Concept hierarchy:-

A concept hierarchy defines a sequence of mappings from a set of lowlevel concepts to higher-level, more general concepts. Consider a concept hierarchy for the dimension location. • City values for location include Vancouver, Toronto, New York and Chicago. Each city, however, can be mapped to the province or state to which it belongs.

“How are concept hierarchies useful in OLAP?”

• In the multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by concept hierarchies. • Let’s look at some typical OLAP operations for multidimensional data. The cube contains the dimensions location, time, and item, where location is aggregated with respect to city values, time is aggregated with respect to quarters, and item is aggregated with respect to item types.

Roll-Up •

The roll-up operation (also called the drill-up operation) performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction.

Drill-down

is the reverse of roll-up. It navigates from less detailed data to more detailed data. Drill-down can be realized by either stepping down a concept hierarchy for a dimension or introducing additional dimensions.

The slice operation

The slice operation performs a selection on one dimension of the given cube, resulting in a subcube.

The dice operation defines a subcube by performing a selection on two or more dimensions.

Pivot (also called rotate) is a visualization operation that rotates the data axes in view in order to provide an alternative presentation of the data.

A Starnet Query Model for Querying Multidimensional Databases: The querying of multidimensional databases can be based on a starnet model. A starnet model consists of radial lines emanating from a central point, where each line represents a concept hierarchy for a dimension

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.
Contains historical data.	Contains current data.
Provides summarized and multidimensional consolidated data.	Provides primitive and highly detailed data.
Number or users is in hundreds.	Number of users is in thousands.
Number of records accessed is in millions.	Number of records accessed is in tens.
Database size is from 100 GB to 1 TB	Database size is from 100 MB to 1 GB.
Highly flexible.	Provides high performance.
Based on Star Schema, Snowflake, Schema and Fact Constellation Schema.	Based on Entity Relationship Model.

Multidimensional model:

• A data cube allows data to be modeled and viewed in multiple dimensions It is defined by dimensions and facts.

A multidimensional model views data in the form of a data-cube. A data cube enables data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.

The dimensions are the perspectives or entities concerning which an organization keeps records.

A multidimensional data model is organized around a central theme, for example, sales. This theme is represented by a fact table. Facts are numerical measures. The fact table contains the names of the facts or measures of the related dimensional tables.

Types of scemmas:

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.

Simplicity and Understandability:

- Star schemas are easy to understand and navigate, even for non-technical users. The structure resembles a star, with a central fact table surrounded by dimension tables, making it intuitive and straightforward.

Scalability:

- Star schemas are scalable. You can easily add new dimensions or fact tables without significantly impacting existing structures or performance. This flexibility is essential for accommodating evolving business requirements.

asier ETL Processes:

- Extract, Transform, Load (ETL) processes are simplified in star schemas. Data can be extracted from various sources, transformed into the required format, and loaded into the fact and dimension tables without the need for complex transformations within the database

Data Quality and Consistency:

- By centralizing dimensions in dimension tables, data quality and consistency can be maintained more effectively. Updates and corrections to dimension attributes only need to be made in one place, ensuring data accuracy.

Disadvantage:

Data Redundancy: In a star schema, dimension tables are denormalized, meaning that some data may be duplicated across multiple dimension tables. This redundancy can lead to increased storage requirements and potential data inconsistencies if not managed properly.

Maintenance Complexity: Because dimension tables are denormalized, any changes to the source data or business requirements may require updates to multiple dimension tables. This can increase the complexity of data maintenance and ETL (Extract, Transform, Load) processes.

Slower ETL Processes: The denormalization of dimension tables can lead to slower ETL processes, as there may be more data to transform and load into the data warehouse.

Snow flake scheam

A snowflake schema is equivalent to the star schema. "A schema is known as a snowflake if one or more dimension tables do not connect directly to the fact table but must join through other dimension tables.

Fact constalllion:

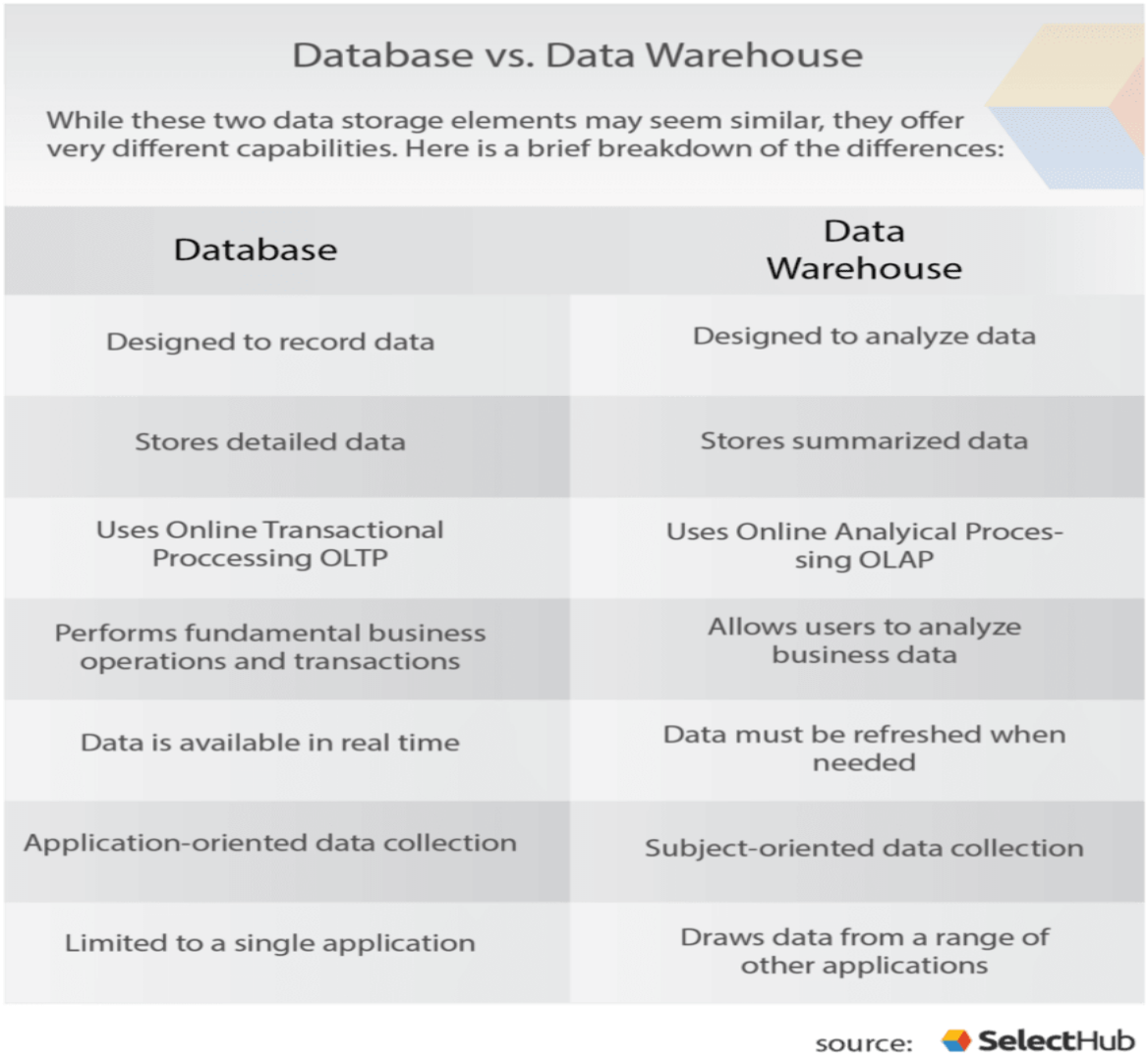
Star Schema	Snowflake Schema
In star schema, fact tables and the dimension tables are present.	In snowflake schema, fact tables, dimension tables, as well as sub-dimension tables are present.
This is a top-down model i.e. seeks to identify the big picture and all of its components	It is a bottom-up model i.e. first focuses on solving the smaller problems at the fundamental level and then integrating them into a whole and complete solution
Star schema requires more space	The space requirement is lesser.
It takes less time for the execution of queries	It takes more time than star schema for the execution of queries
In star schema, normalized data is not used	Both normalized and denormalized data are used
It's design is very simple	It's design is complex.
The complexity of the queries is low.	The complexity of queries is greater.

Data warehouse:

Subject-Oriented Reference No.: R1, R2, R3 • A data warehouse is subject oriented as it offers information regarding a theme instead of companies' ongoing operations. These subjects can be sales, marketing, distributions, etc. • A data warehouse never focuses on the ongoing operations. Instead, it put emphasis on modeling and analysis of data for decision making. It also provides a simple and concise view around the specific subject by excluding data which not helpful to support the decision process.

Intregretaed:

The data also needs to be stored in the Data warehouse in common and universally acceptable manner. • A data warehouse is developed by integrating data from varied sources like a mainframe, relational databases, flat files, etc.



Generally a data warehouses adopts a three-tier architecture. Following are the three tiers of the data warehouse architecture. •Bottom Tier − The bottom tier of the architecture is the data warehouse database server. It is the relational database system. We use the back end tools and utilities to feed data into the bottom tier. These back end tools and utilities perform the Extract, Clean, Load, and refresh functions. •Middle Tier − In the middle tier, we have the OLAP Server that can be implemented in either of the following ways. • By Relational OLAP (ROLAP), which is an extended relational database management system. The ROLAP maps the operations on multidimensional data to standard relational operations. • By Multidimensional OLAP (MOLAP) model, which directly implements the multidimensional data and operations. •Top-Tier − This tier is the front-end client layer. This layer holds the query tools and reporting tools, analysis tools and data mining tools.ss

Data Warehouse Architecture There are three ways you can construct a data warehouse system. These approaches are classified by the number of tiers in the architecture. Therefore, you can have a: •Single-tier architecture •Two-tier architecture •Three-tier architecture

Single-tier Data Warehouse Architecture The single-tier architecture is not a frequently practiced approach. The main goal of having such an architecture is to remove redundancy by minimizing the amount of data stored.

A two-tier architecture includes a staging area for all data sources, before the data warehouse layer. By adding a staging area between the sources and the storage repository, you ensure all data loaded into the warehouse is cleansed and in the appropriate format.

Data preprocessing:

Data preprocessing is an important step in the data mining process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.

Data preprocessing is an important step in the data mining process that involves cleaning and transforming raw data to make it suitable for analysis. Some common steps in data preprocessing include:

Data Cleaning: This involves identifying and correcting errors or inconsistencies in the data, such as missing values, outliers, and duplicates. Various techniques can be used for data cleaning, such as imputation, removal, and transformation.
Data Integration: This involves combining data from multiple sources to create a unified dataset. Data integration can be challenging as it requires handling data with different formats, structures, and semantics. Techniques such as record linkage and data fusion can be used for data integration.

.**data reduction technique**:

Data reduction strategies include ¬ dimensionality reduction, ¬ numerosity reduction ¬ data compression

Data mining is the process of sorting through large data sets to identify patterns and relationships that can help solve business problems through data analysis..

Data Mining in Healthcare:

Data mining in healthcare has excellent potential to improve the health system. It uses data and analytics for better insights and to identify best practices that will enhance health care services and reduce costs.

Data Mining in Market Basket Analysis:

Market basket analysis is a modeling method based on a hypothesis. If you buy a specific group of products, then you are more likely to buy another group of products. This technique may enable the retailer to understand the purchase behavior of a buyer.

Data mining in Education:

Education data mining is a newly emerging field, concerned with developing techniques that explore knowledge from the data generated from educational Environments.

Data Mining Financial Banking:

The Digitalization of the banking system is supposed to generate an enormous amount of data with every new transaction. The data mining technique can help bankers by solving business-related problems in banking and finance by identifying trends, casualties, and correlations in business information and market costs that are not instantly evident to managers or executives because the data volume is too large or are produced too rapidly on the screen by experts.

Data mining task:

Data mining tasks are specific objectives or operations performed using data mining techniques and algorithms to discover valuable and actionable patterns, relationships, or knowledge from large datasets.

Data entries can be associated with classes or concepts.

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts.

what are major issues in data mining system

ChatGPT

Data mining systems are powerful tools for extracting valuable insights and knowledge from large datasets, but they also face several major issues and challenges. Some of the key issues in data mining systems include:

1. **Data Quality:** Poor data quality, which includes missing values, inaccuracies, inconsistencies, and outliers, can significantly impact the results of data mining. Data cleansing and preprocessing are crucial steps in addressing this issue.
2. **Data Privacy and Security:** Privacy concerns arise when dealing with sensitive or personal data. Data mining systems must adhere to privacy regulations and implement secure data handling and access controls to protect sensitive information.
3. **Scalability:** As datasets continue to grow in size, scalability becomes a major challenge. Data mining algorithms and systems need to handle large volumes of data efficiently. Distributed and parallel computing techniques are often employed to address this issue.
4. **High Dimensionality:** High-dimensional data can lead to the "curse of dimensionality," making it challenging to find meaningful patterns and relationships. Dimensionality reduction techniques are used to mitigate this problem.

5. **Complexity of Algorithms:** Many data mining algorithms are computationally intensive and complex. Balancing accuracy with computational efficiency is often a trade-off, and choosing the right algorithm for a specific task can be challenging.
6. Lack of Domain Knowledge: Effective data mining often requires domain expertise to interpret results correctly and make informed decisions. Without domain knowledge, meaningful insights may be missed.
7.

KDD Process

KDD (Knowledge Discovery in Databases) is a process that involves the extraction of useful, previously unknown, and potentially valuable information from large datasets. The KDD process is an iterative process and it requires multiple iterations of the above steps to extract accurate knowledge from the data.The following steps are included in KDD process:

Data Cleaning

Data cleaning is defined as removal of noisy and irrelevant data from collection.

1. Cleaning in case of **Missing values**.
2. Cleaning **noisy** data, where noise is a random or variance error.
3. Cleaning with **Data discrepancy detection** and **Data transformation tools**.

Data Integration

Data integration is defined as heterogeneous data from multiple sources combined in a common source(DataWarehouse). Data integration using **Data Migration tools, Data Synchronization tools and ETL**(Extract-Load-Transformation) process.

Data Selection

Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection. For this we can use **Neural network, Decision Trees, Naive bayes, Clustering**, and **Regression** methods.

Data Transformation

Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure. Data Transformation is a two step process:

1. **Data Mapping**: Assigning elements from source base to destination to capture transformations.
2. **Code generation**: Creation of the actual transformation program.

Data Mining

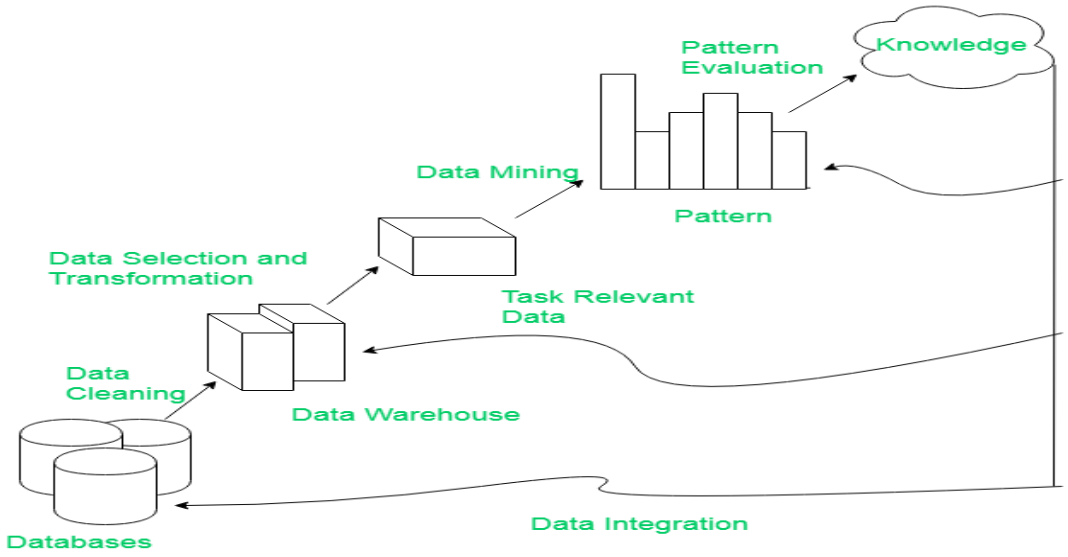
Data mining is defined as techniques that are applied to extract patterns potentially useful. It transforms task relevant data into **patterns, and d**ecides purpose of model using **classification** or **characterization**.

Pattern Evaluation

Pattern Evaluation is defined as identifying strictly increasing patterns representing knowledge based on given measures. It find **interestingness score** of each pattern, and uses **summarization** and **Visualization** to make data understandable by user.

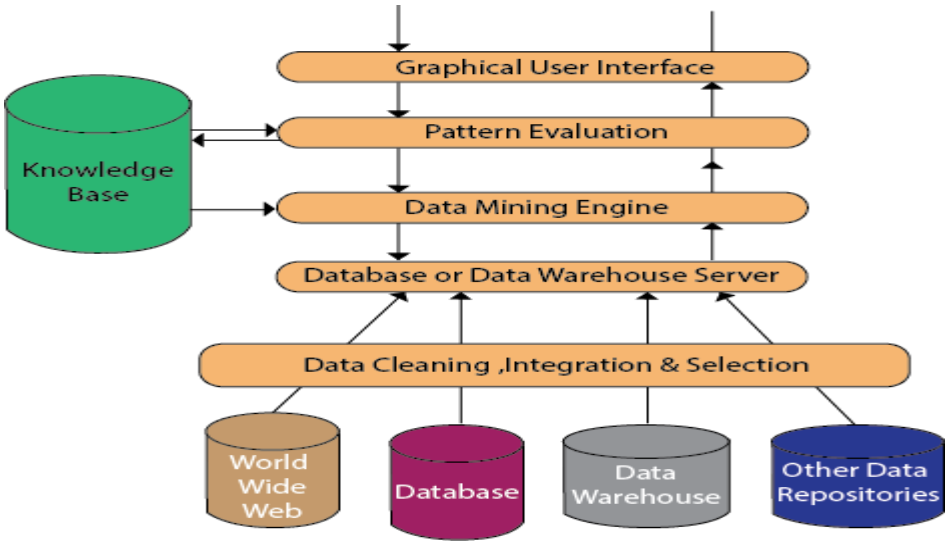
Knowledge Representation

This involves presenting the results in a way that is meaningful and can be used to make decisions.



Data Mining Architecture

The significant components of data mining systems are a data source, data mining engine, data warehouse server, the pattern evaluation module, graphical user interface, and knowledge base.



Data Source:

The actual source of data is the Database, data warehouse, World Wide Web (WWW), text files, and other documents. You need a huge amount of historical data for data mining to be successful. Organizations typically store data in databases or data warehouses. Data warehouses may comprise one or more databases, text files

spreadsheets, or other repositories of data. Sometimes, even plain text files or spreadsheets may contain information. Another primary source of data is the World Wide Web or the internet.

Different processes:

Before passing the data to the database or data warehouse server, the data must be cleaned, integrated, and selected. As the information comes from various sources and in different formats, it can't be used directly for the data mining procedure because the data may not be complete and accurate. So, the first data requires to be cleaned and unified. More information than needed will be collected from various data sources, and only the data of interest will have to be selected and passed to the server. These procedures are not as easy as we think. Several methods may be performed on the data as part of selection, integration, and cleaning.

Database or Data Warehouse Server:

The database or data warehouse server consists of the original data that is ready to be processed. Hence, the server is cause for retrieving the relevant data that is based on data mining as per user request.

Data Mining Engine:

The data mining engine is a major component of any data mining system. It contains several modules for operating data mining tasks, including association, characterization, classification, clustering, prediction, time-series analysis, etc.

In other words, we can say data mining is the root of our data mining architecture. It comprises instruments and software used to obtain insights and knowledge from data collected from various data sources and stored within the data warehouse.

Pattern Evaluation Module:

The Pattern evaluation module is primarily responsible for the measure of investigation of the pattern by using a threshold value. It collaborates with the data mining engine to focus the search on exciting patterns.

This segment commonly employs stake measures that cooperate with the data mining modules to focus the search towards fascinating patterns. It might utilize a stake threshold to filter out discovered patterns. On the other hand, the pattern evaluation module might be coordinated with the mining module, depending on the implementation of the data mining techniques used. For efficient data mining, it is abnormally suggested to push the evaluation of pattern stake as much as possible into the mining procedure to confine the search to only fascinating patterns.

Graphical User Interface:

The graphical user interface (GUI) module communicates between the data mining system and the user. This module helps the user to easily and efficiently use the system without knowing the complexity of the process. This module cooperates with the data mining system when the user specifies a query or a task and displays the results.

Knowledge Base:

The knowledge base is helpful in the entire process of data mining. It might be helpful to guide the search or evaluate the stake of the result patterns. The knowledge base may even contain user views and data from user experiences that might be helpful in the data mining process. The data mining engine may receive inputs from the knowledge base to make the result more accurate and reliable. The pattern assessment module regularly interacts with the knowledge base to get inputs, and also update it.