Data Mining and Warehousing

Sangeetha K V
Ist MCA
Adhiyamaan College of Engineering,
Hosur-635109.
E-mail:veerasangee1989@gmail.com

Rajeshwari P
Ist MCA
Adhiyamaan College of Engineering,
Hosur-635109.
E-mail: raji18390@gmail.com

Abstract- Data mining has become a popular buzzword but, in fact, promises to revolutionize commercial and scientific exploration. Databases range from millions to trillions of bytes of data. 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. 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.

A data warehouse is a relational database that is designed for query and analysis rather than transaction processing. It usually contains historical data that is derived from transaction Data in the warehouse can be seen as materialized views generated from the underlying multiple data sources. Materialized views are used to speed up query processing on large amounts of data. It separates analysis workload from transaction workload and enables an organization to consolidate data from several sources. These views need to be maintained in response to updates in the source data. This is often done using incremental techniques that access data from underlying sources. In the data-warehousing scenario, accessing base relations can be difficult; sometimes data sources may be unavailable, since these relations are distributed across different sources.

This paper provides an introduction to the basic technologies of data mining. As well as a basic description of how data warehouse architectures can evolve to deliver the value of data mining to end-users.

According to my view to install a new information source by Datamining & Warehousing in order to serve the people is introduced as follows.

Keywords: *Data mining* **,** hidden predictive, *warehouse.*

I. Topics Covered in Data Mining and Warehousing

Data Mining

Overview

Data, Information, and Knowledge

How does data mining work?

Crucial Concepts in Data Mining

- Bagging
- Boosting
- Text Mining
- Stacked Generalization
- Data Preparation
- Deployment
- Drill-Down Analysis
  Feature Selection
  Machine Learning

Different levels of analysis

Models for Data Mining

The Foundations of Data Mining

The Scope of Data Mining

Commonly used techniques

Architecture

Profitable Applications

Advantages

Disadvantages

Conclusion

Data Warehousing

Overview

Data warehousing working

Benefits

Information Access Layer

Data Access Layer

Data Directory (Metadata) Layer

Process Management Layer

Application Messaging Layer

Data Warehouse (Physical) Layer

Data Staging Layer

Data Warehouse Options

Data Warehouse Scope

Data Marts

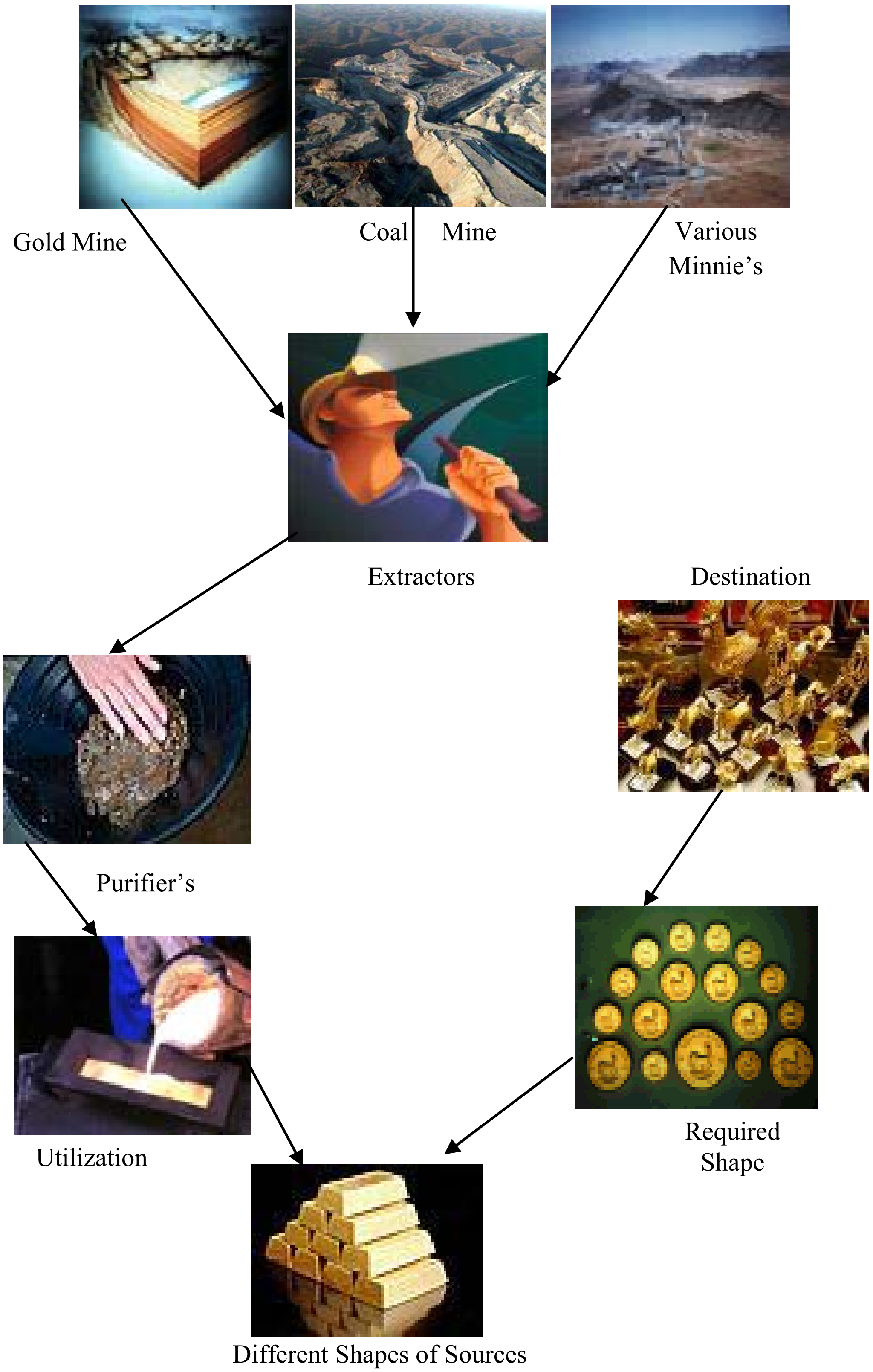
Advantages

Disadvantages

Application

Conclusion

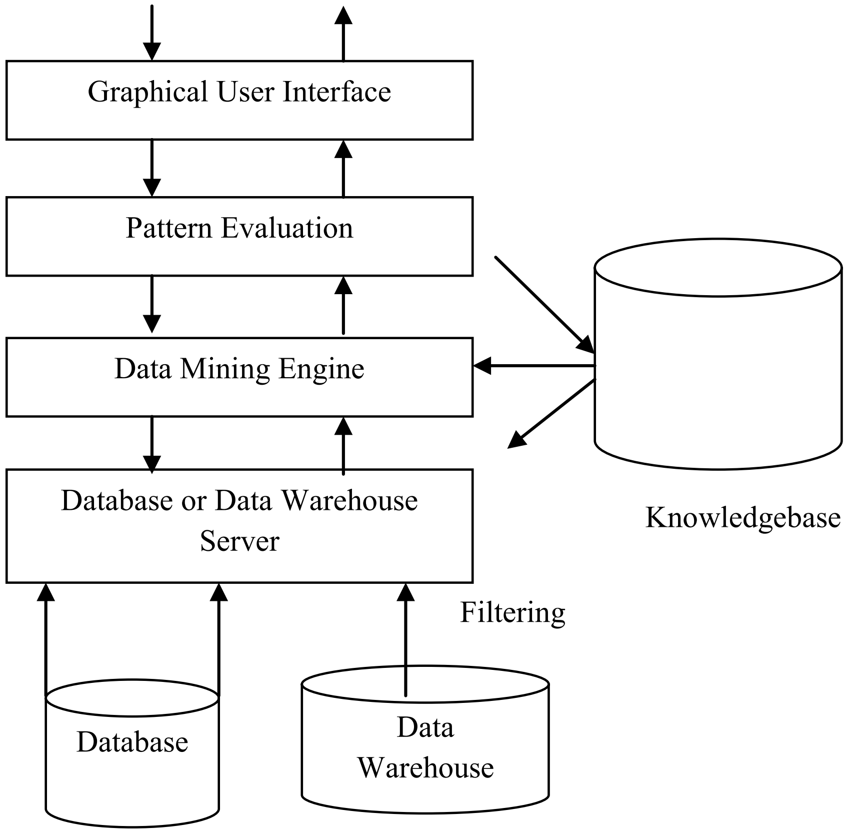
II. DATA MINING COMPARED WITH REAL TIME MINING



III. DATA MINING OVERVIEW

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified.

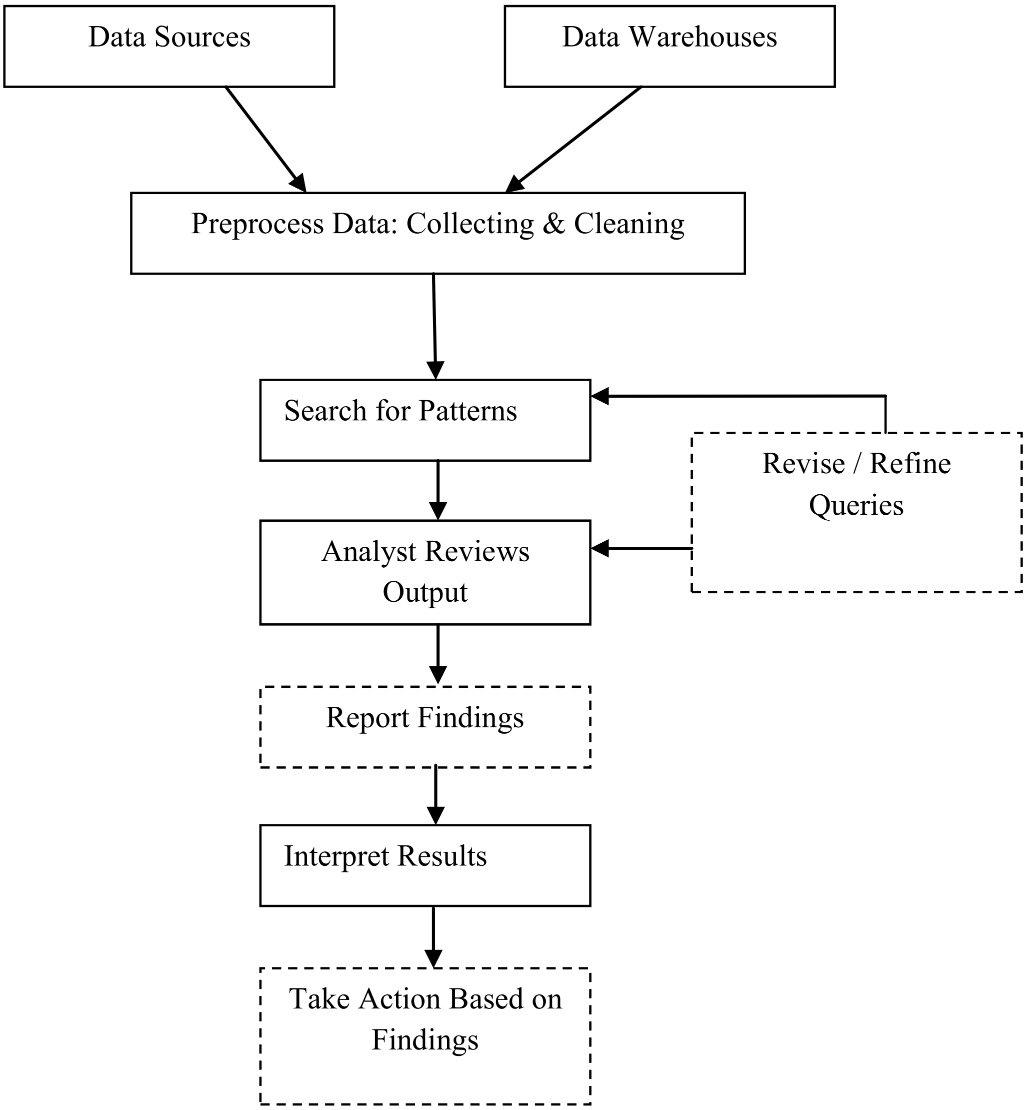
Data Mining Architecture



How Does Data Mining Work?

- Classes: Stored data is used to locate data in predetermined groups.
- Clusters: Data items are grouped according to logical relationships or consumer preferences.
- Associations: Data can be mined to identify associations.
- Sequential patterns: Data is mined to anticipate behavior patterns and trends.

IV. THE DATA MINING PROCESS



Data Mining Consists of Five Major Elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

Different Levels of Analysis are Available:

- Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- Genetic algorithms: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

What Technological Infrastructure is Required?

- Size of the database: The more data being processed and maintained, the more powerful the system required.
- Query complexity: The more complex the queries and the greater the number of queries being processed, the more powerful the system required.

Data Mining Infrastructure:

- Ability to access data from many sources & consolidates
- Ability to score customers based on existing models
- Ability to manage lots of models over time
- Ability to manage lots of model scores over time
- Ability to track model score changes over time
- Ability to reconstruct a customer "signature" on demand
- Ability to publish scores, rules, and other data mining results

The Foundations of Data Mining:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

The Scope of Data Mining:

- Automated prediction of trends and behaviors
- Automated discovery of previously unknown patterns

The Most Commonly Used Techniques in Data Mining are:

- Artificial neural networks
- Decision trees
- Genetic algorithms
- Nearest neighbor method
- Rule induction:

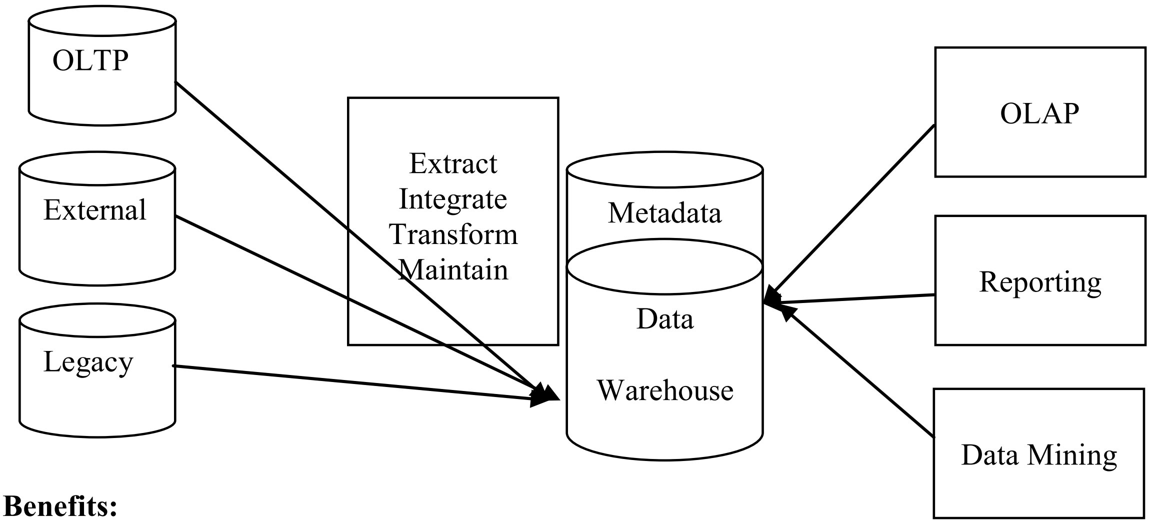
Profitable Applications:

A wide range of companies has deployed successful applications of data mining. While early adopters of this technology have tended to be in information-intensive industries such as financial services and direct mail marketing, the technology is applicable to any company looking to leverage a large data warehouse to better manage their customer relationships. Two critical factors for success with data mining are: a large, well-integrated data warehouse and a well-defined understanding of the business process within which data mining is to be applied (such as customer prospecting, retention, campaign management, and so on).

V. DATA WAREHOUSING

- OLTP (online transaction processing) systems
- Range in size from megabytes to terabytes
- David Nelson
- High transaction throughput
- Decision makers require access to all data
- A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision-making process

VI. DATAWAREHOUSING ARCHITECTURE



Benefits:

- Potential high returns on investment
- Competitive advantage
- Data can reveal previously unknown, unavailable and untapped information
- Increased productivity of corporate decision-makers
- Integration allows more substantive, accurate and consistent analysis five primary unit

Information Flow Processes:

- Inflow - extraction, cleansing and loading of data from source systems into warehouse
- Upflow - adding value to data in warehouse through summarizing, packaging and distributing data
- Downflow - archiving and backing up data in warehouse
- Outflow - making data available to end users
- Metaflow - managing the metadata

Problems of Data Warehousing

- Long duration projects
- Complexity of integration
- Underestimation of resources for data loading
- Hidden problems with source systems
- Required data not captured
- Increased end-user demands

Data Marts

- A subset of a data warehouse that supports the requirements of a particular department or business function
- As data warehouse grows larger, ability to serve needs may be compromised

Extraction, Cleansing & Transformation Tools

- Code generators
- Database data replication tools

Data Warehouse DBMS

- Load performance & Processing
- Data Quality management
- Query performance
- Scalability

Data Mart Issues:

- Functionality
- Size
- Load performance
- Users access to data in multiple data marts
- Internet/intranet access
- Administration

Dimensionality Modelling:

- Similar to E-R modelling but with constraints
- Composed of one fact table with a composite primary key
- Dimension tables have a simple primary key which corresponds exactly to one foreign key in the fact table

Star Schemas:

- The most common dimensional model
- A fact table surrounded by dimension tables
- Fact tables
  - Contains FK for each dimension table
  - Large relative to dimension tables
  - Read-only

- Dimension tables
- Reference data

Applications

- Medicine: disease, treatments
- Molecular or pharmaceutical: new drugs
- Security: face recognition identification
- Judiciary: data on judgment of similar cases
- Biometrics
- Multimedia retrieval
- Scientific data analysis
- Web site or Web store design, and promotion

New information source by Data Mining & Warehousing in order to serve the people is as follows:

All Indian's were not at all favor of all languages and tourist places so for that purpose people always depending up on others on that places this pays way for waste of time and in some conditions (i.e.) viewers from other countries was facing a lot of problems. Due to this reason they got a bad opinion up on Indian's. This problem can be solved by data mining and warehousing model

We will explain the various details regarding this schema with examples during the paper presentation .we are sure that this schema will helps our nation to overcome from lot of disadvantages

VII.CONCLUSION

Comprehensive data warehouses that integrate operational data with customer, supplier, and market information have resulted in an explosion of information. Competition requires timely and sophisticated analysis on an integrated view of the data. However, there is a growing gap between more powerful storage and retrieval systems and the users' ability to effectively analyze and act on the information they contain. Both relational and OLAP technologies have tremendous capabilities for navigating massive data warehouses, but brute force navigation of data is not enough. A new technological leap is needed to structure and prioritize information for specific end-user problems. The data mining tools can make this leap. Quantifiable business benefits have been proven through the integration of data mining with current information systems, and new products are on the horizon that will bring this integration to an even wider audience of users.

REFERENCES

1. www.igoogle.com\data mining
2. www.mtmi.vu.lt
3. www.mein.nagoya-u.ac.jp
4. www.angelfire.com
5. www.msn.encarta.com