EVOLUTION OF DATA WAREHOUSE AND DATA MINING

Reena Hooda*

Nasib S. Gill**

Abstract:

Data warehouse is the requisite of all present competitive business communities' i.e. profitable and non-profitable as well as educational institutions where data is complex, huge and dynamic. The technology advent is not in a day but emerging from late 1980s. Data warehouse is a central repository that merges data from various sources: internal & external along with the web data and upholding the traditional database. Data mining is a tool of data warehousing to extract valuable information, choose a right alternate and make predictions to support the managerial tasks. Though the technologies are still not mature however, due to developments in Information Technologies, high performance softwares & tools and artificial intelligence, it is easier to implement these technologies to face the current challenges. The current paper highlights certain journeys of data warehouse, warehousing and data mining in a gradual and comprehensive way.

Keywords: Data Warehousing, Data Mining. Artificial Intelligence, Statistics, Machine Learning.

* Assistant Professor, Dept. of Computer Sc. & Applications, PDM College of Engineering, Bahadurgarh (Haryana) - India.

** Head, Dept. of Computer Sc. & Applications, Maharshi Dayanand University, Rohtak (Haryana), India.

Introduction:

The concept of data warehousing and data mining dates back to the late 1980s, and has great strides during 1990s with the digital revolutions and dramatic advances in information technology. The data warehousing concept was intended to provide an architectural model for the flow of data from the operational systems to decision support environments that includes data mining too. The historical milieu of data warehouse, data warehousing and data mining is discussed below.

1. Data warehouse and data warehousing

The concept of data warehousing dates back to the late 1980s: when IBM researchers Barry Devlin, and Paul Murphy, developed the "business data warehouse" [1] which was published as "Information Warehouse framework" as early in 1987. Other was Teradata Corporation, which originated the database machine that could handle a terabye of data. It was the first industry-hardened massively parallel computer. Teradata became one of the fastest growing companies at that time [2]. With integrated data warehouse, transactions can be transferred back to the operational systems every day, and this can allow data to be analyzed by companies and organizations. There are a number of devices that will be present in the typical data warehouse. Some of these devices are the source data layer, reporting layer, data warehouse layer, and transformation layer [1] [2].

A number different data sources are availablefor data warehouses. Some popular forms of data sources are Teradata, Oracle database, or Microsoft SQL Server [3]. In essence, concept of data warehousing was intended to provide an architectural model for the flow of data from operational systems to decision support environments and attempted to address the various problems associated with this flow, mainly the high costs associated with it. In the absence of a data warehousing architecture, an enormous amount of redundancy was required to support multiple decision support environments. They were developed to meet a growing demand for management information and analysis that could not be met by operational systems. Operational systems were unable to meet this need for a range of reasons such as the processing load of reporting reduced the response time of the operational systems and development of reports in

operational systems often required a writing specific computer program which was slow and expensive. As a result, separate computer databases came up which were specifically designed to support management information and analysis purposes. These data warehouses were able to bring the data from a range of different data sources, such as mainframe computers, minicomputers, as well as personal computers and office automation software such as spreadsheet, and integrated this information in a single place. This capability, coupled with user-friendly reporting tools and freedom from operational impacts, led to a growth of this type of computer system [4]. Another important concept that is related to data warehouses is called data transformation. As the name suggests, data transformation is a process in which information transferred from a specific source(s), is cleaned and loaded into a repository [3].

In larger corporations, it was typical for multiple decision support environments to operate independently. Each environment serves different users but often required much of the same stored data. The process of gathering, cleaning and integrating data from various sources, usually from long-term existing operational systems (usually referred to as legacy systems), was typically in part replicated for each environment. Moreover, the operational systems were frequently re-examined as new decision support requirements emerged. Often new requirements necessitated gathering, cleaning and integrating new data from "data marts" which were tailored for ready access by users [1]. Data warehousing then became the key trend in corporate computing in the 1990s. Data warehousing is not really a technology trend per se. It was primarily driven by the business environment [2]. Since early 1990s, the data warehouse has become the foundation of advanced decision support applications [5]. Using sophisticated on-line analytical processing (OLAP) and data mining tools, some corporations are able to exploit insights gains from their data warehouse to significantly increase sales [6], reduce costs [7] [6], and offer new and better products or services [7]. The payoff from a well-managed data warehouse can be huge. For instance, a study conducted by IDC, a leading research firm, found the average return on investments in data warehousing projects to be about 400 percent [8]. By the late 1990s, most large corporations had either built or were planning to build a data warehouse [9]. However, the implementation of a data warehouse is both very expensive and highly risky. Due to technological advances (lower cost for more performance), and increased user's requirements (faster data load cycle times and more features), data warehouse has evolved.The various stages of its evolutions [4] [10] [11] are given below.

Offline Operational Databases: Data warehouses in this initial stage are developed by copying the database of an operational system to an off-line server where the processing load of reporting does not impact on the operational system's performance.

Offline Data Warehouse: Data warehouses at 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 structure.

Real Time Data Warehouse: Data warehouses at this stage are updated on a transaction or event basis, whenever an operational system performs a transaction (e.g. an order or a delivery or a booking etc.).

Integrated Data Warehouse: Data warehouses at this stage are used to generate activity or transactions; which are passed back into the operational systems for use in the daily activity of the organization.

Some of the key developments in early years of data warehousing [11] are given below.

1960s: General Mills and Dartmouth College, in a joint research project, developed the terms *dimensions* and *facts*.

1970s: AC Nielsen and IRI provide dimensional data marts for retail sales.

1983: Teradata introduced a database management system specifically designed for decision support.

1988: Barry Devlin and Paul Murphy published the article "*An architecture for a business and information systems* in *IBM Systems Journal"* where they introduced the term "business data warehouse".

1990: Red Brick Systems introduced Red Brick Warehouse, a database management system specifically for data warehousing.

1991: Prism Solutions introduced Prism Warehouse Manager, software for developing a data warehouse.

1991: Bill Inmon published the book "*Building the Data Warehouse"*.

1995: The Data Warehousing Institute, a for-profit organization that promotes data warehousing, is founded.

1996: Ralph Kimball published the book "*The Data Warehouse Toolkit"*.

1997: Oracle 8, with support for star queries, is released.

2. Data mining

Data mining emerged during the late 1980s, and has great treads during 1990s when the work of mathematicians, logicians, and computer scientists combined to create artificial intelligence (AI) and machine-learning [19]. In the 1960s, AI and statistics practitioners developed new algorithms, such as regression analysis, maximum likelihood estimates, neural networks, bias reduction, and linear models of classification [12]. The term 'data mining' was coined during this decade, but it was pejoratively used to describe the practice of wading through data and finding patterns that had no statistical significance [13].

Data mining, in several ways, is fundamentally the adaptation from machine learning techniques to business applications. Statistics are the foundation of most technologies on which data mining is built, e.g. regression analysis, standard distribution, standard deviation, standard variance, discriminant analysis, cluster analysis, and confidence intervals. All of these are used to study data and its relationships. Artificial intelligence (AI), which is built upon heuristics as opposed to statistics, attempts to apply human-thought-like processing to statistical problems. Certain AI concepts which were adopted by some high-end commercial products, such as query optimization modules for relational database management systems (RDBMS). Machine learning is the union of statistics and AI. It could be considered an evolution of AI, because it blends AI heuristics with advanced statistical analysis. Machine learning attempts to let computer programs learn about the data they study, such that programs make different decisions based on the qualities of the studied data, using statistics for fundamental concepts, and adding more advanced AI heuristics and algorithms to achieve its goals [14]. Many of these methods are also frequently used in vision, speech recognition, image processing, handwriting recognition, and natural language understanding. However, the issues of scalability and automated business intelligence solutions drive much of and differentiate data mining from the other applications of

machine learning and statistical modeling [15], [16]. Though data mining is the evolution of a field with a long history, the term itself was only introduced relatively recently, in the 1990s. Its roots are traced back along three family lines:

- Classical Statistics
- Artificial Intelligence
- Machine Learning

The longest of these three lines is classical statistics. Without statistics, there would be no data mining, as statistics is the foundation of the many technologies on which data mining is built. Certainly, within the heart of today's data mining tools and techniques, classical statistical analysis plays a significant role. Because this approach requires vast computer processing power, it was not practical until the early 1980s, when computers began to offer useful power at reasonable prices. AI found a few applications at a very high end scientific/government markets, but the required supercomputers of the era priced AI out of the reach of virtually everyone else. The third family line of data mining is machine learning, which is more accurately described as the union of statistics and AI. Data mining is getting increasing acceptance in science and business areas which need to analyze large amounts of data to discover trends which could not otherwise be revealed [4] [17]. Thus data mining techniques are the result of a long process of research and product development. Although data mining is a relatively new term, the technology is not. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Companies have used powerful computers to filter through volumes of supermarket scanner data and analyze market research reports for years. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost. Data mining algorithms embody techniques that have existed for at least 10 years, but have only recently been implemented as mature, reliable, understandable tools that consistently outperform older statistical methods [18].

In the evolution from business data to business information each new stage has been built upon the previous one. For example, dynamic data access is critical for drill-through in data navigation

applications, and the ability to store large databases is critical to data mining. From the user's point of view, the four steps of evolution of data mining are listed below [18].

1. Data Collection (1960s): Enabling Technologies were Computers, tapes, disks. Service providers were IBM, CDC having retrospective and static type data delivery.
2. Data Access(1980s): Technologies were Relational databases (RDBMS), Structured Query Language (SQL), ODBC. Oracle, Sybase, Informix, IBM, Microsoft provided Retrospective, dynamic data delivery at record level.
3. Data Warehousing & Decision Support(1990s): Enabling Technologies were On-line analytic processing (OLAP), multidimensional databases, data warehouses having features like Retrospective, dynamic data delivery at multiple levels. Providers were Pilot, Comshare, Arbor, Cognos, Microstrategy
4. Data Mining(Emerging Today): Enabling Technologies are Advanced algorithms, multiprocessor computers, massive databases having features like Prospective, proactive information delivery. Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry) are the providers of the product.

These steps are revolutionary because they allowed new business questions to be answered accurately and quickly. With data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments [18].

3. Conclusions

Through the critical analysis of the data warehouse and data mining techniques, it is concluded that the technologies are not a sole discipline rather an umbrella that encompasses various disciplines time to time like traditional statistics methods. Later, it covers advance mathematical methods, management theories & tools, software engineering and economics as well as machine learning i.e. a sub-section of artificial intelligence. Besides the applicability to some general real world problem solutions, like data extraction, predication of unknown values, cluster analysis, anomaly detection in records, decision support, and business intelligence, it is useful to more

complex and useful information extractions like natural language processing, pattern recognitions, pharmaceuticals, deployment of Geographic Information Systems (GIS) specifically, availability of water resources, rock mining. Furthermore Multimedia applications like music, videos, etc. and are enough to increase its scope.

References:

- Inmon, W.H. (2002-04-15.). Tech Topic: What is a Data Warehouse? The Story So Far. *Prism Solutions*. Rretrieved from: http://www.computerworld.com/ databasetopics/data/story/0,10801,70102,00.html.
- Wan, D. (6 January, 2007). History of Data Warehouse. Retrieved January 6, 2007 from: http://dylanwan.wordpress.com/2007/01/06/hostory-of-data-warehouse/
- Data Warehousing Introduction **(**2006, 17 Aug). Retrieved from: www.exforsys.com/.../data-warehousing/data-warehousing-introduction.html
- Data Warehousing: History of Data Warehousing (n.d.). Retrieved from site: http://www.dedupe.com/history.php
- Shim, J.P., Warkentin, M., Courtney, J.F., Power, D.J., Sharda, R., & Carlsson, C. (2002). Past, present, and future of decision support technology. *Decision Support Systems*, *33*(2), 111-126.
- Whiting, R. (1999, May 24). Warehouse ROI. *InformationWeek*, *735*, 99-104.
- Watson, H., & Haley, B. (1998). Managerial considerations. *Communications of the ACM*, *41*(9), 32-37.
- Desai, A. (1999). For pharmaceutical companies, a data warehouse can be just what the doctor ordered. *Health Management Technology*, *20*(2), 20-22.
- Joshi, K. & Curtis, M. (1999). Issues in building a successful data warehouse. *Information Strategy*, *15*(2), 28-35.
- Wikipédia, the free encyclopedia (2008). *History of Data warehousing*. Retrieved from http://www.businesspme.com/uk/articles/technologies/9/History-of-data-warehousing. html
- Wikipedia, the free encyclopedia (n.d.). *Data warehouse*. Retrieved from http://en.wikipedia.org/wiki/Data_warehouse#History,

- Dunham, M.H. (2003). *Data mining introductory and advanced topics*. Upper Saddle River, NJ: Pearson Education, Inc.
- Fayyad, U.M, Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI Magazine, 17(3), pp. 37-54.
- Unc. Edu. *Data minning*. Retrieved from: http://www.unc.edu/~xluan/258/ datamining.html
- Apte, C. (2003). *Data Mining Analytics for Business Intelligence and Decision Support*. Retrieved from: http://www.research.ibm.com/dar/papers/pdf/orms2.pdf, http:// www.research.ibm.com/dar/publications.html. in OR/MS Today, February 2003.
- Williams, G., Hegland, M., & Roberts, S. (1998). A Data Mining Tutorial. *In Proceedings of the Second IASTED International Conference on Parallel and Distributed Computing and Networks (PDCN'98),* 14 December 1998, Copyright c 1998.
- Ayre, L. B. (June, 2006). *Data Mining for Information Professionals*. Retrieved from: http://techessence.info/files/Ayre_DataMiningForInformationProfessionals_June2006.pdf
- Thearling, K. (2009). *Information about data mining and analytic technologies: An Introduction to Data Mining-Discovering hidden value in your data warehouse.* Retrieved from: http://www.thearling.com /text/ dmwhite /dmwhite.htm Copyright © 2009 Kurt Thearling.
- Buchanan, B.G. (2006). Brief History of Artificial Intelligence. Retrieved from: http://www.aaai.org/AITopics/bbhist.html.