FIRST TERM EXAMINATION [SEPT. 2016]
SEVENTH SEMESTER [B.TECH.]
DATA MINING AND BUSINESS INTELLIGENCE
[ETCS–413]

1.30 hrs. **M.M. : 30**

Note: *Attempt any three questions including Q.No 1 which is compulsory.*

Q.1.(a) Define strategic information. Explain its characterstics .

Ans. The types of information needed to make decisions in the formulation and execution of business strategies and objectives are broad-based and encompass the entire organization. All these types of essential information are grouped into one and is called as strategic information. For making decisions about these objectives, executives and managers need this information for the following purposes: to get in-depth knowledge of their company's operations; learn about the key business factors and how these affect one another; monitor how the business factors change over time; and compare their company's performance relative to the competition and to industry benchmarks.

The characteristics of strategic information are:

(*a*) Integrated: Must have a single, enterprise-wide view.

(*b*) Data Integrity: Information must be accurate and must conform to business rules.

(*c*) Accessible: Easily accessible with intuitive access paths, and responsive for analysis.

(*d*) Credible: Every business factor must have one and only one value.

(*e*) Timely: Information must be available within the stipulated time frame.

Q.1. (*b*) Differentiate between operational and informational systems.

Ans.

Operational Systems	Informational Systems
They store only the current values	Archived and summarised values are also stored
They have large number of users	The number of users is relatively small
The access frequency is high	The access frequency is medium to low
The response is fast	It takes minutes to respond
These systems are optimized for transactions only	These systems are used for complex queries
Read, write and update operations can be performed	Only read operation can be performed

Q.1. (c) What is meant by the non-volatility of the data in the data warehouse?

Ans. Every business transaction does not update the data in the data warehouse. The business transactions update the operational system databases in real time. We add, change, or delete data from an operational system as each transaction happens but do not usually update the data in the data warehouse. You do not delete the data in the data warehouse in real time. Once the data is captured in the data warehouse, you do not run individual transactions to change the data there. Data updates are commonplace in an operational database; not so in a data warehouse. The data in a data warehouse is not as volatile as the data in an operational database is.

Q.1. (*d*) Explain the four major sources of data in the data warehouse.

Ans. Source data coming into the data warehouse may be grouped into four broad categories:

(a) Production Data: This category of data comes from the various operational systems of the enterprise. Based on the information requirements in the data warehouse you choose segments of data from the different operational systems.

(b) Internal Data: In every organization, users keep their "private" spreadsheets documents, customer profiles, and sometimes even departmental databases. This is the internal data, parts of which could be useful in a data warehouse.

(c) Archived Data: Operational systems are primarily intended to run the current business. In every operational system, you periodically take the old data and store it in archived files. The circumstances in your organization dictate how often and which portions of the operational databases are archived for storage. Some data is archived after a year. Sometimes data is left in the operational system databases for as long as five years.

(d) External Data: Most executives depend on data from external sources for a high percentage of the information they use. They use statistics relating to their industry produced by external agencies. They use market share data of competitors. They use standard values of financial indicators for their business to check on their performance.

Q.1. (e) Data warehouse is an environment, not a product. Explain.

Ans. A data warehouse is not a single software or hardware product you purchase to provide strategic information. It is, rather, a computing environment where users can find strategic information, an environment where users are put directly in touch with the data they need to make better decisions. It is a user-centric environment.

Q.2.(a) Describe any two requirement gathering methods for building a data warehouse.

Ans. The two major requirement gathering methods are as follows:

(a) Interviews: The interview sessions can use up a good percentage of the project time. Therefore, these will have to be organized and managed well. Before your project team launches the interview process, make sure the following major tasks are completed.

Select and train the project team members conducting the interviews

- Assign specific roles for each team member (lead interviewer/scribe)
- Prepare list of users to be interviewed and prepare broad schedule
- List your expectations from each set of interviews

Most of the users you will be interviewing fall into three broad categories: senior executives, departmental managers/analysts, IT department professionals.

Pre-interview research is important for the success of the interviews. Here is a list of some key research topics:

- History and current structure of the business unit
- Number of employees and their roles and responsibilities
- Locations of the users
- Primary purpose of the business unit in the enterprise
- Relationship of the business unit to the strategic initiatives of the enterprise

(b). JAD: In this method, you are able to get a number of interested users to meet together in group sessions. On the whole, this method could result in fewer group sessions than individual interview sessions. The overall time for requirements gathering may prove to be less and therefore shorten the project. Also, group sessions may be more effective if the users are dispersed in remote locations.

JAD consists of a five-phased approach:

- **Project Definition:** Complete high-level interviews, Conduct management interviews and Prepare management definition guide
- **Research:** Become familiar with the business area and systems, Document user information requirements, Document business processes and Gather preliminary information

•Preparation: Create working document from previous phase, Train the scribes, Prepare visual aids, Conduct presession meetings and Set up a venue for the sessions

• JAD Session: Open with review of agenda and purpose, Review assumptions, Review data requirements and Review business metrics and dimensions

• Final Document: Convert the working document, Map the gathered information, List all data sources and Identify all business metrics

Q.2. (b) Design an information package for an automaker sales company.

Ans.

Information Subject: Automaker sales

Dimensions

Time	Product	Payment Method	Customer Demo-graphics	Dealer	
Year	Model Name	Finance Type	Age	Dealer Name	
Quarter	Model Year	Term (Months)	Gender	City	
Month	Package Styling	Interest Rate	Income Range	State	
Date	Product Line	Agent	Marital Status	Single Brand Flag	
Day of Week	Product Category		House-hold Size	Date First Operation	
Day of Month	Exterior Color		Vehicles Owned		
Season	Interior Color		Home Value		
Holiday Flag	First Year		Own or Rent		
Facts: Actual Sale Price, MSRP Sale Price, Options Price, Full Price, Dealer Add-ons, Dealer Credits, Dealer Invoice, Down Payment, Proceeds, Finance					

Hierarchies/ Categories

Q.3. (a) What is Metadata ? Why it is important to a data warehouse ? Also, explain different types of metadata.

Ans. Metadata in a data warehouse is similar to the data dictionary or the data catalog in a database management system. In the data dictionary, you keep the information about the logical data structures, the information about the files and addresses, the information about the indexes, and so on. The data dictionary contains data about the data in the database. Similarly, the metadata component is the data about the data in the data warehouse.

Significance of metadata is:

• First, it acts as the glue that connects all parts of the data warehouse.

• Next, it provides information about the contents and structures to the developers.

• Finally, it opens the door to the end-users and makes the contents recognizable in their own terms.

Metadata in a data warehouse fall into three major categories:

(a) Operational Metadata: Data for the data warehouse comes from several operational systems of the enterprise. These source systems contain different data structures. The data elements selected for the data warehouse have various field lengths

and data types. In selecting data from the source systems for the data warehouse, you split records, combine parts of records from different source files, and deal with multiple coding schemes and field lengths. When you deliver information to the end-users, you must be able to tie that back to the original source data sets. Operational metadata contain all of this information about the operational data sources.

(b) Extraction and Transformation Metadata: Extraction and transformation metadata contain data about the extraction of data from the source systems, namely, the extraction frequencies, extraction methods, and business rules for the data extraction. Also, this category of metadata contains information about all the data transformations that take place in the data staging area.

(c) End-User Metadata: The end-user metadata is the navigational map of the data warehouse. It enables the end-users to find information from the data warehouse. The end-user metadata allows the end-users to use their own business terminology and look for information in those ways in which they normally think of the business.

Q.3. (b) Differentiate between Data warehouse and data mart .

Ans.

Data Warehouse	Data Mart
It is at the enterprise level	This is at the organisational level
Union of all data marts	It is a single business process
Structure to suit the corporate view of datadata	Structure to suit the departmental view of
Queries on presentation source	Technology optimal for data access and analysis

Q.4. Write short notes on the following :

Q.4. (a) ETL processes

Ans. After you have extracted data from various operational systems and from external sources, you have to prepare the data for storing in the data warehouse. The extracted data coming from several disparate sources needs to be changed, converted, and made ready in a format that is suitable to be stored for querying and analysis.

Three major functions need to be performed for getting the data ready. You have to extract the data, transform the data, and then load the data into the data warehouse storage. These three major functions of extraction, transformation, and preparation for loading take place in a staging area.

(a) *Data Extraction:* This function has to deal with numerous data sources. You have to employ the appropriate technique for each data source. Source data may be from different source machines in diverse data formats. Part of the source data may be in relational database systems. Some data may be on other legacy network and hierarchical data models. Many data sources may still be in flat files. You may want to include data from spreadsheets and local departmental data sets. Data extraction may become quite complex.

(b) *Data Transformation:* In every system implementation, data conversion is an important function. You perform a number of individual tasks as part of data transformation. First, you clean the data extracted from each source. Cleaning may just be correction of misspellings, or may include resolution of conflicts between state codes and zip codes in the source data, or may deal with providing default values for missing data elements, or elimination of duplicates when you bring in the same data from multiple source systems.

Standardization of data elements forms a large part of data transformation. You standardize the data types and field lengths for same data elements retrieved from the various sources. Semantic standardization is another major task. You resolve synonyms and homonyms. When two or more terms from different source systems mean the same thing, you resolve the synonyms. When a single term means many different things in different source systems, you resolve the homonym.

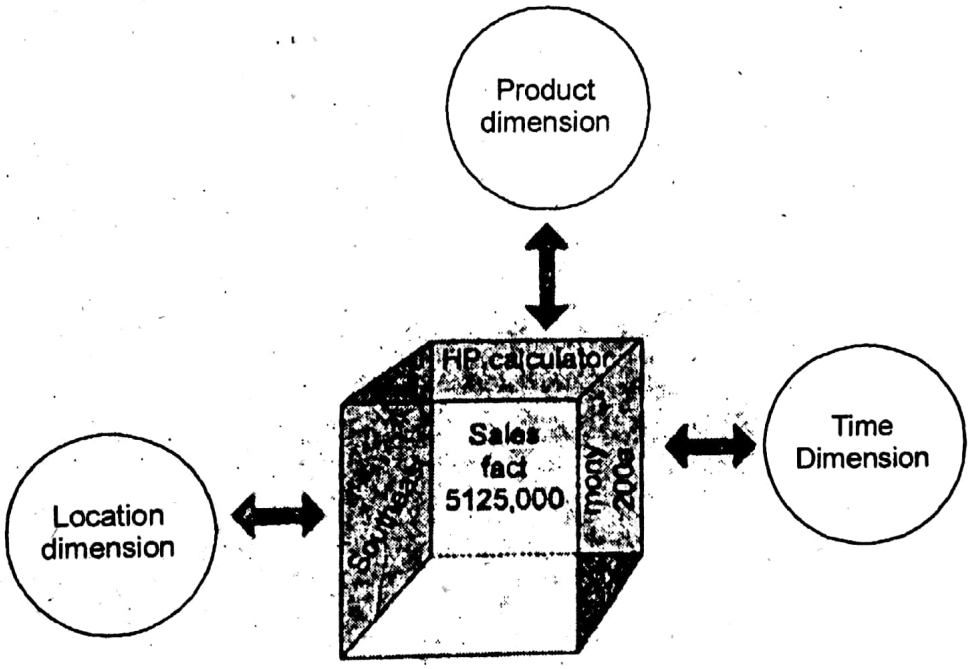
Data transformation involves many forms of combining pieces of data from the different sources. You combine data from single source record or related data elements from many source records. On the other hand, data transformation also involves purging source data that is not useful and separating out source records into new combinations. Sorting and merging of data takes place on a large scale in the data staging area.

(c). Data Loading: Two distinct groups of tasks form the data loading function. When you complete the design and construction of the data warehouse and go live for the first time, you do the initial loading of the data into the data warehouse storage. The initial load moves large volumes of data using up substantial amounts of time. As the data warehouse starts functioning, you continue to extract the changes to the source data, transform the data revisions, and feed the incremental data revisions on an ongoing basis.

Q.4. (b) STAR schema.

Ans. Star Schema and its components:

The star schema is a data-modeling technique used to map multidimensional decision support data into a relational database. In effect,, the star schema creates the near equivalent of a multidimensional database schema from the existing relational database. The star schema was developed because existing relational modeling techniques, ER, and normalization did not yield a database structure that served advanced data analysis requirements well.



The basic star schema has four components: facts, dimensions, attributes, and Attribute hierarchies.

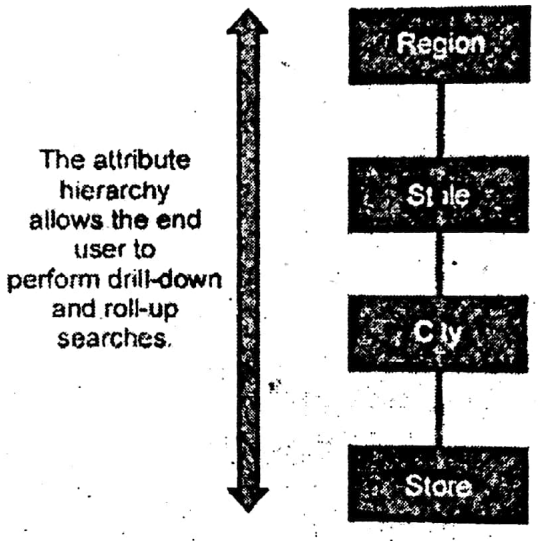
I. Facts: Facts are numeric measurements (values) that represent a specific business aspect or activity. For example, sales figures are numeric measurements that represent product and/or service sales. Facts commonly used in business data analysis are units, costs, prices, and revenues. Facts are normally stored in a fact table that is the center of the star schema. The fact table contains facts that are linked through their dimensions.

Facts can also be computed or derived at run time. Such computed or derived facts are sometimes called metrics to differentiate them from stored facts. The fact table is updated periodically (daily, weekly, monthly, and so on) with data from operational databases.

II. Dimensions: Dimensions are qualifying characteristics that provide additional perspectives to a given fact. Recall that dimensions are of interest because decision support data are almost always viewed in relation to other data. For instance, sales might be compared by product from region to region and from one time period to the next. Dimensions are normally stored in dimension tables. The following diagram depicts a star schema for sales with product, location, and time dimensions.

III. Attributes: Each dimension table contains attributes. Attributes are often used to search, filter, or classify facts. Dimensions provide descriptive characteristics about the facts through their attributes. Therefore, the data warehouse designer must define common business attributes that will be used by the data analyst to narrow a search, group information, or describe dimensions. For example Region, state, city are dimensions of Location, Product type, product ID are dimensions of Product.

***W.* Attribute Hierarchies:** Attributes within dimensions can be ordered in a well-defined attribute hierarchy. The attribute hierarchy provides a top-down data organization that is used for two main purposes: aggregation and drill-down/roll-up data analysis. For example, the following fifrure show how the location dimension attributes can be organized in a hierarchy by region, state, city, and store.



The attribute hierarchy provides the capability to perform drill-down and roll-up searches in a data warehouse. For example, suppose a data analyst looks at the nswer to the query: How does the 2009 month-to-date sales performance compare to the 2010 month-to-date sdales performance? The data analyst spots a sharp sales decline for March 2010. the data analyst might decide to drill down inside the month of March to see how sales by regions compared to the previous year.