# A Review: Analysis on Data Warehousing and Data Mining

**Vaibhav Singh[1],Prof.Ashwini Ghate[2]**

[1]Research Scholar, Department of Information Technology,[2]Assistant Professor, Department of Computer Science and Engineering.
PIGCE, Nagpur(India)

***ABSTRACT—A Review of literature affecting to data warehouse implementations has been undertaken. It was found that the views of data warehouse experts in particular have changed over the period's, to the extent that fewer authors place any emphasis on the need for a clear business purpose before boarding on a data warehouse project. Data warehouse (DW) is key and central to BI applications in that it assimilates some varied data sources, mainly planned transactional databases. However, current studies in the area of BI suggest that, data is no longer always accessible in only to planned folders or format, but they also can be drawing from formless sources to make more power the directors' analysis. So, the ability to manage this current data is critical for the success of the decision making process. The case study review in this paper supports the notion of strategic alignment but it is the mapping of their experiences to the configuration school of strategic management that explains the degree of success.***

***KEYW ORDS —Business Intelligence (BI), Data Warehousing, Data Analysis, Transactional data***

## I.INTRODUCTION

Late 1997, the author embarked on a study to define 'Best Practice for Implementing a Data Warehouse', which was used to explain the experiences of a bank's data warehouse project and ultimate implementation failure Data warehousing is the process of collecting data to be stored in a managed database in which the data are subject-oriented and integrated, time variant, and nonvolatile for the support of decision-making (Inmon, 1993). Data from the different operations of a corporation are reconciled and stored in a central repository (a data warehouse) from where analysts extract information that enables better decision making. The operational data needs of an organization are addressed by the online transaction processing (OLTP) systems which is important to the day-to-day running of its business. Data warehouses support OLAP applications by storing and maintaining data in multidimensional format. Data in an OLAP warehouse is extracted and loaded from multiple OLTP data sources (including DB2, Oracle, SQL Server and flat files) using Extract, Transfer, and Load (ETL) tools.Data can then be aggregated or parsed, and sliced and diced as needed in order to provide information. Most of the practitioners of Data warehouse subscribe to either of the two approaches." Integrated" means that the data are stored in consistent formats, naming conventions, in measurement of variables, encoding structures, physical attributes of data, or domain constraints.

## II.LITERATURE REVIEW:

Eckerson (2003) from the Data warehouse institute did study on the success factor in implementing BI, systems in organizations and the role of data warehouse [1] in this process. Eckerson (2003) views the BI process holistically as a "data refinery" Data from different OLTP systems are integrated, which leads to a new product called information. The data warehouse staging process is responsible for the transformation [2]. Users equipped with program such as specialized reporting tools, OLAP tools and data mining tools transform the information into knowledge. Kimball (1996) includes this as part of the data warehouse. According to Kimball, the aim of the data warehouse is to give end-users (mostly managers) easy access to data in the organization. In order to do this, it is necessary to capture everyday operational data from the operational systems of the organization. These are the OLTP system. The data from the source systems go through a process called data staging to the presentation servers (Kimball et al 1996). The data at the staging process involves four processes namely Extract, Transformation, Loading and finally presentation. It is on the presentation stage that the data marts, which represent business areas in the organization is built on. There is a difference between the data warehouse and business intelligence architecture as advocated by the two known scholars in the industry, (Inmon, 1993) advocates the use of data-driven method. Data Warehouse Design Concepts [6], the design of the database depends on the approaches of the father of data warehouse developers. The two-design processes are referred to as Top-down process, as described by Bill Inmon and Bottom-up as described by Ralph Kimball. These are explained in detail below.

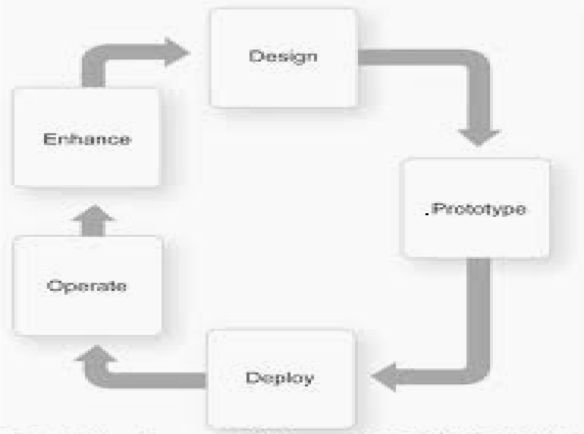

**Fig 1 Development Lifecycle Model**

Top-Down Model These was Introduced by Bill Inmon, the process begins with an Extraction, Transformation, and Loading (ETL) process working from legacy and/or external data sources. Extraction transformation, process data from these sources and output it to a centralized Data Staging Area. Following this, data and metadata are loaded into the Enterprise Data Warehouse and the centralized metadata repository. Once these are constituted, Data Marts are created from summarized data warehouse data and metadata. In the top-down model, integration between the data warehouse and the data marts is automatic as long as the discipline of constituting data marts as subsets of the data warehouse is maintained. 2.7.2 Bottom-Up Model The central idea in Bottom-up model is to construct the data warehouse incrementally over time from independently developed data marts. would be adopted, which is the Kimball's development lifecycle, this states with one data mart (e.g. Sales) later on further data mart are added e.g. Marketing and Collection. Data flows from sources into data marts, then into the data warehouse. It is also implemented in stages (faster) Due to the time constraint and project limitation, it is easier to complete a process for a subset of a company based on the data mart and link it up as the business grows. The stages proposed for the process include Investigation, Data Warehousing: A data ware house is collection of data designed to support management in the decision making process. It is a subject-oriented, integrated, time-variant, non-updatable collection of data used in support of management decision-making processes and business intelligence. A data warehouse is a physical separation of an organization's Online Transaction processing (OLTP) systems from its decision support systems (DSS). It includes a repository of information that is built using data from the distributed, and often departmentally isolated, systems throughout the organization. Data warehousing is the process, where organizations extract meaning and information decision making from their informational assets through the use of data warehouses. It is storing data effectively so that it can be accessed and used efficiently. Different organizations collect different types of data, but many organizations use their data the same way, in order to create reports and analyze their data to make quality business decisions. Data warehousing is usually an organizational wide repository of data.

## III.RELATED WORK

Data warehousing is about turning data into information so that business users have more knowledge with which to make competitive decisions. Data in the data warehouse can be modeled and analyzed to make the organization more competitive. Data in the warehouse are organized by subject rather than application, so the warehouse contains only the information necessary for decision support processing. The data in the warehouse are collected over time and used for comparisons, trends, and forecasting. These data are not updated in real time, but are migrated from operational systems on a regular basis when data extraction and transfer will not adversely affect the performance of the source operational systems.Following are some of the sources where important information about a financial institution can be found:

1. The customers - what they think, what they want, how they see the bank or the financial institution as a provider of service both materially and psychologically.

2. The employees - what they know, their perceptions about the bank or the financial institution

3. The legacy systems

4. The actual data, information and knowledge that flows through the bank

5. The business environment.

**An organization should implement a Data Warehouse because:**

- The primary motivation for a bank to implement a data warehouse usually centers around improving the accuracy of information used in the decision-making process.
- The other important function of data warehouse is to consolidate the rules of business logic practiced by a bank.
- It helps a bank learn about its customers, including their buying habits and patterns.
- The bank or financial institution's functioning can be understood in historical perspective, which allows better tracking and responding to business trends, facilitates forecasting and planning efforts, and thereby leading to strategic business decision.

**The major steps for data warehouse implementation are:**

**a) Subject definition:** It is determining which subjects will be created and populated in the data warehouse

**b) Data capture:** The core of data capture is Data Replication, is defined as 'a set of techniques that provides comprehensive support for copying and transforming data from source to target location in a managed, consistent and well-understood manner'.

**c) Data transformation:** It is used to convert and summarize operational data into a consistent, business-oriented format.

**d) Metadata management:** To access to the data warehouse, it is necessary to maintain some form of data, which describes the data warehouse. This data is called metadata. It masks the complexities of the technology of a Data Warehouse from the users. It acts as a critical aid for navigating the data warehouse.

**e) Loading the warehouse:** This is the periodic loading of static snapshots from the online transaction-processing environment gives the data warehouse its time-variant quality.

## IV.DATA WAREHOUSING

Data warehouse has more than one definitions. The most common one is defined by Bill Inmon we defined it as the following: "A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process" [1].As defined , any data warehouse (DW) should have the following characteristics :

- **Subject-oriented:** DW can be used to analyse any subject.
- **Integrated:** DW integrates current and historical data from different sources.
- **Time-variant:** DW keeps historical data of different time.
- **Non-volatile collection of data:** content of DW should not be changed. It is historical data. Unlike the modeling techniques used to design regular databases - Entity Relationship model, data warehousing is designed by using dimensional modeling techniques [3]. Data warehousing modeling is complex.

It needs:

1) knowledge of the business processes,

2) Understanding the structural and behavioral system's conceptual model, and

3) being familiar with data warehousing techniques

**Exploiting the Data Warehouse:**A data warehouse is incomplete until it provides the exploitation tools that enable end users to view, analyze, and report data in ways that support their decision making. Depending on the end users' requirements, data warehouse exploitation tools may be anything from ready-to-use, simple query and reporting tools to multidimensional analysis tools to advanced Executive Information System (EIS) applications to tools for complex analysis and modeling.

**Benefits of implementing a Data warehouse:** The business benefits derived from implementing a data warehouse are very significant. Data warehousing offers organizations an opportunity to reinvent the tools used for decision making by making large amounts of data collected by business yield copious amounts of useful information about customers and business environment.

**Benefits of Data warehouse in financial services industry:**

- **In financial institution**: A data warehousing solution provides accurate, consistent and comprehensive information, which in turn leads to improved decision support systems. The critical areas in which it helps include Risk management, Asset and liability management, Analyses of customer and market information as well as Profitability Analysis like customer profitability and product profitability.
- **Data Mining:** It is the procedure by which analyst utilize the tools of mathematics and statistical testing
- **applied to business**- relevant, historical data in order to identify relationships, patterns, or affiliations among variables or sections of variables in that data to gain greater insights into the underpinnings of the business process. The broad categories of application of Data Mining and Business intelligence techniques in the banking and financial industry vertical may be viewed as follows:
- **Risk Management:** Credit and financial market risk, liquidity risk, operational risk, or concentration risk can be analyzed using advanced database and data mining technology.
- **Trading:** The goal of this technique is to spot times when markets are cheap or expensive by identifying the factor that are important in determining market returns. The trading system examines the relationship between relevant information and piece of financial assets, and gives you buy or sell recommendations when they suspect an under or overvaluation.
- **Portfolio Management:** Risk measurement approaches on an aggregated portfolio level quantify the risk of a set of instrument or customer including diversification effects. On the other hand, forecasting models give an induction of the expected return or price of a financial instrument. With the data mining and optimization techniques investors are able to allocate capital across trading activities to maximize profit or minimize risk.

## V.DATA MINING

Data Mining (DM) is a combination of Database and Artificial Intelligent used to extract useful data from huge amount of datasets to help the users to make better decisions. It is usually used as a decision support system [5].

### A.Data Mining Usage

Having enormous volume of data, makes it very difficult for human to analyze and get useful info. This causes the importance of using Data Mining techniques. DM is used in different areas to help to extract useful information then make better decisions. For example, DM can be used for marketing purposes. It can help by giving useful information about the best media and time to publish an advertisement which would help to increase the sales of a product.

### B.Data Mining Process

Data mining process is not an easy process. It is complicated and has feedback loops which make it an iterative process. Figure 1 [7] shows the steps of data mining process. It also shows that the steps might be repeated and sometimes it is possible to restart the entire process from the beginning. Actually, the data mining process involves six steps:

1) Problem definition
2) Data Preparation
3) Data Exploration
4) Modeling
5) Evaluation
6) Deployment

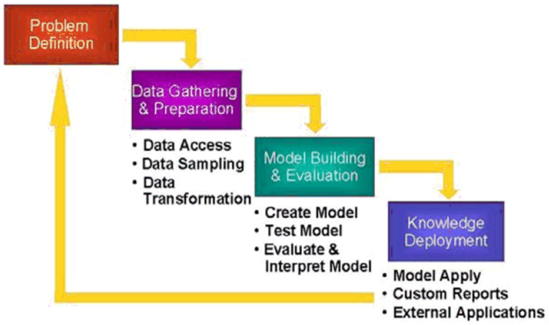

**Fig 2 Data Mining Process Model**

## VI.EXPLANATION/DISCUSSION OF MODEL

### 1.Data Mining Process – Goal

The Data Mining process is not a simple function, as it often involves a variety of feedback loops since while applying a particular technique, the user may determine that the selected data is of poor quality or that the applied techniques did not produce the results of the expected quality.

### 2.Problem Definition

A data-mining project starts with the understanding of the business problem. Data mining experts, business experts, and domain experts work closely together to define the project objectives and the requirements from a business perspective. The project objective is then translated into a data mining problem definition. In the problem definition phase, data.

**3.Data Exploration**

Domain experts understand the meaning of the metadata. They collect, describe, and explore the data. They also identify quality problems of the data. A frequent exchange with the data mining experts and the business experts from the problem definition phase is vital. In the data exploration phase, traditional data analysis tools, for example, statistics are used.

**4.Data Preparation**

Domain experts build the data model for the modeling process. They collect, cleanse, and format the data because some of the mining functions accept data only in a certain format. They also create new derived attributes, for example, an average value. In the data preparation phase, data is tweaked multiple times in no prescribed order.

**5.Modeling**

Data mining experts select and apply various mining functions because you can use different mining functions for the same type of data mining problem. Some of the mining functions require specific data types. The data mining experts must assess each model. In the modeling phase, a frequent exchange with the domain experts from the data preparation phase is required.

**6.Evaluation**

Data mining experts evaluate the model. If the model does not satisfy their expectations, they go back to the modeling phase and rebuild the model by changing its parameters until optimal values are achieved.

## Conclusion:

Nowadays we have enormous volume of data which lead to the necessity of using data warehousing and data mining. Data warehouse is used as a central store of a subject oriented, integrated, time-variant and non-volatile collection of data from different sources (operational databases) [1]. For faster performance, data warehousing organizes data in a different architecture – fact table and dimension tables [4]. For that reason, modeling the data warehouse is unlike modeling the operational database. A dimensional modeling is used to model the data warehouse (star schema, snowflake schema, or galaxy schema) but the operational database uses entity relationships diagram [3].Data mining has become an important tool which can extract useful information from the huge amount of data we have nowadays. It also may help to extract information from the Internet which becomes part of our life. It is a complicated process. It involves six phases: (1) Problem definition, (2) Data Preparation, (3) Data Exploration, (4) Modeling, (5)Evaluation, and (6) Deployment [7]. It is an iterative process which includes feedbacks between the phases and sometimes needs to repeat the entire process from the beginning. The iterations are needed in the mining process in order to provide better answers which will be used by the users to make better decisions.

## VII.REFERENCES

[1]Xiaoyan ,“Data Mining Based Algorithm for Traffic Network Flow Forecasting” , IEEE, 2003.

[2] C. Y. Fang et. al. “ A System to Detect Complex Motion of Nearby Vehicles on Freeways” , IEEE, 2003, pp. 1122 – 1127

[3]J.Han and M.Kamber, “Data Mining: Concepts and Techniques,” Morgan Kaufmann Publishers, San Francisco, CA, 2006. ISBN: 1-55860-489-8.

[4].JemalAbawajy. Comprehensive analysis of big data variety landscape.International Journal of Parallel, Emergent and Distributed Systems.2015,30(1):5-14.

[5].Ana L.C. Bazzan, FranziskaKlügl. Introduction to Intelligent Systems in Traffic and Transportation. Synthesis Lectures on Artificial Intelligence and Machine Learning. 2013,7(3).

[6]. Emad Felemban, Adil A. Sheikh. A Review on Mobile and Sensor Networks Innovations in Intelligent Transportation Systems.JournalofTransportation Technologies.2014,4(3):196-204.

[7]. Wei Shi, Jian Wu, Shaolin Zhou, Ling Zhang. Variable message sign and dynamic regional traffic guidance. Intelligent Transportation Systems Magazine, IEEE. 2009,1(3):15-21.

[8]. EPJ Data Science. Personalized routing for multitudes in smart cities.EPJ Data Science.2015,4(1).

[9]. Yuan Yuan Zhang, Shi Song Yang, Qing Cai, Peng Sun.

Traffic Flow Forecasting Based on Chaos Neural Network. Applied Mechanics and Materials.2010,20-23:1236-1240.

[10].Muhammad Rauf, Ahmed N. Abdalla, AzharFakharuddin;Elisha. Response Surface Methodology in-Cooperating Embedded System for Bus's Route Optimization. Research Journal of Applied Sciences, Engineering and Technology.2013,5(22):5170-5181.

[11]. Cueva-Fernandez, Guillermo, Espada, JordánPascual, etc. An expert system for vehicle sensor tracking and managing application generation. Journal of Network & Computer Applications.2014,42:178-188.

[12]. Filippo, L., Rindt C. R., McNally, M. G. and Ritchie, S. G. (2001). TRICPS /CARTESIUS:An ATMS Testbed Implementation for the Evaluation of Inter-Jurisdictional Traffic Management Strategies. In Proceedings 80th Annual Meeting of TRB (CD-ROM), Washington, D.C.