Brief Analysis on Data Warehousing and Data Mining

[1]Prof. Gayatri Jagnade [2]Prof. Ashwini Ghate [3]Prof. Saleha I. Saudagar
[1,3]Information Technology Department
[2]Computer Engineering Department
Prof. Ram Meghe Institute of Technology
And Research,Badnera, India

***Abstract*— A brief 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 has similar 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 amorphous source 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.**

Keywords— ***Business Intelligence (BI), Data Warehousing, Data Analysis, Transactional data, automation data***

I. INTRODUCTION

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 file in which the data are subject-oriented and integrated, time variation, and nonvolatile for the support of managerial (Inman, 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.

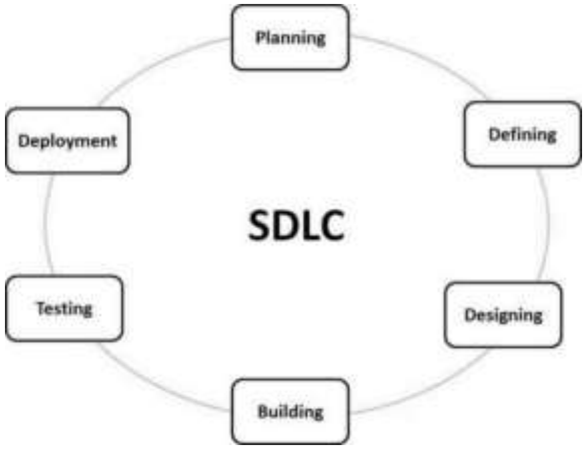
Data can then be aggregated or parsed, and sliced and diced as needed in order to provide information. Most of the practitioners of Data warehouse subscribe to either of the two approaches.

" 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. LITRERATURE REVIEW

Ackerson (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. Ackerson (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 change [2]. Users equipped with program such as specialized reporting tools, OLAP tools and data mining tools transform the in turn 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) trouble-free access to data in the organization. In order to do this, it is needed to capture everyday operational data from the ready systems of the group. These are the OLTP system. The data from the source systems go through a course called data staging to the arrangement server. The data at the staging process involves four processes namely Extract, alteration, 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 disparity between the data warehouse and business intelligence building as advocated by the two known scholars in the industry, (Inman, 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.



Top-Down Model These was Introduced by Bill Inman, the process begins with an Extraction, Transformation, and Loading (ETL) process working from legacy and/otherwise outer data sources. Withdrawal change, process data from these sources and output it to a federal Data Staging Area. Following this, data and metadata are loaded into the Enterprise Data Warehouse and the centralized metadata warehouse. Once these are constitute, Data Marts are created from summarize 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 urbanized 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 compilation. Data flows from sources into data marts, then into the data storehouse. 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 industry 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 result 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, system throughout the group. Data warehousing is the process, where organization 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 organization 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.

An organization should implement a Data Warehouse because:

- The chief inspiration 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 improved 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 occupied in the data warehouse

b) **Data imprison:** The core of data capture is Data duplication, is distinct 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 conversion:** It is used to convert and summarize operational data into a consistent, business-oriented format.

d) **Metadata organization:** 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 episodic loading of static snapshots from the online transaction-processing setting gives the data warehouse its time-variant quality.

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 gung ho. 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:

a) The clientele - what they think, what they want, how they see the bank or the fiscal institution as a source of service both significantly and expressively.

b)The employees – what they know, their perceptions about the bank or the financial institution

b) The inheritance systems.

c) The actual data, in order and knowledge that flows through the bank

d) The commerce environment.

IV. EXPLNATION/DISCUSSION OF MODEL

A. Data Mining Process – Aim

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

B. 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.

C. Data examination

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.

D. Data grounding

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.

E. 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.

F. Assessment

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.

V. CONCLUSION

Now we have huge volume of data which escort to the requirement of using data warehousing and data mining. Data warehouse is used as a inner store of a theme leaning, included, time-variant and non-volatile compilation of data from different sources (operational databases) [1]. For faster presentation, data warehousing organize data in a dissimilar structural design – fact bench and dimension tables [4]. For that reason, modeling the data warehouse is unlike model the operational database. A dimensional modeling is used to model the data warehouse (star plan, 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 vast 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.

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. Journal of Transportation 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:

[1] 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.