Challenges in Data Warehouse Deployment and Maintenance

[1]Avinash Kumar and [2]Pravin S. Metkewar
[1] Student, MBA-IT, [2]Associate Professor,
[1,2]SICSR, Affiliated to Symbiosis International University (SIU), Pune, Maharashtra, India

Abstract-- In the past few years, the world has witnessed the growth of a new technology, namely data warehousing. A data warehouse is a global repository that stores pre-processed queries on data which resides in multiple, possibly heterogeneous, operational or legacy sources. The information stored in the data warehouse can be easily and recently accessed for making effective decisions. The On-Line Analytical Processing (OLAP) tools access data from the data warehouse for complex data analysis, such as multidimensional data analysis, and decision support activities. Current research has led to new developments in every aspect of data warehousing, however, there are still a number of problems that need to be solved for making data warehousing effective. In this paper, we discuss the problems faced while deploying and maintaining a data warehouse.

Keywords-- *Data Warehousing, Deployment, Problems, Maintenance*

I.   INTRODUCTION

A Data Warehouse is a stored collection of diverse data, A single repository of information. It is a subject oriented mechanism used for analysis on historical data as well as data mining. It is different from transaction-oriented databases and aimed at executives who take part in decision making in an organization. A data warehouse contains a large volume of data (GB, TB) and is nonvolatile as it stores historical and time attributed data. A Data warehouse is not frequently updated and may only be append only. It is a specialized database which is generally read only as mentioned prior. The queries executed on a data warehouse are long and complex and uses a lot of scans to showcase summarized and consolidated data. It generally has a small user base consisting of decision makers and analysts only.

Warehousing is a huge business. In 1995, the data warehousing was valued at $2 Billion whereas it dramatically increased in turnover of $8 billion within 3 years by 1998. It is the most important part of a decision support system which utilizes On-Line Analytical Processing, complex queries and aggregations to produce the desired results.

A. Phases of Development and Deployment

Building a data warehouse can be a hectic task and cannot be done by someone alone. Development of a data warehouse requires a fully equipped and expert team of individuals who are experts in their related domains. Data warehouse encompasses both business practices and policies along with information system technologies which require the cooperation of both business and information technology.

There are various Data Warehouse Development Lifecycle frameworks available to be utilized by data warehouse project team. The Kimball DW/BI Lifecycle Methodology is one of the prominent framework developed by Kimball Group in the mid-1980's. Since the origin of Kimball methodology, various project teams have successfully utilized and completed thousands of data warehousing projects.

The Kimball lifecycle approach depicts and overall framework laying out the sequence of task based on priority for successful data warehouse and business intelligence solution.

a. Project Planning

This phase focuses on defining the data warehouse projects and provide the guidelines and advice to get started with the data warehouse initiative from accessing the scope to estimating the cost involved. Detailed project planning activities such as staffing and plan development are conceived during this phase.

b. Business Requirement Definition

Business users and their requirements are very critical and make an impact on the decisions were taken through the data warehouse implementation. The business requirements are the center of the universe for any data warehouse implementation. Requirement helps determine what data should be made available within the scope of the data warehouse. The phase generally consists of techniques that effectively uncover the business requirements, recommending overall approach and techniques for requirement gathering.

c. Dimensional Modeling

After the business requirements are captured and the decision has been taken on the data inclusion, the project team can now focus and start with the logical and physical design of the system. This phase transforms the legacy data resources into data warehouse structures. After the data warehouse structures are created, the plans for data extraction and transformations can be carried out.

d. Data Warehouse Architecture

A typical data warehouse architecture can be defined by the following areas.

1. Data Architecture Area

As per Kimball, Data is what the data warehouse is all about. Data is the object that we work within a data warehouse. The data can be shuffled, transformed and examined. The data area architecture contains the contents of the warehouse which is important for the business. It also includes the logical , physical designs, aggregation and hierarchies concluded from the business requirements.

2. Technical Architecture Area

The processes and tools that are utilized come under the technical area architecture of the data warehouse. The area focuses on how to get the data, transform the data and utilize the same for business purposes. The technical architecture gives life to a warehouse and is made up of tools, utilities, and code.

3. Infrastructure Architecture Area

The infrastructures constitute the environment and platform that will host the data and the processing capabilities of the

data warehouse. It constitutes the operating system and hardware that is required for implementing a data warehouse.

e. Product Selection

The architecture of a data warehouse constitutes of various tools and utilities that are included as per the business requirements. The major part of the architecture is to identify and procure these tools and utilities as per their capabilities and requirements from various vendors. Requirements having high priority are answered first.

f. Deployment

After the completion of data warehouse development tasks, data warehouse deployments require coordination from various variables such as planning. Deployment basically involves convergence of the technology finalized, requirements and the users.

g. Maintenance

After the successful deployment of the warehouse, the focus shifts on the maintenance and growth part. After deployment, the business users are now able to utilize the data from the data warehouse to its full extent. The data warehouse team becomes responsible for providing the users with the required data, views and processing capabilities.

II. CHALLENGES FACED IN DATA WAREHOUSE DEPLOYMENT

Developing and deploying a hectic task as mentioned earlier and issues are complications are bound to arise out of such large scale tasks.

Following are the issues that are faced generally

A. Roles and Responsibility Crisis

A data warehouse project team might take the time to figure out the roles and responsibilities.

B. Data Issues

Issues in mapping the data back to its origin can be a complicating task.

C. Infrastructure Issues

Environmental and Hardware issues are prominent

D. Architecture Issues

Issues involving software, hardware, and the environment.

III. CHALLENGES FACED IN DATA WAREHOUSE MAINTENANCE

Maintenance is a continuous task. Data warehouses are huge in size. An estimated 20% of the time is spent on data extraction, cleansing and loading process. After deployment, the users demand and expectancy increases and it becomes a challenge for the data warehouse team. Following are the issues that are generally faced

A. Change in Business Process/Policy

If there is a change in the business process and policy in the organization, the Data warehouse teams may have to reorganize and the source system may require the data warehouse to change. updates are required in such cases.

B. Taking a decision on the data.

It becomes a hefty task for the data warehouse team to decide on the data sources and also which data to keep and what not.

C. Deciding on the Queries and Access

One of the primary task for the project team is to identify the queries that are required for data extraction. Some of the queries are inbuilt and some require users input. The task of deciding the same can be confusing and might create complications in the future. Providing users access to custom queries is also a challenging task and requires security policies in place.

D. Uncertainty over Warehouse and Operational Store

Users are uncertain in determining if the data should be fetched from the data warehouse or operational store.

E. Updates need to be applied

As mentioned earlier, a change in business policy or a new loophole might require a new update or patch to be applied to the system.

F. Architecture Maintainance

Maintaining the architecture of the data warehouse is more difficult than establishing a new one. Variables such as vendor support and updates play an important role.

G. Security

Security of a data warehouse is as important as protecting your own child. The huge amount of data stored in the data warehouse makes it an easy target for unauthorized access. Loopholes and breaches can be imminent and needs to be addressed during the development phases. A change in security policies may require the data warehouse security to be changed or overhauled.

References

[1] Data Warehouse Maintenance - ISSN 1653-0187 / ISRN LTU-PB-EX--06/38--SE / NR 2006:38
[2] The Data Warehouse Lifecycle Toolkit, Second Edition - ISBN-10: 0470149779
[3] The Data Warehousing Handbook - ISBN-10: 1847286658