Which Data Warehouse Architecture Is Most Successful?

Thilini Ariyachandra and Hugh J. Watson

Thilini Ariyachandra is assistant professor of MIS in the College of Business at the University of Cincinnati.
ariyact@uc.edu

Hugh J. Watson holds a C. Herman and Mary Virginia Terry Chair of Business Administration in the Terry College of Business at the University of Georgia. He is a Fellow of TDWI and the senior editor of the *Business Intelligence Journal*.
hwatson@uga.edu

For over a decade, discussion and even controversy have lingered about which is the best data warehouse architecture. The two "giants" of the data warehousing field, Bill Inmon and Ralph Kimball, are at the heart of the disagreement. Inmon advocates the hub-and-spoke architecture (e.g., the Corporate Information Factory), while Kimball promotes the data mart bus architecture with conformed dimensions. There are other architecture alternatives, but these two options are fundamentally different approaches, and each has strong advocates.

Considering the importance of the architecture choice, surprisingly little research on the topic exists. The literature tends to either describe the architectures, provide case-study examples, or present survey data about the popularity of the various options. There has been little rigorous, empirical research, and this motivated us to investigate the success of the various architectures. Here we highlight and summarize some findings from our research, but we encourage you to read the complete research report.[1]

The Architectures Studied

For our study, we investigated five architectures: (1) independent data marts, (2) data mart bus architecture with linked dimensional data marts, (3) hub-and-spoke, (4) centralized data warehouse (no dependent data marts), and (5) federated. Other architectures are discussed in the literature, but they tend to be variations on these five. Figure 1 depicts the five architectures.

[1] The full research report is available at www.terry.uga.edu/~hwatson/ DW_Architecture_Report.pdf

Data Collection

A Web-based survey—targeted at individuals involved in an organization's data warehouse implementation—was used to collect data. The survey included questions about the respondent, the respondent's company, the company's data warehouse, and the success of the data warehouse architecture. Four hundred fifty-four respondents provided usable information.

The positions of the respondents were distributed relatively evenly among data warehouse managers, data warehouse staff members, IS managers, and independent consultants/system integrators. Respondents in the last group were asked to complete the survey with a particular client in mind.
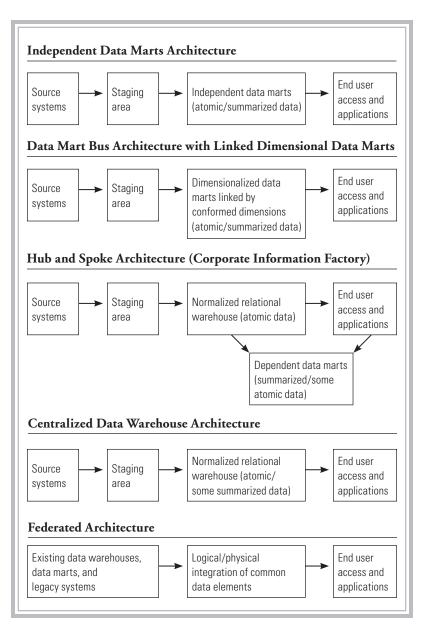

Figure 1. *The five data warehouse architectures studied*

Surveyed companies ranged from small (less than $10 million in revenue) to large (in excess of $10 billion). Most of the companies are located in the United States (60%) and represent a variety of industries, with the financial services industry (15%) providing the most responses.

The predominant architecture was the hub-and-spoke (39%), followed by the bus architecture (26%), centralized (17%), independent data marts (12%), and federated (4%). The most common platform for hosting the data warehouses was Oracle (41%), followed by Microsoft (19%) and IBM (18%). The average (mean) gross revenue varied from $3.7 billion for independent data marts to $6 billion for the federated architecture.

Success of the Architectures

Four measures were used to assess the success of the architectures: (1) information quality, (2) system quality, (3) individual impacts, and (4) organizational impacts.[2] The questions used a seven-point scale, with the higher score indicating a more successful architecture. Figure 2 shows the average scores for the measures across the architectures.

Independent data marts scored the lowest on all measures. This finding

	Independent Data Marts	Bus Architecture	Hub and Spoke	Centralized (No Dependent Data Marts)	Federated
Information Quality	4.42	5.16	5.35	5.23	4.73
System Quality	4.59	5.60	5.56	5.41	4.69
Individual Impacts	5.08	5.80	5.62	5.64	5.15
Organizational Impacts	4.66	5.34	5.24	5.30	4.77

Figure 2. The success of the five architectures

confirms the conventional wisdom that independent data marts are a poor architectural solution.

Next lowest on all measures was the federated architecture. Firms sometimes have disparate decision-support platforms resulting from mergers and acquisitions, and they may choose a federated approach, at least in the short run. The findings suggest that the federated architecture is not an optimal long-term solution. What is interesting, however, is the similarity of the averages for the bus, hub-and-spoke, and centralized architectures. The differences are sufficiently small that no claims can be made for a particular architecture's superiority over the others, at least based on a simple comparison of these success measures.

We also collected data on the domain (e.g., varying from a sub-unit to company-wide) and the size (i.e., amount of data stored) of the warehouses. We found that the hub-and-spoke architecture is typically used with more enterprisewide implementations and larger warehouses.

[2] Information quality considers information accuracy, completeness, and consistency. System quality includes system flexibility, scalability, and integration. An architecture has individual impacts when users can quickly and easily access data; think about, ask questions, and explore issues in ways that were not previously possible; and improve users' decision-making capabilities. It has organizational impacts when it meets the business requirements, facilitates the use of business intelligence, supports the accomplishment of strategic business objectives, enables improvements in business processes, leads to high, quantifiable ROI, and improves communications and cooperation across organizational units. Separate questions were used for the various aspects of information quality, system quality, individual impacts, and organizational impacts.

We also investigated the cost and time required to implement the different architectures. Overall, the hub-and-spoke architecture was the most expensive and time-consuming to implement. This is not surprising, however, because of the larger domain and size of these warehouses. The architecture also requires a considerable commitment to up-front planning, which takes time and money.

Conclusion

The bus, hub-and-spoke, and centralized architectures earned similar scores on the success metrics. This finding helps explain why these competing architectures have survived over time—they are equally successful for their intended purposes. In terms of information and system quality and individual and organizational impacts, no single architecture is dominant.

The similar success of the bus, hub-and-spoke, and centralized architectures is perhaps not all that surprising. In some ways, the architectures have evolved over time and become more similar. For example, the hub-and-spoke architecture often includes dimensional data marts, which is at the heart of the bus architecture. Even the development methodologies (e.g., top down for the hub-and-spoke and centralized architectures, and life cycle or bottom up for the bus architecture) have evolved and become more similar. Each stresses the need to start small and deliver short-term "wins" but have a long-term plan. This evolution is appropriate and good for the industry, but it is also a likely reason that the scores on the success metrics are similar. ∎