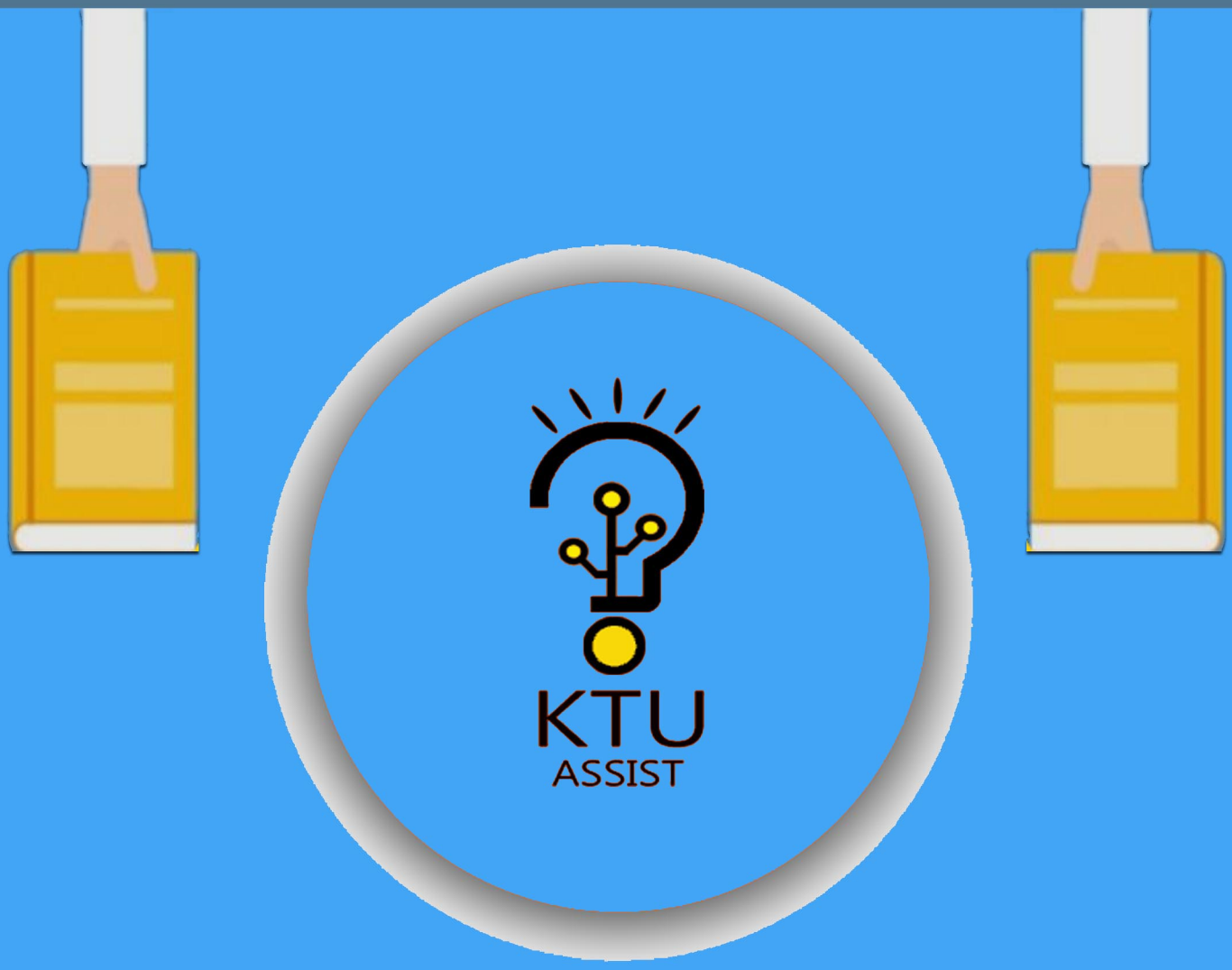
APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

STUDY MATERIALS



Data mining

We live in a world where vast amounts of data are collected daily. Analyzing such data is an important need. “We are living in the information age” is a popular saying; however, we are actually living in the data age. Terabytes or petabytes of data pour into our computer networks, the World Wide Web (WWW), and various data storage devices every day from business society, science and engineering, medicine, and almost every other aspect of daily life

What Is Data Mining?

Data mining refers to extracting or mining knowledge from large amounts of data. Mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material.

IN KNOWLEDGE DISCOVERY FROM DATA

KDD steps – STEPS

Many people treat data mining as a synonym for another popularly used term, knowledge discovery from data, or KDD, while others view data mining as merely an essential step in the process of knowledge discovery. The terms knowledge discovery in databases (KDD) and data mining are often used interchangeably.

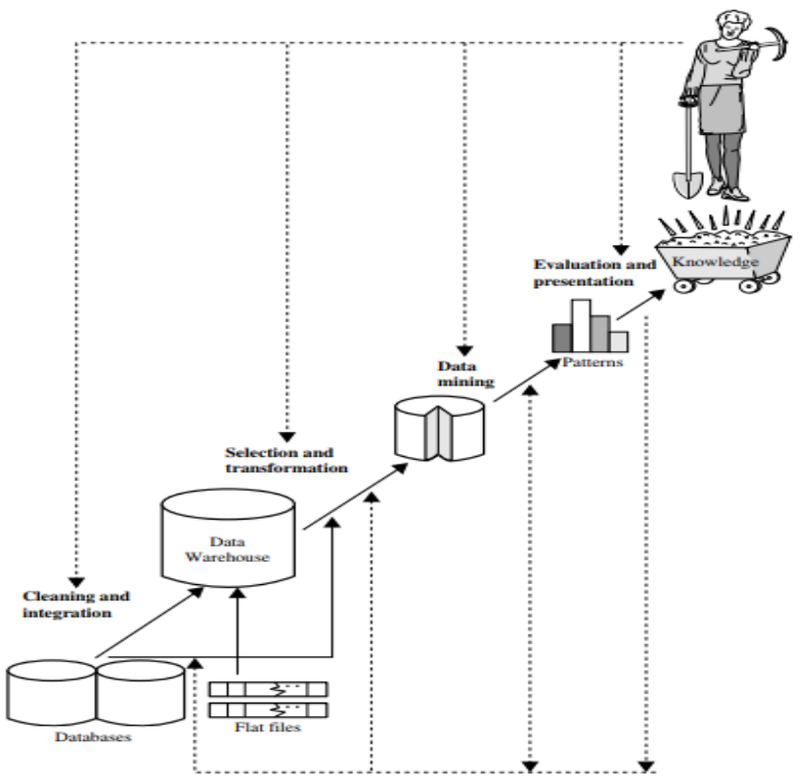


Figure 1.4 Data mining as a step in the process of knowledge discovery.

Over the last few years KDD has been used to refer to a process consisting of many steps, while data mining is only one of these steps.

Knowledge discovery in databases (KDD) is the process of finding useful information and patterns in data. Data mining is the use of algorithms to extract the information and patterns derived by the KDD process

Data mining stages:

1. Data cleaning (to remove noise and inconsistent data)

2. Data integration (where multiple data sources may be combined)

3. Data selection (where data relevant to the analysis task are retrieved from the database)

4. Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)

5. Data mining (an essential process where intelligent methods are applied to extract data patterns)

6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on interestingness measures)

7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)

Steps 1 through 4 are different forms of data preprocessing, where data are prepared for mining. The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base. The preceding view shows data mining as one step in the knowledge discovery process, albeit an essential one because it uncovers hidden patterns for evaluation.

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

Data mining applications:

1. Classification: Eg: In loan database, to classify an applicant as a prospective or defaulter, given his various personal and demographic features along with previous purchase characteristics.
2. Estimation: Predict the attribute of a data instance. Eg: estimate the percentage of marks of a student, whose previous marks are already known.
3. Prediction: Predictive model predicts a future outcome rather than the current behavior. Eg: Predict next week's closing price for the Google share price per unit.
4. Market basket analysis(association rule mining)
5. Business intelligence
6. Business data analytics
7. Bioinformatics
8. Web mining

9. Text mining
10. Social network data analysis

Architecture of a typical data mining system/Major Components

Data mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. Based on this view, the architecture of a typical data mining system may have the following major components:

1. A database, data warehouse, or other information repository, which consists of the set of databases, data warehouses, spreadsheets, or other kinds of information repositories containing the student and course information.

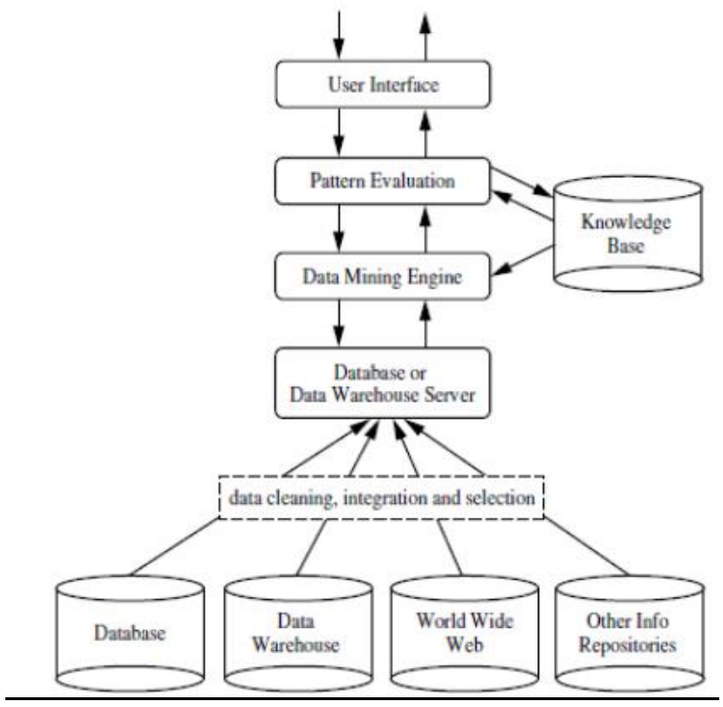
2. A database or data warehouse server which fetches the relevant data based on users' data mining requests.

3. A knowledge base that contains the domain knowledge used to guide the search or to evaluate the interestingness of resulting patterns. For example, the knowledge base may contain metadata which describes data from multiple heterogeneous sources.

4. A data mining engine, which consists of a set of functional modules for tasks such as classification, association, classification, cluster analysis, and evolution and deviation analysis.

5. A pattern evaluation module that works in tandem with the data mining modules by employing interestingness measures to help focus the search towards interestingness patterns.

6. A graphical user interface that allows the user an interactive approach to the data mining system.



Data mining models

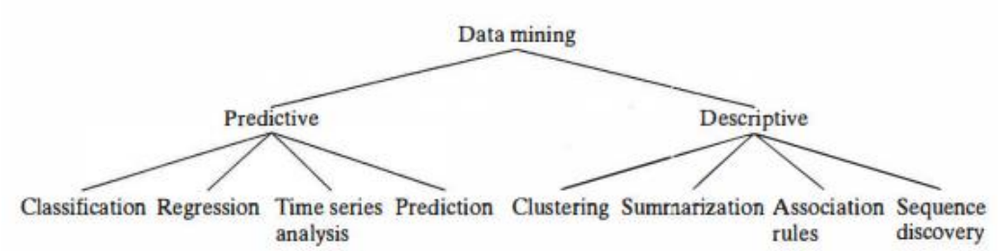


FIGURE 1.5 Data mining models and tasks.

Data mining functionalities/Data mining tasks: what kinds of patterns can be mined?

1. Class/Concept Description: Characterization and Discrimination: Data characterization is a summarization of the general characteristics or features of a target class of data. The data corresponding to the user-specified class are typically collected by a query.

For example, to study the characteristics of software products with sales that increased by 10% in the previous year, the data related to such products can be collected by executing an SQL query on the sales database.

Data discrimination is a comparison of the general features of the target class dataobjects against the general features of objects from one or multiple contrasting classes.The target and contrasting classes can be specified by a user, and the correspondingdata objects can be retrieved through database queries. For example, a user may want tocompare the general features of software products with sales that increased by 10% lastyear against those with sales that decreased by at least 30% during the same period.

2. Mining Frequent Patterns, Associations, and Correlations, Frequent patterns, as the name suggests, are patterns that occur frequently in data. There are many kinds of frequent patterns, including frequent item sets, frequent subsequences (also known as sequential patterns), and frequent substructures. A *frequent item set* typically refers to a set of items that often appear together in a transactional data set for example, milk and bread, which are frequently bought together in grocerystores by many customers. A frequently occurring subsequence, such as the patternthat customers, tend to purchase first a laptop, followed by a digital camera, and thena memory card, is a (*frequent*) *sequential pattern*. A substructure can refer to differentstructural forms (e.g., graphs, trees, or lattices) that may be combined with itemsets or subsequences. If a substructure occurs frequently, it is called a (*frequent*) *structuredpattern*. Mining frequent patterns leads to the discovery of interesting associations andcorrelations within data.

Association analysis:

$$buys(X, \text{“}computer\text{”}) \Rightarrow buys(X, \text{“}software\text{”}) \; [support = 1\%, confidence = 50\%],$$

A **confidence**, or certainty, of 50%means that if a customer buys a computer, there is a 50% chance that she will buysoftware as well. A 1% **support** means that 1% of all the transactions under analysisshow that computer and software are purchased together. This association rule involvesa single attribute or predicate (i.e., *buys*) that repeats. Association rules that contain asingle predicate are referred to as **single-dimensional association rules**. Dropping thepredicate notation, the rule can be written simply as *“computer)* *software* [1%, 50%].”

Adopting the terminology used inmultidimensional databases, where each attribute is referred to as a dimension, theabove rule can be referred to as a **multidimensional association rule**.

3. Classification and Regression for Predictive Analysis:

Classification is the process of finding a **model** (or function) that describes and distinguishesdata classes or concepts. The model is derived based on the analysis of a set of**training data** (i.e., data objects for which the class labels are known). The model is usedto predict the class label of objects for which the the class label is unknown.

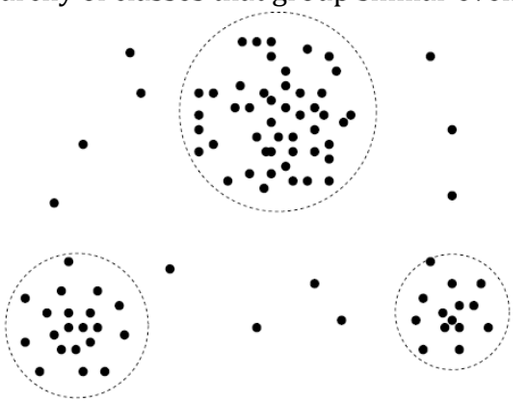
A **decisiontree** is a flowchart-like tree structure,where each node denotes a test on an attribute value, each branch represents an outcomeof the test, and tree leaves represent classes or class distributions.

A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decisiontree, or (c) a neural network. A **neural network**,when used for classification, is typicallya collection of neuron-like processing unitswithweighted connections between theunits. There are many other methods for constructing classificationmodels, such as naıveBayesian classification, support vector machines, and *k*-nearest-neighbor classification.Whereas classification predicts categorical (discrete, unordered) labels, **regression**models continuous-valued functions. That is, regression is used to predict missing orunavailable *numerical data values* rather than (discrete) class labels. The term *prediction*refers to both numeric prediction and class label prediction.

Regression analysis is a statistical methodology that is most often used for numeric prediction, although othermethods exist as well. Regression also encompasses the identification of distribution*trends* based on the available data.
Unlike classification and regression, which analyze class-labeled (training) data sets,

4. Clustering analyzes data objects without consulting class labels. In many cases, classlabeleddata may simply not exist at the beginning. Clustering can be used to generateclass labels for a group of data. The objects are clustered or grouped based on the principle of *maximizing the intraclass similarity and minimizing the interclass*

similarity. That is,clusters of objects are formed so that objects within a cluster have high similarity in comparisonto one another, but are rather dissimilar to objects in other clusters. Each clusterso formed can be viewed as a class of objects, from which rules can be derived. Clusteringcan also facilitate **taxonomy formation**, that is, the organization of observationsinto a hierarchy of classes that group similar events together.

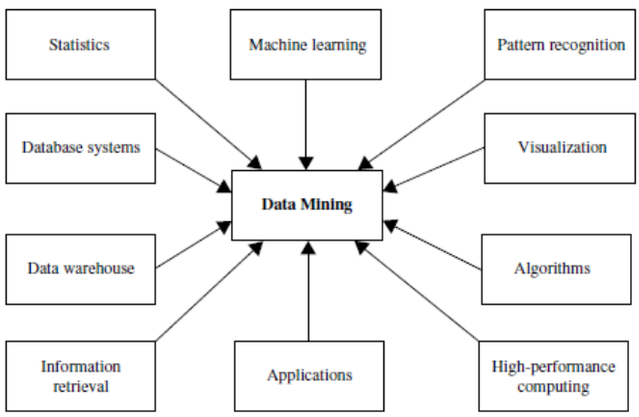


6. Outlier Analysis

A data set may contain objects that do not comply with the general behavior or modelof the data. These data objects are **outliers**. Many data mining methods discard outliersas noise or exceptions. However, in some applications (e.g., fraud detection) the rareevents can be more interesting than the more regularly occurring ones. The analysis ofoutlier data is referred to as **outlier analysis** or **anomaly mining**.

Outlier analysis may uncover fraudulent usage of credit cards bydetecting purchases of unusually large amounts for a given account number in comparisonto regular charges incurred by the same account. Outlier values may also be detectedwith respect to the locations and types of purchase, or the purchase frequency.

<u>Technologies for data mining</u>



Statistics:

Statistics studies the collection, analysis, interpretation or explanation, and presentationof data. Data mining has an inherent connection with statistics.A **statistical model** is a set of mathematical functions that describe the behavior ofthe objects in a

target class in terms of random variables and their associated probabilitydistributions. Statistical models are widely used to model data and data classes.

In other words, such statistical models can be theoutcome of a data mining task. Alternatively, data mining tasks can be built on top ofstatistical models. For example, we can use statistics to model noise and missing datavalues. Then, when mining patterns in a large data set, the data mining process can usethe model to help identify and handle noisy or missing values in the data.

Statistics research develops tools for prediction and forecasting using data and statisticalmodels. Statistical methods can be used to summarize or describe a collectionof data.

Statistical methods can also be used to verify data mining results. For example, aftera classification or prediction model is mined, the model should be verified by statisticalhypothesis testing. A **statistical hypothesis test** (sometimes called *confirmatory dataanalysis*) makes statistical decisions using experimental data. A result is called *statisticallysignificant* if it is unlikely to have occurred by chance. If the classification or predictionmodel holds true, then the descriptive statistics of the model increases the soundness of the model.

Machine learning: Investigates how computers can learn (or improve their performance)based on data. A main research area is for computer programs to *automatically* learn torecognize complex patterns and make intelligent decisions based on data. For example, atypical machine learning problem is to program a computer so that it can automaticallyrecognize handwritten postal codes on mail after learning from a set of examples.Machine learning is a fast-growing discipline. Here, we illustrate classic problems inmachine learning that are highly related to data mining.

Supervised learning is basically a synonym for classification. The supervision in thelearning comes from the labeled examples in the training data set. For example, inthe postal code recognition problem, a set of handwritten postal code images andtheir corresponding machine-readable translations are used as the training examples,which supervise the learning of the classification model.

Unsupervised learning is essentially a synonym for clustering. The learning processis unsupervised since the input examples are not class labeled. Typically, we may useclustering to discover classes within the data. For example, an unsupervised learningmethod can take, as input, a set of images of handwritten digits. Suppose that it finds10 clusters of data. These clusters may correspond to the 10 distinct digits of 0 to9, respectively. However, since the training data are not labeled, the learned modelcannot tell us the semantic meaning of the clusters found.

Semi-supervised learning is a class of machine learning techniques that make useof both labeled and unlabeled examples when learning a model. In one approach,labeled examples are used to learn class models and unlabeled examples are used torefine the boundaries between classes. For a two-class problem, we can think of theset of examples belonging to one class as the *positive examples* and those belongingto the

other class as the *negative examples*. In Figure if we do not consider theunlabeled examples, the dashed line is the decision boundary that best partitionsthe positive examples from the negative examples. Using the unlabeled examples,we can refine the decision boundary to the solid line. Moreover, we can detect thatthe two positive examples at the top right corner, though labeled, are likely noise oroutliers.

Active learning is a machine learning approach that lets users play an active rolein the learning process. An active learning approach can ask a user (e.g., a domainexpert) to label an example, which may be from a set of unlabeled examples orsynthesized by the learning program. The goal is to optimize the model quality byactively acquiring knowledge from human users, given a constraint on how manyexamples they can be asked to label.

Major issues in data mining

1. Mining methodology and user interaction issues
    - *Mining different kinds of knowledge in databases:*

Because different users can be interested in different kinds of knowledge, data mining should cover a wide spectrum of data analysis and knowledge discovery task. These tasks may use the same database in different ways and require the development of numerous data mining techniques.

- *Interactive mining of knowledge at multiple levels of abstraction:*

Interactive mining allows users to focus the search for patterns, providing and refining data mining requests based on returned results. The user can interact with the data mining system to view data and discovered patterns at multiple granularities and from different angles.

- *Incorporation of background knowledge:*

Background knowledge, or information regarding the domain under study, may be used to guide the discovery process and allow discovered patterns to be expressed in concise terms and at different levels of abstraction

2. Mining methodology and user interaction issues
    - *Pattern evaluation—the interestingness problem:*

A data mining system can uncover thousands of patterns. Several challenges remain regarding the development of techniques to assess the interestingness of discovered patterns, particularly with regard to subjective measures that estimate the value of patterns with respect to a given user class, based on user beliefs or expectations. The use of interestingness measures or user-specified constraints to guide the discovery process and reduce the search space is another active area of research.

3. Performance issues
    - *Efficiency and scalability of data mining algorithms:*

To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable. The running time of a data mining algorithm must be predictable and acceptable in large databases.

- *Parallel, distributed, and incremental mining algorithms:*

The huge size of many databases, the wide distribution of data, and the computational complexity of some data mining methods are factors motivating the development of parallel and distributed data mining algorithms. Such algorithms divide the data into partitions, which are processed in parallel. The results from the partitions are then merged.

4. Issues relating to the diversity of database types:
    - *Handling of relational and complex types of data:*

Because relational databases and data warehouses are widely used, the development of efficient and effective data mining systems for such data is important. However, other databases may contain complex data objects, hypertext and multimedia data, spatial data, temporal data, or transaction data. Specific data mining systems should be constructed for mining specific kinds of data.

- *Mining information from heterogeneous databases and global information systems:*

Local- and wide-area computer networks (such as the Internet) connect many sources of data, forming huge, distributed, and heterogeneous databases. The discovery of knowledge from different sources of structured, semi structured, or unstructured data with diverse data semantics poses great challenges to data mining

Data Warehouse

Data warehousing provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions. Data warehouse refers to a database that is maintained separately from an organization's operational databases.A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision making process"

1. **Subject-oriented**: A data warehouse is organized around major subjects, such as customer, supplier, product, and sales.

A data warehouse focuses on the modelling and analysis of data for decision makers(not on day to day transaction).

Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

2. **Integrated**: data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and on-line transaction records.
3. **Time-variant**: Data are stored to provide information from a historical perspective

   Every key structure in the data warehouse contains, either implicitly or explicitly, an element of time.

4. **Non-volatile**: A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment. Due to this separation, a data warehouse does not require transaction processing, recovery, and concurrency control mechanisms. It usually requires only two operations in data accessing: *initial loading of data and access of data.*

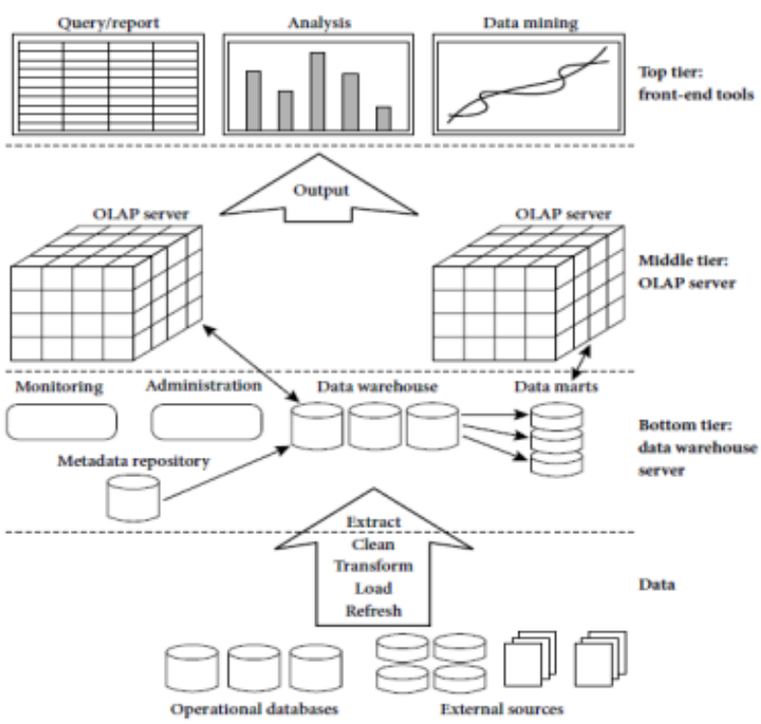
Data warehousing is the process of constructing and using data warehouses.

- The construction of a data warehouse requires data cleaning, data integration, and data consolidation.
- The utilization of a data warehouse often necessitates a collection of decision support technologies. This allows "knowledge workers" (e.g., managers, analysts, and executives) to use the warehouse to quickly and conveniently obtain an overview of the data, and to make sound decisions based on information in the warehouse.

Data warehousing is very useful from the point of view of heterogeneous database integration.

The traditional database approach to heterogeneous database integration was a 'query- driven' approachdata warehousing employs an update-driven approach in which information from multiple, heterogeneous sources is integrated in advance and stored in a warehouse for direct querying and analysis.

A Three Tier Data Warehouse Architecture



Tier-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 (such as 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. A gateway is supported by the underlying DBMS and allows client programs to generate SQL code to be executed at a server.

Examples of gateways include ODBC (Open Database Connection) and OLEDB (Open Linking and Embedding for Databases) by Microsoft and JDBC (Java Database Connection). This tier also contains a metadata repository, which stores information about the data warehouse and its contents.

Tier-2:

The middle tier is an OLAP server that is typically implemented using either a relational OLAP (ROLAP) model or a multidimensional OLAP. OLAP model is an extended relational DBMS that maps operations on multidimensional data to standard relational operations A multidimensional OLAP (MOLAP) model, that is, a special-purpose server that directly implements multidimensional data and operations.

Tier-3:

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

Data Warehouse Models

There are three data warehouse models.

1. Enterprise warehouse:

An enterprise warehouse collects all of the information about subjects spanning the entire organization. It provides corporate-wide data integration, usually from one or more operational systems or external information providers, and is cross-functional in scope. It typically contains detailed data as well as summarized data, and can range in size from a few gigabytes to hundreds of gigabytes, terabytes, or beyond. An enterprise data warehouse may be implemented on traditional mainframes, computer super servers, or parallel architecture platforms. It requires extensive business modeling and may take years to design and build.

2.Data mart: *********

A data mart contains a subset of corporate-wide data that is of value to a specific group of users. The scope is confined to specific selected subjects. For example, a marketing data mart may confine its subjects to customer, item, and sales. The data contained in data marts tend to be summarized. Data marts are usually implemented on low-cost departmental servers that are UNIX/LINUX- or Windows-based. The implementation cycle of a data mart is more likely to be measured in weeks rather than months or years. However, it may involve complex integration in the long run if its design and planning were not enterprise-wide.

Depending on the source of data, data marts can be categorized as independent or dependent. Independent data marts are sourced from data captured from one or more operational systems or external information providers, or from data generated locally within a particular department or geographic area. Dependent data marts are sourced directly from enterprise data warehouses.

3.virtual warehouse is a set of views over operational databases. For efficient query processing, only some of the possible summary views may be materialized. A virtual warehouse is easy to build but requires excess capacity on operational database servers.

Meta Data Repository: *****

Metadata are data about data. When used in a data warehouse, metadata are the data that define warehouse objects. Metadata are created for the data names and definitions of the given warehouse. Additional metadata are created and captured for time stamping any extracted data, the source of the extracted data, and missing fields that have been added by data cleaning or integration processes.

A metadata repository should contain the following: A description of the structure of the data warehouse, which includes the warehouse schema, view, dimensions, hierarchies, and derived data definitions, as well as data mart locations and contents. Operational metadata, which include data lineage (history of migrated data and the sequence of transformations applied to it), currency of data (active, archived, or purged), and monitoring information (warehouse usage statistics, error reports, and audit trails).

The algorithms used for summarization, which include measure and dimension definition algorithms, data on granularity, partitions, subject areas, aggregation, summarization, and predefined queries and reports. The mapping from the operational environment to the data warehouse, which include source databases and their contents, gateway descriptions, data partitions, data extraction, cleaning, transformation rules and defaults, data refresh and purging rules, and security (user authorization and access control). Data related to system performance, which include indices and profiles that improve data access and retrieval performance, in addition to rules for the timing and scheduling of refresh, update, and replication cycles. Business metadata, which include business terms and definitions, data ownership information, and charging policies.

OLAP(Online analytical Processing):

OLAP is an approach to answering multi-dimensional analytical (MDA) queries swiftly. OLAP is part of the broader category of business intelligence, which also encompasses relational database, report writing and data mining. OLAP tools enable users to analyze multidimensional data interactively from multiple perspectives.

OLAP consists of three basic analytical operations:

- Consolidation(Roll-Up)
- Drill-Down
- Slicing
- Dicing
- Pivot

Consolidation involves the aggregation of data that can be accumulated and computed in one or more dimensions.

For example, all sales offices are rolled up to the sales department or sales division to anticipate sales trends.

The drill-down is a technique that allows users to navigate through the details. For instance, users can view the sales by individual products that make up a region’s sales.

Slicing and dicing is a feature whereby users can take out (slicing) a specific set of data of the OLAP cube and view (dicing) the slices from different viewpoints.

Types of OLAP:

1. Relational OLAP (ROLAP):

ROLAP works directly with relational databases. The base data and the dimension tables are stored as relational tables and new tables are created to hold the aggregated information. It depends on a specialized schema design. This methodology relies on manipulating the data stored in the relational database to give the appearance of traditional OLAP's slicing and dicing functionality. In essence, each action of slicing and dicing is equivalent to adding a "WHERE" clause in the SQL statement. ROLAP tools do not use pre-calculated data cubes but instead pose the query to the standard relational database and its tables in order to bring back the data required to answer the question. ROLAP tools feature the ability to ask any question because the methodology does not limit to the contents of a cube. ROLAP also has the ability to drill down to the lowest level of detail in the database.

2. Multidimensional OLAP (MOLAP):

MOLAP is the 'classic' form of OLAP and is sometimes referred to as just OLAP. MOLAP stores this data in an optimized multi-dimensional array storage, rather than in a relational database. Therefore it requires the pre-computation and storage of information in the cube - the operation known as processing. MOLAP tools generally utilize a pre-calculated data

set referred to as a data cube. The data cube contains all the possible answers to a given range of questions. MOLAP tools have a very fast response time and the ability to quickly write back data into the data set.

3. Hybrid OLAP (HOLAP):

There is no clear agreement across the industry as to what constitutes Hybrid OLAP, except that a database will divide data between relational and specialized storage. For example, for some vendors, a HOLAP database will use relational tables to hold the larger quantities of detailed data, and use specialized storage for at least some aspects of the smaller quantities of more-aggregate or less-detailed data. HOLAP addresses the shortcomings of MOLAP and ROLAP by combining the capabilities of both approaches. HOLAP tools can utilize both pre-calculated cubes and relational data sources.

Difference between Operational Database systems and Data Warehouse

- Operational Database systems
  - Main task is to perform on-line transaction and query processing. These systems are called **on-line transaction processing (OLTP)** systems.
  - They cover most of the day-to-day operations of an organization, such as purchasing, inventory, manufacturing, banking, payroll, registration, and accounting.
- Data Warehouse
  - serve users or knowledge workers in the role of data analysis and decision making.
  - Such systems can organize and present data in various formats in order to accommodate the diverse needs of the different users. These systems are known as **on-line analytical processing (OLAP)** systems.

Difference between OLTP and OLAP

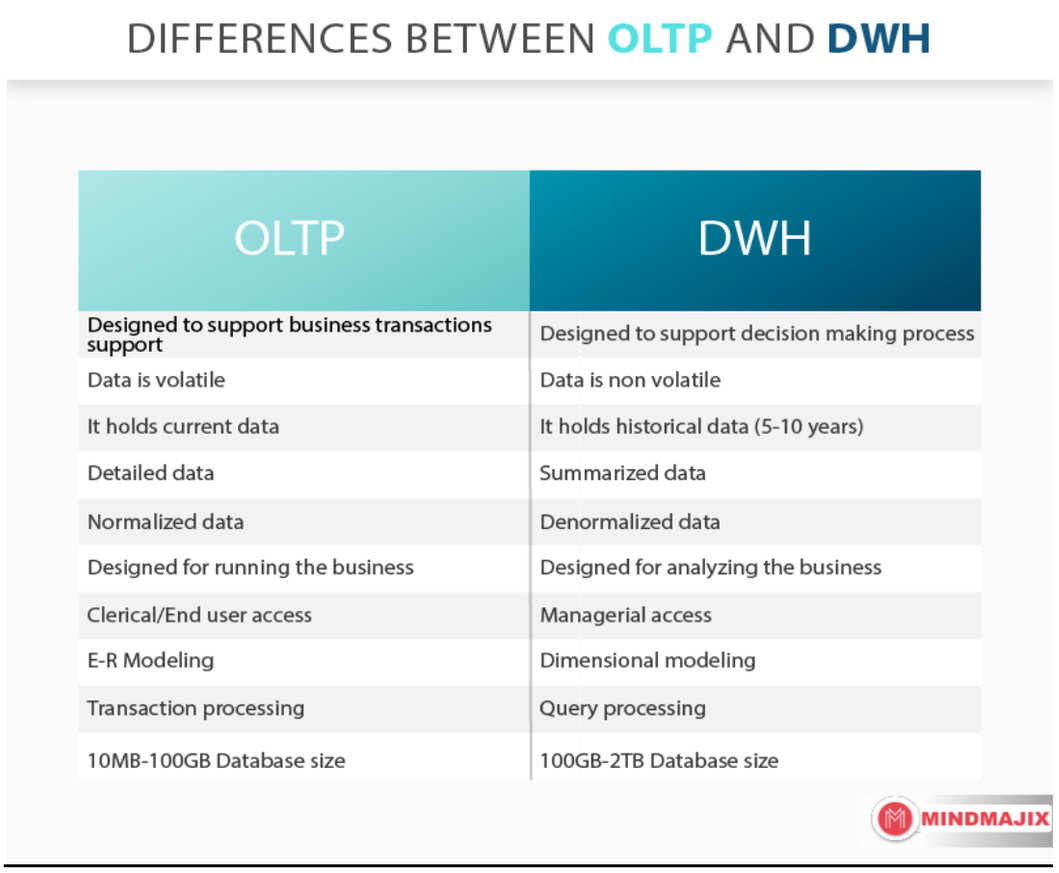
- **Users and system orientation**:
  - OLTP system is *customer-oriented and is used for* transaction and query processing by clerks, clients, and information technology professionals.
  - OLAP system is *market-oriented and is used for data analysis by knowledge* workers, including managers, executives, and analysts.
- **Data contents**:
  - OLTP system manages current data
  - OLAP system manages large amounts of historical data, provides facilities for summarization and aggregation, and stores and manages information at different levels of granularity.
- **Database design**:
  - An OLTP system usually adopts an entity-relationship (ER) data model and an application-oriented database design.

- An OLAP system typically adopts either a *star or snowflake model and a subject oriented* database design.

- **View**:
  - An OLTP system focuses mainly on the current data within an enterprise or department, without referring to historical data or data in different organizations.
  - An OLAP system often spans multiple versions of a database schema, due to the evolutionary process of an organization.
  - OLAP systems also deal with information that originates from different organizations.
  - OLAP data are stored on multiple storage media.
- **Access patterns**:
  - The access patterns of an OLTP system consist mainly of short, atomic transactions. Such a system requires concurrency control and recovery mechanisms.

Accesses to OLAP systems are mostly read-only operations although many could be complex queries.

Feature	*OLTP*	*OLAP*
Characteristic	operational processing	informational processing
Orientation	transaction	analysis
User	clerk, DBA, database professional	knowledge worker (e.g., manager, executive, analyst)
Function	day-to-day operations	long-term informational requirements, decision support
DB design	ER based, application-oriented	star/snowflake, subject-oriented
Data	current; guaranteed up-to-date	historical; accuracy maintained over time
Summarization	primitive, highly detailed	summarized, consolidated
View	detailed, flat relational	summarized, multidimensional
Unit of work	short, simple transaction	complex query
Access	read/write	mostly read
Focus	data in	information out
Operations	index/hash on primary key	lots of scans
Number of records accessed	tens	millions
Number of users	thousands	hundreds
DB size	100 MB to GB	100 GB to TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time



A Multidimensional Data Model

Data warehouses and OLAP tools are based on a multidimensional data model. This model views data in the form of a data cube. A data cube allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts. In general terms, dimensions are the perspectives or entities with respect to which an organization wants to keep records.

For example, AllElectronics shop may create a sales data warehouse in order to keep records of the store's sales with respect to the dimensions time, item, branch, and location. These dimensions allow the store to keep track of things like monthly sales of items and the branches and locations at which the items were sold. Each dimension may have a table associated with it, called a dimension table, which further describes the dimension. For example, a dimension table for item may contain the attributes item name, brand, and type. Dimension tables can be specified by users or experts, or automatically generated and adjusted based on data distributions.

A multidimensional data model is typically organized around a central theme, like sales, for instance. This theme is represented by a fact table. Facts are numerical measures. Think of them as the quantities by which we want to analyze relationships between dimensions. Examples of facts for a sales data warehouse include dollars sold (sales amount in dollars), units sold (number of units sold), and amount budgeted. The

fact table contains the names of the facts, or measures, as well as keys to each of the related dimension tables.

Table 3.3 A 3-D view of sales data for *AllElectronics*, according to the dimensions *time*, *item*, and *location*. The measure displayed is *dollars_sold* (in thousands).

	location = *"Chicago"*				***location*** = *"New York"*				***location*** = *"Toronto"*				***location*** = *"Vancouver"*			
	item				***item***				***item***				***item***			
	home				*home*				*home*				*home*			
time	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*
Q1	854	882	89	623	1087	968	38	872	818	746	43	591	605	825	14	400
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682	680	952	31	512
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728	812	1023	30	501
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784	927	1038	38	580

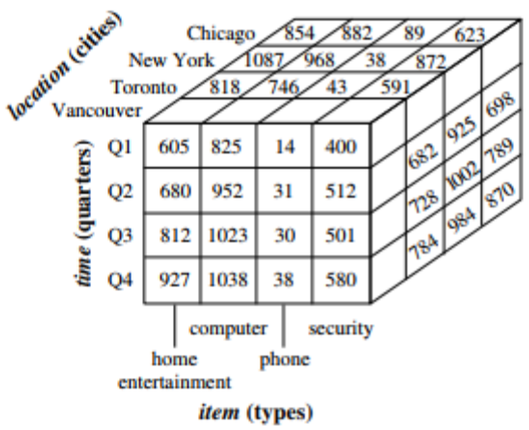


Figure 3.1 A 3-D data cube representation of the data in Table 3.3, according to the dimensions *time*, *item*, and *location*. The measure displayed is *dollars_sold* (in thousands).

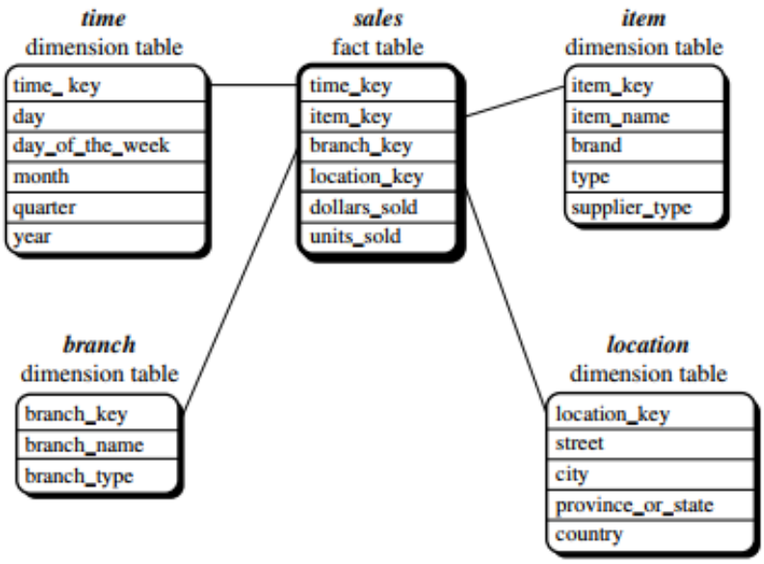
Stars, Snowflakes, and Fact Constellations: Schemas for Multidimensional Databases

The entity-relationship data model is commonly used in the design of relational databases, where a database schema consists of a set of entities and the relationships between them. Such a data model is appropriate for on-line transaction processing.

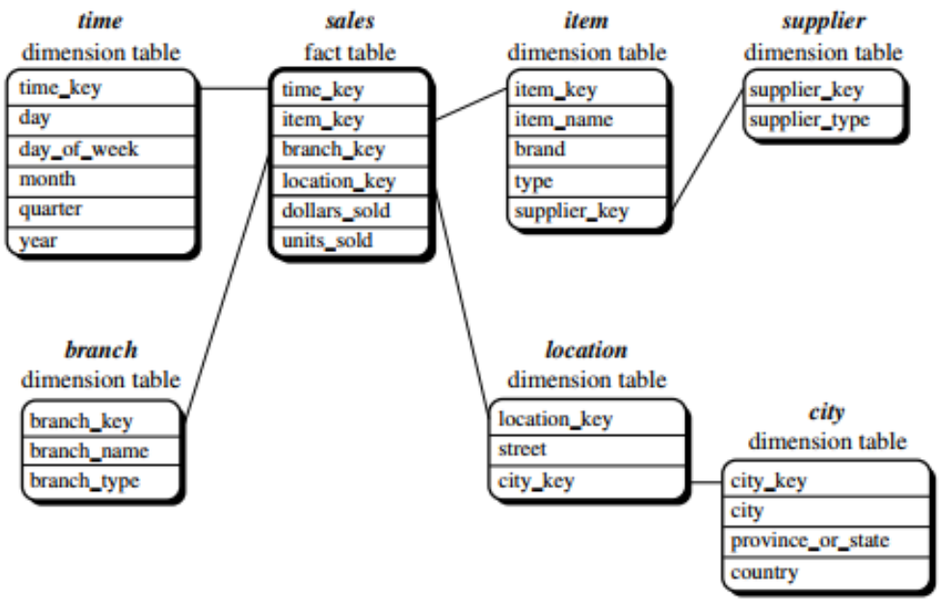
A data warehouse, however, requires a concise, subject-oriented schema that facilitates on-line data analysis. The most popular data model for a data warehouse is a multidimensional model. Such a model can exist in the form of a star schema, a snowflake schema, or a fact constellation schema. Let's look at each of these schema types.

Star schema: The most common modeling paradigm is the star schema, in which the data warehouse contains (1) a large central table (fact table) containing the bulk of the data, with no redundancy, and (2) a set of smaller attendant tables (dimension

tables), one for each dimension. The schema graph resembles a starburst, with the dimension tables displayed in a radial pattern around the central fact table.



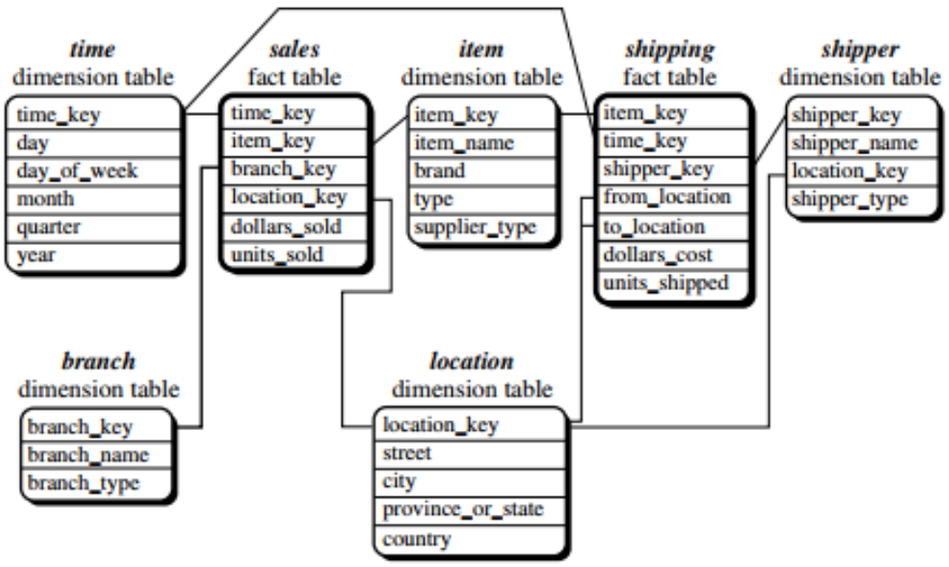
Snowflake schema: The snowflake schema is a variant of the star schema model, where some dimension tables are normalized, thereby further splitting the data into additional tables. The resulting schema graph forms a shape similar to a snowflake.



The major difference between the snowflake and star schema models is that the dimension tables of the snowflake model may be kept in normalized form to reduce redundancies. Such a table is easy to maintain and saves storage space. However, this saving of space is negligible in comparison to the typical magnitude of the fact table. Furthermore, the snowflake structure can reduce the effectiveness of browsing, since more joins will be needed to execute a query. Consequently, the system

performance may be adversely impacted. Hence, although the snowflake schema reduces redundancy, it is not as popular as the star schema in data warehouse design.

Fact constellation: Sophisticated applications may require multiple fact tables to share dimension tables. This kind of schema can be viewed as a collection of stars, and hence is called a galaxy schema or a fact constellation.



Need for Data Warehousing

1. The data ware house market supports such diverse industries as manufacturing, retail, telecommunications, and health care. Think of a personnel database for a company that is continually modified as personnel are added and deleted... If management wishes determine if there is a problem with too many employees quitting. To analyze this problem, they would need to know which employees have left, when they left, why they left, and other information about their employment. For management to make these types of high-level business analyses, more historical data not just the current snapshot are required.

   A data warehouse is a data repository used to support decision support systems

2. The basic motivation is to increase business profitability. Traditional data processing applications support the day-to-day clerical and administrative decisions, while data warehousing supports long-term strategic decisions.
3. For increasing customer focus, which includes the analysis of customer buying patterns (such as buying preference, buying time, budget cycles, and appetites for spending)

4. For repositioning products and managing product portfolios by comparing the performance of sales by quarter, by year, and by geographic regions in order to

7. Data warehouse queries are often complex. They involve the computation of large groups of data at summarized levels, and may require the use of special data organization, access, and implementation methods based on multidimensional views.

8. **Data Quality** – In a data warehouse, data is coming from many disparate sources from all facets of an organization. When a data warehouse tries to combine inconsistent data from disparate sources, it encounters errors. Inconsistent data, duplicates, logic conflicts, and missing data all result in data quality challenges. Poor data quality results in faulty reporting and analytics necessary for optimal decision making.

9. **Understanding Analytics** – When building a data warehouse, analytics and reporting will have to be taken into design considerations. In order to do this, the business user will need to know exactly what analysis will be performed.
10. **Quality Assurance** – The end user of a data warehouse is using Big Data reporting and analytics to make the best decisions possible. Consequently, the data must be 100 percent accurate or a credit union leader could make ill-advised decisions that are detrimental to the future success of their business. This high reliance on data quality makes testing a high priority issue that will require a lot of resources to ensure the information provided is accurate.

11. **Performance** – Building a data warehouse is similar to building a car. A car must be carefully designed from the beginning to meet the purposes for which it is intended. Yet, there are options each buyer must consider to make the vehicle truly meet individual performance needs. A data warehouse must also be carefully designed to meet overall performance requirements. While the final product can be customized to fit the performance needs of the organization, the initial overall design must be carefully thought out to provide a stable foundation from which to start.

12. **Designing the Data Warehouse** – People generally don't want to "waste" their time defining the requirements necessary for proper data warehouse design. Usually, there is a high level perception of what they want out of a data warehouse. However, they don't fully understand all the implications of these perceptions and, therefore, have a difficult time adequately defining them. This results in miscommunication between the business users and the technicians building the data warehouse. The typical end result is a data warehouse which does not deliver the results expected by the user. Since the data warehouse is inadequate for the end user, there is a need for fixes and improvements immediately after initial delivery.

13. **User Acceptance** – People are not keen to changing their daily routine especially if the new process is not intuitive. There are many challenges to overcome to make a data warehouse that is quickly adopted by an organization.

14. **Cost** – A frequent misconception among credit unions is that they can build data warehouse in-house to save money.. The harsh reality is an effective do-it-yourself effort is very costly.

Applications of DWH

There are three kinds of data warehouse applications: information processing, analytical processing, and data mining.

1) Information processing supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts, or graphs. A current trend in data warehouse information processing is to construct low-cost web-based accessing tools that are then integrated with web browsers.

2) Analytical processing supports basic OLAP operations, including slice-and-dice, drill-down, roll-up, and pivoting. It generally operates on historic data in both summarized and detailed forms. The major strength of online analytical processing over information processing is the multidimensional data analysis of data warehouse data.

3) Data mining supports knowledge discovery by finding hidden patterns and associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools.

Different areas are:

- **Banking Industry**
  - In the banking industry, concentration is given to risk management and policy reversal as well analyzing consumer data, market trends, government regulations and reports, and more importantly financial decision making.
  - Certain banking sectors utilize them for market research, performance analysis of each product, interchange and exchange rates, and to develop marketing programs.
  - Analysis of card holder's transactions, spending patterns and merchant classification, all of which provide the bank with an opportunity to introduce special offers and lucrative deals based on cardholder activity.
- **Finance Industry**
  - Revolve around evaluation and trends of customer expenses which aids in maximizing the profits earned by their clients.
- **Consumer Goods Industry**

- They are used for prediction of consumer trends, inventory management, market and advertising research.
- In-depth analysis of sales and production is also carried out.

- **Government and Education**
  - The federal government utilizes the warehouses for research in compliance, whereas the state government uses it for services related to human resources like recruitment, and accounting like payroll management.
  - The government uses data warehouses to maintain and analyze tax records, health policy records and their respective providers.
  - Criminal law database is connected to the state's data warehouse. Criminal activity is predicted from the patterns and trends, results of the analysis of historical data associated with past criminals.
  - Universities use warehouses for extracting of information used for the proposal of research grants, understanding their student demographics, and human resource management.
- **Healthcare**
  - All of their financial, clinical, and employee records are fed to warehouses as it helps them to strategize and predict outcomes, track and analyze their service feedback, generate patient reports, share data with tie-in insurance companies, medical aid services, etc.
- **Hospitality Industry**
  - A major proportion of this industry is dominated by hotel and restaurant services, car rental services, and holiday home services.
  - They utilize warehouse services to design and evaluate their advertising and promotion campaigns where they target customers based on their feedback and travel patterns.
- **Insurance**
  - The warehouses are primarily used to analyze data patterns and customer trends, apart from maintaining records of already existing participants.
  - The design of tailor-made customer offers and promotions is also possible through warehouses.
- **Manufacturing and Distribution Industry**

- A manufacturing organization has to take several make-or-buy decisions which can influence the future of the sector, which is why they utilize high-end OLAP tools as a part of data warehouses to predict market changes, analyze current business trends, detect warning conditions, view marketing developments, and ultimately take better decisions.
- They also use them for product shipment records, records of product portfolios, identify profitable product lines, analyze previous data and customer feedback to evaluate the weaker product lines and eliminate them.
- For the distributions, the supply chain management of products operates through data warehouses.

- **The Retailers**
  - Retailers serve as middlemen between producers and consumers.
  - They use warehouses to track items, their advertising promotions, and the consumers buying trends.
  - They also analyze sales to determine fast selling and slow selling product lines and determine their shelf space through a process of elimination.
- **Telephone Industry**
  - The telephone industry operates over both offline and online data burdening them with a lot of historical data which has to be consolidated and integrated.
  - Analysis of fixed assets, analysis of customer's calling patterns for sales representatives to push advertising campaigns, and tracking of customer queries, all require the facilities of a data warehouse.
- **Transportation Industry**
  - In the transportation industry, data warehouses record customer data enabling traders to experiment with target marketing where the marketing campaigns are designed by keeping customer requirements in mind.
  - To analyze customer feedback, performance, manage crews on board as well as analyze customer financial reports for pricing strategies.

****

try it now

A KTU
STUDENTS
PLATFORM

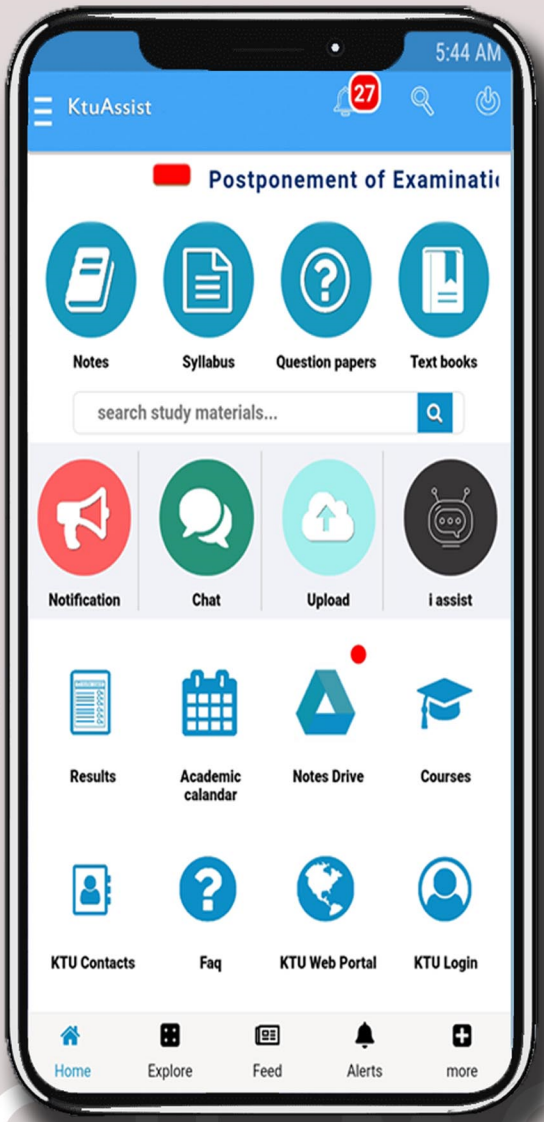
SYLLABUS

NOTES

TEXT BOOKS

QUESTION PAPERS

KTU NOTIFICATION



DOWNLOAD
IT
FROM
GOOGLE PLAY

CHAT

A FAQ

LOGIN

CALENDAR

MUCH MORE

DOWNLOAD APP

ktuassist.in

instagram.com/ktu_assist

facebook.com/ktuassist