Decision Support Systems 49 (2010) 200–212

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Key organizational factors in data warehouse architecture selection

Thilini Ariyachandra [a], Hugh Watson [b,*]

[a] Xavier University, United States
[b] University of Georgia, United States

ARTICLE INFO

Article history:
Received 2 July 2007
Received in revised form 16 February 2010
Accepted 22 February 2010
Available online 4 March 2010

Keywords:
Data warehousing
Data warehouse architecture
Archictecture selection
Independent data marts
Bus architecture
Enterprise data warehouse
Hub and spoke
Federated

ABSTRACT

Even though data warehousing has been in existence for over a decade, companies are still uncertain about a critical decision — which data warehouse architecture to implement? Based on the existing literature, theory, and interviews with experts, a research model was created that identifies the various contextual factors that affect the selection decision. The results from the field survey and multinomial logistic regression suggest that various combinations of organizational factors influence data warehouse architecture selection. The strategic view of the data warehouse prior to implementation emerged as a key determinant. The research suggests an overall model for predicting the data warehouse architecture selection decision.

1. Introduction

Over the past decade, many companies have made data warehouses the foundation of their decision support infrastructures. These data repositories provide a solution to a recurring problem that has plagued companies' decision support initiatives — the lack of clean, accurate, timely, and integrated data. Whether the data is used for queries and reporting, decision support systems (DSS), executive information systems (EIS), online analytical processing (OLAP), or data mining, data warehouses provide the data that "fuels" these applications [36]. Data warehouses are also critical enablers of current strategic initiatives such as customer relationship management (CRM), business performance management (BPM), and supply chain management [18,23,35,85].

Despite the importance and growing experience with data warehousing, there is still a considerable discussion and disagreement over which architecture to use. Bill Inmon, commonly referred to as "the father of data warehousing," advocates what is variously called the hub and spoke architecture (i.e., a centralized data warehouse with dependent data marts), Corporate Information Factory, DW 2.0 (i.e., Inmon's current term), or enterprise data warehouse architecture [23]. Another architecture alternative is the data mart bus architecture with linked dimensional data marts (i.e., bus architecture), advocated by Ralph Kimball, the other preeminent figure in data warehousing [52].[1] These and other architectures (e.g., independent data marts and federated) have fundamental differences and strong advocates.

The selection of an appropriate architecture is an important decision. A study by the Meta Group found that architecture selection is one of the key factors influencing data warehousing success [54]. A Gartner report identified the architecture selection decision as one of the five problem areas associated with data warehouse projects [72]. A poor architecture selection decision can lead to problems such as lack of scalability, performance difficulties, and no "single version of the truth."

The data warehouse architecture selection decision is a subset of IT infrastructure (ITI) design. Currently, the academic literature contains only limited research on ITI design [21,87] and typically presents findings from case studies or offers anecdotal recommendations [10,85]. A recent article [68], one of the few empirical studies on ITI infrastructure, analyzes current practices in data warehouse implementation methodologies but does not explicitly describe factors that influence architecture selection. Another exploratory study [58] on data warehousing identifies a list of organizational factors that influence data warehousing refresh policies but does not examine how these factors influence data warehouse design. In contrast, the IS field is rich with research on the design of IT applications [8,16]. This latter substantial body of literature does suggest factors that

* Corresponding author.
E-mail addresses: ariyachandrat@xavier.edu (T. Ariyachandra), hwatson@uga.edu (H. Watson).

[1] Kimball takes exception with the use of the enterprise data warehouse term to describe the Inmon architecture, arguing that his approach also results in an enterprise-wide solution when fully and correctly implemented. Despite this objection, the term is widely used and is also used in this paper.

0167-9236/$ – see front matter
doi:10.1016/j.dss.2010.02.006

potentially influence ITI design. Many of these factors are theory-based and draw from a variety of rational and social/political theories. However, it is unclear how these factors affect ITI design decisions, and more specifically, how they influence decision support infrastructure design choices.

Considering the millions of dollars that companies typically spend on data warehouses, it is surprising that there is no empirical, theory-based research on warehouse architecture selection. This lack of rigorous academic research motivated our study. As part of our research, we were interested in understanding the factors that affect the architecture selection decision. More specifically, we wanted to explore the following research questions: "What factors are most important to the architecture selection decision?" "What factors influence the selection of a particular architecture?" And finally, we were interested in: "Do the findings on the data warehouse architecture selection decision confirm and extend current understandings about ITI decisions?" In order to investigate these questions, a multi-phase study was conducted using both qualitative and quantitative methods.

This paper is organized as follows. First, the major data warehouse architectures are described. Next, the relevant literature, theory, and expert input used to develop the research model are described. The research model and the hypotheses are presented next, followed by a detailed description of the research methodology that was used. The data analysis section discusses the validity of the research model, testing of hypotheses, and additional analyses. In the concluding sections, the research results are discussed, an overarching model for data warehouse architecture selection is proposed, and the implications for research and practice are presented.

2. Data warehouse architectures

A data warehouse is a specially prepared data repository that is used to support decision making [36]. Data is extracted from data sources, transformed, and loaded (ETL) into data stores. The data is then made available for end user access and decision support applications. It can be considered as decision support data infrastructure that is used for multiple, varied decision support purposes.

Architecture is the structure of something [80]. While the term is usually associated with the kind or style of a building, it is used more broadly, including for example, the structure of human societies or the solar system. A *reference* architecture is a particular kind of architecture. It helps in understanding or explaining what currently exists or serves as a guide for creating something new, such as a building or a data warehouse. In this study we refer to four different data warehouse reference architectures. The reference architectures identify the alternative ways that data is extracted, transformed, loaded, and stored in a data warehouse.

The data warehousing literature provides discussions and examples of different architectures [47,52]. Vendors and consultants promote a variety of reference architectures to guide data warehouse implementation efforts [82]. Let us consider how the major architectures have evolved over time and how they were described in the study. As discussed later, the descriptions were based on the data warehousing literature and interviews with leading experts in the field. Other architectures are also discussed in the literature, but they tend to be variations on those that were studied [63].

The early 1970s ushered in decision support systems, which were fundamentally different from operational or transactional systems. A popular way of conceptualizing a DSS was through a dialog-data-model (DDM) paradigm Sprague et al. [71]. For the data component, it was recognized that a separate data repository was needed that drew data from operational systems and other data sources. In response, independent data marts were developed as the first decision support data infrastructure. This was an application-centric approach to data management because the repositories were designed to support only a single or a few applications [83].

This evolution of decision support management can also be viewed from a maturity model perspective. In Eckerson's [23] six-stage model, which uses a human evolution metaphor, prenatal is the first stage. It features production reporting from operational systems. In the infant stage, there are "spreadmarts" in the form of Excel spreadsheets. Next is the child stage and it features independent data marts.

2.1. Independent data marts (IDM)

Just as companies have legacy operational systems, they also have legacy independent data marts.[2] In addition, some companies choose to create new marts. These marts are independent of other data stores, and while they may meet localized needs, they do not provide "a single version of the truth" for the entire organization. They typically have inconsistent data definitions and use different dimensions and measures (i.e., non-conformed) that make it difficult to analyze data across the marts [41]. Fig. 1 shows the architecture for each independent data mart.

In the late 1980s, a number of companies, especially in the financial services, telecommunications, and retail industries, developed the first data warehouses. The warehouses emerged, in part, because of the siloed nature of independent data marts. Companies wanted an enterprise-wide data repository (often focusing on all interactions with customers) to support a variety of analytical applications (e.g., queries, OLAP, and data mining). This approach represented a data-centric approach to decision support data management. In Eckerson's maturity model, it was a movement to the teenager stage.

Two competing architectures for data warehousing quickly emerged, each advocated by one of the two luminaries in the field: Ralph Kimball for the data mart bus architecture and Bill Inmon for the enterprise data warehouse.

2.2. Data mart bus architecture with linked dimensional data marts (DBA)

The development of the data mart bus architecture begins with the identification of the business requirements for a specific business process, such as orders, deliveries, customer calls, or billing. The first mart is built for a single business process using dimensions and measures (i.e., conformed dimensions) that will be used with other marts [52]. Additional marts are developed using these conformed dimensions, which result in logically integrated marts and an enterprise view of the data. Atomic and summarized data are maintained in the marts and are organized in star schemas to provide a dimensional view of the data. This architecture is illustrated in Fig. 2.

2.3. Enterprise data warehouse architecture (EDW)

An extensive enterprise-level analysis of data requirements provides the basis for this architecture [47]. Attention is focused on building a scalable infrastructure. Using the enterprise view of the data, the architecture is developed in an iterative manner, subject area by subject area. Atomic level data is typically maintained in the warehouse in 3rd normal form. Dependent data marts are created that

[2] Although we include independent data marts as a data warehouse architecture, it is not a formally advocated architecture in the industry. As discussed, independent data marts are present day stovepipe solutions that result from historical organizational efforts to build decision support solutions. Yet they exist and are used in organizations as data warehouse solutions. They are also commonly considered in discussions and surveys of the various data warehouse architectures. Consequently, we discuss independent data marts as an architecture in this study.