UNIT 3 INTRODUCTION TO DATA WAREHOUSING

Structure **Page Nos.**

3.0	Introduction	59
3.1	Objectives	59
3.2	What is Data Warehousing?	60
3.3	The Data Warehouse: Components and Processes	62
	3.3.1 Basic Components of a Data Warehouse	
	3.3.2 Data Extraction, Transformation and Loading (ETL)	
3.4	Multidimensional Data Modeling for Data Warehouse	67
3.5	Business Intelligence and Data Warehousing	70
	3.5.1 Decision Support System (DSS)	
	3.5.2 Online Analytical Processing (OLAP)	
3.6	Building of Data Warehouse	73
3.7	Data Marts	75
3.8	Data Warehouse and Views	76
3.9	The Future: Open Issues for Data Warehouse	77
3.10	Summary	77
3.11	Solutions/Answers	78

3.0 INTRODUCTION

Information Technology (IT) has a major influence on organisational performance and competitive standing. With the ever increasing processing power and availability of sophisticated analytical tools and techniques, it has built a strong foundation for the product - data warehouse. But, why should an organisation consider investing in a data warehouse? One of the prime reasons, for deploying a data warehouse is that, the data warehouse is a kingpin of business intelligence.

The data warehouses provide storage, functionality and responsiveness to queries, that is far superior to the capabilities of today's transaction-oriented databases. In many applications, users only need read-access to data, however, they need to access larger volume of data very rapidly – much more than what can be conveniently handled by traditional database systems. Often, such data is extracted from multiple operational databases. Since, most of these analyses performed do occur periodically, therefore, software developers and software vendors try to design systems to support these functions. Thus, there is a definite need for providing decision makers at middle management level and higher level with information as per the level of details to support decision-making. The data warehousing, online analytical processing (OLAP) and data mining technologies provide this functionality.

This unit covers the basic features of data warehousing and OLAP. Data Mining has been discussed in more details in unit 4 of this Block.

3.1 OBJECTIVES

After going through this unit, you should be able to:

- explain the term data warehouse;
- define key concepts surrounding data warehousing systems;

- compare database warehouse with operational information systems;
- discuss data warehousing architecture;
- identify the main stages in the life cycle of data warehousing, and
- discuss the concepts of OLAP, MOLAP, ROLAP.

3.2 WHAT IS DATA WAREHOUSING?

Let us first try to answer the question: What is a data warehouse? A simple answer could be: A data warehouse is a tool that manage of data *after and outside* of operational systems. Thus, they are not replacements for the operational systems but are major tools that acquires data from the operational system. Data warehousing technology has evolved in business applications for the process of strategic decision-making. Data warehouses may be sometimes considered as the key components of IT strategy and architecture of an organisation. We will give the more formal definition of data warehouse in the next paragraph.

A data warehouse as defined by **W.H. Inmon** is a *subject-oriented, integrated, nonvolatile, time-variant collection* of data that supports decision-making of the management. Data warehouses provide controlled access to data for complex analysis, knowledge discovery, and decision-making.

Figure 1 presents some uses of data warehousing in various industries

S.No.	Industry	Functional Areas of Use	Strategic Uses
1	Banking	Creating new schemes for loans and other banking products, helps in operations, identities information for marketing	Finding trends for customer service, service promotions, reduction of expenses.
2	Airline	Operations, marketing	Crew assignment, aircraft maintenance plans, fare determination, analysis of route profitability, frequent - flyer program design
3	Hospital	Operation optimisation	Reduction of operational expenses, scheduling of resources
4	Investment and Insurance	Insurance product development, marketing	Risk management, financial market analysis, customer tendencies analysis, portfolio management

Figure 1: Uses of Data Warehousing

A data warehouse offers the following advantages:

- It provides historical information that can be used in many different forms of comparative and competitive analysis.
- It enhances the quality of the data and tries to make it complete.
- It can help in supporting disaster recovery although not alone but with other back up resources.

One of the major advantages a data warehouse offers is that it allows a large collection of historical data of many operational databases, which may be heterogeneous in nature, that can be analysed through one data warehouse interface, thus, it can be said to be a ONE STOP portal of historical information of an organisation. It can also be used in determining many trends through the use of data mining techniques.

Remember a data warehouse does not create value of its own in an organisation. However, the value can be generated by the users of the data of the data warehouse. For example, an electric billing company, by analysing data of a data warehouse can predict frauds and can reduce the cost of such determinations. In fact, this technology has such great potential that any company possessing proper analysis tools can benefit from it. Thus, a data warehouse supports Business Intelligence (that is), the technology that includes business models with objectives such as reducing operating costs, increasing profitability by improving productivity, sales, services and decision-making. Some of the basic questions that may be asked from a software that supports business intelligence include:

- What would be the income, expenses and profit for a year?
- What would be the sales amount this month?
- Who are the vendors for a product that is to be procured?
- How much of each product is manufactured in each production unit?
- How much is to be manufactured?
- What percentage of the product is defective?
- Are customers satisfied with the quality? etc.

Data warehouse supports various business intelligence applications. Some of these may be - online analytical processing (**OLAP**), decision-support systems (DSS), data mining etc. We shall be discussing these terms in more detail in the later sections.

A data warehouse has many characteristics. Let us define them in this section and explain some of these features in more details in the later sections.

Characteristics of Data Warehouses

Data warehouses have the following important features:

1) **Multidimensional conceptual view:** A data warehouse contains data of many operational systems, thus, instead of a simple table it can be represented in multidimensional data form. We have discussed this concept in more detail in section 3.3.

2) **Unlimited dimensions and unrestricted cross-dimensional operations:** Since the data is available in multidimensional form, it requires a schema that is different from the relational schema. Two popular schemas for data warehouse are discussed in section 3.3.

3) **Dynamic sparse matrix handling:** This is a feature that is much needed as it contains huge amount of data.

4) **Client/server architecture:** This feature help a data warehouse to be accessed in a controlled environment by multiple users.

5) **Accessibility and transparency, intuitive data manipulation and consistent reporting performance:** This is one of the major features of the data warehouse. A Data warehouse contains, huge amounts of data, however, that should not be the reason for bad performance or bad user interfaces. Since the objectives of data warehouse are clear, therefore, it has to support the following

easy to use interfaces, strong data manipulation, support for applying and reporting of various analyses and user-friendly output.

3.3 THE DATA WAREHOUSE: COMPONENTS AND PROCESSES

A data warehouse is defined as *subject-oriented, integrated, nonvolatile, time-variant collection,* but how can we achieve such a collection? To answer this question, let us define the basic architecture that helps a data warehouse achieve the objectives as given/stated above. We shall also discuss the various processes that are performed by these components on the data.

3.3.1 The Basic Components of a Data Warehouse

A data warehouse basically consists of three components:

The Data Sources
The ETL and
The Schema of data of data warehouse including meta data.

Figure 2 defines the basic architecture of a data warehouse. The analytical reports are not a part of the data warehouse but are one of the major business application areas including OLAP and DSS.

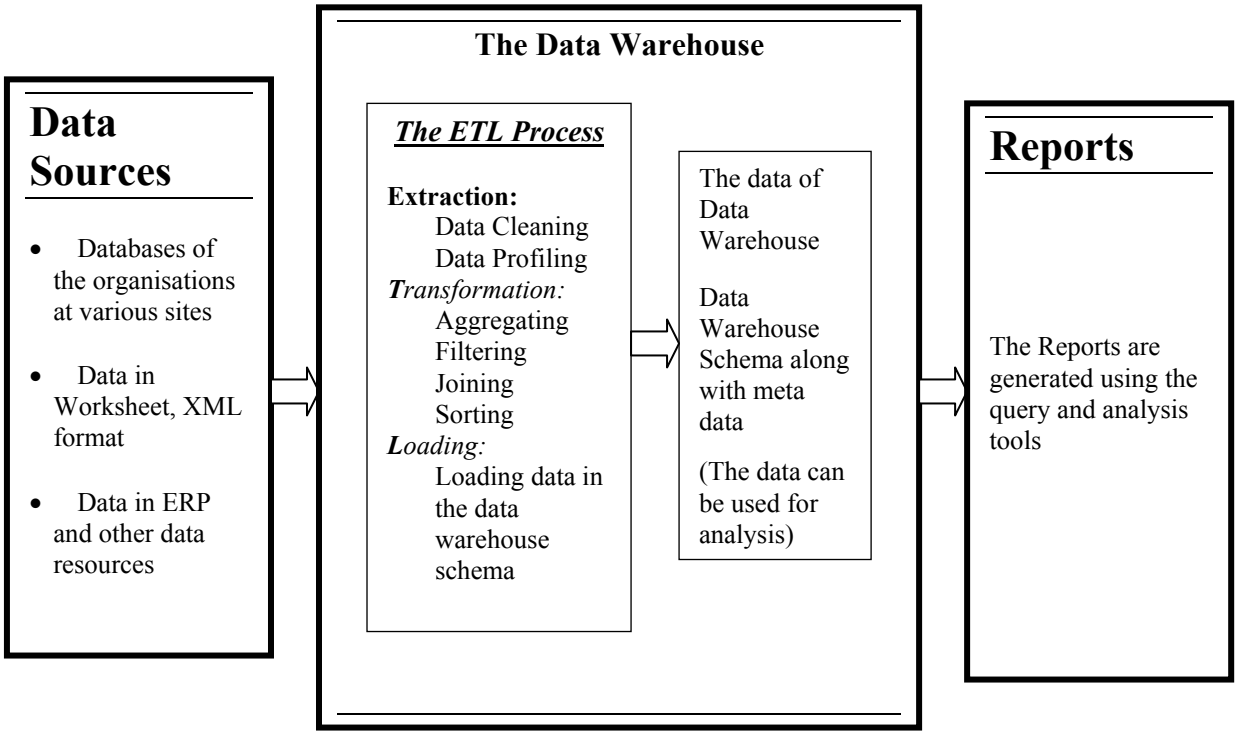

Figure 2: The Data Warehouse Architecture

The Data Sources

The data of the data warehouse can be obtained from many operational systems. A data warehouse interacts with the environment that provides most of the source data for the data warehouse. By the term environment, we mean, traditionally developed applications. In a large installation, hundreds or even thousands of these database systems or files based system exist with plenty of redundant data.

The warehouse database obtains most of its data from such different forms of legacy systems – files and databases. Data may also be sourced from external sources as well as other organisational systems, for example, an office system. This data needs to be integrated into the warehouse. But how do we integrate the data of these large numbers of operational systems to the data warehouse system? We need the help of ETL tools to do so. These tools capture the data that is required to be put in the data warehouse database. We shall discuss the ETL process in more detail in section 3.3.2.

Data of Data Warehouse

A data warehouse has an integrated, "subject-oriented", "time-variant" and "non-volatile" collection of data. The basic characteristics of the data of a data warehouse can be described in the following way:

i) Integration: Integration means bringing together data of multiple, dissimilar operational sources on the basis of an enterprise data model. The enterprise data model can be a basic template that identifies and defines the organisation's key data items uniquely. It also identifies the logical relationships between them ensuring organisation wide consistency in terms of:

> ***Data naming and definition:*** Standardising for example, on the naming of "student enrolment number" across systems.
>
> ***Encoding structures:*** Standardising on gender to be represented by "M" for male and "F" for female or that the first two digit of enrolment number would represent the year of admission.
>
> ***Measurement of variables:*** A Standard is adopted for data relating to some measurements, for example, all the units will be expressed in metric system or all monetary details will be given in Indian Rupees.

ii) Subject Orientation: The second characteristic of the data warehouse's data is that its design and structure can be oriented to important objects of the organisation. These objects such as STUDENT, PROGRAMME, REGIONAL CENTRES etc., are in contrast to its operational systems, which may be designed around applications and functions such as ADMISSION, EXAMINATION and RESULT DELCARATIONS (in the case of a University). Refer to *Figure 3*.

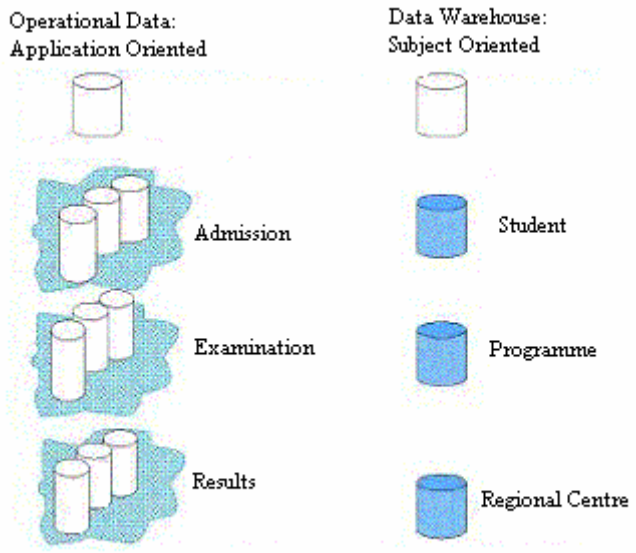

Figure 3: Operations system data orientation vs. Data warehouse data orientation

iii) Time-Variance: The third defining characteristic of the database of data warehouse is that it is time-variant, or historical in nature. The entire data in the data

warehouse is/was accurate at some point of time. This is, in contrast with operational data that changes over a shorter time period. The data warehouse's data contains data that is date-stamped, and which is historical data. *Figure 4* defines this characteristic of data warehouse.

OPERATIONAL DATA	DATA WAREHOUSE DATA
• It is the current value data	• Contains a snapshot of historical data
• Time span of data = 60-90 days	• Time span of data = 5-10 years or more
• Data can be updated in most cases	• After making a snapshot the data record cannot be updated
• May or may not have a timestamp	• Will always have a timestamp

Figure 4: Time variance characteristics of a data of data warehouse and operational data

iv) Non-volatility (static nature) of data: Data warehouse data is loaded on to the data warehouse database and is subsequently scanned and used, but is not updated in the same classical sense as operational system's data which is updated through the transaction processing cycles.

Decision Support and Analysis Tools

A data warehouse may support many OLAP and DSS tools. Such decision support applications would typically access the data warehouse database through a standard query language protocol; an example of such a language may be SQL. These applications may be of three categories: simple query and reporting, decision support systems and executive information systems. We will define them in more details in the later sections.

Meta Data Directory

The meta data directory component defines the repository of the information stored in the data warehouse. The meta data can be used by the general users as well as data administrators. It contains the following information:

i) the structure of the contents of the data warehouse database,
ii) the source of the data,
iii) the data transformation processing requirements, such that, data can be passed from the legacy systems into the data warehouse database,
iv) the process summarisation of data,
v) the data extraction history, and
vi) how the data needs to be extracted from the data warehouse.

Meta data has several roles to play and uses in the data warehouse system. For an end user, meta data directories also provide some additional information, such as what a particular data item would mean in business terms. It also identifies the information on reports, spreadsheets and queries related to the data of concern. All database management systems (DBMSs) have their own data dictionaries that serve a similar purpose. Information from the data dictionaries of the operational system forms a valuable source of information for the data warehouse's meta data directory.

3.3.2 Data Extraction, Transformation and Loading (ETL)

The first step in data warehousing is, to perform data extraction, transformation, and loading of data into the data warehouse. This is called ETL that is Extraction, Transformation, and Loading. ETL refers to the methods involved in accessing and manipulating data available in various sources and loading it into a target data warehouse. Initially the ETL was performed using SQL programs, however, now there are tools available for ETL processes. The manual ETL was complex as it required the creation of a complex code for extracting data from many sources. ETL tools are very powerful and offer many advantages over the manual ETL. ETL is a step-by-step process. As a first step, it maps the data structure of a source system to the structure in the target data warehousing system. In the second step, it cleans up the data using the process of data transformation and finally, it loads the data into the target system.

What happens during the ETL Process?

The ETL is three-stage process. During the *Extraction* phase the desired data is identified and extracted from many different sources. These sources may be different databases or non-databases. Sometimes when it is difficult to identify the desirable data then, more data than necessary is extracted. This is followed by the identification of the relevant data from the extracted data. The process of extraction sometimes, may involve some basic transformation. For example, if the data is being extracted from two Sales databases where the sales in one of the databases is in Dollars and in the other in Rupees, then, simple transformation would be required in the data. The size of the extracted data may vary from several hundreds of kilobytes to hundreds of gigabytes, depending on the data sources and business systems. Even the time frame for the extracted data may vary, that is, in some data warehouses, data extraction may take a few days or hours to a real time data update. For example, a situation where the volume of extracted data even in real time may be very high is a web server.

The *extraction* process involves data cleansing and data profiling. Data cleansing can be defined as the process of removal of inconsistencies among the data. For example, the state name may be written in many ways also they can be misspelt too. For example, the state Uttar Pradesh may be written as U.P., UP, Uttar Pradesh, Utter Pradesh etc. The cleansing process may try to correct the spellings as well as resolve such inconsistencies. But how does the cleansing process do that? One simple way may be, to create a Database of the States with some possible fuzzy matching algorithms that may map various variants into one state name. Thus, cleansing the data to a great extent. Data profiling involves creating the necessary data from the point of view of data warehouse application. Another concern here is to eliminate duplicate data. For example, an address list collected from different sources may be merged as well as purged to create an address profile with no duplicate data.

One of the most time-consuming tasks - data *transformation* and *loading* follows the extraction stage. This process includes the following:

- Use of data filters,
- Data validation against the existing data,
- Checking of data duplication, and
- Information aggregation.

Transformations are useful for transforming the source data according to the requirements of the data warehouse. The process of transformation should ensure the quality of the data that needs to be loaded into the target data warehouse. Some of the common transformations are:

Filter Transformation: Filter transformations are used to filter the rows in a mapping that do not meet specific conditions. For example, the list of employees of the Sales department who made sales above Rs.50,000/- may be filtered out.

Joiner Transformation: This transformation is used to join the data of one or more different tables that may be stored on two different locations and could belong to two different sources of data that may be relational or from any other sources like XML data.

Aggregator Transformation: Such transformations perform aggregate calculations on the extracted data. Some such calculations may be to find the sum or average.

Sorting Transformation: requires creating an order in the required data, based on the application requirements of the data warehouse.

Once the data of the data warehouse is properly extracted and transformed, it is *loaded* into a data warehouse. This process requires the creation and execution of programs that perform this task. One of the key concerns here is to propagate updates. Some times, this problem is equated to the problem of maintenance of the materialised views.

When should we perform the ETL process for data warehouse? ETL process should normally be performed during the night or at such times when the load on the operational systems is low. **Please note** that, the integrity of the extracted data can be ensured by synchronising the different operational applications feeding the data warehouse and the data of the data warehouse.

☞ Check Your Progress 1

1) What is a Data Warehouse?

……

……

……

2) What is ETL? What are the different transformations that are needed during the ETL process?

……

……

……

3) What are the important characteristics of Data Warehousing?

……

……

……

4) Name the component that comprise the data warehouse architecture?

……

……

……

3.4 MULTIDIMENSIONAL DATA MODELING FOR DATA WAREHOUSING

A data warehouse is a huge collection of data. Such data may involve grouping of data on multiple attributes. For example, the enrolment data of the students of a University may be represented using a student schema such as:

Student_enrolment (year, programme, region, number)

Here, some typical data value may be (These values are shown in *Figure 5* also. Although, in an actual situation almost all the values will be filled up):

- In the year 2002, BCA enrolment at Region (Regional Centre Code) RC-07 (Delhi) was 350.
- In year 2003 BCA enrolment at Region RC-07 was 500.
- In year 2002 MCA enrolment at all the regions was 8000.

Please note that, to define the student number here, we need to refer to three attributes: the year, programme and the region. Each of these attributes is identified as the dimension attributes. Thus, the data of student_enrolment table can be modeled using dimension attributes (year, programme, region) and a measure attribute (number). Such kind of data is referred to as a Multidimensional data. Thus, a data warehouse may use multidimensional matrices referred to as a data cubes model. The multidimensional data of a corporate data warehouse, for example, would have the fiscal period, product and branch dimensions. If the dimensions of the matrix are greater than three, then it is called a hypercube. Query performance in multidimensional matrices that lend themselves to dimensional formatting can be much better than that of the relational data model.

The following figure represents the Multidimensional data of a University:

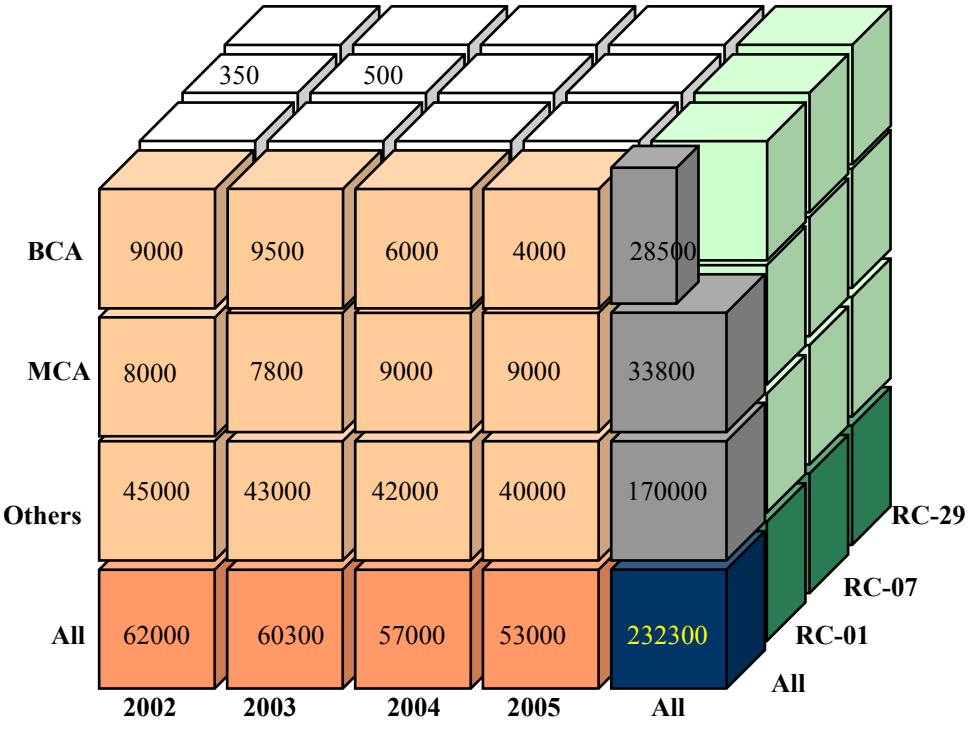

Figure 5: A sample multidimensional data

Multidimensional data may be a little difficult to analyse. Therefore, Multidimensional data may be displayed on a certain pivot, for example, consider the following table:

Region: ALL THE REGIONS				
	BCA	**MCA**	**Others**	**All the Programmes**
2002	9000	8000	45000	62000
2003	9500	7800	43000	60300
2004	6000	9000	42000	57000
2005	4000	9000	40000	53000
ALL the Years	**28500**	**33800**	**170000**	**232300**

The table given above, shows, the multidimensional data in *cross-tabulation*. This is also referred to as a *pivot-table*. **Please note that** cross-tabulation is done on any two dimensions keeping the other dimensions fixed as ALL. For example, the table above has two dimensions Year and Programme, the third dimension Region has a fixed value ALL for the given table.

Please note that, the cross-tabulation as we have shown in the table above is, different to a relation. The relational representation for the data of the table above may be:

Table: Relational form for the Cross table as above

Year	Programme	Region	Number
2002	BCA	All	9000
2002	MCA	All	8000
2002	Others	All	45000
2002	All	All	62000
2003	BCA	All	9500
2003	MCA	All	7800
2003	Others	All	43000
2003	All	All	60300
2004	BCA	All	6000
2004	MCA	All	9000
2004	Others	All	42000
2004	All	All	57000
2005	BCA	All	4000
2005	MCA	All	9000
2005	Others	All	40000
2005	All	All	53000
All	BCA	All	28500
All	MCA	All	33800
All	Others	All	170000
All	All	All	232300

A cross tabulation can be performed on any two dimensions. The operation of changing the dimensions in a cross tabulation is termed as pivoting. In case a cross tabulation is done for a value other than ALL for the fixed third dimension, then it is called *slicing*. For example, a slice can be created for Region code RC-07 instead of ALL the regions in the cross tabulation of regions. This operation is called *dicing* if values of multiple dimensions are fixed.

Multidimensional data allows data to be displayed at various level of granularity. An operation that converts data with a fine granularity to coarse granularity using aggregation is, termed *rollup* operation. For example, creating the cross tabulation for All regions is a rollup operation. On the other hand an operation that moves from a coarse granularity to fine granularity is known as *drill down* operation. For example, moving from the cross tabulation on All regions back to Multidimensional data is a drill down operation. **Please note:** For the drill down operation, we need, the original data or any finer granular data.

Now, the question is, how can multidimensional data be represented in a data warehouse? or, more formally, what is the schema for multidimensional data?

Two common multidimensional schemas are the star schema and the snowflake schema. Let us, describe these two schemas in more detail. A multidimensional storage model contains two types of tables: the dimension tables and the fact table. The dimension tables have tuples of dimension attributes, whereas the fact tables have one tuple each for a recorded fact. In order to relate a fact to a dimension, we may have to use pointers. Let us demonstrate this with the help of an example. Consider the University data warehouse where one of the data tables is the **Student enrolment table.** The three dimensions in such a case would be:

- Year
- Programme, and
- Region

The star schema for such a data is shown in *Figure* 6.

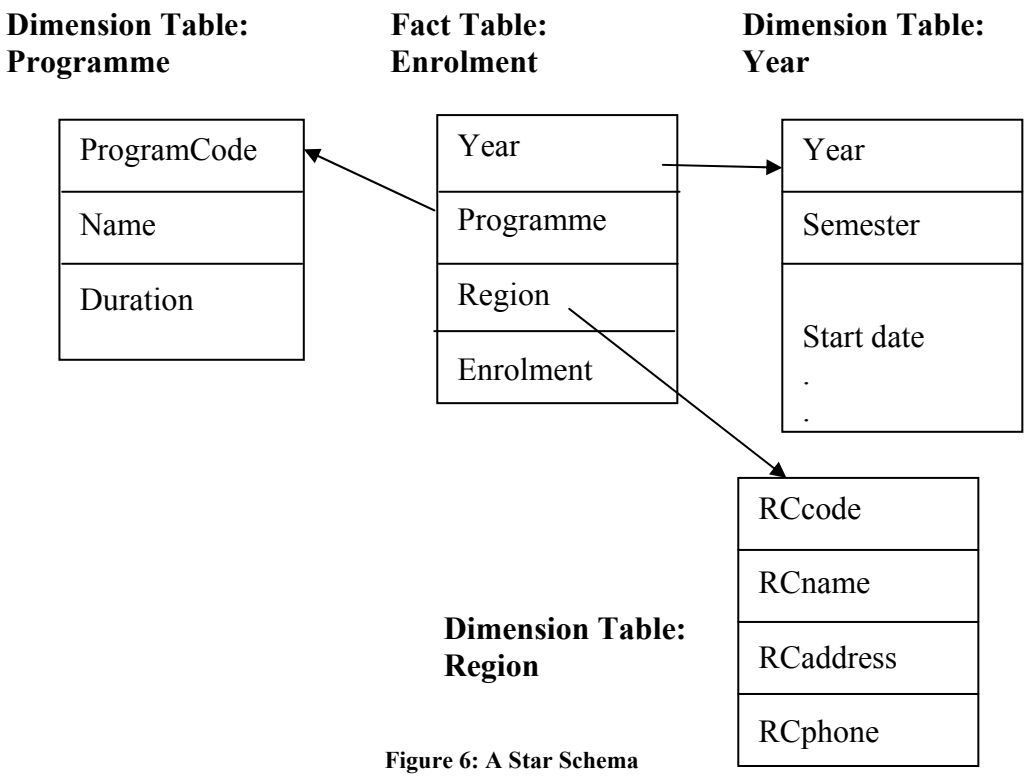

Figure 6: A Star Schema

Please note that in *Figure 6,* the fact table points to different dimension tables, thus, ensuring the reliability of the data. **Please notice that,** each Dimension table is a table for a single dimension only and that is why this schema is known as a star schema. However, a dimension table may not be normalised. Thus, a new schema named the snowflake schema was created. A snowflake schema has normalised but hierarchical dimensional tables. For example, consider the star schema shown in *Figure* 6, if in the Region dimension table, the value of the field Rcphone is multivalued, then the Region dimension table is not normalised.

Thus, we can create a snowflake schema for such a situation as:

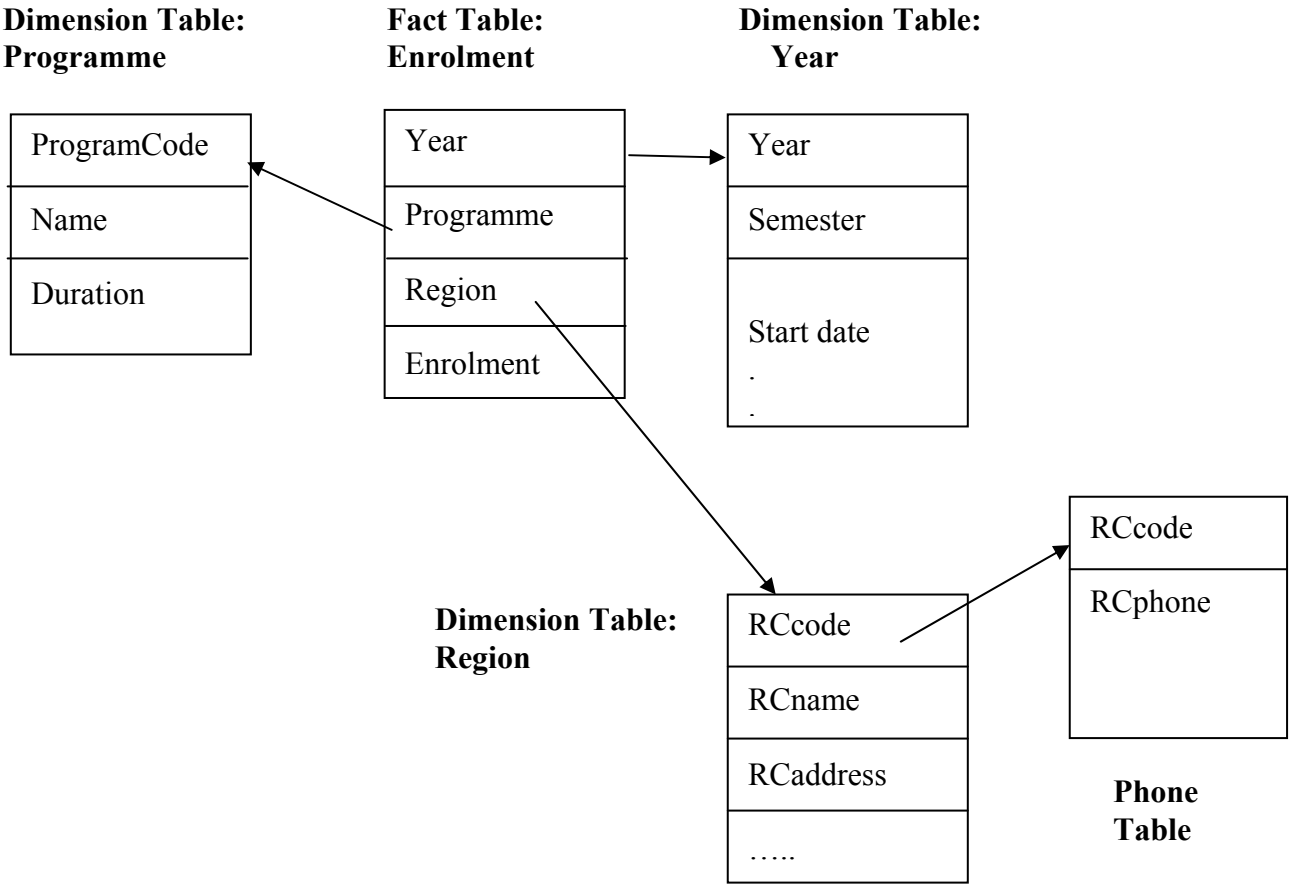

Figure 7: Snowflake Schema

Data warehouse storage can also utilise indexing to support high performance access. Dimensional data can be indexed in star schema to tuples in the fact table by using a join index. Data warehouse storage facilitates access to summary data due to the non-volatile nature of the data.

3.5 BUSINESS INTELLIGENCE AND DATA WAREHOUSING

A data warehouse is an integrated collection of data and can help the process of making better business decisions. Several tools and methods are available to that enhances advantage of the data of data warehouse to create information and knowledge that supports business decisions. Two such techniques are Decision-support systems and online analytical processing. Let us discuss these two in more details in this section.

3.5.1 Decision Support Systems (DSS)

The DSS is a decision support system and NOT a decision-making system. DSS is a specific class of computerised information systems that support the decision-making activities of an organisation. A properly designed DSS is an *interactive* software based system that helps decision makers to compile useful information from raw data, documents, personal knowledge, and/or business models to identify and solve problems and make decisions.

A decision support system may gather or present the following information:

- Accessing current information from data warehouse or legacy databases or other data resources.
- Presenting comparative sales figures of several months for an organisation.
- Creating projected revenue details based on new product sales assumptions.
- Demonstrating the consequences of different decision alternatives based on past experiences.

The DSS assists users in evaluating appropriate analysis or performing different types of studies on the datasets. For example, a spreadsheet can be used to store answers to a series of questionnaires in the form of Excel spreadsheets. This information then, can be passed on to decision makers. More specifically, the feedback data collected on a programme like CIC may be given to subject matter experts for making decisions on the quality, improvement, and revision of that programme. The DSS approach provides a self-assessment weighing tool to facilitate the determining of the value of different types of quality and quantity attributes. Decision support systems are sometimes also referred to as the Executive Information Systems (EIS).

Executive Information System (EIS): Executive information systems (EIS) are created for purpose of providing executives with the information they require to run their businesses. An EIS is intended to facilitate and support information and decision-making at senior executives level by, providing easy access to both *internal and external* information. Of course, this information should be relevant and should help them in establishing and meeting the strategic goals of the organisation.

The emphasis of DSS/EIS is mainly on graphical displays and easy-to-use user interfaces as they are there chiefly, to provide help. They offer strong reporting and drill-down capabilities. In general, EIS are enterprise-wide DSS that help top-level executives analyse, compare, and bring to light trends in important market/operational variables so that, they can monitor performance and identify opportunities and future problems. EIS and data warehousing technologies converge are convergent.

The concept of providing information to the executive management is not a new concept except for the ease with which they can get it. Given that top management has succeeded in acquiring the information till date, they can run their business without direct access to computer-based information systems. So why does one need a DSS/EIS? Well, there are a number of factors in support of DSS/EIS. These seem to be more managerial in nature. Some of these factors are:

- The first factor is a strange but true ‘pull’ factor, that is, executives are suggested to be more computer-literate and willing to become direct users of computer systems. For example, a survey suggests that more than twenty percent of senior executives have computers on their desks but rarely 5% use the system, although there are wide variations in the estimates yet, there is a define pull towards this simple easy to use technology.
- The other factor may be the increased use of computers at the executive level. For example, it has been suggested that middle managers who have been directly using computers in their daily work are being promoted to the executive level.

This new breed of executives do not exhibit the fear of computer technology that has characterised executive management up to now and are quite willing to be direct users of computer technology.

- The last factor is more on the side of technology. Technology is gradually becoming extremely simple to use from the end users point of view and it is now finding more users attracted towards it.

3.5.2 Online Analytical Processing (OLAP)

Data warehouses are not suitably designed for transaction processing, however, they support increased efficiency in query processing. Therefore, a data warehouse is a very useful support for the analysis of data. But are there any such tools that can utilise the data warehouse to extract useful analytical information?

On Line Analytical Processing (OLAP) is an approach for performing analytical queries and statistical analysis of multidimensional data. OLAP tools can be put in the category of business intelligence tools along with data mining. Some of the typical applications of OLAP may include reporting of sales projections, judging the performance of a business, budgeting and forecasting etc.

OLAP tools require multidimensional data and distributed query-processing capabilities. Thus, OLAP has data warehouse as its major source of information and query processing. But how do OLAP tools work?

In an OLAP system a data analyst would like to see different cross tabulations by interactively selecting the required attributes. Thus, the queries in an OLAP are expected to be executed extremely quickly. The basic data model that may be supported by OLAP is the star schema, whereas, the OLAP tool may be compatible to a data warehouse.

Let us, try to give an example on how OLAP is more suitable to a data warehouse rather than to a relational database. An OLAP creates an aggregation of information, for example, the sales figures of a sales person can be grouped (aggregated) for a product and a period. This data can also be grouped for sales projection of the sales person over the regions (North, South) or states or cities. Thus, producing enormous amount of aggregated data. If we use a relational database, we would be generating such data many times. However, this data has many dimensions so it is an ideal candidate for representation through a data warehouse. The OLAP tool thus, can be used directly on the data of the data warehouse to answer many analytical queries in a short time span. The term OLAP is sometimes confused with OLTP. OLTP is online transaction processing. OLTP systems focus on highly concurrent transactions and better commit protocols that support high rate of update transactions. On the other hand, OLAP focuses on good query-evaluation and query-optimisation algorithms.

OLAP Implementation

This classical form of OLAP implementation uses multidimensional arrays in the memory to store multidimensional data. Such implementation of OLAP is also referred to as Multidimensional OLAP (MOLAP). MOLAP is faster as it stores data in an already processed aggregated data form using dimension and fact tables. The other important type of OLAP implementation is Relational OLAP (ROLAP), which stores data directly in the relational databases. ROLAP creates multidimensional views upon request rather than in advance as in MOLAP. ROLAP may be used on complex data with a wide number of fields.

3.6 BUILDING OF DATA WAREHOUSE

The first basic issue for building a data warehouse is to identify the USE of the data warehouse. It should include information on the expected outcomes of the design. A good data warehouse must support meaningful query facility on the attributes of dimensional and fact tables. A data warehouse design in addition to the design of the schema of the database has to address the following three issues:

- How will the data be acquired?
- How will the data be stored?
- What would be the environment of the data warehouse?

Some of the key concerns of the issues above are:

Data Acquisition: A data warehouse must acquire data so that it can fulfil the required objectives. Some of the key issues for data acquisition are:

- Whether the data is to be extracted from multiple, heterogeneous sources? The location of these sources and the kind of data they contain?
- The method of acquiring the data contained in the various sources in a standard data warehouse schema. Remember, you must have consistent data in a data warehouse.
- How will the data be cleaned so that its validity can be ensured?
- How is the data going to be transformed and converted into the data warehouse multidimensional schema model?
- How will the data be loaded in the warehouse. After all, the data is huge and the amount of time the loading will take needs to be ascertained? Here, we need to find the time required for data cleaning, formatting, transmitting, creating additional indexes etc. and also the issues related to data consistency such as, the currency of data, data integrity in multidimensional space, etc.

Data storage: The data acquired by the data is also to be stored as per the storage schema. This data should be easily accessible and should fulfil the query needs of the users efficiently. Thus, designers need to ensure that there are appropriate indexes or paths that allow suitable data access. Data storage must be updated as more data is acquired by the data warehouse, but it should still provide access to data during this time. Data storage also needs to address the issue of refreshing a part of the data of the data warehouse and purging data from the data warehouse.

Environment of the data warehouse: Data warehouse designers should also keep in mind the data warehouse environment considerations. The designers must find the expected use of the data and predict if it is consistent with the schema design. Another key issue here would be the design of meta data directory component of the data warehouse. The design should be such that it should be maintainable under the environmental changes.

DATA WAREHOUSING LIFE CYCLE

The data warehouse technologies use very diverse vocabulary. Although the vocabulary of data warehouse may vary for different organisations, the data warehousing industry is in agreement with the fact that the data warehouse lifecycle model fundamentally can be defined as the model consisting of five major phases – design, prototype, deploy, operation and enhancement.

Let us introduce these terms:

1) **Design**: The design of database is to be done for available data inventories, DSS analyst requirements and analytical needs. It needs to produce a robust star schema or snowflake schema. Key activities in the design phase may include

communication with the end users, finding the available catalogues, defining key performance and quality indicators, mapping of decision-making processes as per the information needs at various end user levels, logical and physical schema design etc.

2) **Prototype**: A data warehouse is a high cost project, thus, it may be a good idea to deploy it partially for a select group of decision-makers and database practitioners in the end user communities. This will help in developing a system that will be easy to accept and will be mostly as per the user's requirements.

3) **Deploy**: Once the prototype is approved, then the data warehouse can be put to actual use. A deployed data warehouse comes with the following processes; documentation, training, maintenance.

4) **Operation**: Once deployed the data warehouse is to be used for day-to-day operations. The operation in data warehouse includes extracting data, putting it in database and output of information by DSS.

5) **Enhancement**: These are needed with the updating of technology, operating processes, schema improvements etc. to accommodate the change.

Please note you can apply any software life cycle model on the warehouse life cycle.

Data Warehouse Implementation

After the design of the data warehouse, the next step for building the data warehouse may be its implementation. Please remember that implementing a data warehouse is a very challenging process. It tests the ability of an organisation to adjust to change. The implementation of the data warehouse may require the following stages:

Implementation of Data Model: The data model that is to be implemented should be checked to ensure that it has the key entities and their interrelationships. It also should see that the system records of the data warehouse must be as per the data warehouse data model and should be possible best matches for the operational system data. The physical design should support the schema design.

Implementation of Data Transformation Programs: Now, the transformation programs that will extract and transform the data from the system of record should be implemented. They should be tested to load the required data in the data warehouse database.

Populating Data Warehouse and Maintenance: Once the data transformation programs are found to be ok, they can be used to populate the data warehouse. Once the data warehouse is operation at it needs to be maintained properly.

Some general Issues for Warehouse Design and Implementation

The programs created during the previous phase are executed to populate the data warehouse's database.

The Development and Implementation Team: A core team for such implementation may be:

A Project Leader responsible for managing the overall project and the one who helps in obtaining resources and participates in the design sessions.
Analysts documents the end user requirements and creates the enterprise data models for the data warehouse.

A Data Base Administrator is responsible for the physical data base creation, and

Programmers responsible for programming the data extraction and transformation programs and end user access applications.

Training: Training will be required not only for end users, once the data warehouse is in place, but also for various team members during the development stages of the data warehouse.

☞ Check Your Progress 2

1) What is a dimension, how is it different from a fact table?

……

……

……

2) How is snowflake schema different from other schemes?

……

……

……

3) What are the key concerns when building a data warehouse?

……

……

……

4) What are the major issues related to data warehouse implementation?

……

……

……

5) Define the terms: DSS and ESS.

……

……

……

6) What are OLAP, MOLAP and ROLAP?

……

……

……

3.7 DATA MARTS

Data marts can be considered as the database or collection of databases that are designed to help managers in making strategic decisions about business and the organisation. Data marts are usually smaller than data warehouse as they focus on some subject or a department of an organisation (a data warehouses combines databases across an entire enterprise). Some data marts are also called dependent data marts and may be the subsets of larger data warehouses.

A data mart is like a data warehouse and contains operational data that helps in making strategic decisions in an organisation. The only difference between the two is

that data marts are created for a certain limited predefined application. Even in a data mart, the data is huge and from several operational systems, therefore, they also need a multinational data model. In fact, the star schema is also one of the popular schema choices for a data mart.

A dependent data mart can be considered to be a logical subset (view) or a physical subset (extraction) of a large data warehouse. A dependent data mart may be isolated for the following reasons.

(i) For making a separate schema for OLAP or any other similar system.

(ii) To put a portion of the data warehouse or a separate machine to enhance performance.

(iii) To create a highly secure subset of data warehouse.

In fact, to standardise data analysis and usage patterns, data warehouses are generally organised as task specific small units the data marts. The data organisation of a data mart is a very simple star schema. For example, the university data warehouse that we discussed in section 3.4 can actually be a data mart on the problem "The prediction of student enrolments for the next year." A simple data mart may extract its contents directly from operational databases. However, in complex multilevel data warehouse architectures the data mart content may be loaded with the help of the warehouse database and Meta data directories.

3.8 DATA WAREHOUSE AND VIEWS

Many database developers classify data warehouse as an extension of a view mechanism. If that is the case, then how do these two mechanisms differ from one another? For, after all even in a database warehouse, a view can be materialised for the purpose of query optimisation. A data warehouse may differ from a view in the following ways:

- A data warehouse has a multi-dimensional schema and tries to integrate data through fact-dimension star schema, whereas views on the other hand are relational in nature.
- Data warehouse extracts and transforms and then stores the data into its schema; however, views are only logical and may not be materialised.
- You can apply mechanisms for data access in an enhanced way in a data warehouse, however, that is not the case for a view.
- Data warehouse data is time-stamped, may be differentiated from older versions, thus, it can represent historical data. Views on the other hand are dependent on the underlying DBMS.
- Data warehouse can provide extended decision support functionality, views normally do not do it automatically unless, an application is designed for it.

3.9 THE FUTURE: OPEN ISSUE FOR DATA WAREHOUSE

The administration of a data warehouse is a complex and challenging task. Some of the open issues for data warehouse may be:

- Quality control of data despite having filtration of data.
- Use of heterogeneous data origins is still a major problem for data extraction and transformation.
- During the lifetime of the data warehouse it will change, hence, management is one of the key issues here.
- Data warehouse administration is a very wide area and requires diverse skills, thus, people need to be suitably trained.
- Managing the resources of a data warehouse would require a large distributed team.
- The key research areas in data warehouse is, data cleaning, indexing, view creation, queries optimisation etc.

However, data warehouses are still an expensive solution and are typically found in large firms. The development of a central warehouse is capital intensive with high risks. Thus, at present data marts may be a better choice.

☞ Check Your Progress 3

1) How is data mart different from data warehouse?

……

……

……

2) How does data warehouse differ from materialised views?

……

……

……

3.10 SUMMARY

This unit provided an introduction to the concepts of data warehousing systems. The data warehouse is a technology that collects operational data from several operational systems, refines it and stores it in its own multidimensional model such as star schema or snowflake schema. The data of a data warehouse can be indexed and can be used for analyses through various DSS and EIS. The architecture of data warehouse supports contains – an interface that interact with operational system, transformation processing, database, middleware and DSS interface at the other end. However, data warehouse architecture is incomplete if, it does not have meta data directory which is extremely useful for each and every step of the data warehouse. The life cycle of a data warehouse has several stages for designing, prototyping, deploying and maintenance. The database warehouse's life cycle, however, can be clubbed with SDLC. Data mart is a smaller version of a data warehouse designed for a specific purpose. Data warehouse is quite different from views. A data warehouse is complex and offers many challenges and open issues, but, in the future data warehouses will be-extremely important technology that will be deployed for DSS. Please go through further readings for more details on data warehouse.

3.11 SOLUTIONS/ANSWERS

Check Your Progress 1

1) A Data Warehouse can be considered to be a "corporate memory". It is a repository of processed but integrated information that can be used for queries and analysis. Data and information are extracted from heterogeneous sources as they are generated. Academically, it is subject – oriented, time-variant, and a collection of operational data.

   Relational databases are designed in general, for on-line transactional processing (OLTP) and do not meet the requirements for effective on-line analytical processing (OLAP). The data warehouses are designed differently from relational databases and are suitable for OLAP.

2). ETL is Extraction, transformation, and loading. ETL refers to the methods involved in accessing and manipulating data available in various sources and loading it into target data warehouse. The following are some of the transformations that may be used during ETL:

   - Filter Transformation
   - Joiner Transformation
   - Aggregator transformation
   - Sorting transformation.

3) Some important characteristics of Data Warehousing are

   i) Multidimensional view
   ii) Unlimited dimensions and aggregation levels and unrestricted cross-dimensional operations.
   iii) Dynamic sparse matrix handling
   iv) Client/server architecture
   v) Accessibility and transparency, intuitive data manipulation and consistent reporting performance.

4) The data warehouse architecture consists of six distinct components that include:

   i) Operational systems
   ii) Transformation processing
   iii) Database
   iv) Middleware
   v) Decision support and presentation processing and
   vi) Meta data directory.

Check Your Progress 2

1) A dimension may be equated with an object. For example, in a sales organisation, the dimensions may be salesperson, product and period of a quarterly information. Each of these is a dimension. The fact table will represent the fact relating to the dimensions. For the dimensions as above, a fact table may include sale (in rupees) made by a typical sales person for the specific product for a specific period. This will be an actual date, thus is a fact. A fact, thus, represents an aggregation of relational data on the dimensions.

2) The primary difference lies in representing a normalised dimensional table.

3)
- How will the data be acquired?
- How will it be stored?
- The type of environment in which the data warehouse will be implemented?

4)
- Creation of proper transformation programs
- Proper training of development team
- Training of data warehouse administrator and end users
- Data warehouse maintenance.

5) The DSS is a decision support system and not a decision-making system. It is a specific class of information system that supports business and organisational decision-making activities. A DSS is an interactive software-based system that helps decision makers compile useful information from raw data or documents, personal knowledge, etc. This information helps these decision makers to identify and solve problems and take decisions.

An Executive Information System (EIS) facilitates the information and decision making needs of senior executives. They provide easy access to relevant information (both internal as well as external), towards meeting the strategic goals of the organisation. These are the same as for the DSS.

6) OLAP refers to the statistical processing of multidimensional such that the results may be used for decision-making. MOLAP and ROLAP are the two implementations of the OLAP. In MOLAP the data is stored in the multidimensional form in the memory whereas in the ROLAP it is stored in the relational database form.

Check Your Progress 3

1) The basic constructs used to design a data warehouse and a data mart are the same. However, a Data Warehouse is designed for the enterprise level, while Data Marts may be designed for a business division/department level. A data mart contains the required subject specific data for local analysis only.

2) The difference may be:

- Data warehouse has a multi-dimensional schema whereas views are relational in nature.
- Data warehouse extracts and transforms and then stores the data into its schema that is not true for the materialised views.
- Materialised views needs to be upgraded on any update, whereas, a data warehouse does not need updation.
- Data warehouse data is time-stamped, thus, can be differentiated from older versions that is not true for the materialised views.