3.6 Data Mart and Data Mining

3.6.1 Data Mart:

A data mart is a simple form of a data warehouse that is focused on a single subject (or functional area), such as sales, finance or marketing. Data marts are often built and controlled by a single department within an organization. Given their single-subject focus, data marts usually draw data from only a few sources. The sources could be internal operational systems, a central data warehouse, or external data.

3.6.2 Dependent and Independent Data Marts

There are two basic types of data marts: dependent and independent. The categorization is based primarily on the data source that feeds the data mart. Dependent data marts draw data from a central data warehouse that has already been created. Independent data marts, in contrast, are standalone systems built by drawing data directly from operational or external sources of data, or both.

The main difference between independent and dependent data marts is how you populate the data mart; that is, how you get data out of the sources and into the data mart. This step, called the Extraction-Transformation-and Loading (ETL) process, involves moving data from operational systems, filtering it, and loading it into the data mart. With dependent data marts, this process is somewhat simplified because formatted and summarized (clean) data has already been loaded into the central data warehouse. The ETL process for dependent data marts is mostly a process of identifying the right subset of data relevant to the chosen data mart subject and moving a copy of it, perhaps in a summarized form.

With independent data marts, however, you must deal with all aspects of the ETL process, much as you do with a central data warehouse. The number of sources is

likely to be fewer and the amount of data associated with the data mart is less than the warehouse, given your focus on a single subject. The motivations behind the creation of these two types of data marts are also typically different.

Dependent data marts are usually built to achieve improved performance and availability, better control, and lower telecommunication costs resulting from local access of data relevant to a specific department. The creation of independent data marts is often driven by the need to have a solution within a shorter time.

3.6.3 Steps in Implementing a Data Mart

Simply stated, the major steps in implementing a data mart are to design the schema, construct the physical storage, populate the data mart with data from source systems, access it to make informed decisions, and manage it over time.

- Designing
- Constructing
- Populating
- Accessing
- Managing

1. **Designing**

The design step is first in the data mart process. This step covers all of the tasks from initiating the request for a data mart through gathering information about the requirements, and developing the logical and physical design of the data mart. The design step involves the following tasks:

- Gathering the business and technical requirements
- Identifying data sources
- Selecting the appropriate subset of data
- Designing the logical and physical structure of the data mart

2. Constructing

This step includes creating the physical database and the logical structures associated with the data mart to provide fast and efficient access to the data. This step involves the following tasks:

 Creating the physical database and storage structures, such as tablespaces, associated with the data mart

 Creating the schema objects, such as tables and indexes defined in the design step

 Determining how best to set up the tables and the access structures

3. **Populating**

The populating step covers all of the tasks related to getting the data from the source, cleaning it up, modifying it to the right format and level of detail, and moving it into the data mart. More formally stated, the populating step involves the following tasks:

 Mapping data sources to target data structures

 Extracting data

 Cleansing and transforming the data

 Loading data into the data mart

 Creating and storing metadata

4. Accessing

The accessing step involves putting the data to use: querying the data, analyzing it, creating reports, charts, and graphs, and publishing these. Typically, the end user uses a graphical front-end tool to submit queries to the database and display the results of the queries. The accessing step requires that you perform the following tasks:

 Set up an intermediate layer for the front-end tool to use. This layer, the meta layer, translates database structures and object names into business terms, so that the end user can interact with the data mart using terms that relate to the business function. Maintain and manage these business interfaces. Set up and manage database structures, like summarized tables, that help queries submitted through the front-end tool execute quickly and efficiently.

5. Managing

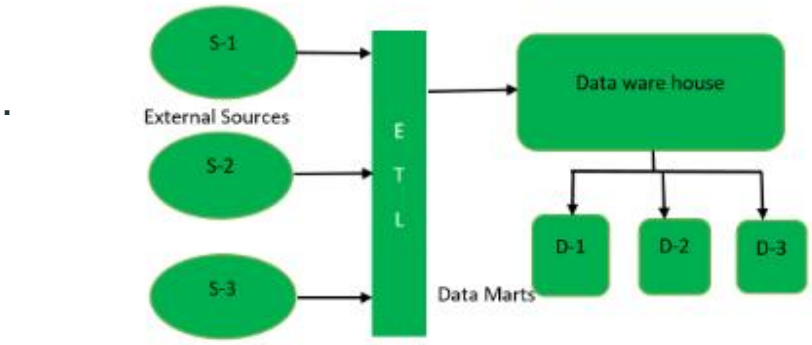
This step involves managing the data mart over its lifetime. In this step, you perform management tasks such as the following: Providing secure access to the data Managing the growth of the data Optimizing the system for better performance Ensuring the availability of data even with system failures

3.6.4 Data Mart issues

 Data mart functionality - the capabilities of data marts have increased with the growth in their popularity

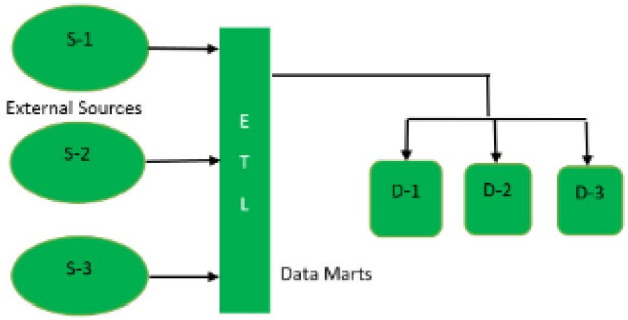
 Data mart size - the performance deteriorates as data marts grow in size, so need to reduce the size of data marts to gain improvements in performance

 Data mart load performance - two critical components: end-user response time and data loading performance - to increment DB updating so that only cells affected by the change are updated and not the entire MDDB structure.



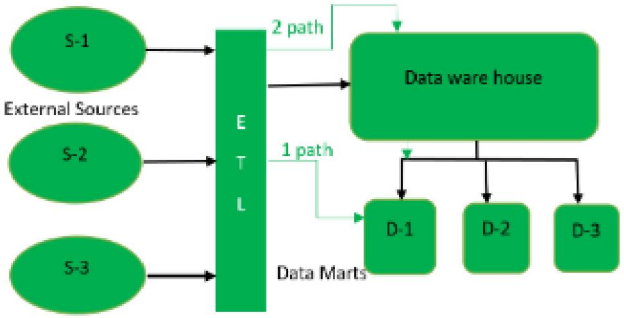
Dependent Data Mart

Dependent Data Mart is created by extracting the data from central repository, Datawarehouse. First data warehouse is created by extracting data (through ETL tool) from external sources and then data mart is created from data warehouse. Dependent data mart is created in top-down approach of data warehouse architecture. This model of data mart is used by big organizations.



Independent Data Mart

Independent Data Mart is created directly from external sources instead of data warehouse. First data mart is created by extracting data from external sources and then datawarehouse is created from the data present in data mart. Independent data mart is designed in bottom-up approach of datawarehouse architecture. This model of data mart is used by small organizations and is cost effective comparatively.



Hybrid Data Mart

This type of Data Mart is created by extracting data from operational source or from data warehouse. 1Path reflects accessing data directly from external sources and 2Path reflects dependent data model of data mart.

3.6.5 Advantages of Data Mart:

1. Data marts are designed to serve the specific needs of a particular business unit

2. Since data marts are smaller subsets of data warehouses, they can often deliver faster query performance
3. Data marts are typically built with the end-user in mind. They are tailored to the specific requirements of a department or team, making it easier for non-technical users to access and analyze the data relevant to their needs
4. Data marts can be implemented more quickly compared to large-scale data warehouses.
5. By focusing on a specific subject area or business process, data marts can enforce data quality standards more effectively.
6. Each data mart can operate independently, allowing departments or business units to have greater control over their data and analytics processes.

3.6.6 Disadvantages of Data Mart:

1. Data marts, if not properly integrated, can lead to data silos within an organization.
2. When different data marts are developed independently, there's a risk of inconsistency in data definitions and metrics.
3. While the focus on a specific business area is an advantage, it can also be a limitation.
4. In a decentralized environment, there's a possibility of redundant data existing across multiple data marts. This redundancy can lead to increased storage requirements
5. Connecting data marts to each other or to a central data warehouse can be complex
6. Data marts may inadvertently duplicate data that already exists in a central data warehouse.

3.6.7 Applications of Data Mart:

Data marts are specialized subsets of data warehouses that are designed to serve the specific analytical needs of a particular business unit or department within an organization. They find applications in various domains, helping organizations make informed decisions and derive valuable insights. Here are some common applications of data marts:

1. **Sales and Marketing Analytics:**
    - Analyzing customer behavior, preferences, and purchasing patterns.
    - Evaluating the effectiveness of marketing campaigns.
    - Monitoring sales performance and forecasting future trends.
2. **Finance and Accounting**:
    - Financial reporting and analysis.
    - Budgeting and forecasting.
    - Compliance reporting and auditing.
3. **Human Resources**:
    - Workforce analytics, including employee performance and retention.
    - HR reporting for monitoring recruitment, training, and employee satisfaction.
    - Benefits administration and compensation analysis.
4. **Supply Chain and Operations**:
    - Inventory management and optimization.
    - Supply chain analytics for improving efficiency and reducing costs.
    - Monitoring production and distribution processes.
5. **Customer Relationship Management (CRM):**
    - Customer segmentation and profiling.
    - Tracking customer interactions and feedback.
    - Analyzing customer lifetime value and churn rates.
6. **Healthcare Analytics:**
    - Patient outcomes analysis and medical research.
    - Healthcare cost analysis and resource optimization.
    - Monitoring and improving patient care quality.
7. **Retail Analytics**:
    - Inventory management and demand forecasting.
    - Point-of-sale (POS) analysis for sales trends.
    - Customer experience and satisfaction analysis.

8. **Risk Management:**
    - Fraud detection and prevention.
    - Credit risk analysis in financial institutions.
    - Compliance reporting to meet regulatory requirements.
9. **E-commerce:**
    - Website analytics and user behavior analysis.
    - Personalized recommendations for customers.
    - Order fulfilment and logistics optimization.
10. **Education Analytics:**
    - Student performance analysis and predictive modelling.
    - Enrolment management and admissions analytics.
    - Educational resource allocation and planning.
11. **Government and Public Sector**:
    - Public safety analytics for law enforcement.
    - Budget and expenditure analysis.
    - Social program effectiveness evaluation.
12. **Manufacturing Analytics:**
    - Quality control and defect analysis.
    - Equipment maintenance and performance monitoring.
    - Production efficiency and yield optimization.

These applications demonstrate the versatility of data marts in addressing specific business needs within different functional areas. By providing a focused and tailored view of data, data marts empower decision-makers in various departments to extract meaningful insights and drive improvements in their respective domains.

3.6.8 DATA MINING:

- Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses.

- Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems.
- Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.
- Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on- line.
- This white paper provides an introduction to the basic technologies of data mining. Examples of profitable applications illustrate its relevance to today‘s business environment as well as a basic description of how data warehouse architectures can evolve to deliver the value of data mining to end users.

3.6.9 The Foundations of Data Mining:

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time.

Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

 Massive data collection

 Powerful multiprocessor computers

 Data mining algorithms

- Commercial databases are growing at unprecedented rates. A recent META Group survey of data warehouse projects found that 19% of respondents are beyond the

50 giga byte level, while 59% expect to be there by second quarter of 1996.In some industries, such as retail, these numbers can be much larger.

- The accompanying need for improved computational engines can now be met in a cost-effective manner with parallel multiprocessor computer technology. Data mining algorithms embody techniques that have existed for at least 10 years, but have only recently been implemented as mature, reliable, understandable tools that consistently outperform older statistical methods.

Table 1. Steps in the Evolution of Data Mining.

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	Prospective, proactive information delivery

- In the evolution from business data to business information, each new step has built upon the previous one. For example, dynamic data access is critical for drill-through in data navigation applications, and the ability to store large databases is critical to data mining. From the user‘s point of view, the four steps listed in Table 1 were revolutionary because they allowed new business questions to be answered accurately and quickly.
- The core components of data mining technology have been under development for decades, in research areas such as statistics, artificial intelligence, and machine learning. Today, the maturity of these techniques, coupled with high-performance relational database engines and broad data integration efforts, make these technologies practical for current data warehouse environments.

3.6.10 The Scope of Data Mining

Data mining derives its name from the similarities between searching for valuable business information in a large database for example, finding linked products in gigabytes of store scanner data and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

- Automated prediction of trends and behaviors. Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

- Automated discovery of previously unknown patterns. Data mining tools sweep through databases and identify previously hidden patterns in one step. An example

of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Data mining techniques can yield the benefits of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed. When data mining tools are implemented on high performance parallel processing systems, they can analyze massive databases in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions.

3.6.11 Databases can be larger in both depth and breadth:

- More columns. Analysts must often limit the number of variables they examine when doing hands-on analysis due to time constraints. Yet variables that are discarded because they seem unimportant may carry information about unknown patterns. High performance data mining allows users to explore the full depth of a database, without preselecting a subset of variables.
- More rows. Larger samples yield lower estimation errors and variance, and allow users to make inferences about small but important segments of a population.

A recent Gartner Group Advanced Technology Research Note listed data mining and artificial intelligence at the top of the five key technology areas that "will clearly have a major impact across a wide range of industries within the next 3 to 5 years. "Gartner also listed parallel architectures and data mining as two of the top 10 new technologies in which companies will invest during the next 5 years.

According to a recent Gartner HPC Research Note, "With the rapid advance in data capture, transmission and storage, large-systems users will increasingly need to implement new and innovative ways to mine the after-market value of their vast stores of

detail data, employing MPP [massively parallel processing] systems to create new sources of business advantage (0.9 probability)."

3.6.12 Techniques in data mining

1. Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
2. Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).
3. Genetic algorithms: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
4. Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset. Sometimes called the k-nearest neighbour technique.
5. Rule induction: The extraction of useful if-then rules from data based on statistical significance.

   Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms. The appendix to this white paper provides a glossary of data mining terms.

3.6.11 How Data Mining Works

How exactly is data mining able to tell you important things that you didn't know or what is going to happen next? The technique that is used to perform these feats in data mining is called modeling. Modeling is simply the act of building a model in one situation where you know the answer and then applying it to another situation that you don't. For instance, if you were looking for a sunken Spanish galleon on the high seas the first thing you might do is to research the times when Spanish treasure had been found by others

in the past. You might note that these ships often tend to be found off the coast of Bermuda and that there are certain characteristics to the ocean currents, and certain routes that have likely been taken by the ship‘s captains in that era. You note these similarities and build a model that includes the characteristics that are common to the locations of these sunken treasures. With these models in hand you sail off looking for treasure where your model indicates it most likely might be given a similar situation in the past. Hopefully, if you've got a good model, you find your treasure.

This act of model building is thus something that people have been doing for a long time, certainly before the advent of computers or data mining technology. What happens on computers, however, is not much different than the way people build models. Computers are loaded up with lots of information about a variety of situations where an answer is known and then the data mining software on the computer must run through that data and distill the characteristics of the data that should go into the model. Once the model is built it can then be used in similar situations where you don't know the answer. For example, say that you are the director of marketing for a telecommunications company and you'd like to acquire some new long-distance phone customers. You could just randomly go out and mail coupons to the general population - just as you could randomly sail the seas looking for sunken treasure. In neither case would you achieve the results you desired and of course you have the opportunity to do much better than random - you could use your business experience stored in your database to build a model.

As the marketing director you have access to a lot of information about all of your customers: their age, sex, credit history and long-distance calling usage. The good news is that you also have a lot of information about your prospective customers: their age, sex, credit history etc. Your problem is that you don't know the long-distance calling usage of these prospects (since they are most likely now customers of your competition). You'd like to concentrate on those prospects that have large amounts of long-distance usage.

- The goal in prospecting is to make some calculated guesses about the information in the lower right hand quadrant based on the model that we build going from

Customer General Information to Customer Proprietary Information. For instance, a simple model for a telecommunications company might be:

- 98% of my customers who make more than $60,000/year spend more than $80/month on long distance. This model could then be applied to the prospect data to try to tell something about the proprietary information that this telecommunications company does not currently have access to. With this model in hand new customers can be selectively targeted.
- Test marketing is an excellent source of data for this kind of modeling. Mining the results of a test market representing a broad but relatively small sample of prospects can provide a foundation for identifying good prospects in the overall market. Table 3 shows another common scenario for building models: predict what is going to happen in the future.

3.6.13 Architecture for Data Mining:

To best apply these advanced techniques, they must be fully integrated with a data warehouse as well as flexible interactive business analysis tools. Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. Furthermore, when new insights require operational implementation, integration with the warehouse simplifies the application of results from data mining. The resulting analytic data warehouse can be applied to improve business processes throughout the organization, in areas such as promotional campaign management, fraud detection, new product rollout, and so on. Figure illustrates architecture for advanced analysis in a large data warehouse.

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.

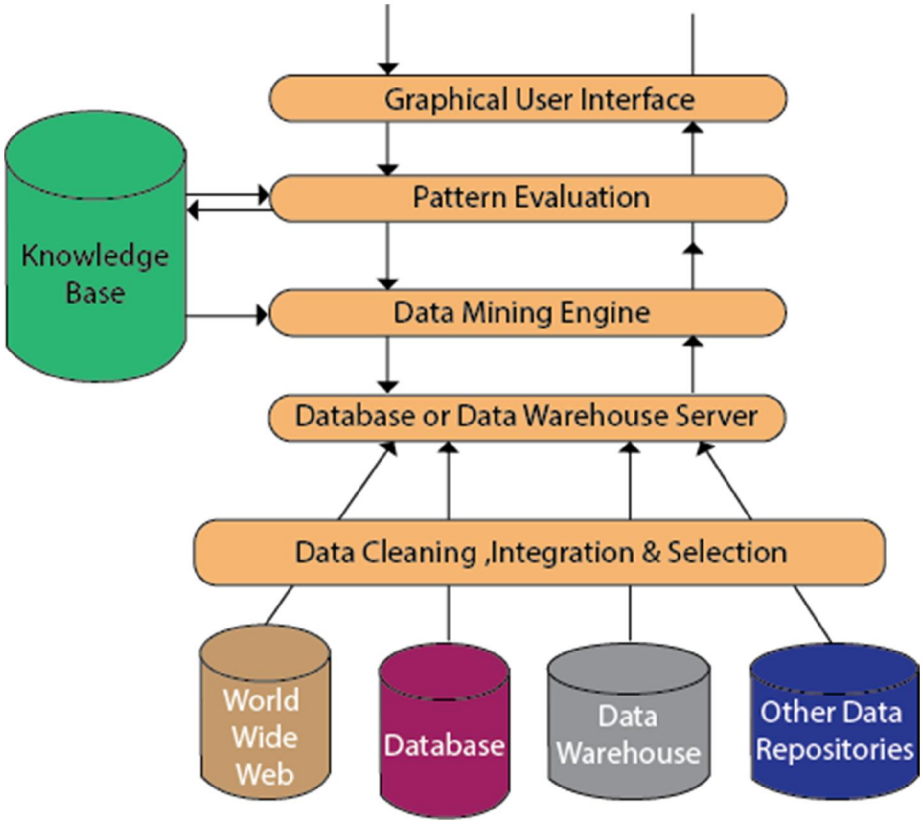
Data Source:

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.

3.6.14 Applications of Data Mining:

Data mining is a process of discovering patterns, trends, correlations, and insights from large datasets. It has numerous applications across various industries, contributing to improved decision-making, strategic planning, and business intelligence. Here are some common applications of data mining:

1. **Marketing and Customer Relationship Management (**CRM):
2. **Customer Segmentation**: Dividing customers into groups based on their behavior, preferences, or demographics for targeted marketing.
3. **Market Basket Analysis**: Identifying associations between products purchased together to enhance cross-selling and promotions.
4. **Customer Churn Prediction**: Predicting which customers are likely to leave or discontinue services.
5. **Finance and Banking**:
6. **Credit Scoring**: Assessing the creditworthiness of individuals or businesses.
7. **Fraud Detection**: Identifying unusual patterns that may indicate fraudulent activities.
8. **Risk Management**: Analyzing data to assess and mitigate financial risks.
9. **Healthcare**:
10. **Disease Prediction**: Predicting disease outbreaks or identifying individuals at risk based on health data.

11. **Clinical Decision Support**: Analyzing patient records to assist in diagnosis and treatment planning.
12. **Fraud Detection in Healthcare Billing**: Identifying fraudulent billing practices in healthcare insurance claims.
13. **Retail and E-commerce**:
14. **Price Optimization**: Analyzing customer behavior and market conditions to optimize pricing strategies.

    Inventory Management: Predicting demand, optimizing stock levels, and reducing excess inventory.

    Personalized Recommendations: Offering personalized product recommendations based on customer preferences and behavior.
15. **Manufacturing and Production**:

    Quality Control: Analyzing production data to identify and prevent defects.

    Predictive Maintenance: Predicting equipment failures to schedule maintenance proactively.

    Process Optimization: Identifying inefficiencies in manufacturing processes for improvement.
16. **Telecommunications**:

    Customer Usage Analysis: Analyzing usage patterns to offer personalized plans and services.

    Network Optimization: Identifying and resolving network issues for improved performance.

    Churn Prediction: Predicting customer churn and implementing strategies for retention.
17. **Education**:

    Student Performance Analysis: Analyzing academic data to identify trends and patterns related to student success.

    Educational Planning: Assisting in course scheduling, resource allocation, and curriculum development.

    Social Media and Web Analytics:

Sentiment Analysis: Analyzing social media and online content to understand public sentiment.

User Behavior Analysis: Understanding user interactions on websites for improving user experience.

Targeted Advertising: Utilizing user data to personalize and target online advertisements.

18. **Government and Public Services**:

    Crime Prediction and Prevention: Analyzing crime data to predict and prevent criminal activities.

    Public Health Surveillance: Monitoring and predicting disease outbreaks for timely intervention.

    Tax Fraud Detection: Identifying patterns indicative of tax fraud and evasion.

19. **Environmental Sciences:**

    Climate Modeling: Analyzing climate data to model and predict environmental changes.

20. **Natural Disaster Prediction**: Using data to predict and mitigate the impact of natural disasters.

    These applications showcase the diverse ways in which data mining techniques contribute to extracting valuable insights and knowledge from large datasets across different domains.

****