horizontally partitioning the fact table into small segments and then physically storing each segment in a different node.

(a) Only (I) above
(b) Only (II) above
(c) Only (III) above
(d) Both (I) and (II) above
(e) All (I), (II) and (III) above.

7. Which of the following statements is/are **true** about the phases in data warehouse delivery process? <Answer>

I. Technical blueprint phase is the stage where the first production deliverable is produced.
II. Build the vision phase must deliver an overall architecture that satisfies the long-term requirements and a definition of the components that must be implemented in the short term in order to derive any business benefit.
III. The purpose of the business case is to identify the projected business benefits that should be derived from using the data warehouse.

(a) Only (I) above
(b) Only (II) above
(c) Only (III) above
(d) Both (I) and (II) above
(e) Both (II) and (III) above.

8. Which of the following is/are **not** produced in the technical blueprint stage of data warehouse delivery process? <Answer>

I. Detailed design of database.
II. Essential components of database design.
III. Server and data mart architecture.
IV. Backup and recovery strategy.

(a) Only (I) above
(b) Only (III) above
(c) Both (I) and (IV) above
(d) (I), (III) and (IV) above
(e) (II), (III) and (IV) above.

9. Which Redundant Array of Inexpensive Disks (RAID) levels use byte wise striping of data with parity disk? <Answer>

(a) Level 1
(b) Level 2
(c) Level 3
(d) Level 4
(e) Level 5.

10. Which of the following statements is/are **false** about query management process? <Answer>

I. Query management process is the system process that manages the queries and speeds them up by directing queries to the most effective data source.
II. Like other system processes, query management process generally operates during the regular load of information into the data warehouse.
III. Query management process must ensure that no single query can affect the overall system performance.

(a) Only (I) above
(b) Only (II) above
(c) Only (III) above
(d) Both (I) and (II) above
(e) Both (II) and (III) above.

<Answer>

11. Which of the following tasks is performed by system management tool in a warehouse manager?

(a) Job control
(b) Create indexes
(c) Backup and archive
(d) Generate star schemas
(e) Query profile analysis.

<Answer>

12. Which of the following statements is/are **true** about fact data and dimension data?

I. Fact data represents a physical transaction that has occurred at a point in time and as such is unlikely to change on an ongoing basis during the life of the data warehouse.
II. In general, dimension data in a star schema or snowflake schema is designed to minimize the cost of change and is typically very low volume data (i.e, under 5GB).
III. Fact data will have only one foreign key whereas reference data will have one primary key.

(a) Only (I) above
(b) Only (II) above
(c) Only (III) above
(d) Both (I) and (II) above
(e) Both (II) and (III) above.

<Answer>

13. The reason(s) for partitioning the fact table is/are

I. To increase the performance.
II. To assist the management of the data.
III. To assist backup/recovery.

(a) Only (I) above
(b) Only (II) above
(c) Only (III) above
(d) Both (I) and (II) above
(e) All (I), (II) and (III) above.

<Answer>

14. Which of the following statements is/are **true** about metadata?

I. As a part of extraction and load process, metadata is used to map data sources to the common view of information within the data warehouse.
II. As a part of the warehouse management process, metadata is used to direct a query to the most appropriate data source.
III. As a part of query management process, metadata is used to automate the production of summary tables.

(a) Only (I) above
(b) Only (II) above
(c) Only (III) above
(d) Both (I) and (II) above
(e) Both (II) and (III) above.

<Answer>

15. Fact table identification process requires four steps. Arrange the following four steps in **correct** sequence.

I. Look for elemental transactions.
II. Check if fact is a dimension.
III. Check if dimension is a fact.
IV. Determine key dimensions.

(a) I-II-III-IV
(b) I-IV-II-III
(c) I-III-II-IV
(d) IV-I-III-II
(e) IV-II-III-I.

<Answer>

16. As with any relational system, foreign keys within a fact table can be structured in two ways: using intelligent keys and using non-intelligent keys. Which of the following statements are **true** about intelligent and non-intelligent keys?

I. In intelligent keys, each key represents the unique identifier for the item in the real world.
II. In non-intelligent keys, each unique key is generated automatically and refers to the unique identifier for the item in the real world.
III. Usage of intelligent key can be costly and time consuming.
IV. Unless it is certain that identifiers will not change, it is safer to use intelligent keys.

(a) Both (I) and (II) above
(b) Both (II) and (III) above
(c) (I), (II) and (III) above
(d) (II), (III) and (IV) above
(e) All (I), (II), (III) and (IV) above.

<Answer>

17. Which of the following statements is/are **false** 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.

<Answer>

18. Which of the following statements is/are **false** about Online Analytical Processing (OLAP) tools?

I. OLAP tools do not learn.
II. OLAP tools create new knowledge.
III. OLAP tools are more powerful than data mining.
IV. OLAP tools 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.

<Answer>

19. NetSol, a reputed Garments 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) Analyzing
(b) Normalizing
(c) Structuring
(d) Randomizing
(e) Actualizing.

<Answer>

20. Which of the following backup software packages is produced by 'Legato'?

(a) OmnibackII
(b) ADSM
(c) Alexandria
(d) Epoch
(e) Networker.

<Answer>

21. In an organization, what is the relationship between DEPARTMENT and EMPLOYEE?

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

<Answer>

22. Which of the following are the data mining techniques?

I. Association rules.
II. Neural networks.
III. Normalization.
IV. Genetic algorithms.

(a) Both (I) and (II) above
(b) Both (I) and (III) above
(c) (I), (II) and (IV) above
(d) (II), (III) and (IV) above
(e) All (I), (II), (III) and (IV) above.

<Answer>

23. Consider the student table with attributes: sname, snum, totalmarks, semester. Write an SQL statement to display the name, total marks and student number whose snum is CB14.

(a) select sname, totalmarks, snum from student where snum=‘CB14’;
(b) select sname, totalmarks, snum from student where snum=CB14;
(c) select sname, totalmarks from student where snum=CB14;
(d) select sname, totalmarks from student where snum=‘CB14’;
(e) select sname, totalmarks where snum=‘CB14’;.

<Answer>

24. According to 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.

<Answer>

25. Which of the following are the stages in Knowledge Discovery Process?

I. Data encapsulation.
II. Data selection.
III. Enrichment.
IV. Reporting.

(a) Both (I) and (II) above
(b) Both (I) and (III) above
(c) (I), (II) and (III) above
(d) (II), (III) and (IV) above
(e) All (I), (II), (III) and (IV) above.

<Answer>

26. In Knowledge Discovery Process, which of the following is a coding operation in which an attribute with cardinality ‘n’ is replaced by ‘n’ binary attributes?

(a) Flattening
(b) Replication
(c) Redundancy
(d) Falsification
(e) Atomicity.

<Answer>

27. In general, data mining algorithms should not have a complexity higher than

(a) n(log n)
(b) (log n)
(c) (n+1)(log n)
(d) n(log (n+1))
(e) 2n(log n).

<Answer>

28. In data mining, which of the following statements are **true** about the various types of knowledge?

I. Shallow knowledge is the information that can be analyzed using Online Analytical Processing (OLAP) tools.
II. Multi-dimensional knowledge is the information that can be easily retrieved from databases using a query tool such as Structured Query Language (SQL).
III. Hidden knowledge is the information that can be found easily by using pattern recognition or machine-learning algorithms.
IV. Deep knowledge is the information that is stored in the database but can only be located if we have a clue that tells us where to look.

(a) Both (I) and (II) above
(b) Both (III) and (IV) above
(c) (I), (II) and (III) above
(d) (II), (III) and (IV) above
(e) All (I), (II), (III) and (IV) above.

<Answer>

29. Which of the following statements is/are **true** about Genetic algorithms?

I. These algorithms can be viewed as a kind of meta-learning strategy.
II. Any programmer can write the basic structure of this algorithm easily.
III. Solutions found by these algorithms are coded symbolically and it is very hard to read as compared to neural networks.

(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.

<Answer>

30. Which of the following statements is/are **true** about the various forms of neural networks?

I. A perceptron consists of a simple three-layered network with input units called photo-receptors, intermediate units called associators and output units called responders.
II. A back propagation network not only has input and output nodes, but also a set of intermediate layers with hidden nodes.
III. A Kohenen self-organizing map is a collection of neurons or units, each of which is connected to a small number of other units called its neighbors.

(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.

Data Warehousing and Data Mining (MB3G1IT): January 2009

Section B : Caselets (50 Marks)

- This section consists of questions with serial number 1 – 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. Do you think ETL tools will solve architectural problems for data warehouses? Discuss. **(7 marks)** <Answer>

2. Critically analyze the features of ETL tools. **(9 marks)** <Answer>

3. TechnoSoft developed data warehouse architecture which will allow the information **(10 marks)** <Answer>

base to be extended and enhanced overtime. Explain the architecture of such a data warehouse.

A health benefits company operating in the United States serves more than 11.9 million customers. Its product portfolio includes a diversified mix of managed care products, including Preferred Provider Organizations (PPOs), Health Maintenance Organizations (HMOs) and Point-Of-Service (POS) plans. The company also offers several specialty products, including group life and disability insurance benefits, pharmacy benefit management and dental, vision and behavioral health benefits services.

Business Need

The 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. The company wanted to outsource the reengineering of the data warehouse and partnered with TechnoSoft to build a modular, scalable architecture so that the processing batch time could be reduced considerably.

After the success of this engagement, TechnoSoft handled the maintenance of the data warehouse, thus freeing the client's internal resources to tackle other pressing tasks, and lending flexibility in operations.

Challenges and Requirements

The task of taking a huge database and reshaping it to improve efficiency is a difficult undertaking. TechnoSoft found that:

The existing ETL mappings were inefficient and this reduced performance while the transformation operation was being performed. TechnoSoft had to reengineer the whole process to derive significantly better performance.

The existing code was not of the highest standards and TechnoSoft was faced with poor code manageability. This means that TechnoSoft' engineers had to carry out tasks to make the programs easier to maintain.

The application took too long to run and this was proving difficult for the client because of the huge amounts of data involved. The client wanted TechnoSoft to ensure that the speed of the operations was boosted by a remarkable factor.

TechnoSoft's Role

Since time was as critical as cost savings, TechnoSoft used its Global Delivery Model and its team of nine personnel completed the project in 16 months. The initial task concerned a detailed study of the existing ETL model to identify the bottlenecks. Once this was completed, TechnoSoft team determined areas of improvement. These changes were implemented in a short duration, resulting in an ETL that processed information at a faster pace.

Once the speed factor had been satisfactorily handled, TechnoSoft had to ensure future maintainability of the ETL. This was done by making the ETL concurrent and scalable, so that the client could confidently ramp up the storage and processing capabilities of the data warehouse at a later date, if required.

The final task before TechnoSoft was maintaining the data warehouse so that the client's key personnel could handle other tasks. This is an ongoing task and is being achieved to the full satisfaction of the client.

A recent engagement for a global corporation with operations in over a hundred nations involved IT innovation in making recommendations for a data mart strategy to complement an existing data warehouse infrastructure. TechnoSoft developed data warehouse architecture which will allow the information base to be extended and enhanced overtime.

Benefits

The new ETL delivered many benefits to the client including:

- The new batch processes ran significantly faster than before thanks to the tightly integrated code. The immense speed ensured a 70% time reduction in batch processes and enhanced the efficiency of the client's processing capabilities.
- Flexibility was also improved and the client could easily add newer data from sources that were not initially supported. This improved the capabilities of the data warehousing solution.
- As part of the improvement process, the client's data warehouse solution started generating more useful output and this helped the company to take key business decisions with high quality data.
- The continuous enhancements that are being carried out to the production environment by automating the load processes, adding reconciliation, and automated balancing processes, are helping the client to improve the satisfaction of its customers.

TechnoSoft Global Delivery Model (GDM) lowered costs drastically and helped the company to focus its IT budget savings on more important tasks.

END OF CASELET 1

Caselet 2

Read the caselet carefully and answer the following questions:

4. Assume that you are Certified Information Security Auditor (CISA) at VennShine, explain what Legal and Audit requirements are needed to be fulfilled by VennShine to ensure the accuracy of end results? **(7 marks)** <Answer>

5. If you are Security Manager at VennShine, explain how security can affect the design of data warehouse. **(7 marks)** <Answer>

6. Assume that you are a warehouse manager, apart from metadata management, what other responsibilities do you undertake at VennShine? Explain. **(10marks)** <Answer>

VennShine incorporated in early 1999, started its operations and concentrating on the formal process of data warehousing after being in the business of informal data warehousing since the department’s inception. As a unit, they possess an enormous amount of business and their operations were transforming, summarizing and reporting data from almost every enterprise data source. A next logical step in their evolution as warehouse manager has been to create metadata management. Requiring a robust, departmental, client/server physical environment, an open architecture and logical design for integration into enterprise solutions and browser based end-user interfaces, they chose the SAS System. As technology changes, industry competition tightens, and our clients become increasingly computer savvy, our department must rapidly seek new technology and methods to meet the needs of our wide and varied list of clients. Once having identified our customer base and business cases, we must identify the issues relating to the growth and performance of a data warehouse because any data warehouse solution will grow over time, sometimes quite dramatically. Transforming and presenting data as information is our primary function in the corporation. We are constantly looking for new and improved ways to accomplish this directive. Source data is crucial to data quality and mining efforts. As each new on-line transactional system and database platform is introduced the complexity of our tasks increases. Also, the more disparate the input data sources, the more complicated the integration.

Firstly, a clear definition of the business need is also required to ensure the accuracy of the end results. Defining a logical view of the data needed to supply the correct information, independent of source data restraints, is necessary. Here clients and analysts get the opportunity to discuss their business needs and solutions proactively. Next, it is important to establish early any legal and audit requirements that will be placed on the data warehouse. It is very difficult to add legal restrictions after the data warehouse has gone live, so it is important to capture the ultimate legal and audit requirements at the beginning. VennShine is responsible to ensure that data remains clean, consistent and integral. Any security restrictions can be seen as obstacles to the business and it becomes constraint on the design of data warehouse. Security can also affect the design of the data warehouse and can also affect any application code that has to be written and also affect development time-scales. By using formal data warehousing practices, tools and methodologies, state-of-the-art data extraction, transformation and summarization tools and thin client application deployment, we want to move beyond “data reporting” to “data mining”. To successfully engage data mining in our processes, the first step is to know who our customers are. We are able to list them by name, job title, function, and business unit, and communicate with them regularly. Next we must be able to identify the appropriate business requirements.

END OF CASELET 2

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. Explain the ten golden rules used for setting up a reliable data mining environment. **(10 marks)** <Answer>

8. Explain the different forms of neural networks. **(10 marks)** <Answer>

END OF SECTION C

END OF QUESTION PAPER

Suggested Answers
Data Warehousing and Data Mining (MB3G1IT): January 2009

Section A : Basic Concepts

Answer **Reason**

1. D Warehouse manager transform and manages the data. < TOP >

2. E Simple transformation task is performed by the management tool in a load manager. < TOP >

3. D MPP machines require the use of a distributed lock manager to maintain the integrity of the distributed resources across the system as a whole. Design and management of a data warehouse on an MPP system is considerably more difficult than on SMP system. The amount of CPU power that can be made available in an MPP system is dependent on the number of nodes that can be connected. < TOP >

4. D In database sizing, if n is the number of concurrent queries allowed and P is the size of the partition, then temporary space (T) is set to $T = (2n+1)P$. < TOP >

Therefore, T = [(2 * 4) + 1] 3

T = (8 + 1) 3 = 27 GB.

5. A OmnibackII backup software package is produced by HP. < TOP >

6. E Vertical partitioning can take two forms: normalization and row splitting, before using a vertical partitioning we need to be very sure that there will be no requirements to perform major join operations between the two partitions. Horizontal hardware partitioning technique spreads the processing load by horizontally partitioning the fact table into small segments and then physically storing each segment in a different node. < TOP >

7. C Technical blueprint phase must deliver an overall architecture that satisfies the long-term requirements and a definition of the components that must be implemented in the short term in order to derive any business benefit. Build the vision is the stage where the first production deliverable is produced. The purpose of the business case is to identify the projected business benefits that should be derived from using the data warehouse. < TOP >

8. A Detailed design of database is not produced at technical blueprint stage. < TOP >

9. C Redundant Array of Inexpensive Disks (RAID) Level 3 uses byte wise striping of data with parity disk. < TOP >

10. B Query management process is the system process that manages the queries and speeds them up by directing queries to the most effective data source. Unlike other system processes, query management does not generally operate during the regular load of information into the data warehouse. Query management process must ensure that no single query can affect the overall system performance. < TOP >

11. C In warehouse manager, backup and archive task is performed by system management tool. < TOP >

12. D Fact data represents a physical transaction that has occurred at a point in time and as such is unlikely to change on an ongoing basis during the life of the data warehouse. Dimension data in a star schema or snowflake schema is designed to minimize the cost of change and is typically very low volume data and the requirements will be constantly refined over the life of the data warehouse. Fact data will have multiple foreign keys, whereas reference data will have one primary key. < TOP >

13. E The reasons for partitioning the fact table are: Partitioning for performance, Partitioning for ease of management, Partitioning to assist backup/recovery. < TOP >

14. A As a part of extraction and load process, metadata is used to map data sources to the common view of information within the data warehouse. As a part of the warehouse management process, metadata is used to automate the production of summary tables. As a part of query management process, metadata is used to direct a query to the most appropriate data source. < TOP >

15. B Steps for fact table identification process are: Look for elemental transactions, Determine key dimensions, Check if fact is a dimension, Check if dimension is a fact. < TOP >

16. C In designing the fact tables: Using intelligent keys, each key represents the unique identifier for the item in the real world, Using non-intelligent keys, each unique key is generated automatically and refers to the unique identifier for the item in the real world, Usage of intelligent key can be costly and time consuming, Unless it is certain that identifiers will not change, it is safer to use non-intelligent keys. < TOP >

17. B A Symmetric Multi-Processing (SMP) machine is a set of tightly coupled CPUs, that share memory and disk. A Massively Parallel Processing (MPP) machine is a set of loosely coupled CPUs, each of which has its own memory and disk. A Non Uniform Memory Architecture (NUMA) machine is basically a tightly coupled cluster of Symmetric Multi-Processing (SMP) nodes. < TOP >

18. D The statements true about OLAP are OLAP tools do not learn, OLAP creates no new knowledge, OLAP cannot search for new solutions and Data mining is more powerful than OLAP. < TOP >

19. B The process of eliminating redundant data by breaking each table into smaller tables is known as normalizing. < TOP >

20. E Networker backup software package is produced by Legato. < TOP >

21. B In an organization, the relationship between DEPARTMENT and EMPLOYEE is a one-to-many relationship. < TOP >

22. C Association rules, neural networks, genetic algorithms are the various data mining techniques. < TOP >

23. A select sname, totalmarks, snum from student where snum=‘CB14’;. < TOP >

24. B In Freud’s theory of psychodynamics, the human brain was described as a neural network. < TOP >

25. D The stages in Knowledge Discovery Process are: Data selection, Enrichment, Reporting. < TOP >

26. A Flattening: A coding operation in which an attribute with cardinality ‘n’ is replaced by ‘n’ binary attributes. < TOP >

27. A In general, data mining algorithms should not have a complexity higher than n(log n). < TOP >

28. B Shallow knowledge is the information that can be easily retrieved from databases using a query tool such as SQL. Multi-dimensional knowledge is the information that can be analyzed using online analytical processing tools. Hidden knowledge is the data that can be found easily by using pattern recognition or machine-learning algorithms. Deep knowledge is the information that is stored in the database but can only be located if we have a clue that tells us where to look. < TOP >

29. C Genetic algorithms can be viewed as a kind of meta-learning strategy. Any programmer can write the basic structure of this algorithm easily. Solutions found by these algorithms are coded symbolically and are therefore often easily readable as compared to neural networks. < TOP >

30. E A perceptron consists of a simple three-layered network with input units called photo-receptors, intermediate units called associators and output units called responders. A back propagation network not only has input and output nodes, but also a set of intermediate layers with hidden nodes. A Kohenen self-organizing map is a collection of neurons or units, each of which is connected to a small number of other units called its neighbors. < TOP >

Data Warehousing and Data Mining (MB3G1IT): January 2009

Section B : Caselets

1. Yes, ETL tools solve several important architectural problems for data warehouses. An important function of these tools is to generate and maintain centralized metadata. < TOP >

   - Provide coordinated access to multiple data sources. Functions supported by ETL tools include the extraction of data from multiple source environments, data cleansing, reorganization, transformation, aggregation, calculation, automatic loading of data into the target database, and automatic generation of executable code to perform parallel processing of transformations on multiple engines.
   - Used to generate and maintain a central metadata repository. The metadata repository provides a "single version of the truth" that can be used to define enterprise-wide source data definitions, data models for target databases, and transformation rules that convert source data into target data. A metadata exchange architecture is used to synchronize central business rules with local business rules, maintained as local metadata by end-user BI tools.
   - Address the dirty data problem—data from source files can be cleansed and inconsistencies in the data resolved as part of the extraction and transformation process, using procedural data cleansing techniques. Name and address correction, deduping, and householding functions require use of an external data cleansing tool.

2. Short for ***Extract, Transform, Load;*** three database functions that are combined into one tool that automates the process to pull data out of one database and place it into another database. < TOP >

   Extract -- the process of reading data from a specified source database and extracting a desired subset of data.

   Transform -- the process of converting the extracted/ acquired data from its previous form into the form it needs to be in so that it can be placed into another database. Transformation occurs by using rules or lookup tables or by combining with other data.

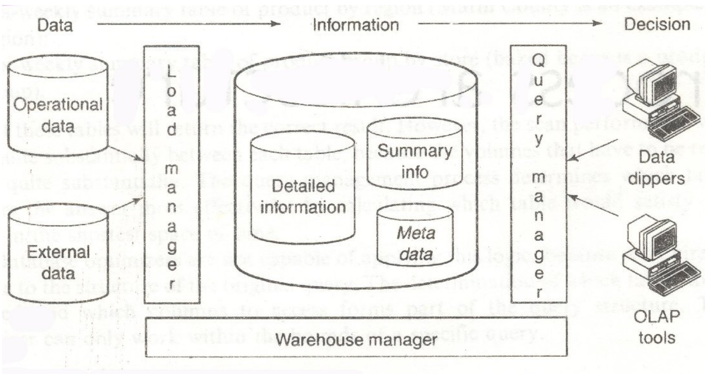
   Load -- the process of writing the data into the target database.

   The features of ETL tools are**:**

   - Standardize data to enable load to conformed target databases.
   - Filter data, convert codes, perform table lookups, and calculate derived values.
   - Incremental aggregation - computation of aggregates by the ETL tool in one pass of the source data.
   - Perform procedural data cleansing functions.
   - Produce audit and operational reports for each data load.
   - Automatic generation of centralized metadata.
   - Automatic generation of data extract programs.
   - Platform independence and scalability to enterprise data warehousing applications.
   - No requirement for intermediate disc files.
   - Support for concurrent processing of multiple source data streams, without writing procedural code.
   - Ability to specify complex transformations using only built-in transformation objects.
   - The goal is to specify transformations without writing any procedural code.
   - Support for change management functions.
   - Automatic generation of central metadata, including source data definitions, transformation objects, target data models, and operational statistics.
   - Central management of distributed ETL engines and metadata using a central console and a global metadata repository.
   - Strong data warehouse administration functions.
   - Support for the analysis of transformations that failed to be accepted by the ETL process.

3.

[< TOP >]

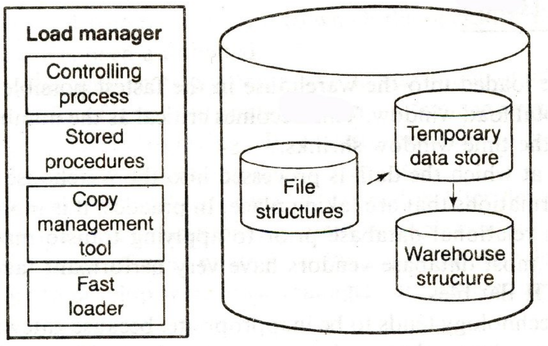


Architecture of a data warehouse

Load Manager Architecture

The architecture of a load manager is such that it performs the following operations:

1. Extract the data from the source system.
2. Fast-load the extracted data into a temporary data store.
3. Perform simple transformations into a structure similar to the one in the data warehouse.



Load manager architecture

Warehouse Manager Architecture

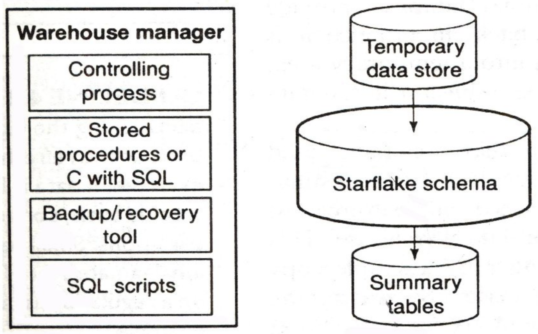
The architecture of a warehouse manager is such that it performs the following operations:

1. Analyze the data to perform consistency and referential integrity checks.

2 Transform and merge the source data in the temporary data store into the published data warehouse.

3. Create indexes, business views, partition views, business synonyms against the base data.
4. Generate denormalizations if appropriate.
5. Generate any new aggregations that may be required.
6. Update all existing aggregation.
7. Back up incrementally or totally the data within the data warehouse.
8. Archive data that has reached the end of its capture life.

In some cases, the warehouse manager also analyzes query profiles to determine which indexes and aggregations are appropriate.

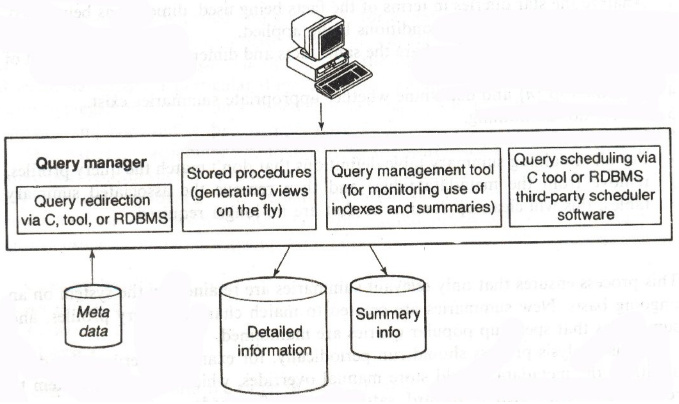


Architecture of a warehouse manager

Query Manager Architecture

The architecture of a query manager is such that it performs the following operations:

1. Direct queries to the appropriate table(s).
2. Schedule the execution of user queries.


Query manager
Query redirection via C, tool, or RDBMS
Stored procedures (generating views on the fly)
Query management tool (for monitoring use of indexes and summaries)
Query scheduling via C tool or RDBMS third-party scheduler software
Meta data
Detailed information
Summary info

Query manager architecture

4. The Legal and Audit requirements are needed to be fulfilled by VennShine to ensure the accuracy of end results are as under:

< TOP >

Legal Requirements

- It is vital to establish any legal requirements on the data being stored. If individual customer data is being held such as customer information for personal phones or accounts details, it may be required by law to enforce certain restrictions.
- Holding data only in summarized form does of course have the downside that detailed drill-down operations are impossible. If the data held online is going to be used for trend analysis and is therefore held in summarized rather than detailed form, legal requirements apply?
- If the customer data to be held on the data warehouse is of a minimal nature legal restrictions apply? So it may be that the data is sufficient to identify an account but not the owner, thereby preserving anonymity and this would be sufficient to allow trending and profiling to be done blindly.
- If the results were to be used to put together a mail shot of a new product to be identified account holders, the account ID numbers could be transferred to an operational machine where the data could be joined to customer data to cross-match account ID to customer details. It is not just individual customer data that can be subject to restrictions.

Audit Requirements

- Auditing is a specific subset of security that is often mandated by organizations.
- Given the volumes of data involved in a data warehouse, auditing can cause extremely heavy overheads on the system. To make up these overheads and to allow operations such as the overnight processing to still finish in time, generally requires more hardware. Thus, where possible or allowed, auditing should be switched off.
- There may be many different audit requirements, but as far as the data warehouse is concerned all audit requirements can basically be categorized into the following list:
  - Connections
  - Disconnections
  - Data access
  - Data change

For each of these categories it may be necessary to audit success, failure or both. For security reasons, the auditing of failures can be particularly important, because they can highlight any attempted unauthorized or fraudulent purchases.

If data is to be audited, it is required to establish whether each access is to be audited separately or whether it is sufficient to audit the fact that a user accessed a specific table during a given session. This can radically affect the information that needs to be held, saving both space and I/O overheads. If changes to the data are being audited, it is sufficient to audit the fact that a change occurred or is it required to capture the actual change that was made? Data warehouses are largely read-only environments except for the data load and therefore data is unlikely to change once it is loaded. If the audit requirement on data change exists purely to prevent data being changed, it may be sufficient to protect the data by making it read-only and then auditing any change from read-only to read-write.

It is imperative to restrict auditable events to a minimum. If necessary, steps may need to be taken to design the audit requirements away. This is one of the reasons why it is so important to understand not just the audit requirements but the reason for them. It is amazing how often audit requirements exist because they are actually required on every system. These sorts of issues can usually be pushed back once the costs of implementation are demonstrated.

5. Security can affect the design of data warehouse and it involves Application Development, Database Design and Testing: < TOP >

- **Application development**: The extra security codes which I propose to VennShine that are needed for each of the process managers are as under:
    - Load manager
    - Warehouse manager
    - Query manager
    - The **load manager** may require checking code to filter records and place them in different locations depending on their contents. It may also need extra transformation rules to modify sensitive fields to hide certain data. Extra metadata may need to be created and maintained to handle any extra objects.
    - The **warehouse manager** may need code to create and maintain all the extra views and aggregations required to enforce the security. Extra checks may need to be coded into the warehouse manager to prevent it from being fooled into moving data into a location where it should not be available.
    - The **query manager** will certainly require changes to handle any access restrictions. It will need to be aware of all the extra views and aggregations and will have to be able to point users at the correct versions of objects.
- **Database design:** If added security increases the number of database objects such as tables and views, this will affect the database layout. If enforcing the security adds to the size of the database, it automatically adds to the complexity of the database management and the backup recovery plan.
- **Testing:** The testing of the design of a data warehouse is a complex and lengthy process. When further security is added to the task it increases the test time in two ways. First, the added complexity of the programs to be tested will increase the time required for integration and system testing. The added complexity will most likely increase the number of errors found in testing and the amount of recording that needs to be performed. Second, there is added functionality to be tested: this will increase the size of the test suite. It means that the testing will take longer to plan and develop which in turn will increase the amount of time taken for testing.

6. As a warehouse manager at VennShine, the responsibilities of warehouse manager are: < TOP >

- Data movement
- Metadata management
- Performance monitoring and tuning
- Data archiving

Data Movement

The warehouse manager is responsible for any data movement within the warehouse, such as aggregation creation and maintenance. Any tables, indexes or other objects required will also be created and maintained by the warehouse manager.

Aggregation creation and maintenance is a process that should be automated. The warehouse manager needs an interface that allows new aggregations to be requested, and old aggregations to be removed. This mechanism can be controlled via tables in the database. Database procedures can be used to insert a row into a table for each new aggregation to be created. The warehouse manager will also need to know what fields are required in the aggregation and what fields to aggregate on. From this information it can build the aggregation and the metadata to describe it.

Most aggregations can be created by a single query. However, the query is often complex and not necessarily the most efficient method of generating the aggregation. It may be possible to generate the aggregation from other aggregations, or by a multiple-stage process with interim result tables. These interim results can often be used as the basis of multiple aggregations, thereby saving time and resource overall. The warehouse manager needs to be capable of taking advantage of these optimizations.

There is no warehouse manager software currently on the market that can do this automatically, which means that human input is required. One suggestion to work around this problem is to allow a program or procedure to be associated with each aggregation. If one exists, the warehouse manager can use it instead of directly generating the aggregation from a single query.

To make the best use of system resources, the warehouse manager may need to use parallelism for any given operation. It may also need to run multiple operations side by side. To ensure that the system is neither underutilized nor swamped with work, this process can be driven via the queuing mechanisms of the schedule manager. To achieve this aggregation may need to be prioritized. They may also need to be sequenced if one aggregation is to be built from another. This means that the warehouse manager needs to integrate with the queuing mechanism of the schedule manager.

The warehouse manager is also responsible for creating and maintaining the indexes on the fact and dimension data and on the aggregations. The indexes on the fact data will need to be created as Soon as the data load completes. Likewise, the indexes on the aggregations will need to be created as soon as a new aggregation is built.

If data marts are being used, the warehouse manager will be responsible for maintaining them. It will schedule any refreshes of the data marts, and may also be responsible for pushing the data out to the different data mart machines.

Performance Monitoring and Tuning

The warehouse manager is responsible for monitoring the performance of any operation it runs. It should also manage the storage of the performance statistics produced by the query manager. These statistics, along with the query history of all queries, should be stored in the database, so that they can be analyzed.

These statistics need to be kept for as long as possible. For each query the statistics could be averaged over time to reduce storage space. However, it is important to keep a time trail for these statistics, so that changing trends can be identified. Therefore, the data should not be aggregated over too large a time period, or it may mask trends. Older data can be aggregated over a greater time period if the averages do not change, but recent data should be averaged over a day or at most a week.

The warehouse manager is responsible for the creation of aggregations and indexes: therefore, any tuning that requires either an aggregation or an index will need to be passed on to the warehouse manager. Ideally the warehouse manager itself should be capable of identifying the need for new aggregations, and automatically creating them. There is no such predictive software on the market today, but software is beginning to appear that can perform that function at a basic level, and it is inevitable that more sophisticated software will follow.

Data Archiving

As data ages, it will need to be purged to clear room for more current data. There is often a desire to hold data online for long periods of time, but usually the practicalities of working with such huge volumes of data dictate a reasonable cutoff point. Most companies have a natural cycle of data that will indicate how long data remains immediately useful.

These times will often conflict with business rules or legal requirements for how long data should actually be maintained. This generally means that data will need to be archived off the data warehouse, either to near line storage, or to tape. Which method should be used will depend on how likely old data is to be retrieved. It is worth mentioning here that data warehouses rarely generate new fact data; they normally load fact data from other sources. This means that the source system will already have the data archived. This is often a sufficient argument to avoid archiving on the data warehouse.

If archiving is required it needs to be factored into the capacity planning and the overnight window design from the beginning. This is important even though archiving may not be required for some years. If it is not designed in from the beginning it may become a serious problem when it is required to run for the first time. It will probably increase the overheads of the overnight process and may even delay the start of the process by having to clean off space to allow new data to be loaded.

If designed properly, the load process will not depend on the archiving process being run first. Sufficient space should be made available to allow the archiving to occur later than immediately required. This allows the archiving to be performed at a more convenient time.

When designing the archiving process there are a number of details that need to be established:

- data life expectancy
- in raw form
- in aggregated form
- data archiving
- start date
- cycle
- workload.

The **data life expectancy** is the amount of time that data is expected to remain in the database. You need to establish the requirement for how long fact data needs to be held online at the fact level. Older fact data can often be held online in aggregated form, reducing the amount of space it requires. This is a useful trick, but it has a cost. Rolling up the data requires processing power, and can require more space in the short term. This needs to be designed into the warehouse manager's processing schedule.

For archiving there are three figures that need to be established. First, you need to establish the start date: that is, when archiving is required to begin. For example, if the aim to keep 3 years' worth of data online, and the data warehouse is to start of with 18 months of historic data loaded, you will need to begin archiving 18 months after the go-live date.

The second figure that needs to be decided is the archive cycle or frequency. How often is data archived? Ellen if data is loaded daily, you may not want to run archiving on a daily cycle. It is often better to use a longer cycle, such as a week, a month, or even a quarter. This will increase the storage requirement to a cycle beyond the life expectancy, but it also makes the archiving more flexible, and gives you room to spread the load over more convenient times.

Finally, you need to estimate the load that archiving will place on the system. This is important to know accurately, as it tells you how much leeway you have in scheduling the archiving of a set of data. It also tells you when the process can be run. For example, if the archiving is to be on a monthly cycle, and archiving a month's worth of data takes several hours, it is not likely that you can do the archiving during a normal overnight run. You may need to put it off to a weekend and use the daily scheduling time to get the job done.

Section C: Applied Theory

7. The ten golden rules for setting of a reliable data mining environment are:

< TOP >

1. **Support extremely large data sets**

   Data mining deals with extremely large data sets consisting in some cases of billions of records, and without proper platforms to store and handle these volumes of data, no reliable data mining is possible. Parallel servers with databases optimized for DSS-oriented queries are useful. Fast and flexible access to large data sets is of vital importance; so too is the possibility of storing intermediate results.

2. **Support hybrid learning**

   Learning tasks can be divided into three areas:

   - Classification tasks
   - Knowledge engineering tasks
   - Problem-solving tasks

   Not all algorithms do equally well in all these areas: in complex data mining projects, such as fraud detection or customer profiling, various pattern recognition and learning algorithms (neural networks, genetic algorithms, statistical techniques, association rules, rule discovery, and so on) are needed.

3. **Establish a data warehouse**

   A data warehouse contains historic data and is subject oriented and static, that is, users do not update the data but it is created on a regular time-frame on the basis of the operational data of an organization. It is thus obvious that a data warehouse is an ideal starting point for a data mining process, since data mining depends heavily on the permanent availability of historic data, and in this sense a data warehouse could be regarded as indispensable.

4. **Introduce data cleaning facilities**

   Even when a data warehouse is in operation, the data is certain to contain all sorts of pollution, and as the data mining process continues more subtle forms of pollution will be identified. Special tools for cleaning data are necessary, and some advanced tools are available, especially in the field of de-duplication of client files. Other cleaning techniques are only just emerging from research laboratories.

5. **Facilitate working with dynamic coding**

   Creative coding is the heart of the KDD process. The environment should enable the user to experiment with different coding schemes. Store partial results, make attributes discrete, create time series out of historic data, select random sub-samples, separate test sets, and so on. A project management environment that keeps track of the genealogy of different samples and tables as well as of the semantics and transformations of the different attributes is vital.

6. **Integrate with DSS**

   Data mining looks for hidden data that cannot easily be found using normal query techniques. Nevertheless, a KDD process always starts with traditional DSS activities, and from there you zoom in on interesting parts of the data set.

7. **Choose extendible architecture**

   New techniques for pattern recognition and machine learning are under development and we also see numerous new developments in the database area. It is advisable to choose an architecture that enables you to integrate new tools at later stages. Object-oriented technology, such as CORBA, typically facilitates this kind of flexibility.

8. **Support heterogeneous databases**
   Not all the necessary data is necessarily to be found in the data warehouse. Sometimes you will need to enrich the data warehouse with information from unexpected sources, such as information brokers, or with operational data that is not stored in your regular data warehouse. In order to facilitate this, the data mining environment must support a variety of interfaces: hierarchical databases, flat files, various relational databases, and object-oriented database systems.

9. **Introduce client/server architecture**
   A data mining environment needs extensive reporting facilities. Some developments, such as data landscapes, point in the direction of highly interactive graphic environments but database servers are not very suitable for this task. Discovery jobs need to be processed by large data mining servers, while further refinement and reporting will take place on a client. Separating the data mining activities on the servers from the clients is vital for good performance. Client/server is a much more flexible system which moves the burden of visualization and graphical techniques from your server to the local machine. You can then optimize your database server completely for data mining. Adequate parallelization of data mining algorithms on large servers is of vital importance in this respect.

10. **Introduce cache optimization**
    Learning and pattern recognition algorithms that operate on database often need very special and frequent access to the data. Usually it is either impossible or impractical to store the data in separate tables or to cache large portions in internal memory. The learning algorithms in a data mining environment should be optimized for this type of database access (storage of intermediate results, low-level access to an underlying database platform, creation of hash tables, and so on). A low-level integration with the database environment is desirable.

8. There are several different forms of neural networks. Only three of them here:

< TOP >

- Perceptrons
- Back propagation networks
- Kohonen self-organizing map

In 1958 Frank Rosenblatt of the Cornell Aeronautical Laboratory built the so-called perceptron, one of the first implementations of what would later be known as a neural network. A **perceptron** consists of a simple three-layered network with input units called photo-receptors, intermediate units called associators, and output units called responders. The perceptron could learn simple categories and thus could be used to perform simple classification tasks. Later, in 1969, Minsky and Papert showed that the class of problem that could be solved by a machine with a perceptron architecture was very limited. It was only in the 1980s that researchers began to develop neural networks with a more sophisticated architecture that could overcome· these difficulties. A major improvement was the introduction of hidden layers in the so-called back propagation networks.

A **back propagation network** not only has input and output nodes, but also a set of intermediate layers with hidden nodes. In its initial stage a back propagation network has random weightings on its synapses. When we train the network, we expose it to a training set of input data. For each training instance, the actual output of the network is compared with the desired output that would give a correct answer; if there is a difference between the correct answer and the actual answer, the weightings of the individual nodes and synapses of the network are adjusted. This process is repeated until the responses are more or less accurate. Once the structure of the network stabilizes, the learning stage is over, and the network is now trained and ready to categorize unknown input. **Figure1** represents a simple architecture of a neural network that can perform an analysis on part of our marketing database. The age attribute has been split into three age classes, each represented by a separate input node; house and car ownership also have an input node. There are four additional nodes identifying the four areas, so that in this way each input node corresponds to a simple yes-no Decision. The same holds for the output nodes: each magazine has a node. It is clear that this coding corresponds well with the information stored in the database.

The input nodes are wholly interconnected to the hidden nodes, and the hidden nodes are wholly interconnected to the output nodes. In an untrained network the branches between the nodes have equal weights. During the training stage the network receives examples of input and output pairs corresponding to records in the database; and adapts the weights of the different branches until all the inputs match the appropriate outputs.

In **Figure 2** the network learns to recognize readers of the car magazine and comics. **Figure 3** shows the internal state of the network after training. The configuration of the internal nodes shows that there isa certain connection between the car magazine and comics readers. However, the networks do not provide a rule to identify this association.

Back propagation networks are a great improvement on the perceptron architecture. However, they also have disadvantages, one being that they need an extremely large training set. Another problem of neural networks is that although they learn, they do not provide us with a theory about what they have learned - they are simply black boxes-that give answers but provide no clear idea as to how they arrived at these answers.

In 1981 Tuevo Kohonen demonstrated a completely different version of neural networks that is currently known as Kohonen's self-organizing maps. These neural networks can be seen as the artificial counterparts of maps that exist in several places in the brain, such as visual maps, maps of the spatial possibilities of limbs, and so on. A Kohonen self-organizing map is a collection of neurons or units, each of which is connected to a small number of other units called its neighbors. Most of the time, the Kohonen map is two- dimensional; each node or unit contains a factor that is related to the space whose structure we are investigating. In its initial setting, the self-organizing map has a random assignment of vectors to each unit. During the training stage, these vectors are incre mentally adjusted to give a better coverage of the space. A natural way to visualize the process of training a self- organizing map is the so-called Kohonen movie, which is a series of frames showing the positions of the vectors and their connections with neighboring cells. The network resembles an elastic surface that is pulled out over the sample space. Neural networks perform well on classification tasks and can be very useful in data mining.

Figure 1

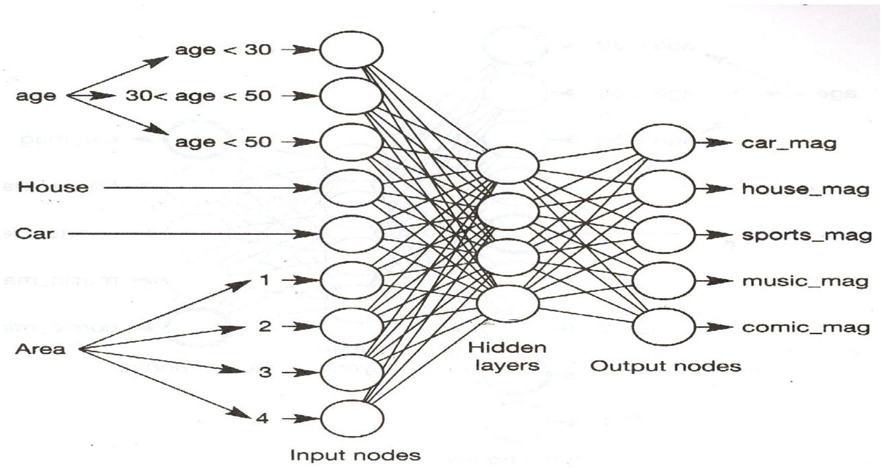


Figure2

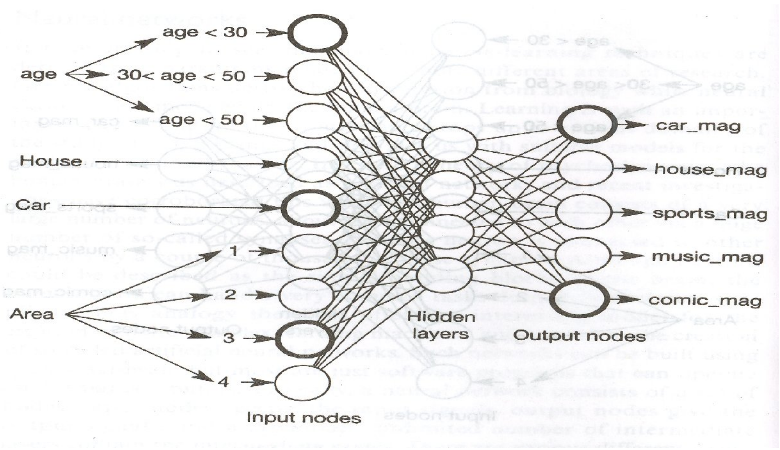
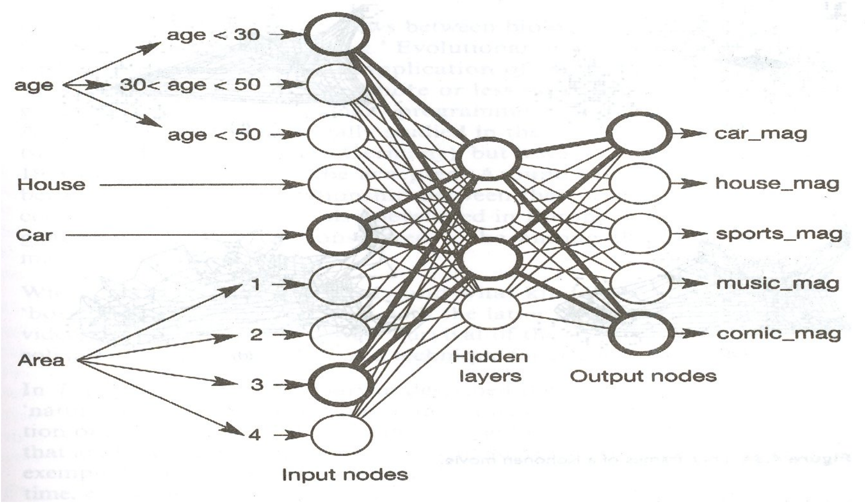


Figure 3



< TOP OF THE DOCUMENT >