ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 3, pp: (229-232), Month: July - September 2016, Available at: www.researchpublish.com

One-Step Ahead Approach for Selection of Data Warehouse Architecture

Jaspreet Kaur[1], Pravin S. Metkewar[2]

[1]MBA-IT, 2nd Year Student, SICSR, affiliated to Symbiosis International University (SIU), Pune, Maharashtra, India
[2]Assoc. Professor, SICSR, affiliated to Symbiosis International University (SIU), Pune, Maharashtra, India

Abstract: **Organizations, nowadays, strive for excellence in their operations. But, with the increasing complexity in business, they are unable to gain significant competitive advantage. And, this has created a need to evaluate and select an appropriate data warehouse architecture. However, deciding the most suitable data warehouse architecture that fits your business needs – is perhaps, the most crucial activity in the Data Warehouse lifecycle. An architecture best describes the requirements and specifics of an organization. In fact, it helps the management in making right decisions. Thus, organizations should aim at selecting an architecture that is robust, flexible and scalable. Any mistake in this area may affect the overall business goals and performance of the organization. The aim of this research paper is to better understand the factors that must be considered before selecting any data warehouse architecture.**

Keywords: **Architecture, Data Warehouse, Data Mart, ETL, Selection Factors.**

1. INTRODUCTION

Over the last few years, millions of rupees have been invested in the field of data warehousing. There is one arena, however, that still causes confusion: Which architecture to select and how?

But, before selecting the architecture, one should know what a Data Warehouse is. Bill Inmon (Father of Data Warehousing) defines Data Warehouse as: "A subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision-making process."[1] Companies, in general, consider the importance of the choice of architecture, but, are not able to figure out the basis for selection. Different types of architectures can be selected on the basis of technical & business requirements, information needs of the organization, vendor characteristics and availability of resources.

1.1 Overview of Data Warehouse Architectures

Data Warehouses and their architectures depend on the requirements and specifications of an organization. Higher the number of requirements, more complex will be the architecture and vice-versa. The two extremes of data warehouse architecture include- Single-layer architecture and N-layer architecture.[2] Small and medium scale organizations (or SMEs) usually adopt single-layer architectures and/or two-tier architectures. Counter to this, large-scale organizations generally opt for high-level architectures with data marts and complex middleware design. This is because large-scale organizations deal with complex operations and thus, require complex design for detailed solutions. The difference in the architectures is the number of middleware(s) between the operational data sources and reporting tools.

1.2 Types of Data Warehouse Architectures

Depending on the number of middleware between the operational data sources and reporting tools, data warehouse architecture can be classified into 3 categories-

- Basic Architecture: In this type of data warehouse architecture, the end users have direct access to data (from multiple operational data sources) through the data warehouse. In the data warehouse, metadata and raw data are present. Summary

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 3, pp: (229-232), Month: July - September 2016, Available at: www.researchpublish.com

data is also essential because data can be easily retrieved for computing long operations. For instance, a company can retrieve its sales figure for the month of January.

- Architecture with a Staging Area: Here, the data coming from various operational sources can be cleaned before it goes into the data warehouse. This is done with the help of staging area. A staging area is used for data cleansing and processing during the ETL (Extract, Transform & Load) process. It also helps in simplifying large summaries of data.[3]

- Architecture with a Staging Area and Data Marts: Organizations who wish to customize their data requirements according to various departments may use such type of architecture. Data Marts are added for handling data needs of a particular line of business. For instance, sales, production and logistics departments can be separated using data marts. Thus, a business analyst might want to analyze the customer trends and behavior from the sales department using its data mart.

All the above architectures are generic in nature, and will vary according to the needs and specifics of the organization.

2. OVERVIEW OF THE PROPOSED SOLUTION

Fig. 1. below diagrams the tree structure of the factors essential for selecting the right data warehouse architecture. Some of the factors include- Technical Factors (example- User Interface, Reporting Tools, Query Capabilities, Response Time, Scalability, Security, Compatibility with the existing systems, Integration with source systems as well as third-party systems, Type of Database used, ETL Functionalities, Administration Capabilities, Metadata Management, Data Quality Checks) and Managerial Constraints (Direct & Indirect Costs, Information Needs of Management, Characteristics of Vendor).

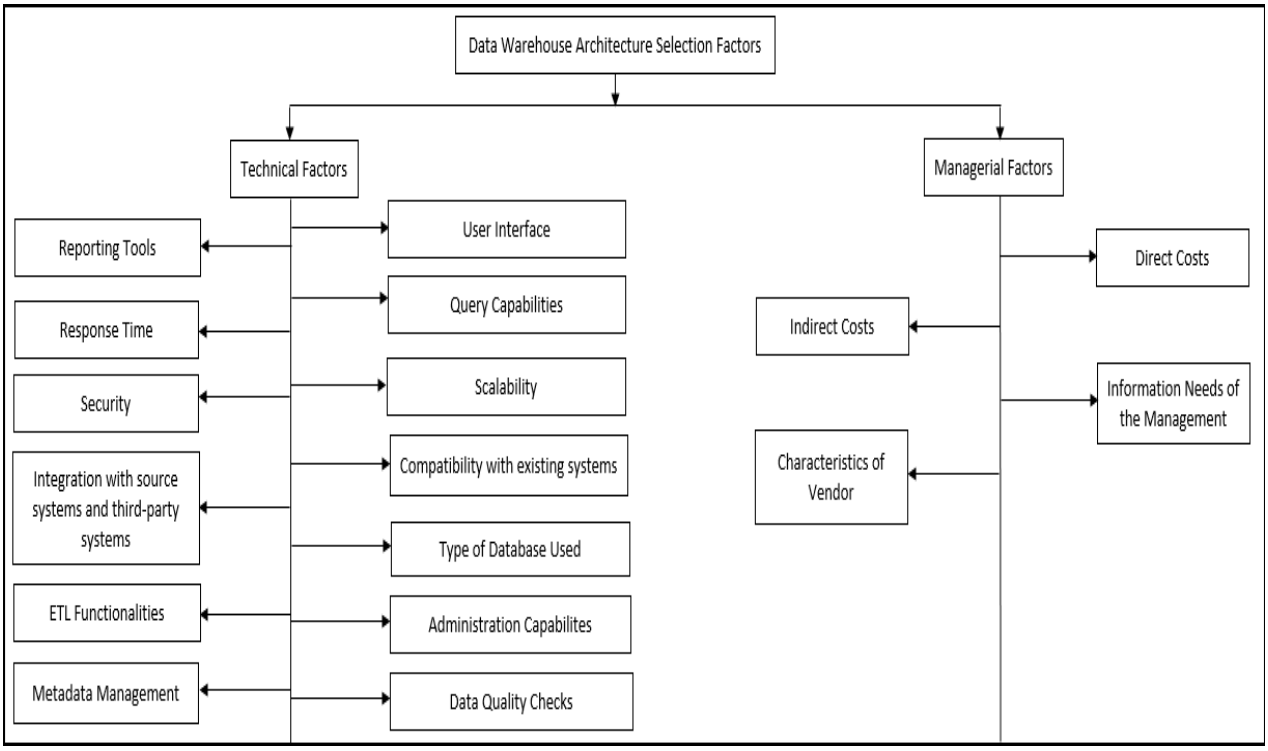

Fig. 1. Tree Structure of Factors associated with Selection of Data Warehouse Architecture

2.1 Technical Factors

Every organization has certain requirements to fulfill. These requirements are the solutions for a particular system. And, these solutions are provided by the technical factors. These factors represent the core capabilities of any type of architecture. Some of the technical factors for data warehouse architecture are as follows-

- User Interface: End-users want access to the right data, at the right time & right place. Thus, a well-defined architecture must have an easy-to-use user interface. Some users might want to access the data warehouse systems remotely. For them, mobile UI can be a better option.

- Reporting Tools: These tools allow the end-users to generate reports from the data and analyze them. The users, who access the data warehouse systems more frequently, can use tools such as- data mining tools, statistical tools & data visualization tools to create reports and map against the queries.
- Query Capabilities: A data warehouse contains a huge amount of data (historic data) and requires a handful of query design tools in its architecture. Thus, any data warehouse architecture must provide the capability of solving complex queries (joins, sort and group-by operations). Features like- slicing & dicing, drill-down & roll-up can be considered.
- Response Time: It is the time taken by a system, in general, to respond to a particular action. For example, a user might give an option of slice & dice in a data warehouse system. Now, the time it takes to understand the command, perform an operation and generate a report is the response time of the data warehouse system.
- Scalability: The increase in the number of operations in an organization has led to an increase in the workload of functions and departments. Thus, any architecture, which is not scalable, may not be able to perform under higher workloads. Scalability is an essential factor when considering data warehouse systems because the amount of data is increasing day-by-day.
- Security: Organizations aim to choose systems which are secure enough to handle any disruptions against the external environment. For example, any data warehouse architecture, containing historic data of past 50 years, can be considered as secure, if it gives alerts on any unauthorized access to steal confidential information.
- Compatibility with the existing systems: Any data warehouse architecture must be compatible with the existing infrastructure of the organization. It includes- hardware (PCs, graphic cards, RAM, motherboard), software (different versions of software) and network components (data link connections).
- Integration with source systems as well as third-party systems: Large-scale organizations, which are heavily dependent on source systems (enterprise architectures, legacy systems) as well as third-party systems (CRM applications, ERP systems), would like to consider architectures which can be integrated with these systems. As far as data warehousing is concerned, organizations would want to integrate their architectures with decision support systems (DSS), management information systems (MIS), executive information systems (EIS).
- Type of Database Used: Data Warehouse Architecture may be implemented on relational databases or OLAP servers, or there could be multi-dimensional databases (MDDBs), depending on what the requirements of the organization are.
- ETL Functionalities: ETL is Extract, Transform & Load. Any data warehouse must be capable of performing the ETL process. This process includes- extracting the data from operational data sources, transforming the data into a consistent state, and loading the transformed data into the data warehouse.[4] For the faster execution of the process, the data warehouse architecture may consider the use of ETL tools (Informatica, CloverETL, and Pentaho).
- Administration Capabilities: These capabilities can help the administrator to manage the data warehouse systems well. Various administration tools can be used for better planning and monitoring of the systems.
- Metadata Management: Metadata clearly shows what data is residing in your data warehouse, like that of a book index. Organizations find managing the metadata as a challenging task. However, it is one of the most important features any data warehouse architecture must consider. It includes managing- business metadata, administrative metadata, and operational metadata.
- Data Quality Checks: Data Quality is one of the most important concerns when implementing any data warehouse architecture. The quality of data must be maintained so as to ensure that the data is consistent, complete and accurate.

2.2 Managerial Factors

These factors are related to the decision-making capabilities in an organization. These are more likely to be generic in nature and which most of the organizations consider it even before taking into account the technical factors for selecting any data warehouse architecture. Some of the managerial constraints for data warehouse architecture are as follows-

- Direct Costs: These costs are incurred during the implementation stages of data warehouse architecture. These include- hardware cost, software cost, and consultant fees.

- Indirect Costs: These costs are not directly related to the implementation of data warehouse architecture. These include- training cost of the employees, maintenance and upgrades cost and personnel cost.
- Information Needs of the Management: In an organization, various types of information may be required by all the levels of management. For instance, the higher level of management (CEOs and other senior executives) may require access to data more frequently as compared to the lower level of management (junior analysts and junior-level employees).
- Characteristics of Vendor: Organizations, these days, consider the reputation and goodwill of the vendors in the market, before making any decision of implementing data warehouse architecture. In fact, it has become a major criterion in selecting any type of architecture. For instance, some vendors provide support services for your architecture.

3. CONCLUSION

Just as we consider various factors before selecting any software, similarly, we must consider a handful of factors before selecting any data warehouse architecture. The factors, identified in this research paper, are an aid for organizations and their people, to choose the best architecture in the field of data warehousing. Organizations must consider, at least, 5-6 factors, before making a choice for data warehouse architecture. Apart from pricing and information needs, they must also look at the major technical factors before making a final choice. These factors can really help the organization make better decisions. The tree structure can help them take a problem systematically and identify the relevant criteria and sub-criteria.

REFERENCES

[1] W. Inmon, “Building the Data Warehouse,” John Wiley Sons, fourth edition, 2005.

[2] Kimball, Ralph, “The Data Warehouse Toolkit”, Wiley Computer Publishing, 1996.

[3] Ralph Kimball, Joe Caserta, Data Warehouse. ETL Toolkit. Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data (Wisely Publication Inc.).

[4] Kimball, R., Reeves, L., Ross, M., & Thornthwaite, W. (1998), “The data warehouse lifecycle toolkit – expert methods for designing, developing and deploying data warehouses”, New York: Wiley & Sons.