CSC 425

DATA MINING AND DATA WAREHOUSING.

INTRODUCTION TO DATA MINING

DATA MINING

Data mining is the process of discovering information (i.e searching for patterns and relationships) hidden in large volume of data. This is meant to enrich an organisation's information support base towards making an effect or informed decision and market strategies that could improve performance in this competitive world of business.

Data mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data. It is the process of automatically discovering useful information in large data repositories.

DATA WAREHOUSE

A data warehouse is a collection of the key pieces of information used to manage and direct business for the most profitable outcome. It is a store in which data is organized in a way to facilitate its viewing and manipulation along various dimensions and to ease information extraction processes.

DATA MINING TECHNOLOGIES

1. **DATA STORE TECHNOLOGY :-** For data storage and retrieval. These include Database Management System and Data warehouse.
2. **DATA ANALYSIS TOOLS :-** For information discovery. The tools are drawn from many subjects areas such as ;
   a) Statistics and mathematics - regression analysis, clustering, sampling, roughest etc.
   b) Artificial Intelligence (AI) (machine learning and pattern recognition) - decision tree, neural networks, genetic algorithm etc

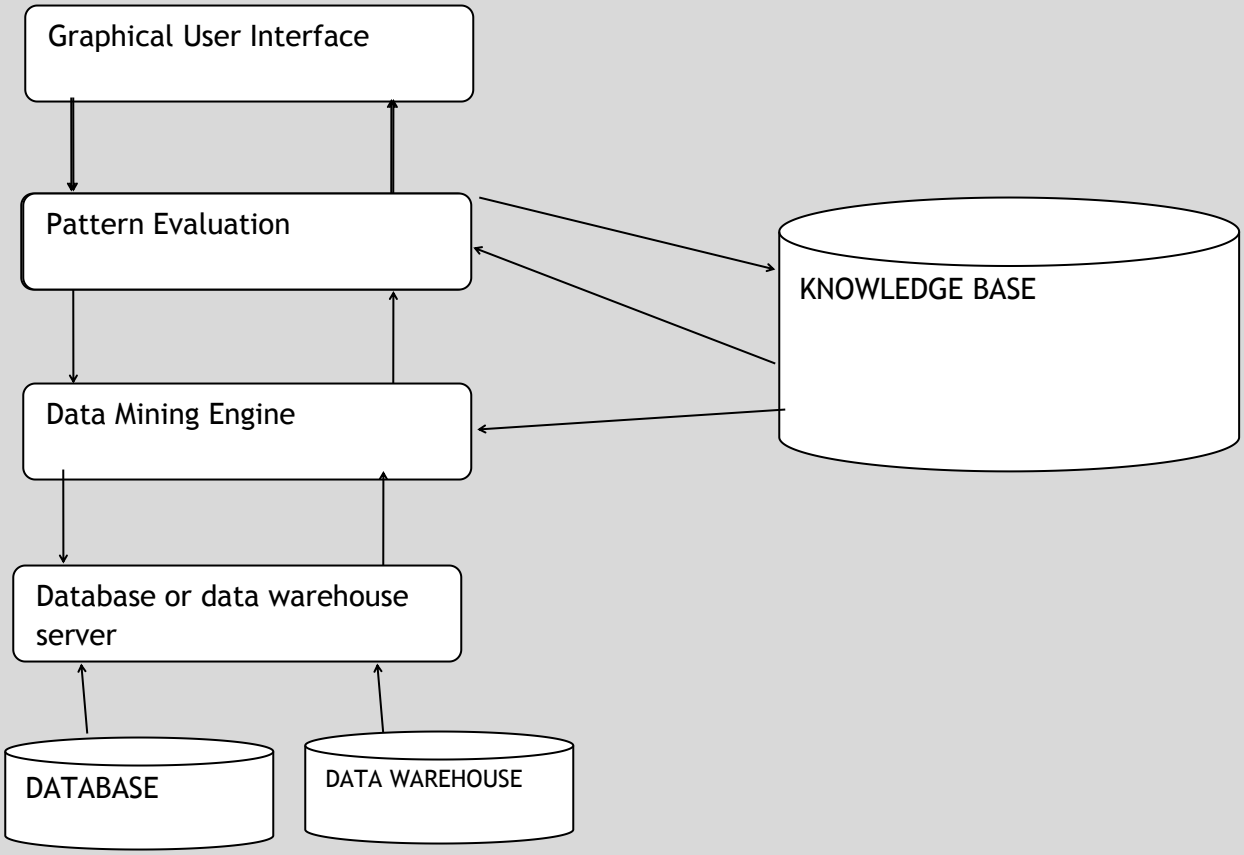
3. **DATA VISUALIZATION** :- Visual (I.e pictorial) representation of data to help users detect patterns and relationships in it. This include graphs, charts (time, line, bar, pie charts), histogram, pictogram, scatter plot, contour, plot, maps are some graphic elements usually used for this.

AREAS OF APPLICATION OF DATA MINING.

These include;

1. Financial Institutions : predicting credit worthiness of customers or detecting credit card fraud.
2. Medical line : finding relationships between diseases, symptoms and drugs, undertaking DNA analysis.
3. Marketing : finding associations in items purchased by customers etc.

DATA MINING ARCHITECTURE



A typical data mining architecture is presented above.

Graphical User Interface : enables interaction between the users and the systems. Supports users queries.

Pattern evaluation : applies measures to ecak8patrwebs and directs the search towards finding interesting or relevant patterns.

Data mining engine : performs the data mining tasks such as classification, clustering etc.

Data or data warehouse server : fetches data relevant to user's query, cleaning, integrating and filtering.

Knowledge base : a repository for domain knowledge to guide, search or evaluate patterns

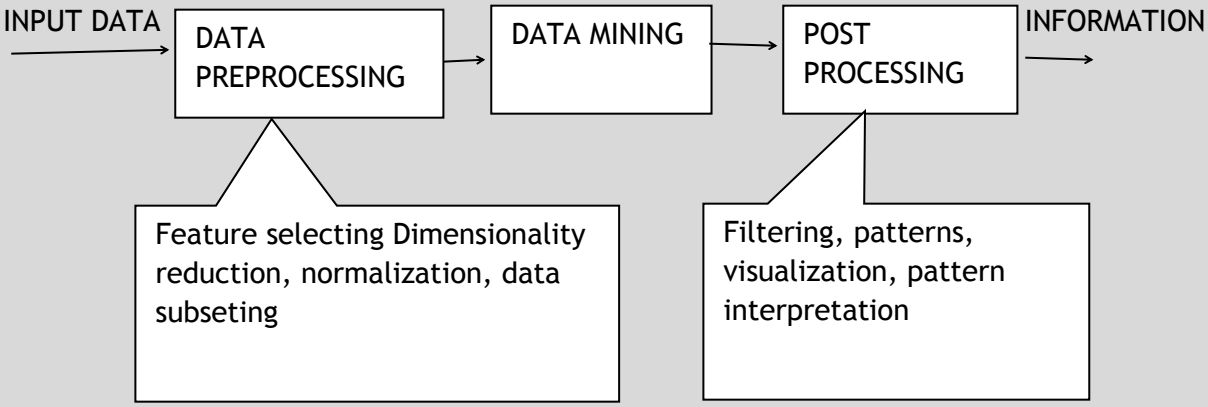
Database : a store for data (usually organized in a relational form).

Data warehouse : store in which data is organized in a way to facilitate its viewing and manipulation along various dimensions and so ease information extraction processes.

DATA MINING AND KNOWLEDGE DISCOVERY

Data mining is an integral part of knowledge discovery in database (KDD), which is the overall process of converting raw data into useful information, as shown in the diagram below. Thus process consists of a series of transformation steps from data processing to post processing of data mining results.

The process of knowledge discovery in database KDD.



MOTIVATING CHALLENGES

Traditional data analysis techniques have often encountered practical difficulties in meeting the challenge posed by new data sets. The following are some of the specific challenges that motivated the development of data mining.

1. Scalability - Because of advances in Sara generation and collection, data sets with sizes of gigabytes, terabytes, or even petabytes are becoming common. If data mining algorithms are to handle these massive datasets, then the must be scalable. Many data mining algorithms employ special search strategies to handle exponential search problems. Scalability may also require the implementation of novel data structures to access individual records in an efficient manner. Scalability can also be improved by using sampling or developing parallel and distributed algorithms.

2. High Dimensionality :- It is now common to encounter datasets with hundreds or thousands of attributes instead of the handful that is common a few decades ago. Data sets with temporal or spatial components also tend to have high dimensionality. For example, consider a dataset that contains measurements of temperature at various locations. If the temperature measurements are taken repeatedly, for an extended period, the number of dimensions (features) increases in proportion to the number of measurements taken. Traditional data analysis techniques that were developed for low dimensional data often do not work well for such high dimensional data. Also, for some data analysis algorithms, the computational complexity increases rapidly as the dimensionality (the number of features) increases.

3. Heterogeneous and complex data :- Traditional data analysis methods often deal with data sets containing attributes of the same types, either continuous or categorical. As the role of data mining in business, science, medicine, and other fields has grown, so has the need for techniques that can handle heterogeneous attributes. Recent years have also seen the emergence of more complex data objects. Examples of such worth traditional types of data include collections of web pages containing semi - structured text, and hyperlinks ; drop data with sequential and three dimensional structure; and climate data that consists of time series measurements (temperature, pressure etc.) of various locations on the earth's surface. Techniques developed for mining such complex objects should take into consideration relationships in the data such as temporal and spatial auto correlation, graph connectivity and parent - child relationships.

4. Data ownership and distribution :- sometimes the data needed for an analysis is not stored in one location or owned by one organization. Instead, the data is geographically distributed among resources belonging to multiple entities. This requires the development of distributed data mining techniques. Among the key challenges faced by distributed data mining algorithms include :

i. How to reduce the amount of communication needed to perform the distributed computation.
ii. How to effectively consolidate the data mining results obtained from multiple sources.
iii. How to address data security issues.

5. Non-traditional Analysis :- the traditional statistical approach is based on a hypothesize - and- test paradigm. In other words, a hypothesis is proposed, an experiment is designed to gather the data, and then the data is analyzed with respect to the hypothesis. Unfortunately, this process is extremely labor - intensive. Current data analysis tasks often require the generation and evaluation of thousands of hypothesis, and consequently, the development of some data mining techniques has been motivated by the desire to automate the process of hypothesis generation and evaluation.

Furthermore, the data sets analyzed in data mining are typically not the result of a carefully designed experiment and often represent opportunistic sample of the data, rather than

random samples. Also, the data sets frequently involve non-traditional types of data and data distributions.

THE ORIGINS OF DATA MINING

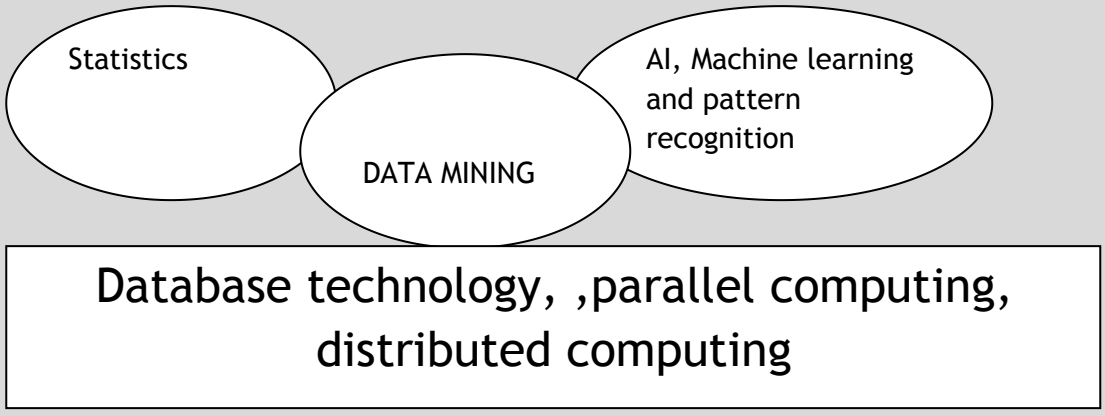
Researchers from different disciplines began to focus on developing more efficient and scalable tools that could handle diverse types of data. This work which culminated in the field of data mining is built upon the methodology and algorithms that researchers have previously used. In particular, data mining draws upon ideas, such as:

1. Sampling, estimation and hypothesis testing from statistics and
2. Search algorithms, modeling techniques, and learning theories from artificial intelligence, pattern recognition and machine learning.

Data mining has also adopted ideas from other areas including optimization, evolutionary computing, information theory, signal processing, visualization and information retrieval.

Database systems are needed to provide support for efficient storage, indexing and query processing. Techniques from high performance (parallel) computing are often important in addressing the massive size of same data sets. Distributed techniques can also help address the issue of size and are essential when the data cannot be gathered in one location.

The figure below shows the relationship of data mining to other areas.



Data mining as a confluence of many discipline

DATA MINING AND CLASSIFICATION

Data mining can be categorized into classes based on many parameters as listed below:

1. Type of data for mining :- whether the data is textual, web documents, multimedia, spatial etc.
2. Data model adopted :- whether it is relational or object - oriented databases and so on.
3. Knowledge sought :- The kind of knowledge (information) to discover.
4. Method employed :- Kind of techniques used do the mining.
5. Application Area:- The problem domain to which the data mining is applied (i.e banking, text mining, fraud analysis, credit worthiness analysis).

DATA MINING TASKS

Data mining tasks are generally divided into two major categories :

1. Predictive tasks : this is to predict the value of a particular attribute based on the values of other attributes. The attribute to be predicted is commonly know as the target or dependent variable while the attribute used for making the prediction are known as the explanatory or independent variables.
2. Descriptive tasks : this is to derive patterns (correlations, trends, clusters, trajectories and anomalies) that summarize the underlying relationships in data. Descriptive data mining tasks are often explanatory in nature and frequently require post processing techniques to validate and explain the results.

Four of the core data mining tasks are described below :

1. Predictive Modeling.
2. Association Analysis.
3. Cluster Analysis.
4. Anomaly Detection.

Predictive Modeling :- refers to the task of building a model for the target variable as a function of the explanatory variables. There are two types of predictive modeling tasks:

i. Classification
ii. Regression.

Classification is used for discrete target variables and regression is used for continuous target variables. For example, predicting whether a web user will make a purchase at an online bookstore is a Classification task because the target variable is binary - valued.

On the other hand, forecasting the future price of a stock is a regression task because price is a continuous - valued attribute. The goal of both tasks is to learn a model that minimizes the error between the predicted and true values of the target variable. E.g predicting the type of a flower.

Association Analysis : is used to describe patterns hat describe strongly associated features in the data. The discovered patterns are typically represented in the form of implication rules or feature subsets. Because of the exponential size of its search space, the goal of association analysis is to extract the most interesting patterns in an efficient manner. Useful applications of association analysis include finding groups of genes that have related functionality, identifying web pages that are accessed together, or understanding the relationship between different elements of earth's climate system.

Cluster Analysis : seeks to find groups of closely related observations so that observations that belong to the same cluster are more similar to each other than observations that belong to other clusters. Clustering has been used to group sets of related customers, find areas of the ocean that have a significant impact on the earth's climate and compress data. E.g Document clustering i.e The collection of news can be grouped based on their respective topics.

Anomaly Detection : is the task of identifying observations whose characteristics are significantly different from the rest of the data. Such observations are known as anomalies or outliers. The goal of an Anomaly Detection algorithm is to discover the real anomalies and avoid falsely labeling normal objects as anomalous. In other words, a good Anomaly detector must have a high detection rate and a low false alarm rate. Applications of Anomaly detection of fraud, network intrusions, unusual patterns of diseases and ecosystem disturbances e.g credit card fraud detection.

SYSTEM PROCESSES

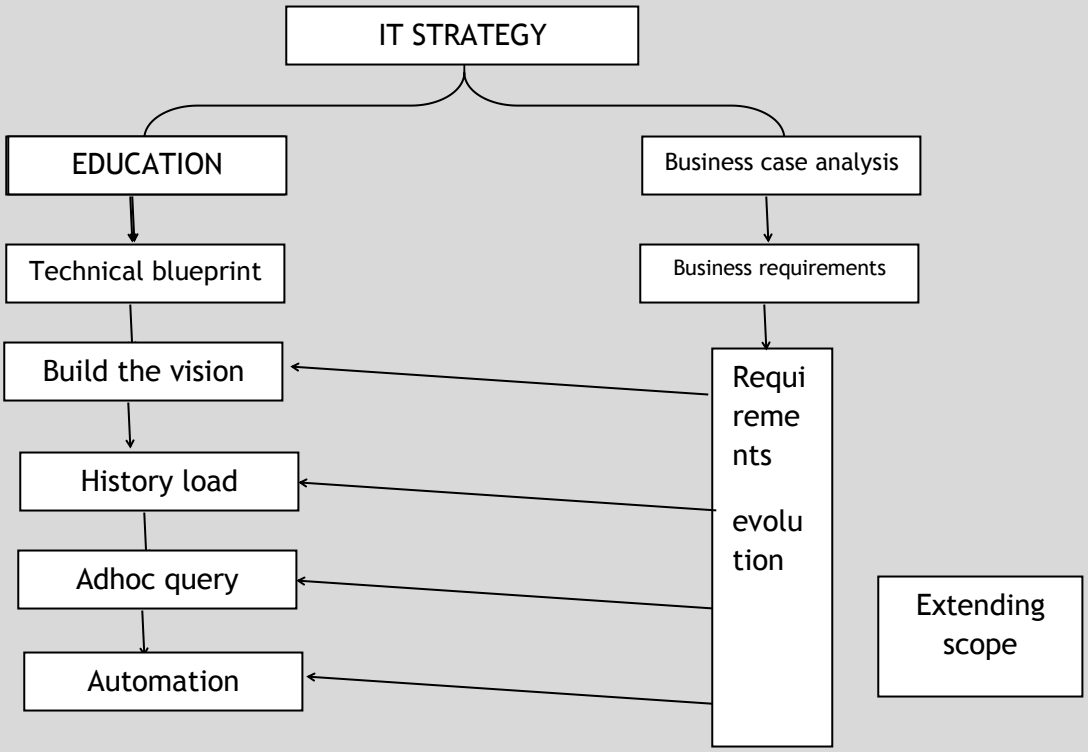
Data warehouses are built to support large data volumes (above 100gb of database) cost effectively. The underlying relational database technology has evolved to satisfy the requirements of smaller online transaction processing (OLTP) systems. The size and complexity of data warehouse systems make them very different from these traditional OLTP systems. They therefore require a different approach to the design and development. It is essential to the success of the data warehouse that sufficient time is set aside to develop an architecture that can evolve as the business requirement evolved. OLAP - Online Analytical Processing Systems.

PROCESS

The architecture of a data warehouse is defined within the technical blueprint stage of the process. The business requirements stage should have identified the initial user requirements

and have developed an understanding of the longer - term business requirements. This is used within the technical blueprint to determine what the overall architecture of the data warehouse should be.

The following diagram shows the stages in the process.



Stages in the process

OVERVIEW

Data warehouses must be architected to support three major driving factors :

1. Populating the warehouse
2. Day to day management of the warehouse
3. Ability to cope with requirements evolution.

These factors are complex, and often require a high degree of cutting edge technology to deliver such facilities. In many cases, a large proportion of the extraction and data load, and the day-to-day management of the data warehouse, can be automated. The processes

required to populate the warehouse focus on extracting the data, cleaning it up and making it available for analysis. This is typically done on a daily basis after the close of business day.

The day-to-day management of the data warehouse is different from the management of an operational system, because the volumes can be larger, and require more active management, such as creating, deleting summaries or rolling data on/off the archive. In essence, a data warehouse is a database that continually changing to satisfy new business requirements.

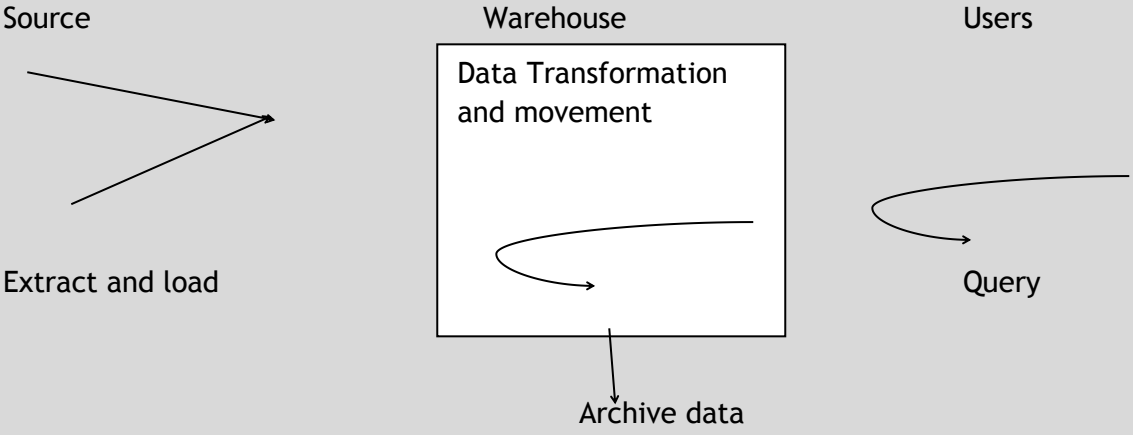
Requirements evolution tend to be the most complex aspect of a data warehouse. This requires the architecture to be structured in such a way as to cope with future changes in query profiles. This evolution will also encompass the addition of completely new subject areas.

TYPICAL PROCESS FLOW WITHIN A DATA WAREHOUSE

Before creating an architecture for a data warehouse, one must first understand the processes that constitute a data warehouse. These processes are depicted in the figure below and correspond to the data flows within a data warehouse.

The processes are ;

1. Extract and load the data.
2. Clean and transform the data into a form that can cope with large data volumes and provide good query performance.
3. Backup and archive data.
4. Manage queries and direct them to the appropriate data sources.



Process flow within a data warehouse

Assignment

State the differences between operational Database systems and data warehouse.

The Operational Database is the source of information for the data warehouse. It includes detailed information used to run the day to day operations of the business. The data frequently changes as updates are made and reflect the current value of the last transactions.

Operational Database Management Systems also called as OLTP (Online Transactions Processing Databases), are used to manage dynamic data in real-time.

Data Warehouse Systems serve users or knowledge workers in the purpose of data analysis and decision-making. Such systems can organize and present information in specific formats to accommodate the diverse needs of various users. These systems are called as Online-Analytical Processing (OLAP) Systems.

Data Warehouse and the OLTP database are both relational databases. However, the goals of both these databases are different.

Operational Database	Data Warehouse
Operational systems are designed to support high-volume transaction processing.	Data warehousing systems are typically designed to support high-volume analytical processing (i.e., OLAP).
Operational systems are usually concerned with current data.	Data warehousing systems are usually concerned with historical data.
Data within operational systems are mainly updated regularly according to need.	Non-volatile, new data may be added regularly. Once Added rarely changed.
It is designed for real-time business dealing and processes.	It is designed for analysis of business measures by subject area, categories, and attributes.
It is optimized for a simple set of transactions, generally adding or retrieving a single row at a time per table.	It is optimized for extent loads and high, complex, unpredictable queries that access many rows per table.
It is optimized for validation of incoming information during transactions, uses validation data tables.	Loaded with consistent, valid information, requires no real-time validation.
It supports thousands of concurrent clients.	It supports a few concurrent clients relative to OLTP.
Operational systems are widely process-oriented.	Data warehousing systems are widely subject-oriented
Operational systems are usually optimized to perform fast inserts and updates of associatively small volumes of data.	Data warehousing systems are usually optimized to perform fast retrievals of relatively high volumes of data.

Data In	*Data Out*
Less Number of data accessed.	*Large Number of data accessed.*
Relational databases are created for on-line transactional Processing (OLTP)	*Data Warehouse designed for on-line Analytical Processing (OLAP)*

Sr.No.	Data Warehouse (OLAP)	Operational Database(OLTP)
1	It involves historical processing of information.	It involves day-to-day processing.
2	OLAP systems are used by knowledge workers such as executives, managers, and analysts.	OLTP systems are used by clerks, DBAs, or database professionals.
3	It is used to analyze the business.	It is used to run the business.
4	It focuses on Information out.	It focuses on Data in.
5	It is based on Star Schema, Snowflake Schema, and Fact Constellation Schema.	It is based on Entity Relationship Model.
6	It focuses on Information out.	It is application oriented.
7	It contains historical data.	It contains current data.
8	It provides summarized and consolidated data.	It provides primitive and highly detailed data.
9	It provides summarized and multidimensional view of data.	It provides detailed and flat relational view of data.
10	The number of users is in hundreds.	The number of users is in thousands.
11	The number of records accessed is in millions.	The number of records accessed is in tens.
12	The database size is from 100GB to 100	The database size is from 100 MB to 100 GB.

	TB.	
13	These are highly flexible.	It provides high performance.

PROCESS ARCHITECTURE

Architecture of a data warehouse.

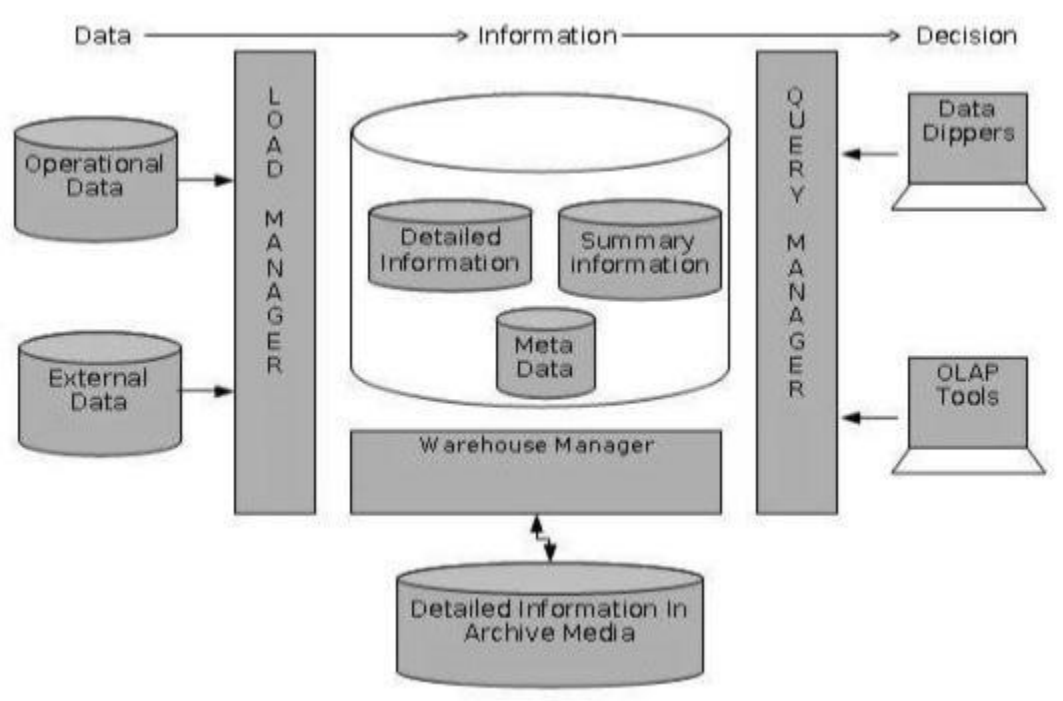


Table 1: mapping processes to system.

Process	Function	System Manager.
Extract and Load	Extracts and loads data performing simple transformations before and during load	Load Manager
Clean and transform data	Transforms and manages the data	Warehouse manager

Backup and archive	Backs up and archive the data warehouse	Warehouse manager
Query management	Directs and manages queries. The query management process is the system process that manages the queries and speeds them up by directing queries to the most effective data source	Query manager.

The data warehouse architecture shown in figure 1 is only an architecture and not a solution. The complexity of each manager will vary from data warehouse to data warehouse. It is important that the functionality detailed for each manager is present somewhere in the solutions.

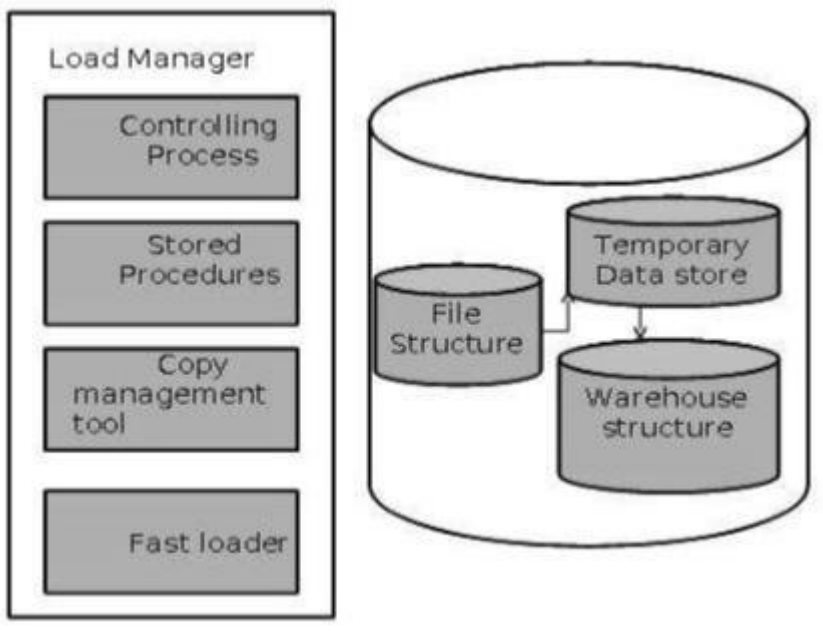
LOAD MANAGER

The load manager is the system component that performs all the operations necessary to support the extract and load process. This system may be constructed using a combination of off-the-shelf tools, coding, C programs and shell scripts.

The size and complexity of the load manager will vary between specific solutions from data warehouse to data warehouse. However, it would be noted that third party tools will probably contribute a maximum of 20-25% of the total functionality.

LOAD MANAGER ARCHITECTURE

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

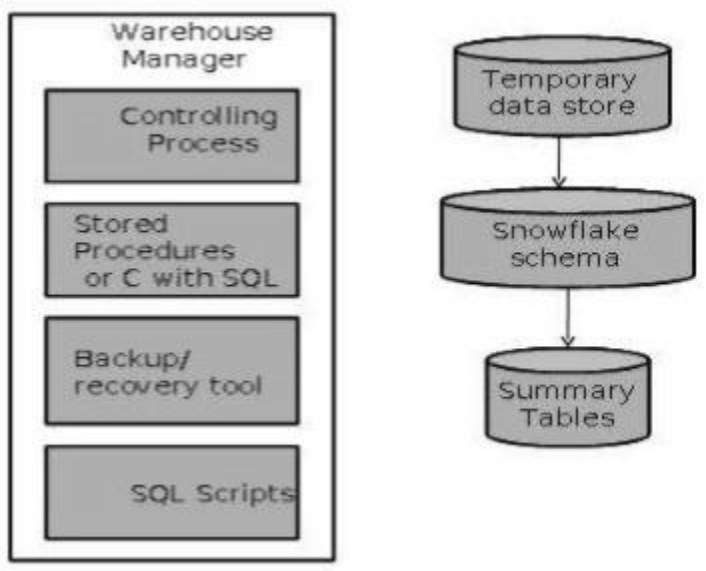


1. Extract the data from the service systems.
2. Fast-load the extracted data into a temporary data store.
3. Perform simple transformations to the one in the data warehouse.

WAREHOUSE MANAGER

The warehouse manager is the system component that performs all the operations necessary t support the warehouse management process. This system is typically constructed using a combination of third party systems management software; coding and C programs and shell scripts. As wiry the load manager, the size and complexity of the warehouse manager will vary between specific solutions.

Warehouse Manager Architecture.



1. Analyze the data to perform consistency and referential integrity checks.
2. Transform and merge the source data in the temporary data store into the published data warehouse.
3. Create indexes, business views, partition views, business synonyms against the base data.
4. Generalize denormalizations if appropriate.
5. Generate any new aggregations that may be required.
6. Update all existing aggregations.
7. Backup incrementally or totally the data within the data warehouse.
8. Archive data that has reached the end of its capture life.

QUERY MANAGER

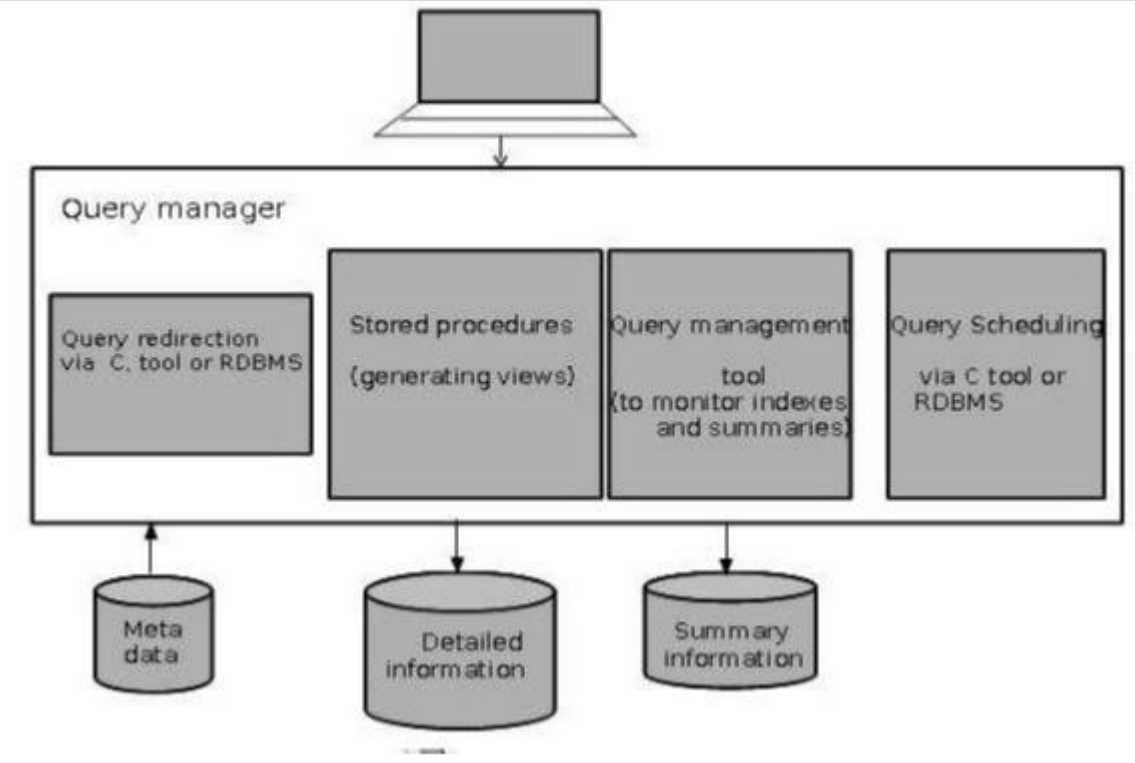
The query manager is the system component that performs all the operations necessary to support the query management process. This system is usually constructed using a combination of user access tools, specialist data warehousing monitoring tools, native database faculties, C programs and shell scripts. As with the load manager, the size and complexity of the query manager will vary between specific solutions.

Query Manager Architecture.

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

1. Direct queries to the appropriate tables.
2. Schedule the execution of user queries.

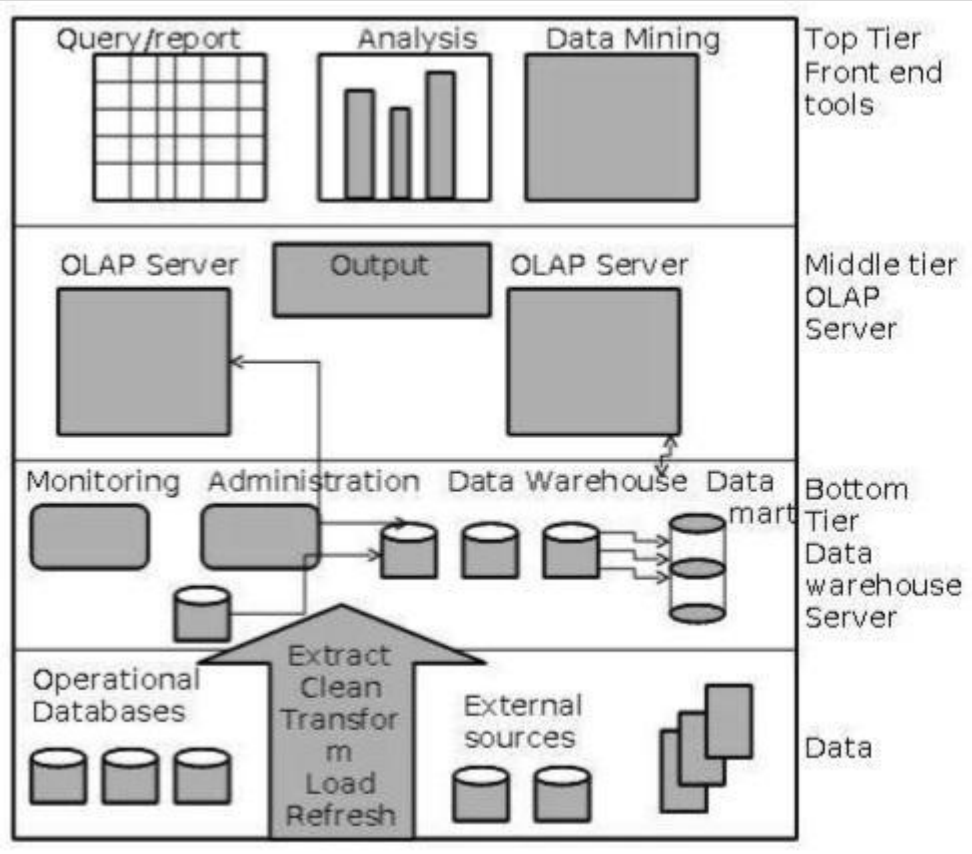
3. In some cases, the query manager also stores query profiles to allow the warehouse manager to determine which indexed and aggregations are appropriate.



Query Manager Architecture

Data warehousing: A multi-tiered architecture.

Generally a data warehouses adopts a three-tier architecture.



a three-tier data warehouse architecture

1. The bottom tier is a warehouse database server that is almost always a relational database system. Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources (e.g customer profile information provided by external consultants). These tools and utilities perform data extraction, cleaning and transformation (e.g, to merge similar data from different sources into a unified format, as well as load and refresh functions to update the data warehouse. The data are extracted using application program interfaces known as gateways. The gateway is supported by the underlying DBMS and allows client programs to generate SQL codes to be executed by a server.
2. The middle tier OLAP server that is implemented using either; relational OLAP (ROLAP) model (i.e an extended relational DBMS that maps operators on multidimensional data to standard relational operations) or (ii) A multi dimensional OLAP (MOLAP) model (i.e

a special purpose server that directly implements multi dimensional data and operations).

3. The top tier is a front-end client layer, which contains query reporting tools, analysis tools, and/or data mining tools (e.g trend analysis, prediction and so on).