Unit 4
New database applications and environments

By

Bhupendra Singh Saud

Data Mining

Data mining refers to the mining or discovery of new information in terms of patterns or rules from vast amounts of data. It is also defined as the process of finding interesting structure in data. Data mining employs one or more computer learning techniques such as machine learning, statistics, neural networks, and genetic algorithms to automatically analyze and extract knowledge from data. To be practically useful, data mining must be carried out efficiently on large files and databases

The process of Discovering meaningful patterns & trends often previously unknown, by shifting large amount of data, using pattern recognition, statistical and Mathematical techniques is called **data mining**. It is also defined as a group of techniques that find relationship that have not previously been discovered.

Data mining is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their business.

Three steps involved are:

1. Exploration
2. Pattern identification
3. Deployment

Exploration: In the first step of data exploration data is cleaned and transformed into another form, and important variables and then nature of the data based on the problem are determined.

Pattern identification: Once data is explored, refined and defined for the specific variables the second step is to form pattern identification. Identify and choose the patterns which make the best prediction.

Deployment: Patterns are deployed for desired outcome.

Functions of data mining

The types of information obtainable from data mining include associations, sequences, classifications, clusters, and forecasts.

- **Association:** Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction. That is the reason why association technique is also known as relation technique. The association technique is used in market basket analysis to identify a set of products that customers frequently purchase together. For instance, books that tends to be bought together. If a customer buys a book, an online bookstore may suggest other associated books. If a person buys a camera, the system may suggest accessories that tend to be bought along with cameras.
- **Prediction:** The prediction, as it name implied, is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables. For instance, the prediction analysis technique can be used in sale to predict profit for the future if we consider sale is an independent variable, profit could be a dependent variable. For instance, when a person applies for a credit card, the credit-card company wants to predict if the person is a good credit risk. The prediction is to be based on known attributes of the person, such as age, income, debts, and past debt repayment history.
- **Classification:** Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. For example, we can apply classification in application that "given all records of employees who left the company; predict who will probably leave the company in a future period." In this case, we divide the records of employees into two groups that named "leave" and "stay". And then we can ask our data mining software to classify the employees into separate groups.
- **Clustering:** Clustering is a data mining technique that makes meaningful or useful cluster of objects which have similar characteristics using automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes. For

example in a library, there is a wide range of books in various topics available. The challenge is how to keep those books in a way that readers can take several books in a particular topic without hassle. By using clustering technique, we can keep books that have some kinds of similarities in one cluster or one shelf and label it with a meaningful name.

Applications of Data Mining

Data mining is widely used in diverse areas. There are a number of commercial data mining system available today and yet there are many challenges in this field. In this tutorial, we will discuss the applications and the trend of data mining.

Data Mining Applications

Here is the list of areas where data mining is widely used −

- Financial Data Analysis
- Retail Industry
- Telecommunication Industry
- Biological Data Analysis
- Other Scientific Applications
- Intrusion Detection

Financial Data Analysis

The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining. Some of the typical cases are as follows −

- Design and construction of data warehouses for multidimensional data analysis and data mining.
- Loan payment prediction and customer credit policy analysis.
- Classification and clustering of customers for targeted marketing.
- Detection of money laundering and other financial crimes.

Retail Industry

Data Mining has its great application in Retail Industry because it collects large amount of data from on sales, customer purchasing history, goods transportation, consumption and services. It is natural that the quantity of data collected will continue to expand rapidly because of the increasing ease, availability and popularity of the web.

Data mining in retail industry helps in identifying customer buying patterns and trends that lead to improved quality of customer service and good customer retention and satisfaction. Here is the list of examples of data mining in the retail industry −

- Design and Construction of data warehouses based on the benefits of data mining.

- Multidimensional analysis of sales, customers, products, time and region.
- Analysis of effectiveness of sales campaigns.
- Customer Retention.
- Product recommendation and cross-referencing of items.

Telecommunication Industry

Today the telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, internet messenger, images, e-mail, web data transmission, etc. Due to the development of new computer and communication technologies, the telecommunication industry is rapidly expanding. This is the reason why data mining is become very important to help and understand the business.

Data mining in telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Here is the list of examples for which data mining improves telecommunication services −

- Multidimensional Analysis of Telecommunication data.
- Fraudulent pattern analysis.
- Identification of unusual patterns.
- Multidimensional association and sequential patterns analysis.
- Mobile Telecommunication services.
- Use of visualization tools in telecommunication data analysis.

Biological Data Analysis

In recent times, we have seen a tremendous growth in the field of biology such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining is a very important part of Bioinformatics. Following are the aspects in which data mining contributes for biological data analysis −

- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.
- Association and path analysis.
- Visualization tools in genetic data analysis.

Other Scientific Applications

The applications discussed above tend to handle relatively small and homogeneous data sets for which the statistical techniques are appropriate. Huge amount of data have been collected from scientific domains such as geosciences, astronomy, etc. A large amount of data sets is being generated because of the fast numerical simulations in various fields such as climate and ecosystem modeling, chemical engineering, fluid dynamics, etc. Following are the applications of data mining in the field of Scientific Applications −

- Data Warehouses and data preprocessing.
- Graph-based mining.
- Visualization and domain specific knowledge.

Intrusion Detection

Intrusion refers to any kind of action that threatens integrity, confidentiality, or the availability of network resources. In this world of connectivity, security has become the major issue. With increased usage of internet and availability of the tools and tricks for intruding and attacking network prompted intrusion detection to become a critical component of network administration. Here is the list of areas in which data mining technology may be applied for intrusion detection −

- Development of data mining algorithm for intrusion detection.
- Association and correlation analysis, aggregation to help select and build discriminating attributes.
- Analysis of Stream data.
- Distributed data mining.
- Visualization and query tools.

Trends in Data Mining

Data mining concepts are still evolving and here are the latest trends that we get to see in this field −

- Application Exploration.
- Scalable and interactive data mining methods.
- Integration of data mining with database systems, data warehouse systems and web database systems.
- Standardization of data mining query language.
- Visual data mining.
- New methods for mining complex types of data.
- Biological data mining.
- Data mining and software engineering.
- Web mining.
- Distributed data mining.
- Real time data mining.
- Multi database data mining.

- Privacy protection and information security in data mining.

Web mining

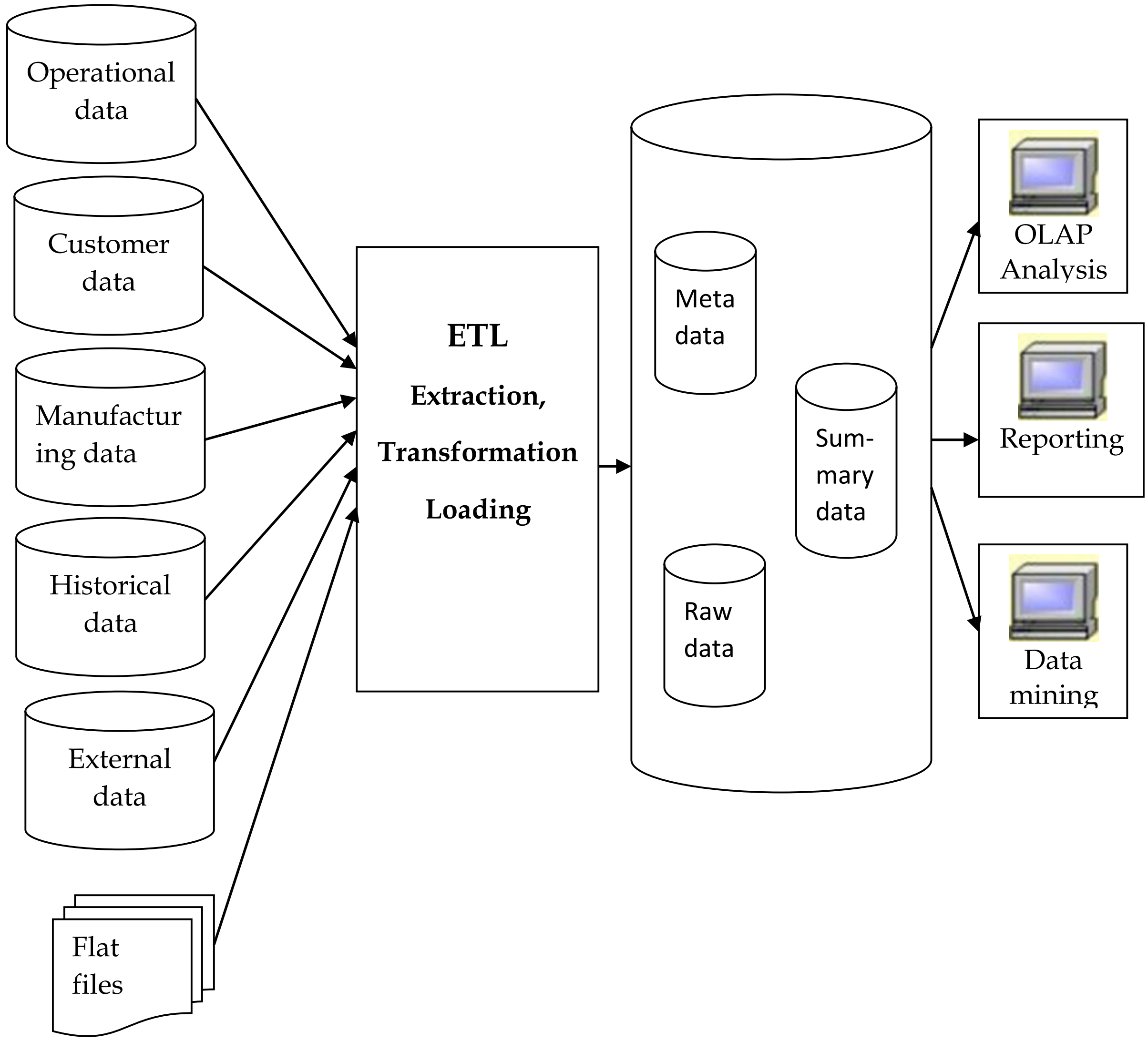
The discovery and analysis of useful patterns and information from the World Wide Web or simply web is called web mining. Web mining is the application of data mining technique to find interesting and potentially useful knowledge from web data. So web mining is the application of data mining technique to extract knowledge from web data, including web documents, hyperlinks between documents, usage logs of web sites etc. Businesses might turn to Web mining to help them understand customer behavior, evaluate the effectiveness of a particular Web site, or quantify the success of a marketing campaign. For instance, marketers use Google Trends and Google Insights for Search services, which track the popularity of various words and phrases used in Google search queries, to learn what people are interested in and what they are interested in buying.

Data Warehouse

A data warehouse is a repository of multiple heterogeneous data sources organized under a unified schema at a single site to facilitate management decision making. A data warehouse is a subject-oriented, integrated, time-variant and nonvolatile collection of data in support of management's decision-making process.

a. **Subject-Oriented**: A data warehouse can be used to analyze a particular subject area. For example, "sales" can be a particular subject.
b. **Integrated**: A data warehouse integrates data from multiple data sources. For example, source A and source B may have different ways of identifying a product, but in a data warehouse, there will be only a single way of identifying a product.
c. **Time-Variant**: Historical data is kept in a data warehouse. For example, one can retrieve data from 3 months, 6 months, 12 months, or even older data from a data warehouse. This contrasts with a transactions system, where often only the most recent data is kept. For example, a transaction system may hold the most recent address of a customer, where a data warehouse can hold all addresses associated with a customer.
d. **Non-volatile**: Once data is in the data warehouse, it will not change. So, historical data in a data warehouse should never be altered.

A **data warehouse** is a repository of current and historical data of an organization that are organized to facilitate reporting and analysis. The data originate in many core operational transaction systems, such as systems for sales, customer accounts, and manufacturing, and may include data from Web site transactions. The data warehouse consolidates and standardizes information from different operational databases so that the information can be used across the enterprise for management analysis and decision making. Figure below illustrates how a data warehouse works. The data warehouse makes the data available for anyone to access as needed, but it cannot be altered. A data warehouse system also provides a range of ad hoc and standardized query tools, analytical tools, and graphical reporting facilities. Many firms use intranet portals to make the data warehouse information widely available throughout the firm.



Features of Data warehouse

- It is separate from operational database.
- Integrates data from heterogeneous systems
- Store huge amount of data, more historical data than current data
- Does not require data to be highly accurate
- Queries are generally complex
- Provides an integrated and total view of the enterprise
- Makes the enterprise's current and historical information easily available for decision making
- Makes decision support transaction possible without hindering operational systems
- Renders the organization's information consistent
- Presents a flexible and interactive source of strategic information

Need of data warehouse

Data warehouse is needed for the following reasons:

1. **Business users:** Business users require data warehouse to view summarized data from past. Since these people are non-technical, the data may be presented to them in a very simple form.
2. **Make strategic decision:** Some strategies may be depending upon the data in data warehouse. So data warehouse contribute in making strategic decisions.
3. **Store historical data:** Data warehouse is required to store the time variable data from past. This data is made to be used for various purposes.
4. **For data quality and consistency:** Bringing the data from different sources at a common place, user can efficiently undertake to bring the uniformity and consistency in data.
5. **High response time**: Data warehouse has to be ready for fairly unexpected loads and types of queries, which demands a high degree of flexibility and quick response time.

How does a data warehouse differ from a database?

There are a number of fundamental differences which separate a data warehouse from a database. The biggest difference between them is that most database place an emphasis on a single application, and this application will generally be one that is based on transaction. If the data is analyzed, it will be done within a single domain. In contrast, data warehouses deal with multiple domains simultaneously.

Because data warehouse deals with multiple subject areas, the data warehouse finds connections between them. This allows the data warehouse to show how the company is performing as a whole, rather than in individual areas.

Another powerful aspect of data warehouse is their ability to support the analysis of trends. They are not volatile, and the information stored in them doesn't change as much as it would in a common database. Some of the major differences between them are listed below:

Database	Data Warehouse
1. In database tables and joins of different tables are complex since they are normalized for RDBMS. This is done to reduce redundant data and to save storage space. 2. Entity Relational modeling techniques are used for RDBMS database design. 3. Performance is low for analysis queries. 4. Database is the place where the data is taken as a base and managed to get available fast and efficient access. 5. Optimized for write operation. 6. Used for Online Transaction Processing (OLTP) but can be used for other purpose such as data warehousing. This records the data from the user for history.	1. In data warehouse tables and joins are simple since they are de-normalized. This is done to reduce the response time for analytical queries. 2. Data modeling techniques are used for Data Warehouse design. 3. High performance for analytical queries 4. Data warehouse is the place where the application data is managed for analysis and reporting purpose. 5. Optimized for read operations. 6. Used for Online Analytical Processing (OLAP). This reads the historical data for the users for business decision.

Meta Data

Meta data is the data about data or documentation about the data that is needed by the users. Another description of metadata is that it is structured data which describes the characteristics of a resource. Several examples of metadata are:

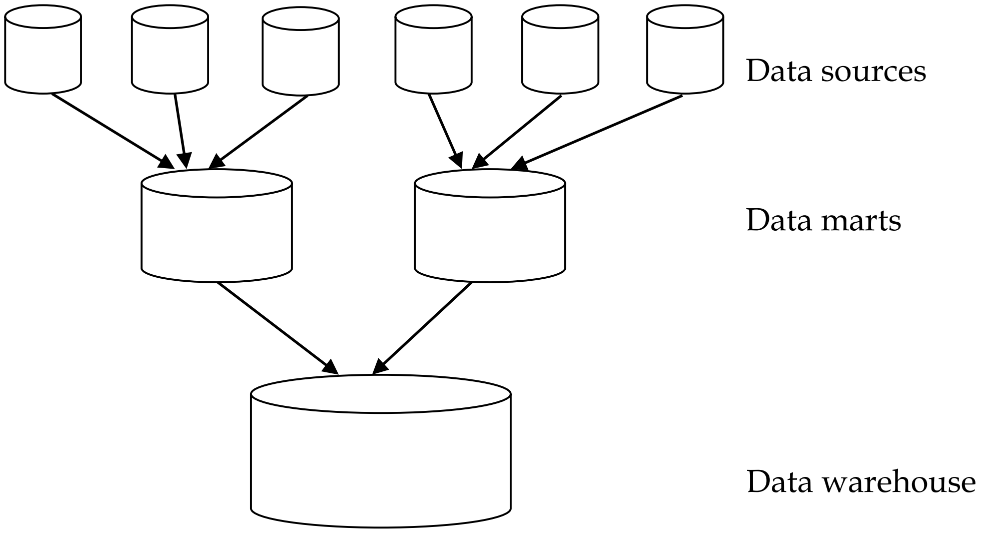
1. The table of contents and the index in a book may be considered metadata for the book.

2. A library catalogue may be considered metadata. The catalogue metadata consists of a number of predefined elements representing specific attributes of a resource, and each element can have one or more values.
3. Another example of metadata is data about the tables and figures in a document. A table has a name and there are column names of the table that may be considered metadata. The figures also have titles or names.

Data Marts

Data mart is a database that contains a subset of data present in a data warehouse. Data marts are created to structure the data in a data warehouse according to issues such as hardware platforms and access control strategies. We can divide a data warehouse into data marts after the data warehouse has been created. The implementation cycle of the data mart is likely to be measured in weeks rather than months or years.

Companies often build enterprise-wide data warehouses, where a central data warehouse serves the entire organization, or they create smaller, decentralized warehouses called data marts. A **data mart** is a subset of a data warehouse in which a summarized or highly focused portion of the organization's data is placed in a separate database for a specific population of users. For example, a company might develop marketing and sales data marts to deal with customer information. A data mart typically focuses on a single subject area or line of business, so it usually can be constructed more rapidly and at lower cost than an enterprise-wide data warehouse.



Reasons for creating a data mart

- Creates collective view by a group of users
- Easy access to frequently needed data
- Ease of creation
- Improves end-user response time
- Lower cost than implementing a full data warehouse
- Potential users are more clearly defined than in a full data warehouse
- Contains only business essential data and is less cluttered

Tools for business Intelligence

Once data have been captured and organized in data warehouses and data marts, they are available for further analysis using tools for business intelligence. Business intelligence tools enable users to analyze data to see new patterns, relationships, and insights that are useful for guiding decision making. Principal tools for business intelligence include software for database querying and reporting, tools for multidimensional data analysis (online analytical processing), and tools for data mining.

The key general categories of business intelligence tools are:

- Spreadsheets
- Reporting and querying software: tools that extract, sort, summarize, and present selected data
- OLAP: Online analytical processing
- Digital dashboards
- Data mining
- Data warehousing
- Local information systems

Online analytical processing (OLAP): Multidimensional data analysis

OLAP supports multidimensional data analysis, enabling users to view the same data in different ways using multiple dimensions (data cube). Multidimensional data models are designed expressly to support data analyses. The goal of multidimensional data models is to support analysis in a simple and faster way by executives, managers and business professionals. These people are not interested in the overall architecture.

Suppose your company sells five different products—Laptops, Computers, TVs, Camera and Mobiles—in the East, West, North and Central regions. If you wanted to ask a fairly straightforward question, such as how many Computers were sold in the last week, you could easily find the answer by using sales database. But what if you wanted to know how many Computers sold in each of your sales regions and compare actual results with projected sales, then the querying becomes complicated. In such a case OLAP is used.

Each aspect of information—product, pricing, cost, region, or time period—represents a different dimension. So, a product manager could use a multidimensional data analysis tool to learn how many Computers were sold in the East reason in this week, how that compares with the previous week, and how it compares with the sales forecast. OLAP enables users to obtain online answers to ad hoc questions such as these in a fairly rapid amount of time, even when the data are stored in very large databases, such as sales figures for multiple years.

Time	Product	Location	Sales
2072-01-01	Computer	East region	30
2072-01-01	Laptop	West region	20
2072-01-01	Camera	Central region	50
2072-01-07	Mobile	East region	11
2072-01-07	TV	North region	23
2072-01-15	Computer	West region	54
2072-01-15	Laptop	Central region	09
2072-01-25	Laptop	East region	32
2072-01-25	TV	West region	19

Fig: Tabular representation

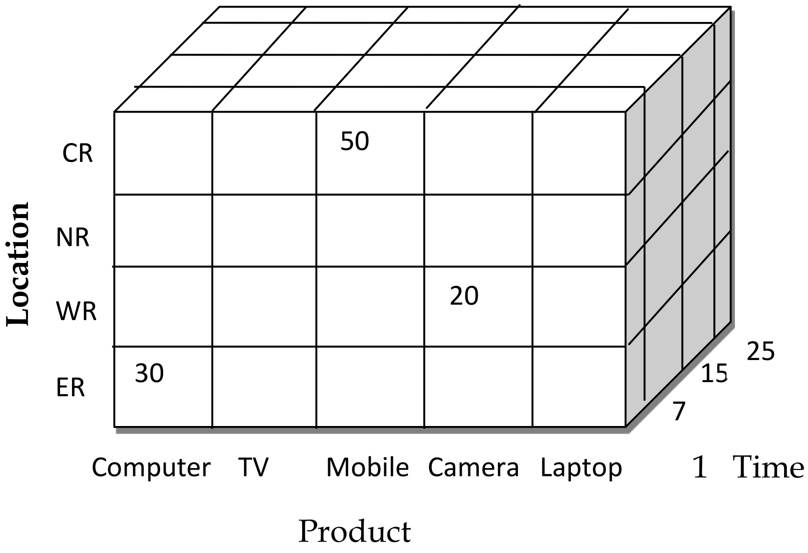


Fig: multidimensional representation

Figure above shows a multidimensional model that could be created to represent products, regions, time, and sales.

OLAP consists of four basic analytical operations: consolidation (roll-up), drill-down, slicing and dicing, and pivoting.

- **Slicing**: The slice operation is the act of picking a rectangular subset of a cube by choosing a single value for one of its dimensions, creating a new cube with one fewer dimension. For instance, the analyst may wish to see a cross-tab on *item-name* and *color* for a fixed value of *size*, for example, large, instead of the sum across all sizes.
- **Dicing:** The dice operation produces a sub-cube by allowing the analyst to pick specific values of multiple dimensions. For example, the analyst may wish to see information about medium and large sized shorts and pants in dark and pastel color.
- **Drill Down**: Drill-down is the reverse of roll-up operation. It navigates from less detailed data to more detailed data. It allows the user to navigate among levels of data ranging from the most summarized (up) to the most detailed (down). It means it is the operation of moving from finer-granularity data to a coarser granularity. For example, analyst may move from summary of pant sales to color wise sales of pant.
- **Roll-up:** A roll-up involves summarizing the data along a dimension. The summarization rule might be computing totals along a hierarchy or applying a set of formulas such as "profit = sales - expenses". . For example, analyst may move from color wise sales of pant sales to total sales of pant.
- **Pivoting or rotating:** ***Pivoting is t***he process of selecting or changing the dimensions used in a cross-tab**.** Each cross-tab is a two-dimensional view on a multidimensional data cube. For instance the analyst may select a cross-tab on *item-name* and *size*, or a cross-tab on *color* and *size*.

OLAP vs OLTP (online transaction processing)

OLAP is an acronym for Online Analytical Processing. **OLAP** performs multidimensional analysis of business data and provides the capability for complex calculations, trend analysis, and sophisticated data modeling.

OLTP (online transaction processing) is a class of software programs capable of supporting transaction-oriented applications on the Internet. Typically, **OLTP** systems are used for order entry, financial transactions, customer relationship management (CRM) and retail sales.

OLAP	OLTP
Involves historical processing of information	Involves day-to-day processing
OLAP systems are used by knowledge workers such as executives, managers and analysts.	OLTP systems are used by clerks, DBAs, or database professionals.
Useful in analyzing the business.	Useful in running the business.
Based on Star Schema, Snowflake, Schema and Fact Constellation Schema.	Based on Entity Relationship Model.
Provides summarized and consolidated data.	Provides primitive and highly detailed data.
Highly flexible.	Provides high performance.