DATA WAREHOUSING & DATA MINING IN AN ORGANIZATION

Name of Author: Harminder Singh
Designation of Author: Assistant Professor
Name of Department: Computer Applications
Name of organization: Gulzar Group of Institutes, City: Khanna, Country: India

Abstract: *The* Cloud-based technologies have revolutionized the business world, allowing companies to easily retrieve and store valuable **data** about their customers, products and employees. This **data** is used to inform **important** and effective business decisions. Many global corporations have turned to data warehousing to organize data that streams in from corporate branches and operations centers around the world. It's essential for IT students to understand how data warehousing helps businesses remain competitive in a quickly evolving global marketplace. The goal of a data warehousing is to effectively support business decision making. Data warehousing provides leverage for management in an organization. Effective decision making is the major function of every management in an organization; data warehouses facilitate meaningful research which facilitates effective management processes. With data warehouse in place, each department in an organization can share data and though the costs of operations will be reduced, this also allows users or management to perform extensive analysis across all departments in the organization. The three tier architecture of data warehousing describes the processing of various operational databases, data marts and queries. OLAP (Online Analytical Processing) server is the technology behind many Business Intelligence (BI) applications. Business Intelligence refers to a set of methods and techniques that are used by organizations for tactical and strategic decision making.
Data mining can be used in conjunction with a **data warehouse** to help with certain types of decisions. **Data mining** helps in extracting meaningful new patterns that cannot be found in the **data warehouse**. Data mining helps to convert data into knowledge by the process of knowledge discovery. The important aspects and roll of the data warehousing and data mining in an organization is discussed in detail as follows.

D**ata warehousing** - A subject oriented integrated, non volatile, time-variant collection of data in support of management decisions. Data warehouses provide access to data for complex analysis, Knowledge discovery, and decision-making. A collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager, and analyst) to make better and faster decisions. D**ata warehouse** is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site. Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing. In computing, a **data warehouse** (**DW** or **DWH**), also known as an **enterprise data warehouse** (**EDW**), is a system used for reporting and data analysis, and is considered a core component of business intelligence.
Data warehouses are widely used within the largest and most complex businesses in the world. Use within moderately large organizations, even those with more than 1,000 employees remains surprisingly low at the moment.
Data is collected periodically from the applications that support business processes and copied onto special dedicated computers. There it can be validated, reformatted, reorganized, summarized, restructured, and supplemented with data from other sources. The resulting data warehouse becomes the main source of information for report generation, analysis, and presentation through ad hoc reports, portals, and dashboards. Data warehousing systems are designed to support online analytical processing (OLAP).

Example of Data warehousing – Facebook

A great example of data warehousing is what Facebook does. Facebook gathers all your accounts data such as your friends, your likes, your groups and WhatsApp accounts data etc. All these data are stored into one central repository. Although Facebook is storing all these information into separate databases, they store the most relevant and significant information into one central aggregated database. This is because of many

reasons like to make sure that you see the most relevant ads that you are most likely to click on or the friends that they suggest are the most relevant to you.

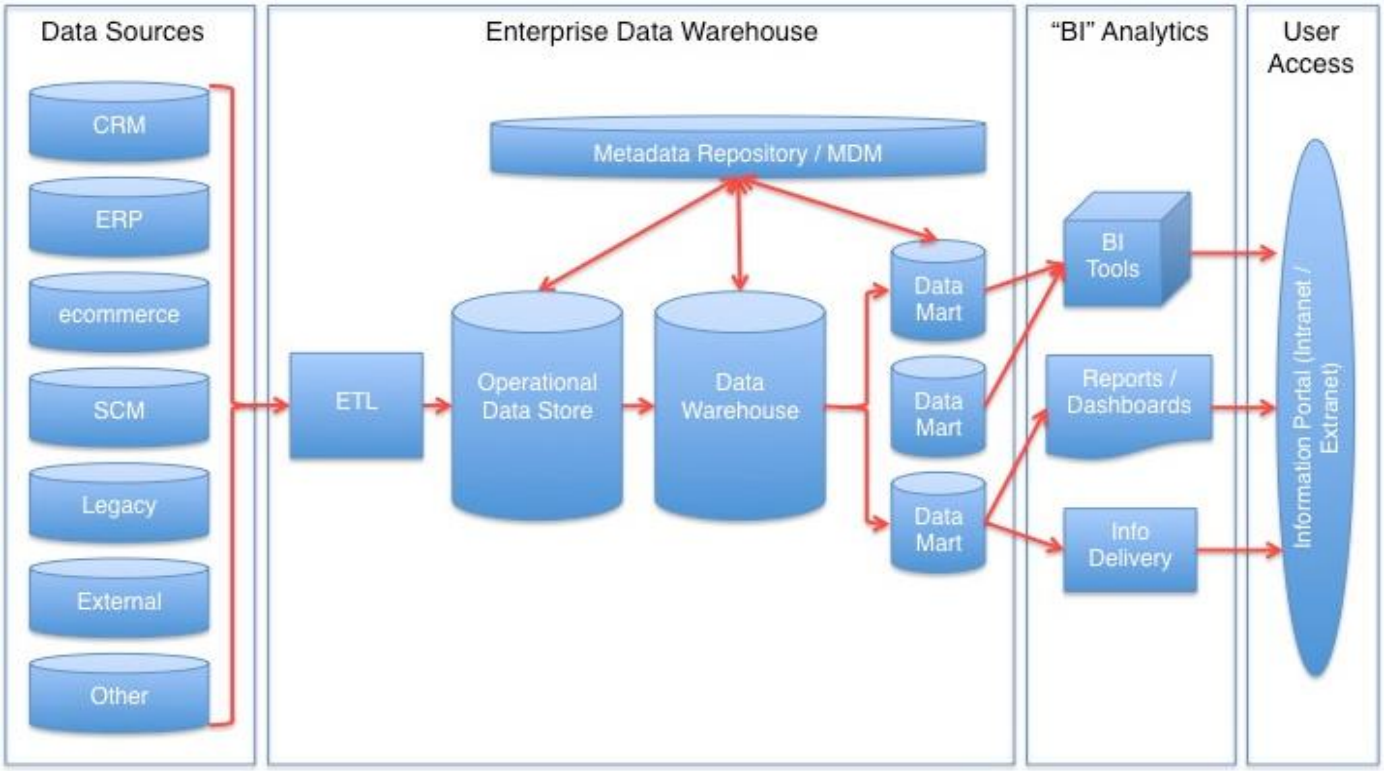

Figure: A data warehousing architecture

Using Data Warehouse Information:

There are decision support technologies that help utilize the data available in a data warehouse. These technologies help executives to use the warehouse quickly and effectively. They can gather data, analyze it, and take decisions based on the information present in the warehouse. The information gathered in a warehouse can be used in any of the following domains −

- **Tuning Production Strategies** − The product strategies can be well tuned by repositioning the products and managing the product portfolios by comparing the sales quarterly or yearly.
- **Customer Analysis** − Customer analysis is done by analyzing the customer's buying preferences, buying time, budget cycles, etc.
- **Operations Analysis** − Data warehousing also helps in customer relationship management, and making environmental corrections. The information also allows us to analyze business operations.

Functions of Data Warehouse Tools and Utilities

The following are the functions of data warehouse tools and utilities −

- **Data Extraction** − Involves gathering data from multiple heterogeneous sources.
- **Data Cleaning** − Involves finding and correcting the errors in data.
- **Data Transformation** − Involves converting the data from legacy format to warehouse format.
- **Data Loading** − Involves sorting, summarizing, consolidating, checking integrity, and building indices and partitions.
- **Refreshing** − Involves updating from data sources to warehouse.

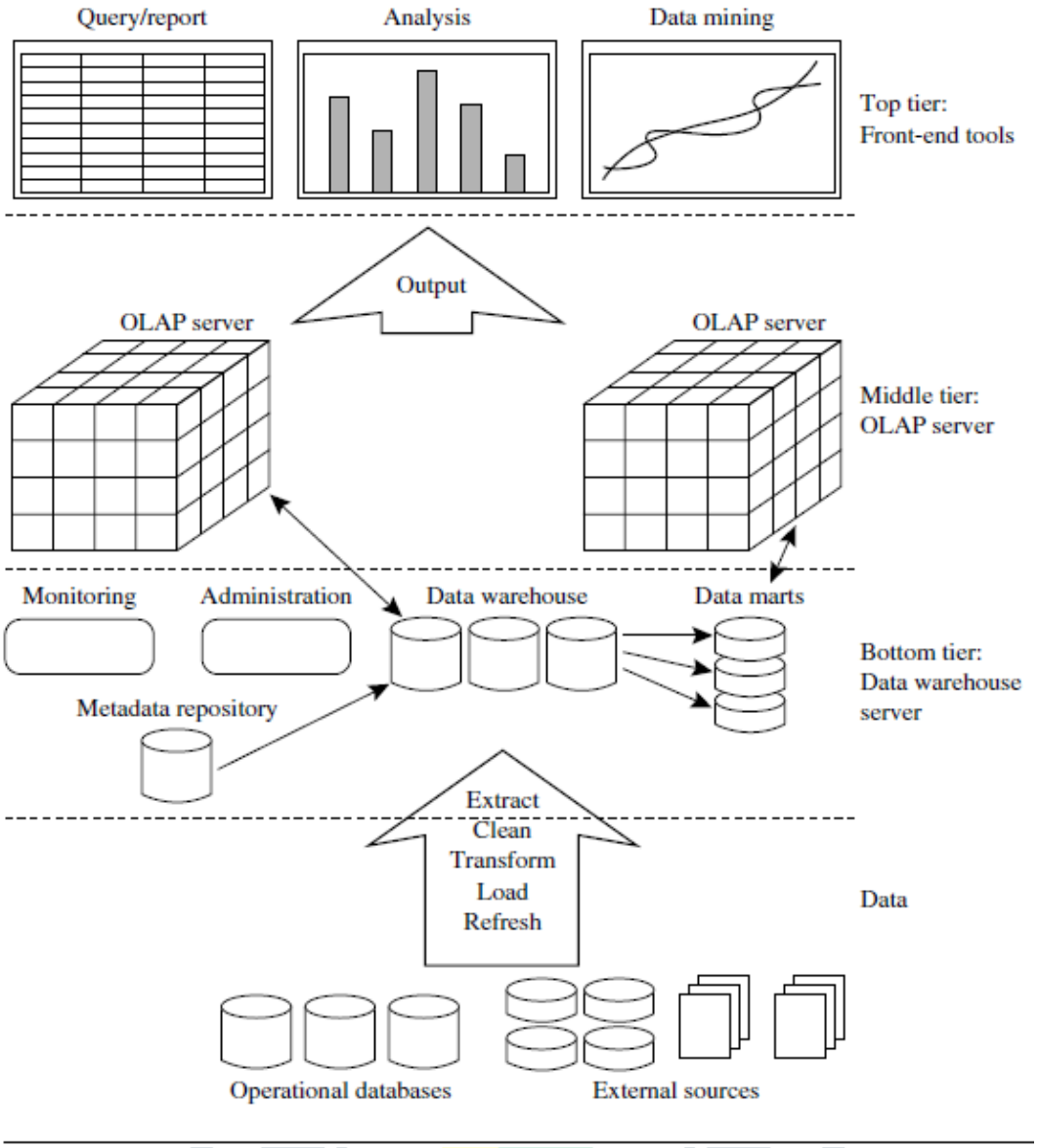

Figure: A three-tier data warehousing architecture

A key to this response is the effective and efficient use of **data** and information by analysts and managers.

These terms refer to the level of sophistication of a data warehouse:

Offline operational data warehouse

Data warehouses in this stage of evolution are updated on a regular time cycle (usually daily, weekly or monthly) from the operational systems and the data is stored in an integrated reporting-oriented data

Offline data warehouse

Data warehouses at this stage are updated from data in the operational systems on a regular basis and the data warehouse data are stored in a data structure designed to facilitate reporting.

On time data warehouse

Online Integrated Data Warehousing represent the real time Data warehouses stage data in the warehouse is updated for every transaction performed on the source data

Integrated data warehouse

These data warehouses assemble data from different areas of business, so users can look up the information they need across other systems

Understanding the **characteristics of a data warehouse** expound what a data warehouse really is, and its importance. There are four fundamental characteristics of data warehouse which are listed below:

Subject Oriented: In data warehouse, data are categorized by subjects such as products and sales which provide the decision makers with an all-encompassing picture of an organization and distinguish it from operational database which is product oriented and primarily deals with transactions that modify the database.

Integration: Data warehouse is expected to be completely integrated as it is a place where data from various places are stored. Thus, all data must be in a consistent format.

Time Variant: Time dimension is very important since data warehouse contains historical data that help with the forecast and decision making.

Non Volatile: End users cannot update or change data, once they are keyed to the warehouse. Data in the data warehouse are loaded and refreshed from operational systems. Thus, data warehouse mostly deals with data access. With these characteristics in mind, it is apparent that data warehouse is the decision support tool that organizations can make use of. The next section is going to address what role a data warehouse play in an organization.

Functionality of data warehouse:

- Roll-up
- Roll-down
- Pivot
- Slice and Dice
- Sorting
- Selection
- Derived (computed) attributes

On the basis of architecture, there are three data warehouse models:

(a) **Enterprises Warehouse**: An Enterprises warehouse collects all of the information about subjects concerning the entire organization. It provides corporate wide data integration.

(b) **Data Mart**: Data marts are usually implemented on low cost departmental servers. The implementation cycle of data mart is generally measured in weeks rather than months or year.

(c) **Virtual ware house**: A virtual warehouse is a set of views over operational databases. For efficient query processing, only some of the possible summary views may be materialized.

Data warehouse usage:

Data warehouse contains integrated and processed data to perform data analysis at the time of decision making and planning. It is a very important tool for business executives.

(a) Information Processing: It supports querying basic statistical analysis, and reporting using crosstabs, tables, charts, or graphs.

(b) Analytical Processing: It is helpful in multidimensional analysis of data warehouse data and support basis OLAP (On-Line Analytical Processing) operations (Slice-dice, drilling, pivoting).

(c) Data Mining: The data mining is a process of intelligent pattern discovery from data warehouse. It supports associations, constructing analytical models, performing classification and predication, and presenting the mining results using crosstabs, graphs, and other visualization tools.

THE IMPORTANCE OF DATA WAREHOUSES IN ORGANIZATIONS:

Data warehousing is an increasingly important business intelligence tool. The importance of data warehouse in an organization tends to explain why data warehouse in needed in an organization. Since a data

warehouse reflects the business model of an enterprise that make is an important aspect of an organization. The following are some importance of data warehouse:

1. Repository for historical information for comparative and competitive analysis.
2. Ability to enhanced data quality, consistency and completeness.
3. Real-time consolidation of financial data becomes practical.
4. The IT costs and staff dedicated to reporting are greatly reduced.
5. Allow business process redesign and align with business strategy.
6. Give end users freedom to carry out wide-ranging analysis in various manners.
7. Simplify the process of data access.
8. Identify market trends.
9. Reduce operation costs.
10. Allow business process redesign and align with business strategy.

The ability of a data warehouse to analyze and execute business decisions based on data from multiple sources is of utmost (most extreme) importance. For example, an organization has collected valuable data and stored it in 10 databases. A data warehouse is not only a convenient way to analyze and compare data in all the databases, but it can also give historical data and perspective. Using data warehouse, one can look at past trends, whether they be product sales or customers or whatever and may be do some predictions of what is going to happen in the future. Also data retrieved from multiple databases is not constrained by the tables in each of those databases. A data warehouse by itself does not create value, but value comes from the use of the data in the warehouse. In support of a low cost strategy, the data warehouse can provide savings in billing processes, reduce fraud losses, and reduce the cost of reporting. The data warehouses can provide analysts with pre-calculated reports and graphs. This increases the productivity of business analysts. Most companies can benefit from a data warehouse when the proper tools are in place and users are trained in analysis of results.

Organizational challenges of data warehousing:

An analysis of data warehousing projects in large Swiss and German service companies (Meyer 2000, for project details refer to the competence center intranet) shows that the following issues can be regarded as the most important organizational challenges of data warehousing:

Alignment with regard to company goals: Different units involved in a data warehouse project usually have different - maybe even conflicting - goals. Decision makers want their specific information requirements to be covered flexibly and in real time. Operations units want to manage their daily business efficiently and securely. Information management wants to build a common platform that effectively decouples as many decision support systems and operational systems as possible. In an early phase of a data warehouse projects, the involved units have to agree on common project goals and have to solve obvious or latent conflicts between goals.

Clear responsibilities for data: In contrast to business processes which have usually been assigned to a process owner, it is often not clear whether certain data are owned by information processing units, data management units, or non-IS business units. Have high-quality data to be delivered by some service unit, or are decision makers responsible for collecting and cleansing of data? How should responsibilities be assigned for data that go through a multi-step derivation sequence from operational systems through various staging areas, the enterprise-wide data warehouse, various aggregation/selections stages, data marts, and finally data pools of decision support systems? The concept of data ownership is intended to create a sound foundation for the definition of roles, responsibilities, and data warehousing processes.

Data quality management: Data quality relates not only to the correctness of data, but also to the way data is provided and used. Since operational systems are intended to support business transactions and not to support business decisions, data quality can be sufficient (for information processing units) and insufficient (for non-IS business units) at the same time. For a specific decision, often not all necessary data that decision makers need (or believe to need) can be provided. But quality requirements for decision support often conflict operational systems management tools (e.g. time-consuming integrity checks). For an effective data quality management, business specialists (who only are able to assess data quality from a usage perspective) have to be made responsible for sourcing and transformation even if operational systems are affected.

Integration management: Better coordination between business processes and management processes allows data warehousing to be implemented more effectively because the right, critical information can be provided more accurately. Integration management, therefore, has to identify performance indicators that are both useful for specifying effective management processes and can be provided accurately and timely by operational systems.

Multi-level structure: A physically centralized, enterprise-wide data integration layer can only be implemented in small companies. For a large company, several layers of data integration between operational systems and decision support systems have to be implemented. The most common solution is to differentiate at least an enterprise data warehouse layer and a data mart layer. If *n* layers exist, coordination processes between data-providing systems (and responsibilities) and data-consuming systems (and responsibilities) have to be implemented on *n-1* levels (see the following figure).

Sustainability: Traditional development projects can be characterized by restricted resources, restricted running time, and uniqueness. In contrast, data warehousing is a permanent process (Gardner 1998, p. 54): After an initial development phase, not only stable operations and reliable data supply must be provided, but also continuous improvements and adjustments are needed to reflect changing decision support needs and modifications of operational systems. For a sustainable implementation of data warehousing, dedicated permanent roles and structures must be created.

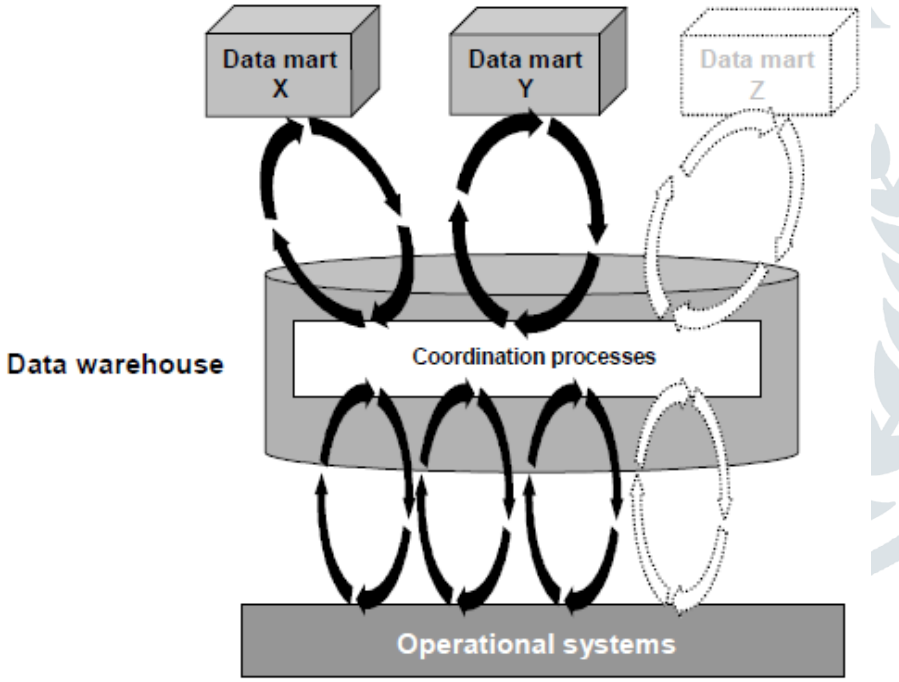

Figure: Multi-level coordination in a multi-layer data warehousing infrastructure

KEYS TO A SUCCESSFUL WAREHOUSING PROJECT

1. Identified and involved warehouse users.
2. Strong and committed leadership.
3. Diversified project team.
4. Established partnerships with all key source data holders.
5. Incremental project plan the produces fast results.
6. Correct design philosophy.

DATA MINING

Data mining has opened a world of possibilities for business. In the meantime, information continues to grow and grow. A 2017 research on big data reveals that 90% of world data is from after 2014 and its volume

doubles every 1.2 years. In this context, data mining is a strategic practice considered important by almost 80% of organizations that apply business intelligence, according to Forbes.

An *information* extraction activity whose goal is to discover hidden facts contained in **databases.** Data mining refers to the mining or discovery of new information in the term of pattern or rules from vast amount of data. Data mining helps in extracting meaningful patterns that cannot be found necessarily by merely querying or processing data or metadata in the data warehouse.

Data mining turns a large collection of data into knowledge. A search engine (e.g. Google) receives hundreds of millions of queries every day. Each query can be viewed as a transaction where the user describes her or his information need.

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.

Data mining involves six common classes of tasks:

- Anomaly detection (outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.
- Association rule learning (dependency modelling) – Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
- Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
- Regression – attempts to find a function which models the data with the least error that is, for estimating the relationships among data or datasets.
- Summarization – providing a more compact representation of the data set, including visualization and report generation.

Data mining is a computational process used to discover patterns in large **data** sets. All commercial, government, private and even Non-governmental **organizations** employ the use of both digital and physical **data** to drive their business processes. **Data mining** is widely used to gather knowledge in all industries.
The benefits of mining data covers almost all facets of life which include; gaming, policing, business, science, engineering, human rights organizations and surveillance.
When dealing with a plethora of data sources, confusion usually arises for the incompetent researcher who has neither the tools nor experience to handle bulky projects. Outsourcing data mining projects reduces the risk of dredging up ineffective data.

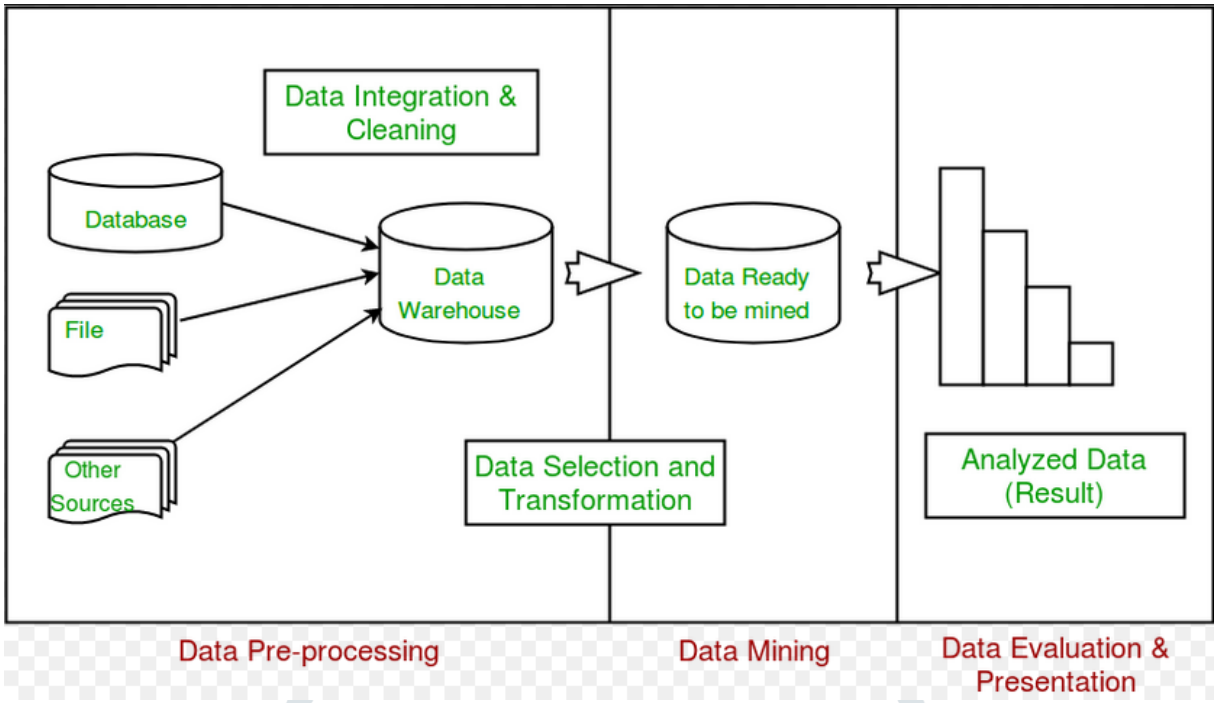

Figure: A data mining process

Data mining helps organizations get the necessary information needed to handle different processes as quickly as possible. For the surveillance and policing industries, the need to meet deadlines and process information quickly is of the utmost importance.

KDD (Knowledge Discovery in Database)

The **KDD** is in the expanding process. The term Knowledge is very broadly interpreted as involving some degree of intelligence. The knowledge is classified often as (a) inductive, and (b) deductive.

Business Intelligence:

Business intelligence usually refers to the information that is available for the enterprise to make decisions on. A data warehousing (or data mart) system is the backend, or the infrastructural, component for achieving business intelligence. Business intelligence also includes the insight gained from doing data mining analysis, as well as unstructured data (thus the need for content management systems). For our purposes here, we will discuss business intelligence in the context of using a data warehouse infrastructure.

The Knowledge discovery process comprises six phases:

- Data selection
- Encoding
- Data cleaning
- Enrichment
- Data Mining, and
- Reporting

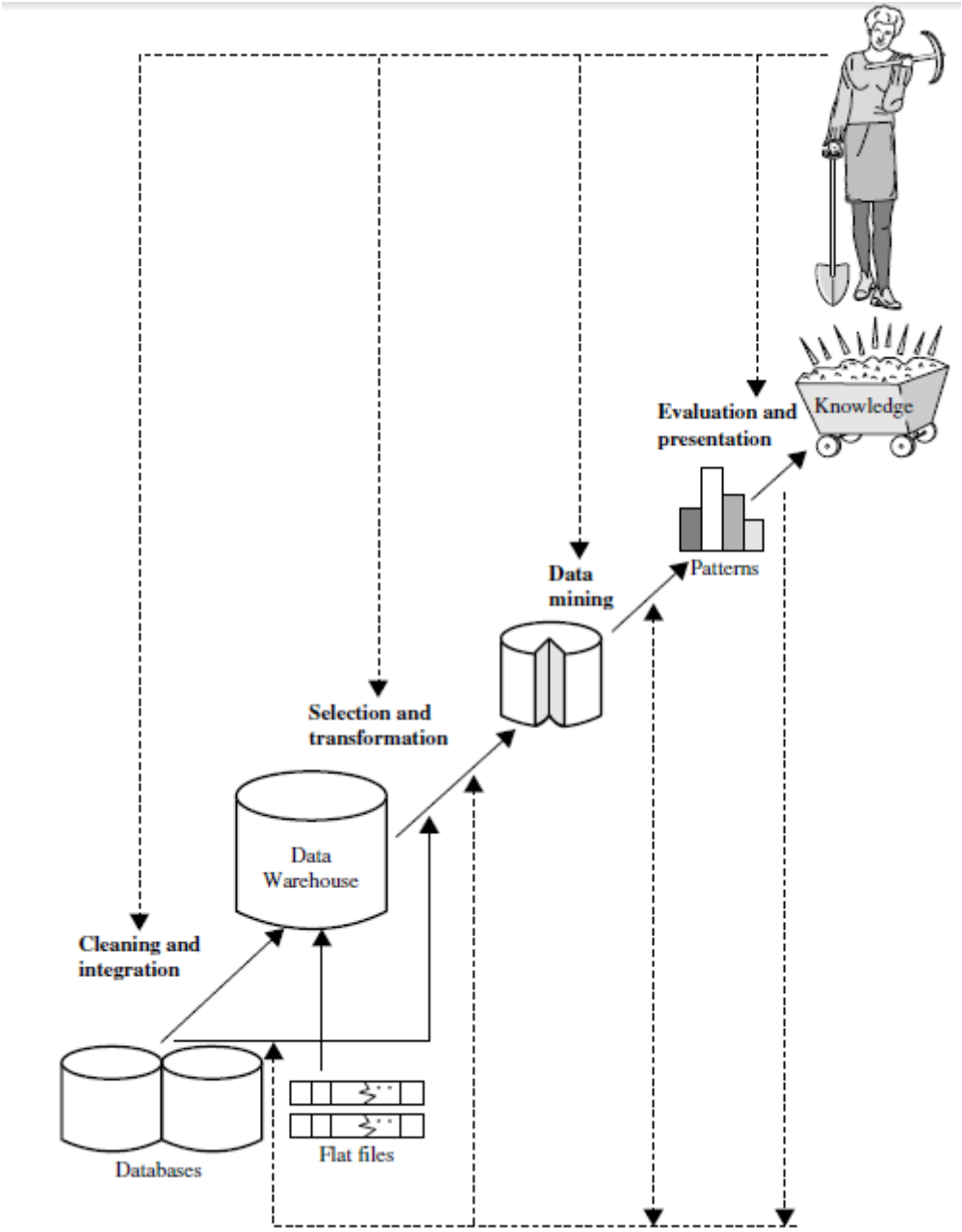

Figure: Data mining as a step in the process of knowledge discovery

Real life example of Data Mining:

Amazon also **uses data mining** for marketing of their products in various aspects to have a competitive advantage. Smart retailers as **Amazon** make effective **use** of **data** gathered through effective sources and **use** the outcomes more reasonably. Also the customers have control over information they want to share or not. It helps the organization to identify the relationship between the internal and external factors affecting the marketing sales, profits and customer satisfaction. This approach is very useful for marketing as it helps retailer to check the sales record of purchases and give promotion to customer reviewing his purchase history. The warranty card data can be used for creating promotion strategies to attract more customers. It thus increases customer loyalty and profitability of company.

***Data Mining Applications**:*

Data mining does not replace skilled business analysts or managers, but rather gives them powerful new tools to improve the job they are doing. It is a something out from traditional tracks of decision making and business planning. It offers great promises in helping organizations to uncover patterns hidden in their *data* that can be used to predict the behavior of customers, products and processes.

1. **Marketing:** Data mining is used to explore increasingly large databases and to improve market segmentation. By analysing the relationships between parameters such as customer age, gender, tastes, etc., it is possible to guess their behaviour in order to direct personalised loyalty campaigns. Data mining in marketing also predicts which users are likely to unsubscribe from a service, what interests them based on their searches, or what a mailing list should include to achieve a higher response rate.
2. **Retail:** Supermarkets, for example, use joint purchasing patterns to identify product associations and decide how to place them in the aisles and on the shelves. Data mining also detects which offers are most valued by customers or increase sales at the checkout queue.
3. **Banking:** Banks use data mining to better understand market risks. It is commonly applied to credit ratings and to intelligent anti-fraud systems to analyse transactions, card transactions, purchasing patterns and customer financial data. Data mining also allows banks to learn more about our online preferences or habits to optimise the return on their marketing campaigns, study the performance of sales channels or manage regulatory compliance obligations.
4. **Medicine:** Data mining enables more accurate diagnostics. Having all of the patient's information, such as medical records, physical examinations, and treatment patterns, allows more effective treatments to be prescribed. It also enables more effective, efficient and cost-effective management of health resources by identifying risks, predicting illnesses in certain segments of the population or forecasting the length of hospital admission. Detecting fraud and irregularities, and strengthening ties with patients with an enhanced knowledge of their needs are also advantages of using data mining in medicine.
5. **Television and Radio:** There are networks that apply real time data mining to measure their online television (IPTV) and radio audiences. These systems collect and analyse, on the fly, anonymous information from channel views, broadcasts and programming. Data mining allows networks to make personalised recommendations to radio listeners and TV viewers, as well as get to know their interests and activities in real time and better understand their behaviour. Networks also gain valuable knowledge for their advertisers, who use this data to target their potential customers more accurately.
6. **Telecommunication Industry**: Telecommunication industries are backbone of any organization. The mismanagement in communication industry can spoil many business organizations, industries, universities, military systems etc, because it does not carry only normal data but also confidential data. In telecommunication industry data mining is used for identifying telecommunication patterns, catching fraudulent activities, making better use of recourses, and improving quality of services.
7. **Biomedical and DNA data analysis**: The genetic engineering is the young discipline of engineering which is totally based on the structure of genes. There are 1065 genes are present in human body and a pair of gene is responsible to control any specific characteristics. The genes are present in DNA (Deoxyribo Nuclic Acid) which is made from nucleotides: Adenine (A), Cytocine (C), Guanine (G), and Thymine (T). The gene engineering is boon for person suffering from hereditary disease. After fertilization, sequence of diseases carrying gene in zygote is changed.
8. **Image processing**: Data mining provides efficient tools for image processing.
9. **Financial data analysis**: The bank and business organizations are often based on data mining for collection, high quality accuracy, better customer service and satisfaction, loan payment, credit rating etc.
10. **Retail Industry:** The customers are major objective for any business organization. The products and services are designed to focusing customers. Data mining is helpful in prediction of behavior of customers in market. It is used to identify customer buying behavior, improve customer service, enhance customer and goods ratio, design more effective goods and discover cost effective transportation methods etc.
11. **Manufacturing sectors**: Manufacturing section of any organization is dependent on data mining for designing of most acceptable products. The market is the name of competition, if there is no any competition your monopoly help you to obtain high profit, but now a days monopoly can exists not for long times. The data mining helps executive to design customer oriented products.

Web Data Mining

The World Wide Web provides rich sources for **data mining**. It is a too huge for effective data warehousing and data mining, and too complex and heterogeneous because it has no standard and structure. The WWW is huge, widely distributed, global information service center for:

- Information services: news, advertisements, consumer information, financial management, education, government, e-commerce, etc.
- Hyper-link information.
- Access and usage information.

Data warehousing and data mining:

- The goal of a data warehouse is to support decision making with data. Data mining can be used in conjunction with a *data warehouse* to help with certain types of decisions.
- Data mining can be applied to operational databases with individual transactions. To make data mining more efficient, the data warehouse should have an aggregated or summarized collection of data. Data mining helps in extracting meaningful new patterns that cannot be found in the data warehouse.
- *Data mining* applications should therefore be strongly considered early, during the design of data warehouse.
- **Data mining** tools should be designed to facilitate their use in conjunction with data warehouses.

Outsourcing bulky data mining projects ultimately helps an organization manage its human resources, its capital expenditures and focus on core business areas.

CONCLUSION:

Data warehousing and data mining talks about the change in business trends, decision support systems and business intelligence applications. All the small and big industries are collecting and using data from various sources to identify their own business trends. Organizations understand the strengths and the weaknesses of their competitor improve their progressing speed towards the goal and expand their business empire. A data warehouse is a solution to a business problem not a technical problem. The data warehousing and data mining need to constantly overcome obstacles that are yet undefined and help the organization in decision making and improves the goodwill of organization. Data mining helps in securing and processing the data into understandable chunks, where warehousing helps in analyzing the data and put it in such a way as to facilitate comparison between trends, analyzing the data for the business predictions and accelerate decision making. The data warehousing and data mining are the best tools for business intelligence applications.

FUTURE SCOPE

Data mining offers an important approach to achieving values from the data ware house for use in decision support. Data warehousing becomes a standard part of an organization, there will be efforts to find new ways to use the data. Data warehousing and data mining will bring several new challenges in future like
1. Regulatory constraints may limit the ability to combine sources of disparate data.
2. These disparate sources are likely to contain unstructured data which is hard to store.
3. The internet makes it possible to access data from virtually “anywhere”. This just increases the disparity.

The Drive for a New Kind of Data Warehousing:

A new kind of data warehousing is essential to new BI deployment, as much of the inefficiency in older BI deployments lies in the time and energy wasted in data movement and duplication. A few factors are driving the development and future of data warehousing, including:

- **Agility** – To succeed today, businesses must use collaboration more than ever. Instead of having separate departments, teams, and implementations for things like data mining and analysis, IT, BI, business, etc., the new model involves cross-functional teams that engage in adaptive planning for

continuous evolution and improvement. This kind of model cannot function with old forms of data warehousing, with just a single server (or set of servers) where data is stored and retrieved.

- **The Cloud** – More and more, people and businesses are storing data on the cloud. Cloud-based computing offers the ability to access more data from different sources without the need for massive amounts of data movement and duplication. Thus, the cloud is a major factor in the future of data warehousing.
- **The Next Generation of Data** – We are already seeing significant changes in data storage, data mining, and all things relate to big data, thanks to the Internet of Things. The next generation of data will (and already does) include even more evolution, including real-time data and streaming data.

Today the challenge is to design data warehousing and data mining applications that are reliable, easy to use and supports effective decision making. As the amount of data increases in the future, data mining and data warehousing will become a valuable tool in industries/business. Data mining will be helpful in finding new quality products, predict the benefits from that quality data, and can help optimize use of sales resources like manpower and marketing. In the future the data warehousing and data mining will be very helpful BI tools for the organizations. Business intelligence (BI), big data and data analytics to analyze raw data and create faster, more effective business solutions.

Companies like SAP are working on that right now. With the launch of the BW/4HANA data warehousing solution running on premise and Amazon Web Services (AWS) and others like it, we can see how businesses can combine historical and streaming data for better implementation and deployment of new BI strategies. This system and others like it work with Apache Spark and Apache Hadoop, as well as other programming frameworks to bring data and systems of insight into the 21st century and beyond.

REFERENCES

[1]The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.

[2] Nwakanma Ifeanyi *et al*, International Journal of Computer Science and Mobile Computing, Vol.3 Issue.10, October 2014, pg. 451-455.

[3] International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-3, Issue-1, March 2013.

[4] Monika Pathak et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (6) , 2013, 995-999.