Final Term

Dataware housing

Name: Muhammad Abdullah Minhas
I.D : 13864
Teacher : Sir Zain
Date : 29th June 2020

Question no 1

Differentiate between OLAP and OLTP.

Ans) Online Transaction Processing system **OLTP.**

OLTP is an Online Transaction Processing system. The main focus of OLTP system is to record the current Update, Insertion and Deletion while transaction. The OLTP queries are simpler and short and hence require less time in processing, and also requires less space.

OLTP database gets updated frequently. It may happen that a transaction in OLTP fails in middle, which may effect data integrity. So, it has to take special care of data integrity. OLTP database has normalized tables.

Example.

Example for OLTP system is an ATM, in which using short transactions we modify the status of our account. OLTP system becomes the source of data for OLAP.

Online Analytical Processing system **OLAP.**

OLAP is an Online Analytical Processing system. OLAP database stores historical data that has been inputted by OLTP. It allows a user to view different summaries of multi-dimensional data. Using OLAP, you can extract information from a large database and analyze it for decision making.

OLAP also allow a user to execute complex queries to extract multidimensional data. In OLTP even if the transaction fails in middle it will not harm data integrity as the user use OLAP system to retrieve data from a large database to analyze. Simply the user can fire the query again and extract the data for analysis.

Example

Example for OLAP is to view a financial report, or budgeting, marketing management, sales report, etc.

Key Differences Between OLTP and OLAP

1) The point that distinguishes OLTP and OLAP is that OLTP is an online transaction system whereas, OLAP is an online data retrieval and analysis system.

2) Online transactional data becomes the source of data for OLTP. However, the different OLTPs database becomes the source of data for OLAP.

3) OLTP's main operations are insert, update and delete whereas, OLAP's main operation is to extract multidimensional data for analysis.

4) OLTP has short but frequent transactions whereas, OLAP has long and less frequent transaction.

5) Processing time for the OLAP's transaction is more as compared to OLTP.

6) OLAPs queries are more complex with respect OLTPs.

7) The tables in OLTP database must be normalized (3NF) whereas, the tables in OLAP database may not be normalized.

8) As OLTPs frequently executes transactions in database, in case any transaction fails in middle it may harm data's integrity and hence it must take care of data integrity. While in OLAP the transaction is less frequent hence, it does not bother much about data integrity

Question no 2

Differentiate between expert system and dss.

Ans) Expert System (ES).

ES is based on simple rule-based logic.Problem is completely defines.There is clear way for the solution method.

ES represent precisely what is needed, the extraction of the expertise from those who know and making that knowledge available to those who don't know, with very positive additional connotations of top-down technology transfer within organizations.

Example:

locating critical areas for non-point leakage of nitrogen and phosphorus. The principle is to use GIS and expert systems to integrate landscape concept which consider hydrological and hydrochemical processes into account.

Decision Support System (DSS)

The problem is open-ended. The evaluation required to solve it is also incompletely defined/ill-defined problems.

Characteristics:

The solution involving a mixture of methods and dependent on the perspective of the user. One way of method: Multi Criteria Evaluation in IDRISI Andes software. DSS has flexibility in the form of choices of data, procedures, and displays.

Example:

Use MCE in ADRISI Andes to determine the best site location for ecovillage.

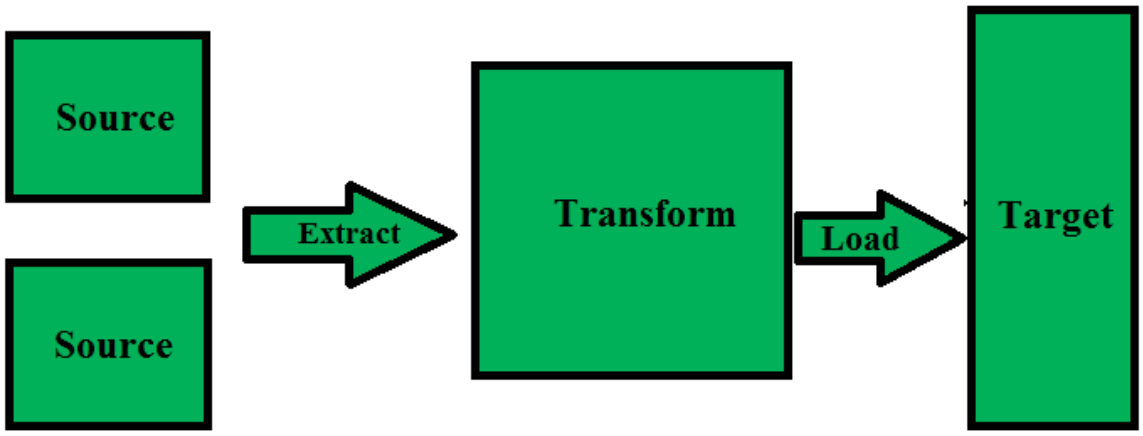
Question no 3

Differentiate between Data Warehousing and Data Mining.

Ans) Data Warehousing:

It is a technology that aggregates structured data from one or more sources so that it can be compared and analyzed rather than transaction processing. A data warehouse is designed to support management decision-making process by providing a platform for data cleaning, data integration and data consolidation. A data warehouse contains subject-oriented, integrated, time-variant and non-volatile data.

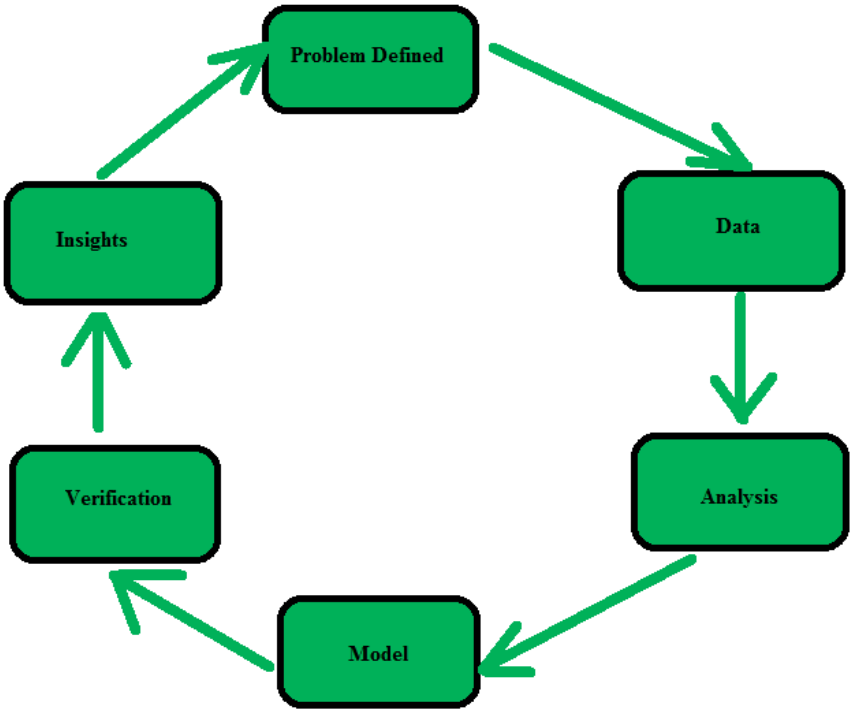
Data warehouse consolidates data from many sources while ensuring data quality, consistency and accuracy. Data warehouse improves system performance by separating analytics processing from transnational databases. Data flows into a data warehouse from the various databases. A data warehouse works by organizing data into a schema which describes the layout and type of data. Query tools analyze the data tables using schema.



Dataware Housing Process

Data mining.

It is the process of finding patterns and correlations within large data sets to identify relationships between data. Data mining tools allow a business organization to predict customer behavior. Data mining tools are used to build risk models and detect fraud. Data mining is used in market analysis and management, fraud detection, corporate analysis and risk management.



Anomaly detection (outlier/change/deviation detection)

The identification of unusual data records, that might be interesting or data errors that require further investigation.

Association rule learning (dependency modeling)

Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

Clustering

The task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

Classification

The task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

Regression

The attempts to find a function that models the data with the least error that is, for estimating the relationships among data or datasets.

Summarization

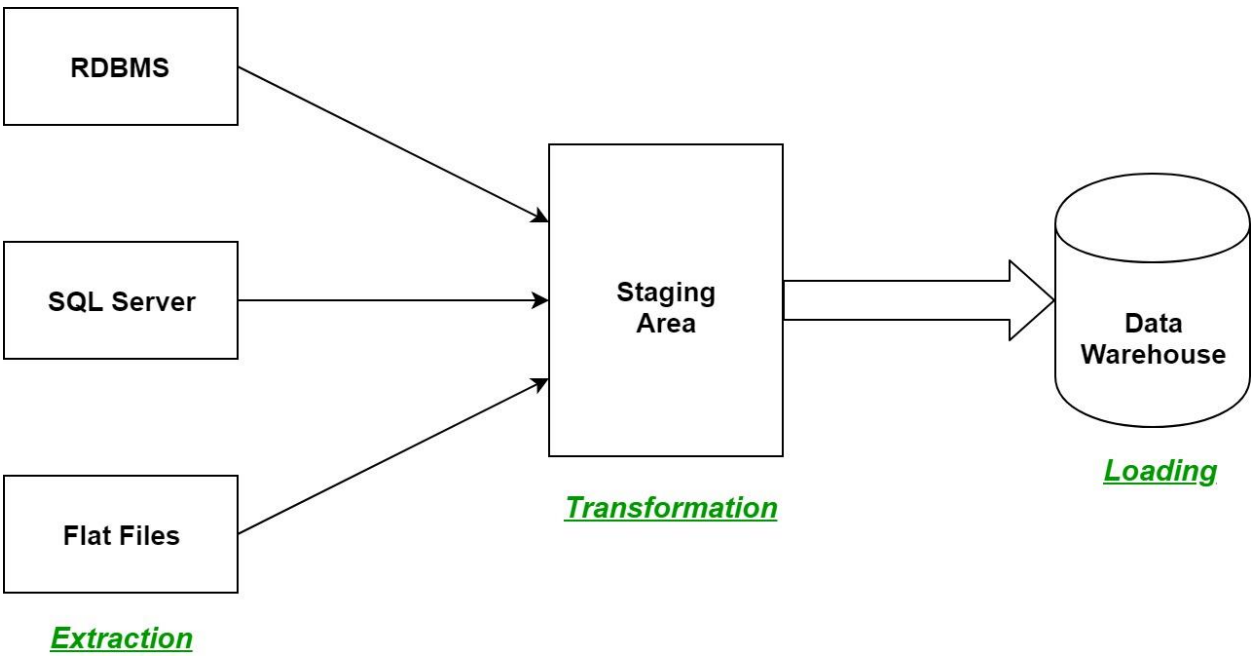
Providing a more compact representation of the data set, including visualization and report generation

Question no 4

Explain Etl process.

Ans) ETL Process in Data Warehouse:

ETL is a process in Data Warehousing and it stands for Extract, Transform and Load. It is a process in which an ETL tool extracts the data from various data source systems, transforms it in the staging area and then finally, loads it into the Data Warehouse system.



Steps of the ETL process

Extraction:

The first step of the ETL process is extraction. In this step, data from various source systems is extracted which can be in various formats like relational databases, No SQL, XML and flat files into the staging area. It is important to extract the data from various source systems and store it into the staging area first and not directly into the data warehouse because the extracted data is in various formats and can be corrupted also. Hence loading it directly into the data warehouse may damage it and rollback will be much more difficult. Therefore, this is one of the most important steps of ETL process

Transformation:

The second step of the ETL process is transformation. In this step, a set of rules or functions are applied on the extracted data to convert it into a single standard format. It may involve following processes/tasks:

Filtering – loading only certain attributes into the data warehouse.

Cleaning – filling up the NULL values with some default values

Joining – joining multiple attributes into one.

Splitting – splitting a single attribute into multipe attributes.

Sorting – sorting tuples on the basis of some attribute (generally key-attribbute)

Loading:

The third and final step of the ETL process is loading. In this step, the transformed data is finally loaded into the data warehouse. Sometimes the data is updated by loading into the data warehouse very frequently and sometimes it is done after longer but regular intervals. The rate and period of loading solely depends on the requirements and varies from system to system.

ETL process can also use the pipelining concept i.e. as soon as some data is extracted, it can transformed and during that period some new data can be extracted. And while the transformed data is being loaded into the data warehouse, the already extracted data can be transformed. The block diagram of the pipelining of ETL process is shown below;