THE TASK OF MANAGING DATA WAREHOUSE LIFECYCLE IN MODERN TECHNOLOGY

Pedro Imiefoh
Computer Science Department,
University of Benin,
P.M.B 1154. Benin.

Abstract
Major corporations around the world have come to rely totally upon the essential information asset stored in their corporate databases and data warehouses. This is informed by the fact that the modern technology, computers, is gradually shifting its research approach from computation to information and knowledge based management. That is, if there is incomplete information available for any reason, particularly for a protracted period, the business may grind to a halt and suffer serious financial consequences. Making a data warehouse available is not easy. Corporate data warehouses range from one terabyte to ten terabytes or more, with the intention of giving users global access on a 24 x365 basis. Data warehouses can take months to set up yet can fail in seconds. And to react to changing business requirements, the data warehouse will need to change in design, content and physical characteristics on a timely basis. Hence the paper describes the characteristics of data warehouses and how to make a data warehouse lifecycle safe, available, perform well and be manageable.

Introduction

Companies set up data warehouses when it is perceived that a body of data is critical to the successful running of their business. Such data may come from a wide variety of sources. Typically, they are made available via a coherent database mechanism, such as an Oracle database. By their very nature, they tend to be very large in size, used by a larger percentage of key employees, and may need to be accessible across the nation, or even the globe. The perceived value of a data warehouse is that executives and others can gain real competitive advantage by having instant access to relevant corporate information. Infact, leading companies now realize that information has become the most potent corporate asset in today's demanding digital markets (Laudon, 2005).

Depending on the business, a data warehouse may contain very different things, ranging from the traditional financial, manufacturing, order and customer data, through document, legal and project data, on to the brave new world of market data, press, multi-media, and links to Internet and Intranet websites. They all have a common set of characteristics, such as size, interrelated data from many sources, and access by a lot of employees. Therefore, in planning the organisation's essential data in such a single central location, care must be taken to ensure that the information contained within it is

highly available, easily managed, secure, backed up and recoverable to any point in time and that its performance meets the demanding needs of the modern user.

It is against this background that this paper attempts to describe the characteristics of data warehouses and how to make a data warehouse lifecycle safe, available, perform well, and be manageable. To achieve this ultimate goal, the paper has a six-item structure and this includes: the introduction; background of data warehousing, which discusses brief overview of data warehousing in modern technology; data warehouse lifecycle change; data warehouse lifecycle optimization; management strategies and tools; and a conclusion that summarizes the central arguments of the paper.

Research Objective

The purpose of this paper is to determine those characteristics of data warehouse lifecycle management in modern technology and their relevance. This will involve the analysis of the following:

i. identifying and analyzing those management tools and strategies in the data warehouse lifecycle; and
ii. the impact of the strategies and tools both for the safe availability of data warehouse and proper management.

Background of Data Warehousing in Modern Technology

Data warehouses became a distinct type of computer databases in recent past. They are developed to meet a growing demand for management information and analysis that could not be met by operational systems. As a result, separate computer databases on departmental basis in the firm began to be built. They were specifically designed to support management information and analysis purposes. These databases were able to bring in data from a range of different data sources. They include, mainframe computers, microcomputers, as well as personal computers and office automation software, such as spreadsheets and integrate the information in a single place. This capability, coupled with user-friendly reporting tools, and freedom from operational impacts has led to the growth of this type of computer system infrastructure, called data warehouse. A data warehouse is an integrated data repository containing historical data of a corporation for supporting decision making processes. It provides a basis for online analytical processing of enterprise (Mattson, 2004).

The modern technologies for building web-based data warehouses are still evolving. Currently, two leading systems are being developed; namely: a Java-based approach known as J2EE and Microsolt. Net platform. Data warehouses are the heart of both systems and they are both server-based technologies designed to build interactive websites. Unfortunately, the two systems are completely independent and incompatible. That is, if you build an application to run with one approach, you would have to completely rewrite it to use the other method. However, the Microsoft approach offers a little more flexibility with its support for multiple languages. It also provides a complete development environment with several easy-to-use tools that make it relatively easy for

beginners to create Object Relational Database Management Systems (ORDBMS) websites (Gerald, and Anderson, 2006).

The data warehouse modeler would be concerned with creating web script pages that would interact with the data warehouse. When customers request a page, the server executes the associated program script. The script sends queries to the data warehouse and retrieves the desired data. For example, the script might retrieve product descriptions, prices, and in-stock status. The data and images are added to the Web page and sent to the customer who sees only the simple results. The data warehouse runs a separate server from the web server. This is to reduce the load on the web server and makes it easier to handle backups and other warehouse maintenance issues. But with increasing powerful servers, one server can handle smaller applications (data marts) See illustrative Figure below.

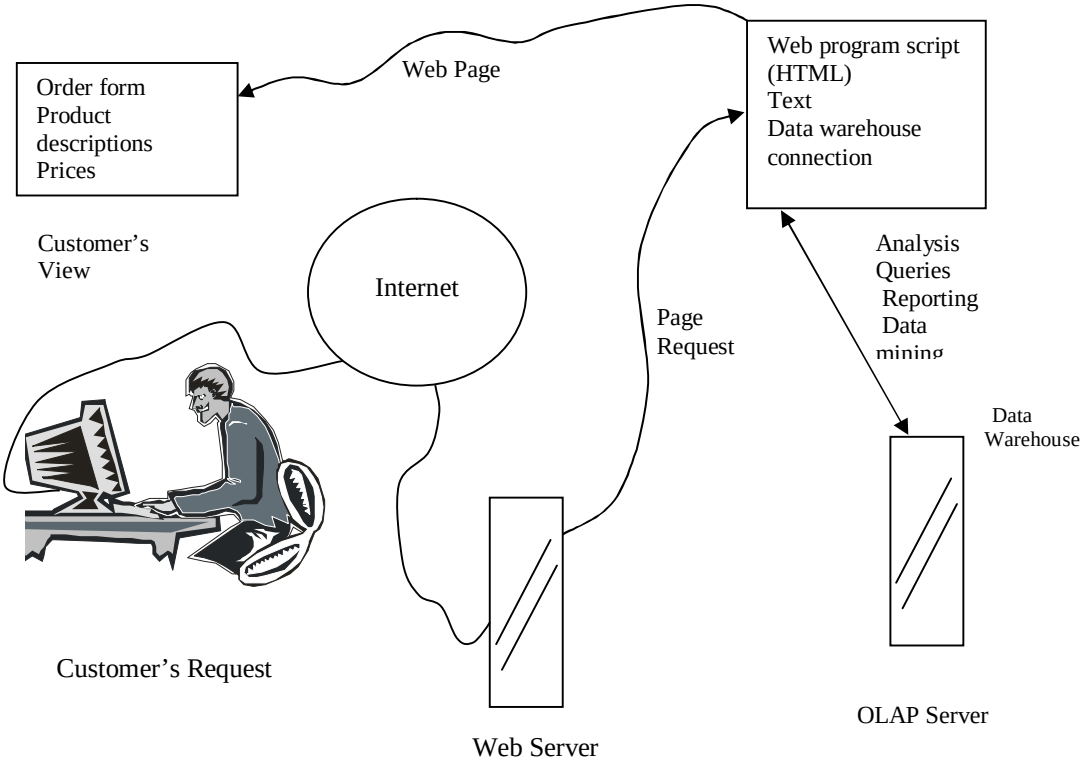

Figure1: Web-Based Data Warehousing Model

Data Warehouse Lifecycle Change

Change is a constant and often frustrating agent in managing a data warehouse. How to manage that change in light of technology that improves over the years can be a daunting task, yet the key to providing improvement in the data warehouse can be accomplished if one understands how to recognize and pro-actively manage the business

challenges. Change in the data warehouse lifecycle can occur in many areas, though logically, we can categorize change in three constructs: organizational, technological, and process. The changes that occur can adversely affect the quality, cost, schedule, resources and outcome of any data warehouse initiative.

Organisational Change

For the vast majority of enterprises in today's economy, acquisition is a normal strategy for growth, but the impacts of an acquisition or merger are many. For a data warehouse initiative, it can mean a total re-scoping of the warehouse development, expected delay on releases, and havoc on a resource plan. The integration of new resources, technology, platforms, data sources, and more importantly, business objectives and strategic goals will require a fair amount of planning and project management acumen to manage this type of change. Attrition is another force of change every company must face. This is especially acute for the Information Technologies (IT). It is shown in Rockart and Short, (2003) that the entire IT staff of 15 resources quit for jobs acquired in less than one week. With a negative employment rate for IT, the resource planning for a data warehouse initiative must be dynamic and anticipate the expected attrition in a market where high salaries, promises of training in bleeding-edge technologies, better benefits, and increased bonus structures are a lure for even the most loyal employee.

Technological Change

It is now a well known fact that technology improves rapidly. Electronic business (e-business) is a reality. Data warehousing and data mining is becoming mainstream. Knowledge management will become mainstream. This is due to the driving improvements in the technological infrastructures that support these realities. Portal technology will make universal access to data a reality. Unlimited bandwidth, will give new power and speed to e-business and improve Business-to-Consumer (B2C) and Business-to-Business (B2B) communications. Although recent movements to standardize models (DIM-Open Information Model) and languages (XML-Extended Markup Language) are ongoing, it is apparent that no matter how a company stores information or does business, everyone can gain secure access in an electronic marketplace. This is where economies of scale will be meaningless and the sharing of information will occur regardless of platform, data type, or orgnaisation. Organizations of all sizes need such knowledge-sharing technology if they are to compete with faster, nimbler dot-com rivals (Bakos and Yannis, 2004). That is, perpetual learning and knowledge management will be the key to organizational effectiveness in modern technology. Predictive management as realized through improving data mining technology will continue to help enterprises pro-actively manage change and drive business before market conditions actually occur.

Process Change

Changes in process are a natural evolution of every enterprise. The more effective a data warehouse is, the more intelligent the enterprise is in managing and

evaluating the effectiveness of the business processes. For an enterprise, it then becomes imperative to build a library of best practices in its processes and methodologies. This means that repository based tools and integrated models are vital in creating a continuous process improvement model for data warehouse development and delivery. This is made possible for the introduction of Open Information Model (OIM) and the Extended Markup Language (XML). Great strides are taking place in a short period of time to make integrated open tools a reality. Thus, knowledge-sharing technologies are fast becoming a reality for optimizing data warehouse lifecycle. That is, implementing a continuous process improvement strategy to facilitate a smooth development effort, and an effective data warehouse that helps drive competitiveness, realizes returns on investment and reduces total cost of ownership by streamlining costs and increasing revenue.

Data Warehouse Lifecycle Optimisation

Managing change in the data warehouse lifecycle is primarily reactive in nature so optimizing the warehouse should not be difficult to realize a balance. The lifecycle of any data warehouse system will test the will of many a manager in its implementation, then deployment in the organization and finally in the day-to-day operational aspects of running a data warehouse. Each of these phases in the lifecycle brings its own challenges. Each phase requires unique skills and the correct mix of tools. In addition to skills and tools, one must possess the wisdom to drive value in the warehouse itself. This can best be understood if the proper techniques that utilize the skills and tools are practiced.

There are many steps that can be taken to utilize the data warehouse lifecycle. The traditional techniques of project management, such as adding appropriate resources to reduce data warehouse tasks to promote cross training and sandboxing, are commonplace techniques. Another means is to extend timelines to manage a data warehouse project. It is a process of building time into the project plan itself. Extending time adds cost to the total effort, and makes this perhaps the least attractive means.

The results of the utility building tasks are primarily to introduce basic utility and compliance with business requirements to support mandatory business intelligence. In data warehousing, this can extend to business intelligence from automating standard reporting and introducing a preliminary level of adhoc queries for detailed and summary reports.

The utility is then extended to improve the efficiency of the data warehouse itself. The best means of improving the efficiency of the development lifecycle is to introduce concurrency of development. Defining concurrent tasks to improve operational efficiency can be easily accomplished by performing unrelated tasks concurrently with other tasks. This can be an effective means of cross-training as well. An example of concurrency can be to define the data warehouse testing strategy concurrent to the warehouse design process. This can ensure that test cases are closely aligned with the requirement elicited during the planning and design phases. Efficiency can also help the data warehouse realize compliance with standards as well as delivering improved operation of the warehouse and improved customer satisfaction.

The most overstated and least implemented practice in data warehousing management is the implementation of enterprise metadata. Metadata is changing from application and warehouse systems to total enterprise knowledge management systems. Metadata can offer more than just a directory of knowledge for knowledge workers and IT warehouse developers. It ensures reuse, data consistency, and confidence in the knowledge shared across an organization. The inclusion of structured and unstructured information (e.g. graphical images, documents, or e-mail) are just the beginning of information that exists within an enterprise. The use of third party information, such as demographical information, is becoming increasingly important in driving a customer relationship model. Another means of discovering effectiveness in the data warehouse lifecycle is to introduce technology accelerators that speed up the time for implementation and therefore, time to market. This is vital to e-business ventures where e-business can mean increased revenue with the little total cost of ownership to an enterprise. An example of this is to utilize enterprise management tools to remotely deploy data warehouse access packages. Another example is the introduction of an Internet portal as a primary means of data warehouse access as well as operational access that serves as the standard interface for business intelligence. That is, discovering effectiveness in the data warehouse lifecycle will provide value for the enterprise by reducing time to revenue, improving customer service and support, delivery of new product or services and most importantly for IT, the speed of delivery of the warehouse itself.

Management Strategies/Tool

It can take six months or more to create a data warehouse, but only a few minutes to lose it. Data 'accidents' happen. Planning is essential. A well planned operation has few 'accidents,' and when they occur, recovery is far more controlled and timely.

Backup and Restore Strategy

The fundamental level of safety that must be put in place for data warehouse lifecycle management is a backup system that automates the process and guarantees that the database can be restored with full data integrity in a timely manner. The first step is to ensure that all of the data sources from which the data warehouse is created are themselves backed up. Even a small field that is used to help integrate larger data sources may play a critical part. Where a data source is external, it may be expedient to 'cache' the data to disk, to be able to back it up as well. Then there is the requirement to produce weekly backup of the entire warehouse itself which can be restored as a coherent whole with full data integrity. But many companies do not attempt this. They rely on a mirrored system not failing, or recreating the warehouse from scratch. Sometimes, they do not even practice the recreation process, so when (not if) the system breaks, the business impact will be enormous. Backing up the data warehouse itself is fundamental. Oracle database with a huge number of related files is found to be very useful in this regard.

Disaster Recovery Strategy

From a business perspective, it is important to have a disaster recovery site set up, to which copies of all the key systems are sent regularly. Several techniques can be used, ranging from manual copies to full automation. The simplest mechanism is to use a backup product that can automatically produce copies for a remote site. For more complex environments. Particularly where there is a hybrid of database management systems and conventional files, an HSM (Hierarchical Storage Management) system can be used to copy data to a disaster recovery site automatically. Policies can be set so that files or backup files can be migrated automatically from media type to media type, and from site to site. For example, disk to optical, to tape to an off-site vault. With a redundant hardware and very-high-bandwidth (fiber channel) communications wide-area network, volume replication can be used to retain a secondary remote site identical to the primary data warehouse site.

Reliability and High Availability Strategy

A reliable data warehouse needs to depend a lot less upon restore and recovery. After choosing reliable hardware and software, the most obvious thing is to use redundant disk technology. Dramatically, this improves both reliability and performance, which are often key in measuring end-user availability. Most data warehouses have kept the database on raw partitions rather than on top of a file system, perhaps to gain performance. The Veritas file system is a journaling file system and recovers in seconds, should there be a crash. This then enables the database administrator to get all the usability benefit of a file system with no loss of performance.

After disks, the next most sensible way of increasing reliability is to use a redundant computer, along with event management and High Availability Software (HAS). Example is First Watch. The event management software should be used to monitor all aspects of the data warehouse-operating system, files, database, and applications. Should a failure occur, the software should deal with the event automatically or escalated instantly to an administrator. In the event of a major problem that cannot be fixed, the HA software should face the system over to the secondary computer within a few seconds.

Management Tools

The task of managing a potential complexity and making decisions about which options to use can become a nightmare. That is, being reactive rather than proactive, means that the resources supporting the data warehouse are not properly deployed. This results in poor performance, excessive number of problems, slow reaction to them, and over buying of hardware and software systems. Management tools must therefore address enabling administrators to switch from reactive to proactive management by automating normal and expected conditions against policy. The tools must encompass all of the storage being managed. These include: database, files, file system, volumes, disk and tape arrays, intelligent controllers etc. That is, to assist the proactive management, the tools must collect data about the data warehouse and how it is and will be used. Such data could be high level, such as number of users, size, growth

online/offline mix, access trends from different parts of the world. The tools should ideally suggest or recommend better ways of doing things.

A simple example would be to analyse disk utilization automatically and recommend moving the backup job to an hour later and to stripe the data on a particular partition to improve performance of the system without adversely impacting other things. The tools can automatically manage and advise on the essential growth of the data warehouse. That is, pre-emptively, it can advise on the problems that will otherwise soon occur using threshold management and trend analysis. The tools can then be used to execute the change envisaged and any subsequent fine tuning that may be acquired.

Veritas is developing a set of management tools that address these issues. They include: *storage manager (*manages all storage objects such as the database, file systems, tape and disk arrays, network-attached intelligent devices etc. It also automates many administrative processes, manager exceptions, collects data about data, enables online performance monitoring and lets you see the health of the data warehouse at a glance); *storage analyst* (collects and aggregates further data and enables analysis of the data over time.); *storage optimizer* (recommends sensible actions to remove hot spots and otherwise improve the performance or reliability of the online/offline storage based on historical usage patterns); and *storage planner* (enables capacity planning of online offline storage, focusing on very large global databases and data warehouses).

The use of these tools and tools from other vendors should ideally be started during the 'design and predict' phase of development of a data warehouse. But in most cases, they will have to be used retrospectively to manage situations that have become difficult to control, perhaps to regain the initiative with these key corporate assets.

Conclusion

Data warehouses, datamarts and other larger database systems are now critical to most global organizations. Their management starts with a good lifecycle process that concentrates on the operational aspects of the system. Their success is dependent on the availability, accessibility and performance of the system. That is, the operational management of a data warehouse should ideally focus on these success factors. Also, managing a data warehouse lifecycle lies in defining utility, improving efficiency, and discovering effectiveness in the data warehouse lifecycle. This means that the warehouse manager as well as the development team must understand the business that drives the value in the warehouse. Continuous process improvement means continuous value in the warehouse.

References

Bakos & Yannis, J. (2004). The emerging role of electronic marketplaces on the Internet communications of the ACM 41.

Bischoff, J. (2002). Data Warehouse: Practical advice from the experts, Prentice Hall. Computerworld Data Management Links www.computerworld.com/itresources /riclinks/04167.KEY241-RL142,00 .html.

Devlin, B. (2002).Data Warehousing: From architecture to implementation. Computerworld.

Erickson, C. (2003). Multidimensionalism and the data warehouse the data warehouse conference UK.

Gerald, V. & Anderson, (2006). *Management information systems: Solving business problems with information technology.* George Werthman, McGraw – HillCompanies. Inc. New York.

Herreman, (2004). How much data do large corporations manage? Computerworld

Hinrichs, H. (2003). Data quality management in data warehouse. Berlin Marz Pp.87-98.

Immon, W. (1996). *Building the operational data store.* John Wiley and Sons Inc.

Kalakota, R. & Whinston, A. (2000). Corporation's use of the internet in development countries. International Finance Corporation Discussion Paper, World BankWashington, D.C.

Kimball, R. (1999).*The data warehouses tooikit: Practical techniques for building data warehouses.* New York: John Wiley and Sons. Inc.

Laudon, K. & Laudon, J. (2005). *Management information systems: Managing the digital firm.* Pearson Education, Inc. Pearson Prentice Hall WC2RORL UK.

Mattson, R. (2004). *Data warehousing strategies, technologies and techniques.* McGraw Hill.

Melntyre, (2005). Data warehousing environment: End-to-End Blueprint presentation material, BM UK. Ltd.

Microsoft (2003). Electronic business issues for world trade. *Microsoft Corporation White Paper,* USA.

Roche, M (2004). Planning for competitive use of information technology in multinational corporations. ALB UK Region. Conference Paper.

Steve Ulfelder (2006). Plumb your clickstream data. Computerworld.

Wisdom, B. (2005). *Research problems in data warehousing*. Proc. Intl. CIKM.