THE CORE PRINCIPLES FOR DEVELOPING A PROFICIENT DATA WAREHOUSE

Rajeev Kumar Singh

Assistant Professor, (Department of IT)
L. N. Mishra College of Business Management, Muzaffarpur
(An autonomous institute runs under BRA Bihar University)

Abstract

A Data warehouse is a repository of information gathered from multiple sources, stored under a unified schema, at a single site. We know that Data warehousing is the most leading, reliable & important technology which is used today by more companies for planning, forecasting, and management to use the minimum resources. Data Warehouse is the centralized store of detailed data from all relevant source systems, allowing for ad hoc discovery and drill-down analysis by multiple user groups. Various implementation factors play critical role to successful data warehouse (DW) project implementation. DW has unique characteristics that need to consider during implementation. There is little empirical research about implementation of DW to get success. Determining factors affecting DW success are important in the deployment of this DSS technology by organizations. A major research has been done in this field regarding design and development of data warehouses and a minor research still needs to be done for better utilization the data warehouse. An area which needs special attention from research community is data warehouse implementation & maintenance.

We know that, the project may be failure due to the poor management of designing the data warehouse. This is the major factor for data warehouse project in the company. Without proper management desired results are nearly impossible to attain from a data warehouse. Unlike operational systems data warehouses need a lot more management and a support team of qualified professionals is needed to take care of the issues that arise after its deployment including data extraction, data loading, network management, training and communication, materialized view and some other related tasks. So, that the topic of my research is to explore the impact of the selected factors, under organizational, project-related and environmental dimensions, on data warehouse projects.

Keywords:

RDBMS; Data warehousing; materialized view; network management; training and communication;

1. Introduction

A data warehouse is a database designed to enable business intelligence activities: it exists to help users understand and enhance their organization's performance. It is designed for query and analysis rather than for transaction processing, and usually contains historical data derived from transaction data, but can include data from other sources. Data warehouses separate analysis workload from transaction workload and enable an organization to consolidate data from several sources. This helps in:

- Maintaining historical records
- Analyzing the data to gain a better understanding of the business and to improve the business

In addition to a relational database, a data warehouse environment can include an extraction, transportation, transformation, and loading (ETL) solution, statistical analysis, reporting, data mining capabilities, client analysis tools, and other applications that manage the process of gathering data, transforming it into useful, actionable information, and delivering it to business users.

To achieve the goal of enhanced business intelligence, the data warehouse works with data collected from multiple sources. The source data may come from internally developed systems, purchased applications, third-party data syndicates and other sources. It may involve transactions, production, marketing, human resources and more. In today's world of big data, the data may be many billions of individual clicks on web sites or the massive data streams from sensors built into complex machinery.

This enables far better analytical performance and avoids impacting your transaction systems. A data warehouse system can be optimized to consolidate data from many sources to achieve a key goal: it becomes your organization's "single source of truth". There is great value in having a consistent source of data that all users can look to; it prevents many disputes and enhances decision-making efficiency.

A data warehouse usually stores many months or years of data to support historical analysis. The data in a data warehouse is typically loaded through an extraction, transformation, and loading (ETL) process from multiple data sources. Modern data warehouses are moving toward an extract, load, and transformation (ELT) architecture in which all or most data transformation is performed on the database that hosts the data warehouse. It is important to note that defining the ETL process is a very large part of the design effort of a data warehouse. Similarly, the speed and reliability of ETL operations are the foundation of the data warehouse once it is up and running.

Traditionally, the information systems Departments are the sole interface to firms' data stored in computer systems. Executives from various departments rely on the IS department to satisfy their needs for information that is necessary for decision making. The turnaround time is usually long and information is often not delivered on time to users, which in turn reduces the value of the information.

3. Purpose of the Research

Complex & Large organizations have different data source to manage operation, so faced significant problem to build single view of their business from different data sources. During the mid-to-late 1990s, "DW became one of the most important developments in the information systems field" [1]. It is estimated that 95 percent of the Fortune 1000 companies either have a DW in place or are planning to develop one [2]. "In2002, the Palo Alto Management Group predicted that the DW market would grow to a $113.5 billion market, including the sales of systems, software, services, and in-house expenditures" [2]. About 3,000 data warehousing projects are undertaken each year and if the lowest perceived data warehouse failure rate(70%) is accurate, then each year there are 2,100 failures [3].DW project is an expensive and risky undertaking [4]. The typical project costs over $1 million in the first year alone and it is estimated that one-half to two-thirds of all initial DW efforts fail [5]. There is a common perception that the failure rate of data warehousing projects is 70 to 80 percent (In mon 2001) and one study reported a 90 percent failure rate(Conning 2000) [3]. "According to a 2003 Gartner report, more than 50 percent of data warehouse projects failed, in a 2007study, Gartner predicted once more that 50 percent of data warehouse projects would have limited acceptance

or be outright failures as a result of lack of attention to data quality issues" [6]. According to [7], the average time for the construction of a data warehouse is 12 to 36 months and the average cost for its implementation is between $1 million to $1.5 million. This study was undertaken to perform a detailed analysis of these cases to determine whether the presence (or absence) of any specific factors or combination of factors might be correlated with instances of failures.

4. Core Principles

A very good discussion on the problems of data warehousing projects is found in above. The paper mentions the logical fact that nobody really speaks about data warehousing failures and goes on to group the reasons for the failure of a data warehousing project into four categories, namely design, technical, procedural and socio-technical factors.

While the implementation of a specific phase of the data warehouse may be completed, but the data warehouse program needs to be continued progress monitoring needed to be continued against the agreed on success criteria. The data warehouse team must ensure that the existing implementations remain on track and continue to address the needs of business. Performance issues in data warehousing are centralized around access performance for running queries and incremental loading of snapshot changes from the source systems. The following concepts can be considered for a better performance:

- **High Data Quality**

High quality of data produces quality of information. Poor quality of data badly affects all company to run smoothly. Often the cause of business problem such as faulty analysis, operational inefficiency and dissatisfied customer are because of inaccurate, inconsistent, incomplete data. The operation cost increase due to poor quality data. The poor quality of data creates the problem data integration.

"Many enterprises fail to recognize that they have an issue with data quality. They focus only on identifying, extracting and loading data to the data warehouse, but do not take the time to assess quality, said Ted Friedman, principal analyst at Gartner". "Consistency and accuracy of data is critical to success with BI, and data quality must be viewed as a business issue and responsibility, not just an IT problem.". "New federal regulations and corporate governance requirements have greatly increased the pressure for improved data quality. Enterprises must eliminate multiple data silos, assign stewardship to critical data, and implement a process for continuous monitoring and measurement of data quality." According to Gartner Inc ,through 2007 more than 50 percent DW project will be outright failed because lack of attention to data quality issue.

- **Proper Communication within Organization**

A Data ware house system is actually about strongly combine different business functions, so the close co-operation and communication across disparate business functions would be a natural prerequisite in a Data ware house project. Some authors have described the co-ordination and communication between departments as the oil that keeps everything working properly in these contexts. An effective communications program keeps people informed and generates interest in using the data warehouse. Organizational rules to rationalize inconsistencies had to be established. Clear and consistent communication of company-wide warehouse goals and policies fosters employee participation on three critical fronts: First, it reinforces the front-line employees' contribution of information to the warehouse. Second, it encourages information sharing to support ongoing business activities. Third, it inspires mid-level managers to use the data warehouse to inform key stakeholders regarding decisions, and new projects. The cooperation between the departments in an organization has a large effect on the smooth flow of the required information and expertise among the departments, which strongly influences the successful adoption of data warehouse technology. Interdepartmental cooperation and communication is a must for any project.

- **IT Initiate better DW Project**

In some organization IT used to provide different report to business to understand the business performance. Due to some reason (data volume increase, improper SQL write, Improper Application design, System over utilized, etc) business sometimes does not received the report right time. After that IT started talked about DW to give report to business on right time without proper analysis. Business agreed because they need report on time to analysis. The DW project is very costly and business talked about because of these report delay this type of costly project they will not finance. IT says to business you can do business analysis if we have DW, but business says we want report as business people no aware about business analysis with DW.

The IT Initiated DW/BI project may pay more attention on technology rather than business. If get sucked into the technology, then missing the whole point. Developed technically great DW/BI system but less importance on business, this DW project will be treating as fail. Need to start with business value.

- **The Management of Metadata**

The term Metadata is defined as “data about data". Metadata help a person to locate and understand data. Metadata is often generally described as “information about data.” More precisely, metadata is the description of the data itself, its purpose, how it is used, and the systems used to manage it. Metadata play important role in DW development. Not only does it shape the data integration process but it also enables the business users to locate, understand and use the data once it is loaded into the data warehouse. Metadata is very valuable to the business because it facilitates the understanding of data. Without this understanding the data would be useless. Metadata need to integrate from different source and developed metadata repository. “At the conceptual level the structure of the repository is described by a meta model or informational model. To develop Meta model at least 4 modeling level are required. To start with level, on the lowest level, level 0 there are actual data item (e.g., the customer data). The level above contains the metadata information: level 1 contains metadata (e.g. database schema), level 2 specifies the schema used to store the metadata (the so called Meta model, information model or metadata schema).”

- **No Gap between Researchers and Practitioners**

“The gap between researchers and practitioners is widely discussed in the IT community”. The situation regarding data warehousing seems to follow the general pattern where practitioners complain that their practical problems are overlooked by research and researchers are general unsatisfied by the acceptance of their ideas in industry.

- **Better Education, Training & Documentation**

Training and education of the employees are required in a successful data warehouse project. A data warehouse is not a simple project or an easy-to-learn system. It demands time to educate and transfer the knowledge to users by setting up training courses and distributing related-material. In most computing projects, management identifies the need for training, but does not always fund training. With every new database there is a need for another training course, complete with reference materials. Every enhancement or change to the warehouse must be documented and communicated to warehouse users.

Training expands the communication process by maintaining a level competence in both the business and IT community as to the tools and mechanisms of the data warehouses. The quality of employees and their development through training and education are major factors in determining long-term profitability of a small business. If you hire and keep good employees, it is good policy to invest in the development of their skills, so they can increase their productivity. Training often is considered for new employees only.

This is a mistake because ongoing training for current employees helps them adjust to rapidly changing job requirements. Training and updating the employees' knowledge of data ware house is a major challenge. Data ware house implementation requires a huge mass of knowledge to enable people to use, cope and solve problems within the framework of the system. Training employees to use ERP is not as simple as training them in any other packaged-software such as a Microsoft package.

- **Do not over budget of the project**

Every company used to declare annual budget, if the project required more budget, the project may be treated a failure. "The DW is costly project so the inadequate budget might be the result of not wanting to tell management the bitter truth about the costs of a data warehouse and expensive consulting help may have been needed. Performance or capacity problems, more users, more queries or more complex queries may have required more hardware to resolve the problems. The scope of project may require updating on the middle of the project running or other factors may have resulted in additional expenses."[12]

- **Support System**

Another role of the data warehouse support and protection group is the problem resolution when some problem is encountered in the data warehouse. In the case of data warehouses, the expensive and the risky nature of data warehouses have forced the potential adopters to pay extra attention in selecting appropriate vendors to increase the possibility of having successful data warehouse initiatives.

The support Center is an important division for any organization as it serves as the primary interaction point between customers and the company. In many situations, it is the only interaction point and therefore, responsible for the customer's experience and satisfaction. Due to this heightened level of importance, it is critical that the contact handling process is conducted both efficiently and effectively. This process specifies how to collect, document, answer and/or escalate calls, requests, and queries related to issues with the data warehousing environment. Problem documentation can be completed either by the support Center representative and/or in conjunction with a form completed by the end user or IT support person requesting a service or action.

All inquiries, no matter how trivial should be logged, especially during the start of a new data warehouse or mart. These bits of information can form clues to taking proactive action to bigger problems before they emerge.

- **Strong Network Management**

If there is a heterogeneous group of platforms for the data warehouse implementation, network management is going to be one of the most demanding tasks. Modern communication networks create large amounts of operational data, Including traffic and utilization statistics and alarm/fault data at various levels of detail. These massive collections of network-management data can grow the order of several Tera bytes per year, and typically hide "knowledge" that is crucial to some of the key tasks involved in effectively managing a communication network (e.g., capacity planning and traffic engineering).

Besides providing easy access to people and data around the globe, modern communication networks also generate massive amounts of operational data throughout their life span. As an example, Internet Service Providers (ISPs) continuously collect traffic and utilization information over their network to enable key network- management applications.

Not only are users coming constantly on-line, but users and equipment are invariably moving to new locations. The networking hardware is proliferating with LANs, WANs, hubs, routers, switches and multiplexers. Leaving behind all this is the next stage – users wanting to access internet based data sources along with the corporate data, requiring even greater bandwidth and network management resources. Managing this environment is one big challenge; capacity planning for the future is another. If the data warehouse team is not quite good in networking technology than there should be at least one person in the organization who understands technology.

- **Complete Requirement of the User**

Before project start DW user compile the requirement. User expect that they will get their requirement fulfill. If the users not get their requirement form get unhappy and due to this the project should be considered a failure. Sometimes users expect more than they got. "Users may be unhappy about the cleanliness of their data, response time, availability, usability of the system, anticipated function and capability, or the quality and availability of support and training."

- **Tight Quality of the Reports**

If the data is not clean and accurate, the queries and reports will be wrong, In which case users will either make the wrong decisions or, if they recognize that the data is wrong, will mistrust the reports and not act on them. Users may spend significant time validating the report figures, which in turn will impact their productivity. This impact on productivity puts the value of the data warehouse in question."

- **User friendly tools**

Should not expect that all the user who will use BI tools for business analyses are IT expert. Some of the user may heard about the BI tools is first time. If the tools is not user friendly people will not used much and start blame on this tools even though form technical and business point of view tools is great. User may start asking ask IT to download the report and provide them. The DW is not only some sort of report, it is more that that where user will build up their own query and design report to analysis the business performance. In this scenario the project will be treating as failed, because the purpose of DW is deviated.

- **ETL Process**

The ETL process is much more than code written to move data. The ETL architect also serves as the central point for understanding the various technical standards that need to be developed if they don't already exist. These might include limits on file size when transmitting data over the company intranet, requirements for passing data through firewalls that exist between internal and external environments, data design standards, standards for usage of logical and physical design tools and configuration management of source code, executables and documentation. The ETL architect must also ensure that the ETL design process is repeatable, documented and put under proper change control.

Extract, transform and load (ETL) is the core process of data integration and is typically associated with data warehousing. ETL tools extract data from a chosen source, transform it into new formats according to business rules, and then load it into target data structure. ETL which stands for extract, transform, and load is a three-stage process in database usage and data warehousing. It enables integration and analysis of the data stored in different databases and heterogeneous formats. After it is collected from multiple sources (extraction), the data is reformatted and cleansed for operational needs (transformation). Finally, it is loaded into a target database, data warehouse or a data mart to be analyzed.

A key consideration for the ETL architect is to recognize the significant differences that the design and implementation methods for a business intelligence system have from an online transaction processing (OLTP) system approach. One last role for the ETL architect must be to ensure that the various software tools needed to perform the different types of data processing are properly selected ETL is one of the most important sets of processes for the sustenance and maintenance of Business Intelligence architecture and strategy [12].

If source data taken from various sources is not cleanse, extracted properly, transformed and integrated in the proper way, the extracted data will often be stored in a central staging area where it will cleanse and otherwise transformed before loading into the warehouse. An alternative approach to information integration is that of mediation: data is extracted from original data sources on demand when a query is posed, with transformation to produce a query result.

5. Conclusion

A data warehouse solution is not only a software package. It is a complex process to establish sophisticated and integrated information systems. The adoption of this technology requires massive capital expenditure, utilizes a certain deal of implementation time and has a very high likelihood of failure. Therefore, many adoption-related factors must be carefully assessed before the real adoption is actualized.

I do study on BSNL its shows that the first and the most important part of a data warehouse maintenance program is the training of its users. The study shows that most business users are reluctant to adopt technology to carry out their work, therefore pursuing a business user to use data warehouse is inevitable. To pursue business users in using data warehouse the communication and training program is a must. The training program gives the users of data warehouse an insight into the qualities and capabilities of a data warehouse and teaches them.

The communication process keeps the business users and IT users in contact with each other to have exchange of views, suggestions and any guidance towards enhanced performance of a data warehouse. The services of help desk and problem management play an important role in taking valuable output from the data warehouse. Support is always required in any information system, same is the case with a data warehouse but here the support is needed 24 hours a day. Some of the processes like ETL are carried out during the night, which require presence of support staff to rectify any problem.

The support team also points out if there is any loop hole or problem area within the data warehouse that should be addressed. Apart from help desk each data warehouse support team develops its own problem management process. The process defines necessary routines and instructions to counter any problem found in the warehouse. If the problems found in the data warehouse are not addressed at the right time, this leads to performance shortfalls, and usability and availability issues in near future. Thus help desk and problem management play a key role in improving data warehouse performance and getting the desired out put from it. Network management also plays its part in improving data warehouse performance. From the case study we concluded that by having a fast and reliable network user queries get a much shorter response time especially in a distributed data warehouse.

References

[1]. Database System Concepts, Henry F. Korth & Abraham Silberschatz, Page No – 889 - 893

[2]. Eckerson, W.W. (2003). Evolution of Data Warehousing: The Trend toward Analytical Applications. Journal of Data Warehousing, 25(1), pp.1-8

[3]. S. Atre, Rules for data cleansing, Computerworld _1998.69–72, March 8.

[4]. N. Alur, Missing links in data warehousing, Database Programming and Design _1995. 21–23, September.

[5]. An investigation of the factors affecting data warehousing success by Roger L. Hayen et. al. in journal of Issues in Information Systems, Volume VIII, No. 2, 2007.

[6]. Developing a Data Warehouse Process that responds to the needs of the Enterprise,Peter R. Welbrock Smith-Hanley Consulting Group Philadelphia, PA

[7]. Keith Lindsey, Mark N. Frolick. CURRENT ISSUESIN DATA WAREHOUSING. 2002 ,Eighth Americas Conference on Information Systems.http://aisel.aisnet.org/amcis2002/7

[8]. Watson, H. J., and Haley, B. J. (2004). Data Warehousing: A Framework and Survey of Current Practices, Journal of Data Warehousing,2(1), pp. 10-17.

[9]. The data warehouse toolkit. 2nd edition. Ralph Kimball, Margy Ross. 2002 Wiley computer publishing..

[10]. Data Warehouse Management Handbook by Richard Kachur. 2000 Prentice Hall .

[11]. Arnott, D. and Pervan, G. (2005). A Critical Analysis of Decision Support Systems Research. Journal of Information Technology, 20(2), pp. 67 – 85.

[12]. S. Deck, Data warehouses: plan well, start small, Computer world _1998. 9, August 3.

[13]. Building, using, and managing the data warehouse. Ramon Barquin, George Zagelow,Katherine hammer, Mark sweiger, George Burch, Dennis Berg, Christopher Heagele,Katherine Gl assey-Edholm, David Menninger, Paul Barth, J.D. Welch, Narsim ganti, Herb Edelstein, Bernard Boar, Robert Small.

[14]. D.P. Ballou, G.K. Tayi, Enhancing data quality in data warehouse environments, Communications of the ACM 42 (1) (1999) 73–78.

[15]. Data warehouse, Practical advice from the experts. 1997. Prentice hall by Joyce Bischoff & Ted Alexander

[16]. Akkermans and Helden, Vicious and virtuous cycles in ER P implementation: A case study of interrelations between critical success factors, European journal of information systems, 2002, Vol.11 Iss. 1, p35..

[17]. Nah et al., Critical factors for successful implementation of enterprise systems, Business process management, Bradford: 2001, Vol.7 Iss.3, p285.:

[18]. J. Bischoff, Achieving warehouse success, Database Programming and Design (1994) 27–33, July.

[19]. Bingi et al., Critical issues affecting an ERP implementation, Information systems management, 1999, Vol.16 Issue3.

[20]. M. Demarest. The politics of data warehousing.http://www.noumenal.com/marc/dwpoly.html 9/23/201

[21]. J. Foley, Data warehouse pitfalls, Information week _1997. 93–96, May 19.

[22]. Lessons from a successful data warehouse implementation. Dr. John D Porter and john. J Rome. Arizona State University.

[23]. Grim, R., and Thorton, P. (2001). P. A Customer for Life: The warehouse Approach. Journal of Data Warehousing, 2(1), pp. 73-79.

[24]. Building a data warehouse for decision support. 1996 Prentice Hall. By Vidette Poe with contributions from Laura L. Reeves.

[25]. Fundamentals of database systems. 4th Edition. Persons international and Addison Wesley. Ramez Elmasri and Shamkant B. Navathe

[26]. Analysts to Show How To Implement a Successful Business Intelligence Program During the Gartner Business Intelligence Summit, March 7-9 in Chicago,ILhttp://www.gartner.com/press_releases/asset_121817_11.html [5/28/2010 12:28:20 AM]

[27]. E. Appleton, The right server for your data warehouse, Datamation 41 _5. _1995. 56–58, March.