DATA WAREHOUSING &MINING

www.Technicalpapers.co.nr

DATA WAREHOUSING AND MINING

ABSTRACT:

A **Data warehouse** is a central repository or '**storehouse**' for data that an enterprise's various business systems collect. Data from various online applications and other sources is selectively extracted and organized in the data warehouse.**Data mining** is the **principle of sorting** or extracting through **large amounts of data** and picking out relevant information. The main purpose of using data mining is to have a quick access over the Database for useful analysis and access**.**

ETL is an important concept in Data Warehousing and as it is the way data actually gets loaded into the warehouse.

- Extracting data from outside sources,
- Transforming it to fit business needs (which can include quality levels), and ultimately
- Loading it into the end target, i.e. the data warehouse.

Business Intelligence (BI) plays a vital role in Data Warehousing. BI systems provide historical, current, and predictive views of business operations, most often using data that has been gathered into a data warehouse or a data mart and occasionally working from operational data. There are many types of **Business tools** that help us to have an interactive examination of the organizations data.

DATA WAREHOUSING AND MINING

SCOPE OF THE PAPER

- *Introduction*
- *Need for Data Warehousing and Mining.*
- *Relationship between Data Mining and Warehousing*
- *Business Intelligence in Data Warehousing*
- *Tools in Data Warehousing and Mining.*
- *ETL concept*
- *Parallel Processing*
- *Process involved in Data Mining.*
- *Advantages*
- *Conclusion*

INTRODUCTION:

A data warehouse is the main repository of an organization's historical data, its corporate memory. It contains the raw material for management's decision support system. Extracting data from legacy systems and other resources; cleaning, scrubbing and preparing data for decision support; maintaining data in appropriate data stores; accessing and analyzing data using a variety of end user tools; and mining data for significant relationships are the main process involved in Data Warehousing. The critical factor leading to the use of a data warehouse is that a data analyst can perform complex queries and analysis, such as data mining, on the information without slowing down the operational systems. The Data Warehousing has the following qualities:

Subject-oriented

The data in the database is organized so that all the data elements relating to the same real-world event or object are linked together.

Time-variant

The changes to the data in the database are tracked and recorded so that reports can be produced showing changes over time.

Non-volatile

Data in the database is never over-written or deleted - once committed, the data is static, read-only, but retained for future reporting.

Integrated:

The database contains most of the organization’s operational application and that this data is made consistent.

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

NEED FOR DATA WAREHOUSING:

- The processing load of reporting reduced the response time of the operational systems.
- The database designs of operational systems were not optimized for information analysis and reporting.
- Most organizations had more than one operational system, so company-wide reporting could not be supported from a single system
- Development of reports in operational systems often required writing specific computer programs which was slow and expensive

As a result separate computer database began to be build that were specifically designed to support mathematical information and analysis purpose.

TYPES OF DATA WAREHOUSE:

Offline Data Warehouse:

Data Warehouse in this stage of evolution are updated on a regular time cycle(monthly or weakly) from the operating system and the data is stored in a integrated reporting oriented data structure.

Real time Data Warehouse:

Data Warehouse at this stage are updated on a transaction are event basis every time an O.S performs the transaction.(e.g. A delivery or a booking)

Integrated Data Warehouse:

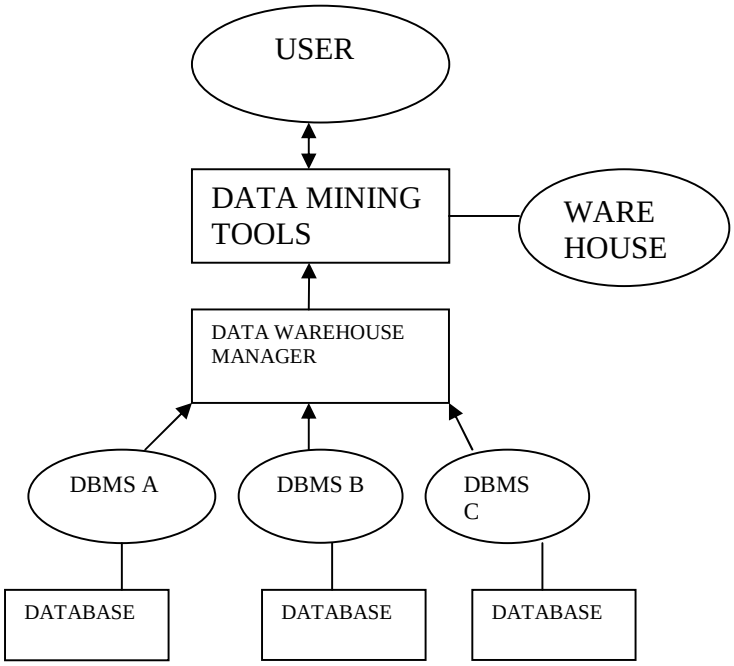
Data Warehouse at this stage are used to generate activity or transaction that are passed back into the O.S for use in the daily activity of an organization.

RELATIONSHIP BETWEEN DATA WAREHOUSING AND MINING:

A **data warehouse** assembles the data from the heterogeneous database. It does **not attempt to extract information** from the data into the warehouse. Data warehouse formats and **organize** the data and support management functions.

Data Mining in the other hand **attempts to extract** useful information from the database.

The following diagram shows the relationship between Data mining and Data warehousing.



BUSINESS INTELLIGENCE:

The term business intelligence (BI) refers to technologies, applications, and practices for the **collection, integration, analysis, and presentation** of business information and also sometimes to the information itself. The purpose of business intelligence is to support better business decision making.

For a ***BI technology*** system to work effectively, a company should have a **secure computer system** which can specify different levels of user access to the data 'warehouse,' depending on whether the user is a junior staffer, a manager, or an executive. Also, a BI system should have sufficient data capacity and a plan for how long data will be stored (data retention). Analysts should set benchmark and performance targets for the system.

Business is a collection of activities carried on for whatever purpose, be it science, technology, commerce, industry, law, government, defense, et cetera. The communication facility serving the conduct of a business may be referred

to as an intelligence system. Intelligence is also defined here, in a more general sense, as "the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal."

TOOLS IN DATA WAREHOUSING:

- **Digital Dashboards**

Also known as Business Intelligence Dashboards, or Executive Dashboards, that gives the visually-based summaries of business data .It will show at-a-glance understanding of business conditions through metrics and Key Performance Indicators (KPIs). A very popular BI tool that has arisen in the last few years.

- **Online Analytical Processing(OLAP):**

A capability of some management, decision support, and executive information systems that supports interactive examination of large amounts of data from many perspectives.

- **Reporting software**

Generates aggregated views of data to keep the management informed about the state of their business.

- **Data mining**

Extraction of consumer information from a database by utilizing software that can isolate and identify previously unknown patterns or trends in large amounts of data. There are a variety of data mining techniques that reveal different types of patterns.. Some of the techniques that belong here are Statistical methods (particularly Business statistics) and Neural networks as very advanced means of analyzing data.

- **Business performance management (BPM)**

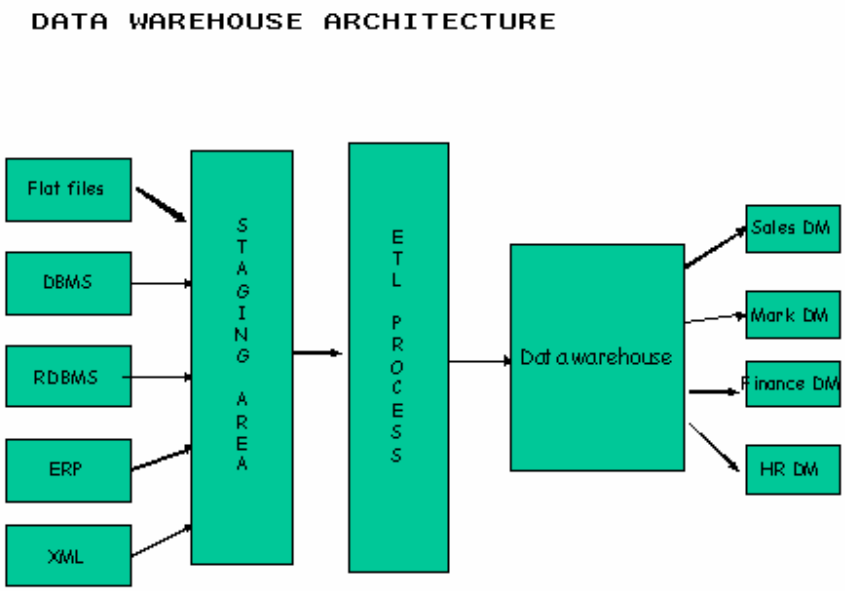
Business performance management (BPM) or Operational performance management is a set of **processes** that help organizations **optimize their business performance**. BPM involves **consolidation of data** from various sources, querying, and analysis of the data, and putting the results into practice.

E .T. L CONCEPT:

It is a process in data warehousing that involves

- **Extracting** data from outside sources,
- **Transforming** it to fit business needs (which can include quality levels), and ultimately
- **Loading** it into the end target, i.e. the data warehouse.

ETL is important, as it is the way data actually gets loaded into the warehouse.



Extract:

The first part of an ETL process is to extract the data from the source systems. Most data warehousing projects consolidate data from different source systems. Each separate system may also use a different data organization / format. Common data source formats are relational databases and flat files, but may include non-relational database structures. **Extraction converts the data into a format for transformation processing**.

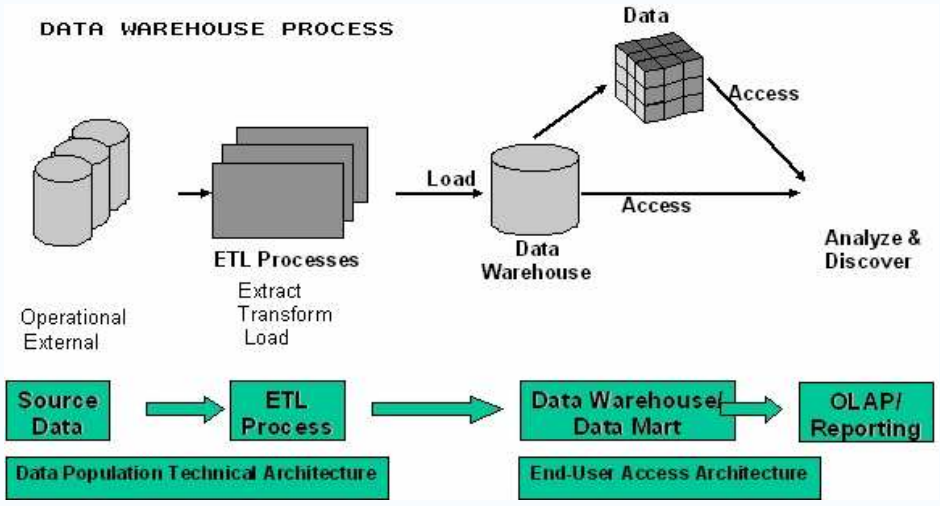
An intrinsic part of the extraction is the parsing of extracted data, resulting in a check if the data meets an expected pattern or structure. If not, the data is rejected entirely.

Transform

The transform stage applies a series of rules or functions to the extracted data from the source to derive the data to be loaded to the end target. Some data sources will require very little or even no manipulation of data. In other cases, one or more of the following transformations types to meet the business and technical needs of the end target may be required:

- Selecting only certain columns to load (or selecting null columns not to load)
- Translating coded values (*e.g.*, if the source system stores 1 for male and 2 for female, but the warehouse stores M for male and F for female), this is called automated data cleansing; no manual cleansing occurs during ETL
- Encoding free-form values (*e.g.*, mapping "Male" and "1" and "Mr" into M)
- Deriving a new calculated value (*e.g.*, sale_amount = qty * unit_price)

- Joining together data from multiple sources (*e.g.*, lookup, merge, *etc.*)
- Summarizing multiple rows of data (*e.g.*, total sales for each store, and for each region)
- Generating surrogate key values
- Transposing or pivoting (turning multiple columns into multiple rows or vice versa)
- Splitting a column into multiple columns (*e.g.*, putting a comma-separated list specified as a string in one column as individual values in different columns)
- Applying any form of simple or complex data validation; if failed, a full, partial or no rejection of the data, and thus no, partial or all the data is handed over to the next step, depending on the rule design and exception handling. Most of the above transformations itself might result in an exception, e.g. when a code-translation parses an unknown code in the extracted data.



Load

The load phase loads the data into the end target, usually being the data warehouse (DW). Depending on the requirements of the organization, this process ranges widely. Some data warehouses might weekly overwrite existing information with cumulative, updated data, while other DW (or even other parts of the same DW) might add new data in a historized form, e.g. hourly. The

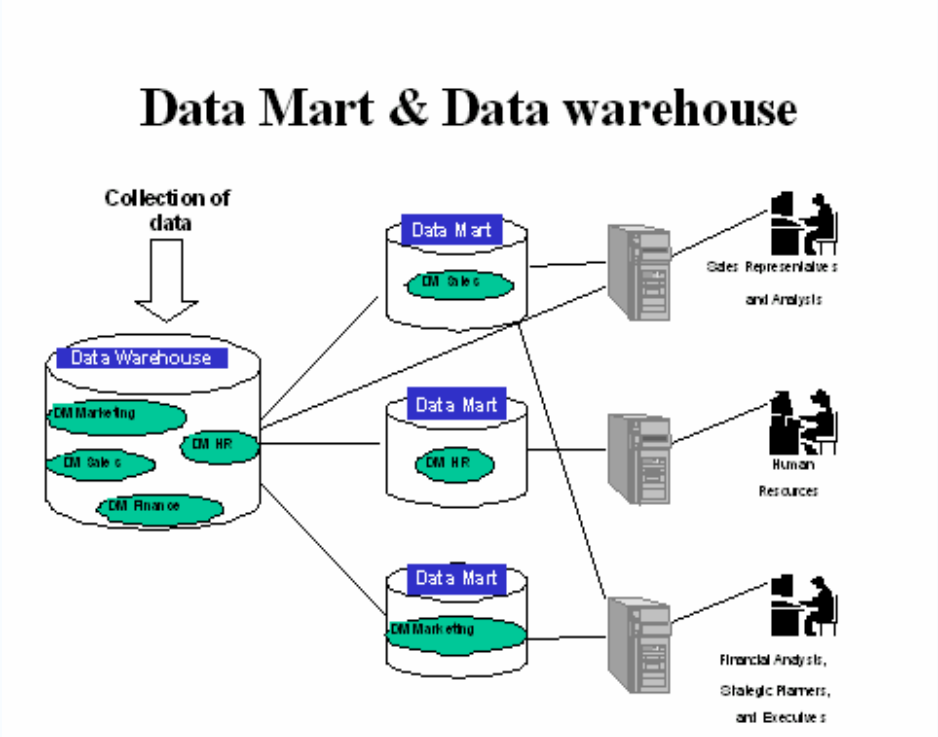
timing and scope to replace or append are strategic design choices dependent on the time available and the business needs.

As the load phase interacts with a database, the constraints defined in the database schema as well as in triggers activated upon data load apply (e.g. uniqueness, referential integrity, mandatory fields), which also contribute to the overall data quality performance of the ETL process.

DATA MART:

It is a database of data gathered from operational data and other sources that is designed to serve a particular group/department of an organization.

Data Mart is typically a **Sub-set of Data Warehouse.** Data Warehouse is a database of data gathered from operational data and other sources that is designed to serve the whole organization.



PARALLEL PROCESSING:

A recent development in ETL software is the implementation of parallel processing. This has enabled a number of methods to improve overall performance of ETL processes when dealing with large volumes of data.

There are 3 main types of parallelisms as implemented in ETL applications:

Data:

By splitting a single sequential file into smaller data files to provide parallel access.

Pipeline:

Allowing the simultaneous running of several components on the same data stream. An example would be looking up a value on record 1 at the same time as adding together two fields on record 2.

Component:

The simultaneous running of multiple processes on different data streams in the same job. Sorting one input file while performing a de-duplication on another file would be an example of component parallelism.

All three types of parallelism are usually combined in a single job.

PROCESS INVOLVED IN DATA MINING:

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries.. Generally, any of four types of relationships are sought:

- **Classes**: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

- **Clusters**: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

- **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

- **Sequential patterns**: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

ANALYSIS OF DATA:

Analysis of data is an important process in data Mining.Different levels of analysis is 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 process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

- **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). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

- **Nearest neighbour 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 (where *k* 1). Sometimes called the *k*-nearest neighbor technique.

- **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.
- **Data visualization**:

The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships

ADVANTAGES:

There are many advantages to using a data warehouse, some of them are:

- Data warehouses enhance end-user access to a wide variety of data.
- Decision support system users can obtain specified trend reports, e.g. the item with the most sales in a particular area within the last two years.

- Data warehouses can be a significant enabler of commercial business applications, particularly customer relationship management (CRM) systems.
- Data warehouse data are organized around major subject areas such as sales, claims, shipments, and enrollments. For example, a data warehouse for sales contains historical records of sales over specific time.

- A data warehouse provides the facility for integration in a heterogeneous, fragmented environment of independent application systems, where the data is stored in multiple, incompatible formats. For example, a department store may have information about the same customers stored in several databases using different format representations. The data warehouse brings the data together into a single/representation
  - The data warehouse organizes and stores the data needed for informational and analytical processing over an extended historical time range. For example, a marketing analyst can analyze the sales history of five years from the information that was collected at the end of each year.

  - Changes to the data warehouse environment occur in a controlled and scheduled manner, unlike the more volatile OLTP environment in which updates continually occur. A similar query run in five minute intervals in an OLTP environment may yield different results, while the same query run within the data warehouse should remain stable and consistent. For example, an airline may capture frequent flyer information in its data warehouse. During check-in for a flight, the additional mileage for a specific passenger is immediately updated in the OLTP system, but is not yet reflected in the data warehouse until its next scheduled load.

CONCERNS:

- Extracting, transforming and loading data consumes a lot of time and computational resources.
- Security could develop into a serious issue, especially if the data warehouse is web accessible.

CONCLUSION:

Thus Data Warehousing and Mining is now an emerging field and plays a vital role in our day today life. Thus the efficiency of storing an organizations data and retrieving them will be greatly developed. The copies of all the databases in a company are maintained in one location and can be accessible by the employees in any location. This plays a major role in **sharing of data** and other resources easily. The real time systems will be very much developed by using this Data Warehousing technique.

Reference:

www.Technicalpapers.co.nr