INTRODUCTION

Data warehousing is a collection of tools and techniques using which more knowledge can be driven out from a large amount of data. This helps with the decision-making process and improving information resources.

Data warehouse is basically a database of unique data structures that allows relatively quick and easy performance of complex queries over a large amount of data. It is created from multiple heterogeneous sources.

Data warehouses and databases both are relative data systems, but both are made to serve different purposes. A data warehouse is built to store a huge amount of historical data and empowers fast requests over all the data, typically using Online Analytical Processing (OLAP). A database is made to store current transactions and allow quick access to specific transactions for ongoing business processes, commonly known as Online Transaction Processing (OLTP).

Important Features of Data Warehouse

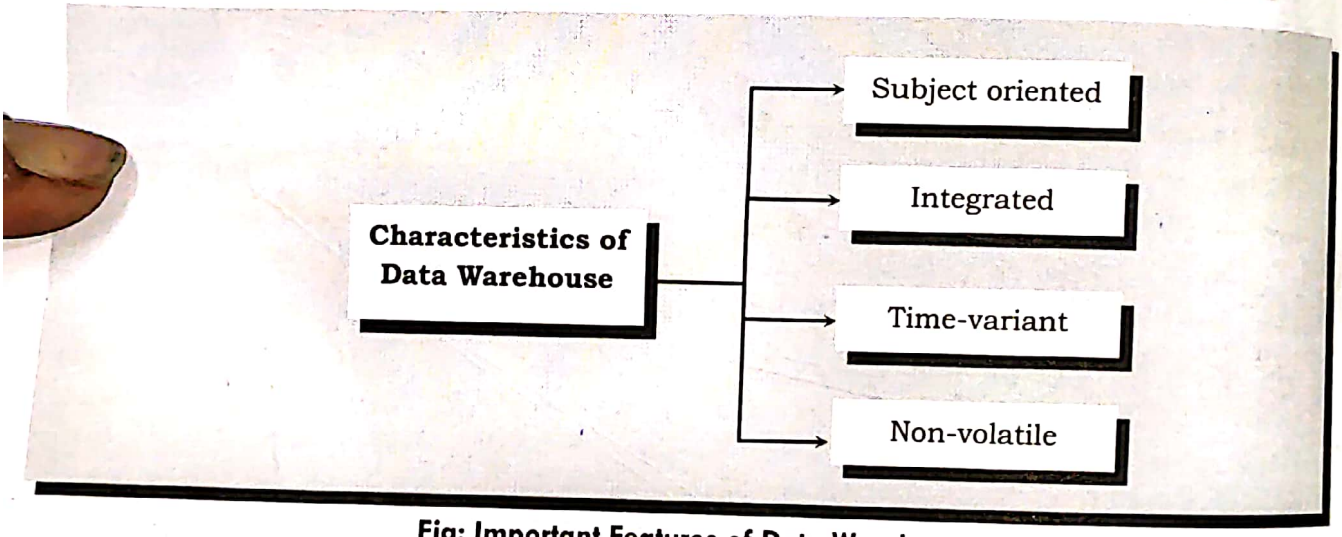


Fig: Important Features of Data Warehouse

The Important features of Data Warehouse are given below:

- ***Subject Oriented***

  A data warehouse is subject-oriented. It provides useful data about a subject instead of the company's ongoing operations, and these subjects can be customers, suppliers, marketing, product, promotion, etc. A data warehouse usually focuses on modeling and analysis of data that helps the business organization to make data-driven decisions.

Time-Variant

The different data present in the data warehouse provides information for a specific period. Historical data is kept in a data warehouse. For example, one can retrieve data from 3 months, 6 months, 12 months, or even older data from a data warehouse. This

contrasts with a transactions system, where often only the most recent data is kept. For example, a transaction system may hold the most recent address of a customer, where a data warehouse can hold all addresses associated with a customer.

3. **Integrated**

A data warehouse is built by joining data from heterogeneous sources, such as social databases, level documents, etc. A data warehouse is constructed by integrating data from heterogeneous sources such as relational databases, flat files, etc. This integration enhances the effective analysis of data.

4. **Non- Volatile**

It means, once data entered into the warehouse cannot be change. The data resided in data warehouse is permanent. It also means that data is not erased or deleted when new data is inserted. It includes the mammoth quantity of data that is inserted into modification between the selected quantity on logical business. It evaluates the analysis within the technologies of warehouse.

Advantages of Data Warehouse

1. **Delivers enhanced business intelligence**

By having access to information from various sources from a single platform, decision makers will no longer need to rely on limited data or their instinct. Additionally, data warehouses can effortlessly be applied to a business's processes, for instance, market segmentation, sales, risk, inventory, and financial management.

2. **Saves times**

A data warehouse standardizes, preserves, and stores data from distinct sources, aiding the consolidation and integration of all the data. Since critical data is available to all users, it allows them to make informed decisions on key aspects. In addition, executives can query the data themselves with little to no IT support, saving more time and money.

3. **Enhances data quality and consistency**

A data warehouse converts data from multiple sources into a consistent format. Since the data from across the organization is standardized, each department will produce results that are consistent. This will lead to more accurate data, which will become the basis for solid decisions.

4. **Generates a high Return on Investment (ROI)**

Companies experience higher revenues and cost savings than those that haven't invested in a data warehouse.

5. **Provides competitive advantage**

Data warehouses help get a holistic view of their current standing and evaluate opportunities and risks, thus providing companies with a competitive advantage.

6. **Improves the decision-making process**

Data warehousing provides better insights to decision makers by maintaining a cohesive database of current and historical data. By transforming data into purposeful information, decision makers can perform more functional, precise, and reliable analysis and create more useful reports with ease.

7. **Enables organizations to forecast with confidence**

Data professionals can analyze business data to make market forecasts, identify potential KPIs, and gauge predicated results, allowing key personnel to plan accordingly.

8. **Streamlines the flow of information**

Data warehousing facilitates the flow of information through a network connecting all related or non-related parties.

Applications of Data Warehousing

Data warehouse is widely used in the following fields.

- Financial services
- Banking services
- Consumer goals
- Retail sectors
- Controlled manufacturing
- Information Processing
- Analytical Processing
- Data Mining
- Real Life
- Various Industries
- Statistical Analysis
- Decision Making
- Mailing Box Applications

Database VS Data Warehouse

Database: Databases are real-time repositories of information, which are usually tied to specific applications. Data Warehouse: Data warehouses pull information from various sources (including databases), with a focus on the storage, filtering, retrieval and, specifically, analysis of huge volumes of structured data.

Data Warehouse (OLAP)	Operational Database (OLTP)
Online Analytical Processing	Online Transactional Processing
OLAP systems are used by knowledge workers such as executives, managers, and analysts.	OLTP systems are used by clerks, DBAs, or database professionals.
The number of users is in hundreds.	The number of users is in thousand.
It provides summarized and multidimensional view of data.	It provides detailed and flat relational view of data.
The database size is from 100GB to 100 TB.	The database size is from 100 MB to 100 GB.
It is based on Star Schema, Snowflake Schema, and Fact Constellation Schema.	It is based on Entity Relationship Model.
It contains historical data.	It contains current data.

DATA MINING

Data mining refers to extracting knowledge from large amounts of data. The data sources can include databases, data warehouse, web etc.

Data mining refers to the analysis of data. It is the computer-supported process of analyzing huge sets of data that have either been compiled by computer systems or have been downloaded into the computer. In the data mining process, the computer analyzes the data and extract useful information from it. It looks for hidden patterns within the data set and try to predict future behavior. Data mining is primarily used to discover and indicate relationships among the data sets.

Data mining aims to enable business organizations to view business behaviors, trends relationships that allow the business to make data-driven decisions. It is also known as knowledge Discover in Database (KDD). Data mining tools utilize AI, statistics, databases, and machine learning systems to discover the relationship between the data. Data mining tools can support business-related questions that traditionally time-consuming to resolve any issue.

Data mining is also called **Knowledge Discovery in Database (KDD).** The knowledge discovery process includes Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation.

Here is the list of steps involved in the knowledge discovery process:

- Data Cleaning: In this step, the noise and inconsistent data is removed.
- Data Integration: In this step, multiple data sources are combined.
- Data Selection: In this step, data relevant to the analysis task are retrieved from the database.

- Data Transformation: In this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- Data Mining: In this step, intelligent methods are applied in order to extract data patterns.
- Pattern Evaluation: In this step, data patterns are evaluated.
- Knowledge Presentation: In this step, knowledge is represented.

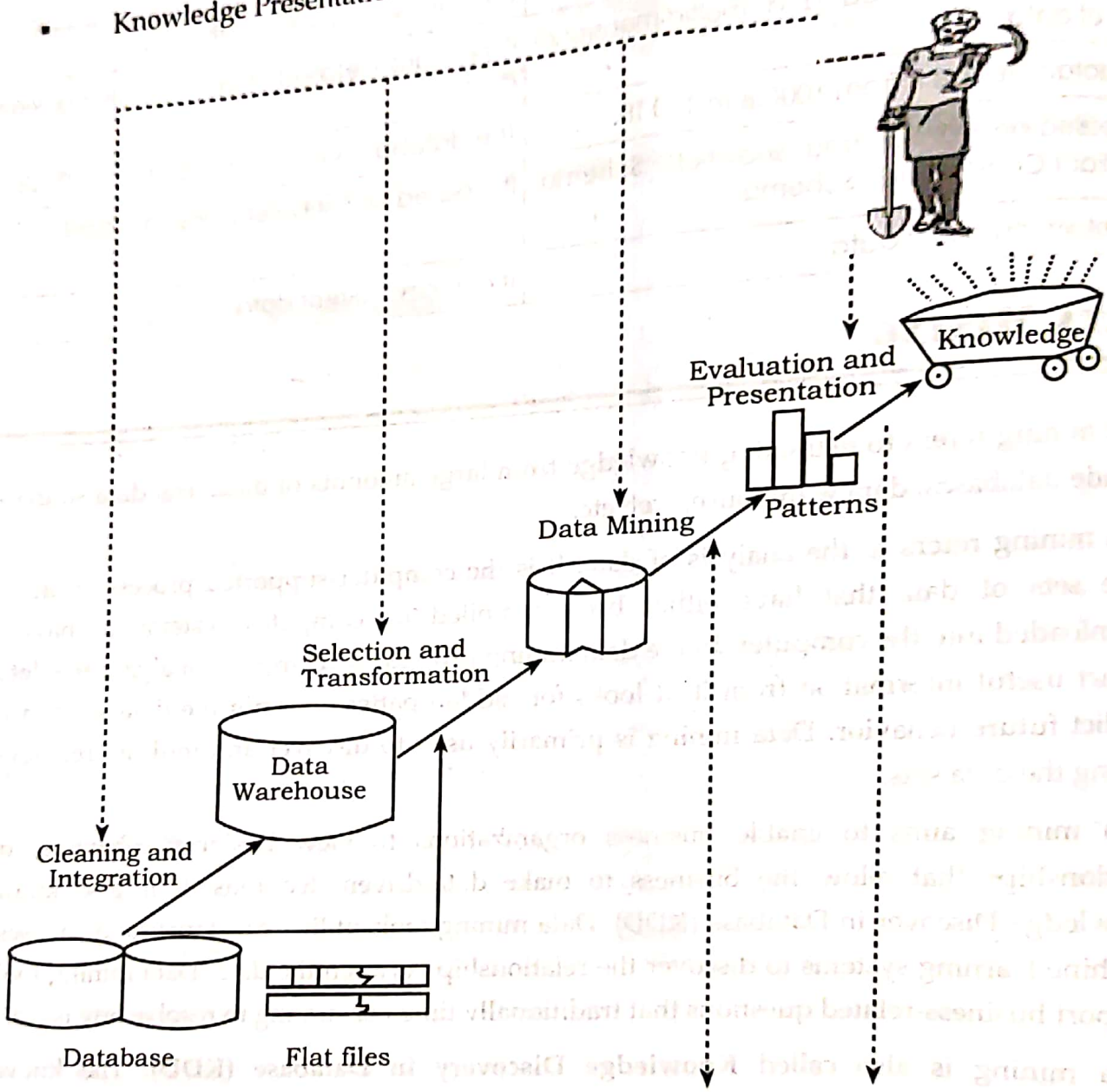


Fig: Data mining as a step in the process of knowledge discove

Important Features of Data Mining

The important features of Data Mining are given below:

- It utilizes the Automated discovery of patterns.
- It predicts the expected results.
- It focuses on large data sets and databases
- It creates actionable information.

Advantages of Data Mining

i. **Market Analysis:** Data Mining can predict the market that helps the business to make the decision. For example, it predicts who is keen to purchase what type of products.

ii. **Fraud detection:** Data Mining methods can help to find which cellular phone calls, insurance claims, credit, or debit card purchases are going to be fraudulent.

iii. **Financial Market Analysis:** Data Mining techniques are widely used to help Model Financial Market

iv. **Trend Analysis:** Analyzing the current existing trend in the marketplace is a strategic benefit because it helps in cost manufacturing process as per market demand.

Disadvantages of Data Mining

i. There is a probability that the organizations may sell useful data of customers to other organizations for money. As per the report, American Express has sold credit card purchases of their customers to other organizations.

ii. Many data mining analytics software is difficult to operate and needs advance training to work on.

iii. Different data mining instruments operate in distinct ways due to the different algorithms used in their design. Therefore, the selection of the right data mining tools is a very challenging task.

iv. The data mining techniques are not precise, so that it may lead to severe consequences in certain conditions.

Application of Data Mining

These are the following areas where data mining is widely used:

1. **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. Analysts use data mining approaches such as Machine learning, Multi-dimensional database, Data visualization, Soft computing, and statistics. Data Mining can be used to forecast patients in each category. The procedures ensure that the patients get intensive care at the right place and at the right time. Data mining also enables healthcare insurers to recognize fraud and abuse.

2. **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. This data may assist the retailer in understanding the requirements of the buyer and altering the store's layout accordingly. Using a different analytical comparison of results between various stores, between customers in different demographic groups can be done.

3. **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. EDM objectives are recognized as affirming student's future learning behavior, studying the impact of educational support, and promoting learning science. An organization can use data mining to make precise decisions and also to predict the results of the student. With the results, the institution can concentrate on what to teach and how to teach.

4. **Data Mining in Manufacturing Engineering**

Knowledge is the best asset possessed by a manufacturing company. Data mining tools can be beneficial to find patterns in a complex manufacturing process. Data mining can be used in system-level designing to obtain the relationships between product architecture, product portfolio, and data needs of the customers. It can also be used to forecast the product development period, cost, and expectations among the other tasks.

5. **Data Mining in CRM (Customer Relationship Management)**

Customer Relationship Management (CRM) is all about obtaining and holding Customers, also enhancing customer loyalty and implementing customer-oriented strategies. To get a decent relationship with the customer, a business organization needs to collect data and analyze the data. With data mining technologies, the collected data can be used for analytics.

6. **Data Mining in Fraud detection**

Billions of dollars are lost to the action of frauds. Traditional methods of fraud detection are a little bit time consuming and sophisticated. Data mining provides meaningful patterns and turning data into information. An ideal fraud detection system should protect the data of all the users. Supervised methods consist of a collection of sample records, and these records are classified as fraudulent or non-fraudulent. A model is constructed using this data, and the technique is made to identify whether the document is fraudulent or not.

NATIONAL DATA WAREHOUSES

A large number of national data warehouse can be identified from the existing data resources within the central government ministries. These potential subject areas on which data warehouse nay be developed at present and in future.

Census Data

The register general and census commissions of India decennially compiles information of all individuals, villages, population groups etc. This information is wide ranging such as the individual slip, a compilation of information individual households of which a data base of 5%. Sample is maintain for analysis. Data mining can be performed for analysis and knowledge discovery. A village level data base was originally develop by national informatics center at Hyderabad under general information services terminal of national informatics center (GISTNIC) for the 1999 census.

Primary census abstract and village amenities. Subsequently a data ware house was also developed for village amenities for Tamil Nadu. This enables multi-dimensional analysis of the village level data in such as education, Health and infrastructure. As the census compilation is performed once in ten years, the data is quasi- static and therefore no refreshing of the ware house needs to be done on a periodic basis. Only the new data needs to be either appended to the data warehouse or alternative a new data warehouse can be built.

Prices of Essential Commodities

The ministry of food and civil supplies, government of India , compiles daily data for about 300 observation centers in the entire country on the prices of essential commodities such as rice , edible oils etc. The data is compiled at the district level by the respective state government agencies and transmitted online on Delhi for aggregation and storage.

A data warehouse can be built for this data and OLAP technique can be applied for its analysis. A data mining and forecasting technique can be applied for advance forecasting of the actual prices of these essential commodities. The forecasting model can be strengthened for more accurate forecasting by taking into account the external factors such as rain fall, growth rate of population and inflation.

OTHER AREAS FOR DATA WAREHOUSE AND DATA MINING

Other possible areas for data ware housing and data mining in central government sectors are:

Agriculture

The agricultural census performed by the ministry of Agriculture, government of India, compiles a large number of agricultural parameters at the national level. District wise

agricultural production area and yield of crops is compiled, analysis, mining and forecasting statistics on consumption of fertilizers can be turned into a data merge.

Data on agricultural inputs such as seeds and fertilizers can also be effectively analyzed in a data ware house. Data from livestock census can be turned into a data ware house. Land use pattern statistics can also be analyzed in a warehousing environment. Thus, there is substantial scope for application of data warehouse housing and data mining techniques in agricultural sector.

Rural Development

Data on individuals below poverty line can be built into a data ware house. Drinking water census data (from drinking water mission) can be effectively utilized by OLAP and data mining technologies. Monitoring and analysis of progress made on implementation of rural development programs can also be made using OLAP and data mining technologies.

Health

Community needs assessment data, immunization data, data from national programs on controlling blindness, leprosy, malaria can all used for data warehousing implementation, OLAP and data mining applications. Generate patient, employee, and financial records share data with other entities, like insurance companies, NGOs, and medical aid services Use data mining to identify patient trends Provide feedback to physicians on procedures and tests.

Planning

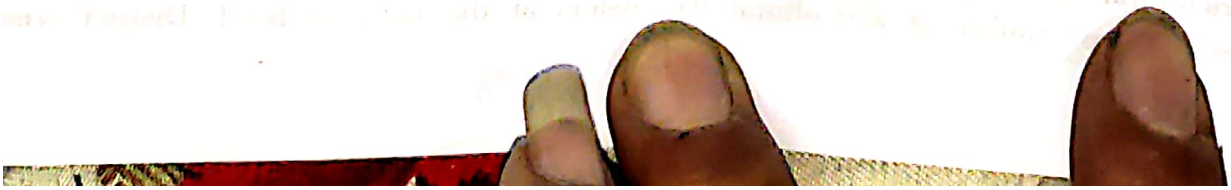
At the planning commission , data warehouses can be built for the state plan data on all sectors , labour ,energy , education , trade and industry, five year plan etc.

Education

The sixth all India educational survey data has been converted into a data ware house. Various types of analytical queries and reports can be answered. Store and analyze information about faculty and students maintain student portals to facilitate student activities extract information for research grants and assess student demographics integrate information from different sources into a single repository for analysis and strategic decision-making.

Commerce and Trade

Data link on trade can be analyzed and converted into a data ware house. World price monitoring system can be made to perform better by using data warehousing and data mining technologies.



Tourism

Tourist arrival behavior and performances, Tourism products data, foreign exchange earnings data and Hotels , Travel and transportation data. Conventionally, the government departments have largely been satisfied with developing single management information system or limited cases a few database which were used online for limited purposes . The techniques used for analysis were conventional statistical techniques on largely batch mode processing and data mining technology, no body was aware of any better techniques for this activity , in fact ,data ware housing and data mining technologies could lead to the most significant advancements in the government functionally, if properly applied and used in the government activities.

Other Sectors

1. **Insurance**
   - Analyze data patterns and customer trends
   - Maintain records of all internal and external sources, including existing participants
   - Design customized offers and promotions for customers
   - Predict and analyze changes in the industry
2. **Manufacturing**
   - Predict market changes and analyze current business trends
   - Analyze previous and current market data
   - Track customer feedback and identify opportunities for improvement
   - Gather, standardize, and store data from various internal and external sources
   - Identify profitable product lines and required product features
3. **Retail**
   - Maintain records of producers and consumers
   - Track items, their promotion strategies, and consumer buying trends (trend analysis)
   - Analyze sales to determine shelf space
   - Understanding the patterns of complaints, claims, and returns