CMR INSTITUTE OF TECHNOLOGY		USN							
Improvement Test									
Sub:	Data warehousing & Data mining							Code:	10IS74
Date:	16 / 11 / 2016	Duration:	90 mins	Max Marks:	50	Sem:	7	Branch:	ISE
Answer Any FIVE FULL Questions									

		Marks	OBE	
			CO	RBT
1(a)	Explain the data ware house design? What is data ware house? Explain the Architecture of data ware house .	[10]	CO1	L1
2(a)	Write about OLAP –FASMI in detail and OLAP software	[10]	CO2	L1
3(a)	Generate some strong rules using Apriori algorithm as an example.	[10]	CO4	L3
4(a)	What is data cube? How you will construct a data cube. List out the OLAP operations involved in Data cube process.	[10]	CO2	L1
5(a)	Discuss in detail about the Decision Tree with hunt algorithm?	[10]	CO5	L3
6(a)	Write in detail about Bayesian classifier with an example? Explain the bayes Method using Binary data attributes as an example.	[10]	CO3	L1
7(a)	What is web mining and text mining? Explain in detail about text clustering and unstructured text.	[10]	CO6	L1

Course Outcomes		PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
CO1:	Discuss the role of data warehousing and Data ware house Architecture.	1	-	2	-	-	-	-	-	-	1	-	-
CO2:	Describe the importance of Data cube design with OLAP operations.	1	2	1	-	-	-	-	-	-	1	-	-
CO3:	Identify the scope and necessity of Data Mining and their data sets, data attributes.	1	1	-	-	-	-	-	-	-	1	-	-
CO4:	Apply data mining technique of association analysis and cluster technique to design models for solving real world problems	2	3	1	-	-	-	-	-	-	1	-	-
CO5:	Apply classification techniques to solve real world problems.	2	3	-	-	-	-	-	-	-	-	-	-
CO6:	Describe the significance of Web mining, Text mining and temporal aspects of Data mining.	1	-	-	-	-	-	-	-	-	-	-	-

Cognitive level	KEYWORDS
L1	List, define, tell, describe, identify, show, label, collect, examine, tabulate, quote, name, who, when, where, etc.
L2	summarize, describe, interpret, contrast, predict, associate, distinguish, estimate, differentiate, discuss, extend
L3	Apply, demonstrate, calculate, complete, illustrate, show, solve, examine, modify, relate, change, classify, experiment, discover.
L4	Analyze, separate, order, explain, connect, classify, arrange, divide, compare, select, explain, infer.
L5	Assess, decide, rank, grade, test, measure, recommend, convince, select, judge, explain, discriminate, support, conclude, compare, summarize.

PO1 - *Engineering knowledge*; PO2 - *Problem analysis*; PO3 - *Design/development of solutions*; PO4 - *Conduct investigations of complex problems*; PO5 - *Modern tool usage*; PO6 - *The Engineer and society*; PO7- *Environment and sustainability*; PO8 – *Ethics*; PO9 - *Individual and team work*; PO10 - *Communication*; PO11 - *Project management and finance*; PO12 - *Life-long learning*

IAT3
SCHEME AND SOLUTION
DATAWAREHOUSING AND DATA MINING

1. **Explain the data ware house design? What is data ware house? Explain the Architecture of data ware house** .

Business Analysis Framework

The business analyst get the information from the data warehouses to measure the performance and make critical adjustments in order to win over other business holders in the market. Having a data warehouse offers the following advantages:

- Since a data warehouse can gather information quickly and efficiently, it can enhance business productivity.

- A data warehouse provides us a consistent view of customers and items, hence, it helps us manage customer relationship.

- A data warehouse also helps in bringing down the costs by tracking trends, patterns over a long period in a consistent and reliable manner.

To design an effective and efficient data warehouse, we need to understand and analyze the business needs and construct a business analysis framework. Each person has different views regarding the design of a data warehouse. These views are as follows:

- The top-down view - This view allows the selection of relevant information needed for a data warehouse.

- The data source view - This view presents the information being captured, stored, and managed by the operational system.

- The data warehouse view - This view includes the fact tables and dimension tables. It represents the information stored inside the data warehouse.

- The business query view - It is the view of the data from the viewpoint of the end-user.

Three-Tier Data Warehouse Architecture

Generally a data warehouses adopts a three-tier architecture. Following are the three tiers of the data warehouse architecture.

- Bottom Tier - The bottom tier of the architecture is the data warehouse database server. It is the relational database system. We use the back end tools and utilities to feed data into the bottom tier. These back end tools and utilities perform the Extract, Clean, Load, and refresh functions.

- Middle Tier - In the middle tier, we have the OLAP Server that can be implemented in either of the following ways.

  - By Relational OLAP (ROLAP), which is an extended relational database management system. The ROLAP maps the operations on multidimensional data to standard relational operations.

  - By Multidimensional OLAP (MOLAP) model, which directly implements the multidimensional data and operations.

- Top-Tier - This tier is the front-end client layer. This layer holds the query tools and reporting tools, analysis tools and data mining tools.

The following diagram depicts the three-tier architecture of data warehouse:

Data Warehouse Models

From the perspective of data warehouse architecture, we have the following data warehouse models:

- Virtual Warehouse
- Data mart
- Enterprise Warehouse

Virtual Warehouse

The view over an operational data warehouse is known as a virtual warehouse. It is easy to build a virtual warehouse. Building a virtual warehouse requires excess capacity on operational database servers.

Data Mart

Data mart contains a subset of organization-wide data. This subset of data is valuable to specific groups of an organization.

In other words, we can claim that data marts contain data specific to a particular group. For example, the marketing data mart may contain data related to items, customers, and sales. Data marts are confined to subjects.

Points to remember about data marts:

- Window-based or Unix/Linux-based servers are used to implement data marts. They are implemented on low-cost servers.

- The implementation data mart cycles is measured in short periods of time, i.e., in weeks rather than months or years.

- The life cycle of a data mart may be complex in long run, if its planning and design are not organization-wide.

- Data marts are small in size.

- Data marts are customized by department.

- The source of a data mart is departmentally structured data warehouse.

- Data mart are flexible.

Enterprise Warehouse

- An enterprise warehouse collects all the information and the subjects spanning an entire organization

- It provides us enterprise-wide data integration.

- The data is integrated from operational systems and external information providers.

- This information can vary from a few gigabytes to hundreds of gigabytes, terabytes or beyond.

Load Manager

This component performs the operations required to extract and load process.

The size and complexity of the load manager varies between specific solutions from one data warehouse to other.

Load Manager Architecture

The load manager performs the following functions:

- Extract the data from source system.
- Fast Load the extracted data into temporary data store.
- Perform simple transformations into structure similar to the one in the data warehouse.

Extract Data from Source

The data is extracted from the operational databases or the external information providers. Gateways is the application programs that are used to extract data. It is supported by underlying DBMS and allows client program to generate SQL to be executed at a server. Open Database Connection(ODBC), Java Database Connection (JDBC), are examples of gateway.

Fast Load

- In order to minimize the total load window the data need to be loaded into the warehouse in the fastest possible time.

- The transformations affects the speed of data processing.

- It is more effective to load the data into relational database prior to applying transformations and checks.

- Gateway technology proves to be not suitable, since they tend not be performant when large data volumes are involved.

Simple Transformations

While loading it may be required to perform simple transformations. After this has been completed we are in position to do the complex checks. Suppose we are loading the EPOS sales transaction we need to perform the following checks:

- Strip out all the columns that are not required within the warehouse.
- Convert all the values to required data types.

Warehouse Manager

A warehouse manager is responsible for the warehouse management process. It consists of third-party system software, C programs, and shell scripts.

The size and complexity of warehouse managers varies between specific solutions.

Warehouse Manager Architecture

A warehouse manager includes the following:

- The controlling process
- Stored procedures or C with SQL
- Backup/Recovery tool
- SQL Scripts

Operations Performed by Warehouse Manager

- A warehouse manager analyzes the data to perform consistency and referential integrity checks.

- Creates indexes, business views, partition views against the base data.

- Generates new aggregations and updates existing aggregations. Generates normalizations.

- Transforms and merges the source data into the published data warehouse.

- Backup the data in the data warehouse.

- Archives the data that has reached the end of its captured life.

Note: A warehouse Manager also analyzes query profiles to determine index and aggregations are appropriate.

Query Manager

- Query manager is responsible for directing the queries to the suitable tables.

- By directing the queries to appropriate tables, the speed of querying and response generation can be increased.

- Query manager is responsible for scheduling the execution of the queries posed by the user.

Query Manager Architecture

The following screenshot shows the architecture of a query manager. It includes the following:

- Query redirection via C tool or RDBMS
- Stored procedures
- Query management tool
- Query scheduling via C tool or RDBMS
- Query scheduling via third-party software

Detailed Information

Detailed information is not kept online, rather it is aggregated to the next level of detail and then archived to tape. The detailed information part of data warehouse keeps the detailed information in the starflake schema. Detailed information is loaded into the data warehouse to supplement the aggregated data.

The following diagram shows a pictorial impression of where detailed information is stored and how it is used.

Note: If detailed information is held offline to minimize disk storage, we should make sure that the data has been extracted, cleaned up, and transformed into starflake schema before it is archived.

2. **Write about OLAP –FASMI in detail and OLAP software**

The FASMI test summarizes the OLAP definition in just five key words F ast A nalysis of S hared M ultidimensional I nformation • it was first used in early 1995 and has now been widely adopted and is cited in over 120 Web sites in about 30 countries.

F –FAST

A- A nalysis

S- S hared

M- M ultidimensional

I -Information

3. **Generate some strong rules using Apriori algorithm as an example.**

Apriori Itemset Generation

- A frequent itemset is an itemset whose support is greater than some user-specified minimum support (denoted L_k, where k is the size of the itemset)
- A candidate itemset is a potentially frequent itemset (denoted C_k, where k is the size of the itemset)

Apriori Algorithm

A Java applet which combines DIC, Apriori and Probability Based Objected Interestingness Measures can be found here.

Apriori Algorithm: (by Agrawal et al at IBM Almaden Research Centre) can be used to generate all frequent itemset

```
Pass 1
    1. Generate the candidate itemsets in $C_1$
    2. Save the frequent itemsets in $L_1$
Pass k
    1. Generate the candidate itemsets in $C_k$ from the frequent
       itemsets in $L_{k-1}$
        1. Join $L_{k-1}$ p with $L_{k-1}$q, as follows:
           insert into $C_k$
           select p.$item_1$, p.$item_2$, . . . , p.$item_{k-1}$, q.$item_{k-1}$
           from $L_{k-1}$ p, $L_{k-1}$q
           where p.$item_1$ = q.$item_1$, . . . p.$item_{k-2}$ = q.$item_{k-2}$, p.$item_{k-1}$ < q.$item_{k-1}$
        2. Generate all (k-1)-subsets from the candidate itemsets in $C_k$
        3. Prune all candidate itemsets from $C_k$ where some (k-1)-subset of the candidate
           itemset is not in the frequent itemset $L_{k-1}$
    2. Scan the transaction database to determine the support for each candidate itemset in $C_k$
    3. Save the frequent itemsets in $L_k$
```

Implementation: A working Apriori Itemset Generation program can be found on the Itemset Implementation page.

Example 1: Assume the user-specified minimum support is 50%

- Given: The transaction database shown below

TID	A	B	C	D	E	F
T_1	1	0	1	1	0	0
T_2	0	1	0	1	0	0
T_3	1	1	1	0	1	0
T_4	0	1	0	1	0	1

- The candidate itemsets in C_2 are shown below

Itemset X	supp(X)
{A,B}	25%
{A,C}	50%
{A,D}	25%
{B,C}	25%
{B,D}	50%
{C,D}	25%

- The frequent itemsets in L_2 are shown below

Itemset X	supp(X)
{A,C}	50%
{B,D}	50%

Example 2: Assume the user-specified minimum support is 40%, then generate all frequent itemsets.

Given: The transaction database shown below

TID	A	B	C	D	E
T_1	1	1	1	0	0
T_2	1	1	1	1	1
T_3	1	0	1	1	0
T_4	1	0	1	1	1
T_5	1	1	1	1	0

Pass 1

C_1

Itemset X	supp(X)
A	?
B	?
C	?
D	?
E	?

L_1

Itemset X	supp(X)
A	100%
B	60%
C	100%
D	80%
E	40%

Pass 2

C_2

Itemset X	supp(X)
A,B	?