Data Warehouse Architectures: Factors in the Selection Decision and the Success of the Architectures

Hugh J. Watson
Terry College of Business
University of Georgia
Athens, Georgia 30602
Phone: (706) 542-3744
Email: hwatson@terry.uga.edu

Thilini Ariyachandra
College of Business
University of Cincinnati
Cincinnati, Ohio 45221
Phone: (513) 556-7036
Email: ariyact@uc.edu

July 2005

About the Study

This research was conducted to (1) better understand the factors that influence the selection of a data warehouse architecture and (2) the success of the various architectures. The academic and data warehousing literature and industry experts were used to identify architecture selection factors and success measures and then to create questions for a Web-based survey that was used to collect data from 454 companies about the respondents, their companies, their data warehouses, the architectures they use, and the success of their architectures. The experts and selected survey respondents were then contacted to help understand and interpret the survey data. The study findings provide interesting and useful insights about topics of long-standing importance to the data-warehousing field.

The study was conducted as academic research and was not financially supported by vendors or consultants. However, vendors and consultants played important roles in fine-tuning the ideas, language, and wording used in the research; promoting the study; and interpreting the research findings. Ultimately, however, the authors are responsible for the materials in this report. We believe that they fairly represent the various perspectives on the topics studied.

About the Authors

Hugh J. Watson (hwatson@terry.uga.edu) is a Professor of MIS in the Terry College of Business at the University of Georgia and a holder of a C. Herman and Mary Virginia Terry Chair of Business Administration. He is a leading scholar and authority on BI and data warehousing and has authored 22 books and over 100 scholarly articles in journals such as *MIS Quarterly, Journal of MIS, Management Science,* and the *Academy of Management Journal*. Hugh is a Fellow of The Data Warehousing Institute and the Association for Information Systems. He is the Senior Editor of the *Business Intelligence Journal* and is the consulting series editor for John Wiley & Sons' MIS series.

Thilini Ariyachandra (ariyact@uc.edu) is an Assistant professor of MIS in the College of Business at the University of Cincinnati. Her main research area is BI and data warehousing. She has published in *Decision Support Systems, Information Systems Management,* and the *Business Intelligence Journal.*

Executive Summary

For over a decade, there has been ongoing discussion and even controversy over which is the best data warehouse architecture to use. The two "giants" of the data-warehousing field, Bill Inmon and Ralph Kimball, are at the heart of the disagreement. Inmon advocates the use of 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 with strong advocates for each.

The Study

We conducted the research to answer two questions:
1. What factors lead companies to select a particular architecture and
2. How successful are the various architectures?

A multi-phased research method was used. First, the academic and practitioner literature were reviewed and 20 leading experts in the field were interviewed. These sources were used to identify the most important architectures, the factors that potentially affect the architecture selection decision, and the metrics to assess the success of the various architectures.

Five architectures were identified for study: independent data marts, data mart bus architecture with conformed dimensions (bus architecture), hub and spoke, centralized, and federated.

Eleven factors that potentially affect the architecture selection decision surfaced: information interdependence between organizational units, upper management's information needs, the urgency of need for a data warehouse, the nature of end user tasks, constraints on resources, the strategic view of the data warehouse prior to implementation, expert influence, compatibility with existing systems, the perceived ability of the in-house IT staff, source of sponsorship, and technical issues.

The success metrics identified were information quality, system quality, individual impacts, organizational impacts, development time, and development cost. Information quality considers information accuracy, information completeness, and information consistency. System quality includes system flexibility, system scalability, and system 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. Development time issues include the time required to implement the first business process(es) or subject area(s) and whether the project was on schedule. Cost issues include the costs of implementing the first business process(es) or subject area(s), the annual maintenance costs, and whether the project was on budget.

In the study's second phase, a Web-based survey was used to collect data about the respondents, their organizations, their data warehouses, the architectures used, the factors that led to the architecture selection decisions, and the success metrics. The experts helped develop the survey instrument and the questions used. A number of people and organizations promoted the study through emails, newsletters, announcements, and website notices. Four hundred and fifty four respondents provided usable survey data.

In the study's third phase, the findings from the data analyses were shared with the experts for their reactions, questions, and interpretations. Also, some of the survey respondents were contacted to gain qualitative insights.

The Respondents, Companies, and Data Warehouses

Individuals who were involved in the implementation of the data warehouse in organizations were the target respondents for the survey. The positions of the respondents are relatively evenly distributed over data warehouse managers, data warehouse staff members, IS managers, and independent consultants/system integrators. The latter were asked to complete the survey with a particular client in mind. The companies included in the survey range from small (i.e., less than $10M in revenues) to large (i.e., in excess of $10B). 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 is the hub and spoke (39 %), followed by the bus architecture (26%), centralized (17 %), independent data marts (12%), and federated (4%). The respondents indicated that their architectures matched the reference architectures closely, especially in the case of the bus, hub and spoke, and centralized architectures. The most common platform for hosting the data warehouses is Oracle (41%), followed by Microsoft (19%) and IBM (18%). The average (i.e., mean) gross revenue of the companies varies across the architectures from $3.7B for independent data marts to $6B for the federated architecture. The average age of the warehouses ranges from 39 months for independent data marts to 56 months for the hub and spoke. Most of the warehouses support either several business units (38%) or the entire company (36%). Fewer than 12 percent of the data warehouses are for a single functional area or sub unit. Companies do not always stay with their initial architecture. In fact, one third of the companies switched architectures at some point.

Selecting an Architecture

When asked to assess the importance of the various architecture selection factors, the responses showed that all of the factors have some influence. However, the most important ones are information interdependence between organizational units, the strategic view of the warehouse prior to implementation, and upper management's information needs. In general, the selection factors for independent data marts receive lower average scores than the other architectures, suggesting that independent data marts are employed more by happenstance than the others. The bus, hub and spoke, and centralized architectures have similar scores for most of the selection factors.

An advanced data analysis technique, stepwise binary logistic regression analysis, was used to investigate why a particular architecture was selected. For this analysis, the hub and spoke and centralized architectures were combined into a single category and the federated architecture was excluded. In the first data analysis, three factors were identified that lead to the selection of the independent data marts architecture: constraints on resources, the view of the data warehouse prior to implementation, and the perceived ability of the in-house IT staff. Four factors were found to influence the selection of the bus architecture: information interdependence between organizational units, urgency of need for a data warehouse, constraints on resources, and source of sponsorship. The analysis for the hub and spoke/centralized architecture revealed that information flow between organizational units, the view of the data warehouse prior to implementation, and the perceived ability of the in-house IT staff significantly influence the selection of this architecture. The final analysis compared the bus architecture against the hub and spoke/centralized architectures. Three factors were found to differentiate between the selection of the bus and hub and spoke/centralized architectures: information flow between organization units, the urgency of need for a data warehouse, and the view of the warehouse prior to implementation.

To analyze the potential affect that the domain of the data warehouse implementation has on the architecture selection decision, the data was grouped and analyzed on the basis of domain. Analysis of variance (ANOVA) was used to examine which selection factors are statistically different for large and small domains. It was found that information interdependence between organizational units is more important as a selection factor for large domains.

Some of the survey respondents and experts provided personal insights into the architecture selection decision. Several of the factors mentioned as important include cost considerations; a need to implement quickly; politics; a belief in the merits of E-R data modeling techniques and normalization principles; expert influence; security and performance reasons; the need for a high level of data accuracy, consistency, and control; the “dumbing down” of requirements; the practicalities of moving to a different architecture; the complexity of existing decision support environments; the ability to meet new information requirements; and the ubiquitous nature of independent data marts.

The Success of the Architectures

To assess the success of the architectures, averages were computed for the various success metrics. This analysis revealed a similar pattern for most of the architectures. Independent data marts architecture rank lowest on system quality, information quality, individual impacts, and organizational impacts. The next lowest rated is the federated architecture. The bus, hub and spoke, and centralized architectures have very similar averages. They also have the lowest percentage of warehouses (about 10%) that are potentially in trouble, and many of them are a runaway success.

Depending on the criticality of the data warehouse, the time to rollout the first subject area(s) or business process(es) can be very important. On average, the hub and spoke architecture requires the most time (11 months), followed by the federated architecture (11 months). Independent data marts, the bus architecture, and the centralized architecture are next in terms of time, and all take about the same amount (9 months) for the initial rollout. Overall, the majority (i.e., at least 50%) of implementations are on or ahead of schedule for all of the architectures. The centralized architecture is the most likely (14%) to be ahead of schedule, while the bus (47%) and the hub and spoke (47%) architectures are the most likely to be behind schedule.

Cost data was collected for rolling out the first subject area(s) or business process(es) and the annual maintenance cost. The cost pattern for rolling out the initial version is similar to that for development time. The hub and spoke architecture is the most expensive (i.e., $2.4M on average), followed by the federated architecture (i.e., $2.1M). The centralized architecture is slightly more expensive on average than the bus architecture (i.e., $1.5 versus $1.4M). The independent data marts are the least expensive (i.e., $1.3M).

The average annual maintenance costs show a similar but slightly different pattern. The hub and spoke architecture is still the most costly (i.e., $1.5M) followed by the federated architecture (i.e., $1.4M). However, the bus architecture costs slightly more on average (i.e., $.9M) than the centralized architecture (i.e., $.8M). The independent data marts are the least expensive (i.e., $.4) to maintain annually.

To investigate the potential effect that domain has on the success metrics, the company-wide implementations were analyzed separately. Overall, the average scores for information quality, system quality, individual impacts, and organizational impacts change very little when adjusted for domain. It was found that the initial development time and costs for independent data marts, the bus architecture, and the centralized architecture are very similar to one another. The hub and spoke architecture takes the longest time to initially develop and is the most costly, but is also associated with warehouses that are larger in size.

Two advanced statistical analysis techniques – multivariate analysis of variance (MANOVA) and structural equation modeling (SEM) – were used to further analyze the information quality, system quality, individual impacts, and organizational impacts. As before, the hub and spoke and centralized architectures were combined for data analysis purposes and the federated architecture was dropped from the analysis. Confirming the visual data analysis, it was found that the independent data marts architecture scores significantly lower than the others. When the bus architecture was compared to the hub and spoke/centralized architectures, no statistically significant differences were found, indicating that these architectures are equally successful.

Analysis of variance (ANOVA) was used to investigate the potential effect that the domain of the data warehouse implementation has on the success metrics. It was found that the time and cost to implement the first subject area(s) or business process(es) are significantly higher for larger domains. There were no statistically significant differences for the other success metrics based on domain.

Some of the survey respondents and experts commented on the success of the various architectures. Data can be flexible, scalable, integrated, and complete, but not accurate or consistent if the source systems are weak. Placing data in cubes makes the data easy to access, is intuitive for users, and improves query performance. When data is easily accessible, understandable, and provides high performance, users perceive the data to be consistent and well integrated. The various architectures may all score high in terms of meeting their objectives, but it should be kept in mind that the warehousing objectives may be different. To be successful, data warehousing projects should use a methodology that is complimentary to the architecture that is selected. Regardless of the underlying architecture, warehousing projects should always allow for something new to go into production in 60 to 120 calendar days.

Conclusion

Based on this research, an overall architecture selection model is proposed. It takes the various selection factors and organizes them into a causal-flow model. In this model, the need for information interdependence between organizational units and the nature of end user tasks combine to create the information requirements for the data warehouse. The information processing requirements and the source of sponsorship then combine to determine the view of the data warehouse. The perceived ability of the IT staff, the availability of resources, and the urgency of need for the data warehouse combine as facilitating conditions for the selection of a particular architecture. And finally, the view of the warehouse and the facilitating conditions influence the architecture selection decision.

Perhaps the most interesting study finding is how similar the bus, hub and spoke, and centralized architectures score on many of the metrics. It helps explain why these competing architectures have survived over time – they are equally successful for their intended purposes! No single architecture is dominant in terms of information and system quality and individual and organizational impacts. Perhaps the similarity of the success of the bus, hub and spoke, and centralized architectures should not be surprising. Much like the development methodologies have converged, so too have the architectures, in some ways.

Table of Contents

Introduction: Which Data Warehouse Architecture Is Best?

Over the past decade, companies have spent billions of dollars on data marts and warehouses. From their experiences, a substantial body of knowledge has been created. We know, for example, the importance of thoroughly understanding source systems before building, starting with only a few subject areas or business processes but having an enterprise-wide goal in mind, and giving end users data access tools and applications that are appropriate for their needs.

There is one area, however, that still causes considerable confusion and disagreement: *Which architecture to use?* There are multiple options. The most common is the hub and spoke architecture (i.e., centralized data warehouse with dependent data marts) that is advocated by Bill Inmon, who is commonly referred to as "the father of data warehousing." Inmon refers to this architecture as the Corporate Information Factory (CIF). Another prevalent choice is the data mart bus architecture with linked dimensional data marts (bus architecture), advocated by Ralph Kimball, the other preeminent figure in data warehousing. Each has strong proponents.

Considering the importance of the choice of architecture, there is surprisingly little research on the topic. The literature tends to either discuss the architectures, provide case study examples, or present survey data about the popularity of the various options. There is little rigorous, empirical research and this omission motivated our research.

Studying the Architectures

A three-phase study was conducted to provide answers to two research questions:

1. *What factors lead companies to select a particular architecture* and
2. *How successful are the various architectures?*

The answers to these questions are important to companies, vendors, and consultants.

The study's first phase identified the factors that potentially affect the selection of a data warehouse architecture and metrics to use in assessing the success of an architecture. The factors and metrics were chosen based on a review of the academic and data warehousing literature and interviews with 20 leading authorities in the field. These same sources were used in developing the survey instrument that was employed in the study's second phase. Appendix A lists the experts who participated. Ultimately, however, the researchers are responsible for the study and its findings and conclusions.

In the study's second phase, a Web-based survey instrument was used to collect data. It asked questions about the data warehouse in the respondent's company, the architecture that was implemented, factors that affected the selection of the architecture, the success of the architecture, the respondent's company, and the respondent. Four hundred and fifty four respondents provided information about their company's data warehousing initiative. Many individuals and organizations helped promote the study in emails, newsletters, announcements, and on websites. They are identified in Appendix B. Without their assistance, the number of respondents would have been much smaller. A

copy of the survey instrument is available at:
http://www.terry.uga.edu/~hwatson/DW_Architecture_Survey.doc.

In the study's third phase, the experts and selected survey respondents were contacted and asked to help interpret the survey data. They asked questions, raised possibilities, and provided examples that helped to understand and bring the survey data alive. Their input was very helpful.

The Five Architectures

The data warehousing literature provides discussions and examples of a variety of architectures. For our study, we investigated five: (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 the five that were studied.

Independent Data Marts

It is common for organizational units to develop their own data marts. These marts are independent of other marts, and while they may meet the needs for which they were created, they do not provide "a single version of the truth." 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. Figure 1 shows the architecture for independent data marts.

Figure 1. The Independent Data Marts Architecture

Data Mart Bus Architecture with Linked Dimensional Data Marts

A business requirements analysis for a specific business process such as orders, deliveries, customer calls, or billing is the foundation for this architecture. The first mart is built for a single business process using dimensions and measures (i.e., conformed dimensions and conformed facts) that will be used with other marts. Additional marts are developed using these conformed dimensions, which results in logically integrated marts and an enterprise view of the data. Atomic and summarized data are maintained in the marts and are organized in a star schema to provide a dimensional view of the data. This architecture is illustrated in Figure 2.

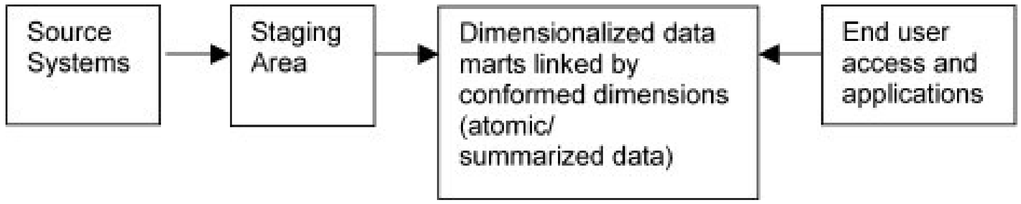

Figure 2. The Data Mart Bus Architecture with Linked Dimensional Data Marts

Hub and Spoke Architecture

An extensive enterprise-level analysis of data requirements provides the basis for this architecture. Attention is also focused on building a scalable and maintainable 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 maintained in the warehouse in 3rd normal form. Dependent data marts are created that source data from the warehouse. The dependent data marts may be developed for departmental, functional area, or special purposes (e.g., data mining) and may have normalized, denormalized, or summarized/atomic dimensional data structures based on user needs. Most users query the dependent data marts. Figure 3 shows this architecture.

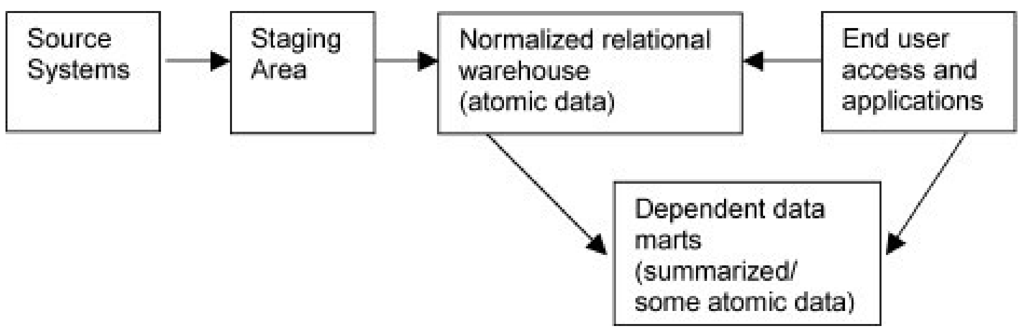

Figure 3. The Hub and Spoke Architecture

Centralized Data Warehouse (No Dependent Data Marts)

This architecture is similar to the hub and spoke architecture except that there are no dependent data marts. The warehouse contains atomic level data, some summarized data, and logical dimensional views of the data. Queries and applications access data from both the relational data and the dimensional views. This architecture is typically a logical rather than a physical implementation of the hub and spoke architecture; see Figure 4.

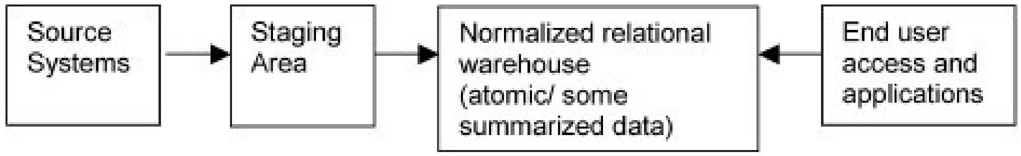

Figure 4. The Centralized Data Warehouse Architecture

Federated

This architecture leaves existing decision support structures (e.g., operational systems, data marts, and data warehouses) in place. Based on business requirements, data is accessed from these sources. The data is either logically or physically integrated using shared keys, global metadata, distributed queries, and other methods. This architecture is advocated as a practical solution for firms that have a preexisting, complex decision support environment and do not want to rebuild. This architecture is shown in Figure 5.

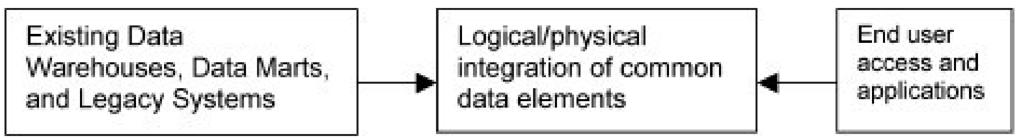

Figure 5. The Federated Architecture

Architecture Is Different than Methodology

It is important to recognize that data warehouse *architecture* identifies component parts, their characteristics, and the relationships among the parts, while *methodology* identifies the activities that have to be performed and their sequencing. Too often, the architecture and methodology terms are used interchangeably, which creates confusion. The architecture is the end product while a methodology is the process for developing an end product. But while architecture and methodology are different, they should be compatible. It is important to use a methodology that is consistent with the architecture that is being implemented.

Sometimes the hub and spoke architecture (e.g., Corporate Information Factory) is referred to as a *top down* approach and the bus architecture as *bottom up.* The reason for this is that the hub and spoke architecture places considerable emphasis on initially putting the infrastructure and processes in place to create an enterprise data warehouse and the bus architecture focuses on delivering a solution that addresses a current business need. These are methodologies rather than architectures because they describe development processes.

Over time, the top down and bottom up approaches have become increasingly similar. Advocates of the top down approach agree on the importance of developing incrementally and delivering early "wins." The bottom-up proponents recognize the importance of having an enterprise plan for integrating the incrementally developed data marts. As a result, the two methodologies are not as different as many people believe.

The Factors that Affect the Selection of an Architecture

No two organizations are the same, and consequently, companies may differ on their architecture selection decisions. There isn't a single architecture that is best for all situations and companies. If it were that simple, there wouldn't be disagreements over architecture selection.

From the literature and the experts, eleven factors were identified that potentially affect the architecture selection decision. Some of the factors relate to rational theory, such as the information processing theory of the firm, while others are based on social/political theories, such as power and politics. Below are the factors that were included in this study.

1. Information Interdependence between Organizational Units

There is a high level of information interdependence when the work of one organizational unit is dependent upon information from one or more other organizational units. In this situation, the ability to share consistent, integrated information is important. It is likely that firms with high information interdependence select an enterprise-wide architecture.

2. Upper Management's Information Needs

In order to carry out their job responsibilities, senior management often requires information from lower organizational levels. It may need to monitor progress on meeting company goals, drill down into areas of interest, aggregate lower-level data, and be confident that the company is in compliance with regulations such as the Sarbanes-Oxley Act. To the extent that this capability is important, so too is having an architecture that supports it.

3. Urgency of Need for a Data Warehouse

An organization can have an urgent need for a data warehouse (or a data mart) and the urgency of the business need may dictate a fast implementation. Some architectures are more quickly implemented than others, which can influence the architecture that is selected.

4. Nature of End User Tasks

Some users perform non-routine tasks. Structured queries and reports are insufficient for their needs. They have to analyze data in novel ways. These users require an architecture that provides enterprise-wide data that can be analyzed "on the fly" in creative ways.

5. Constraints on Resources

Some data warehouse architectures require more resources to develop and operate than others. As a result, the availability of IT personnel, business unit personnel, and monetary resources can impact the selection of the architecture.

6. View of the Data Warehouse Prior to Implementation

Organizations differ in their view or plans for the warehouse (or mart). Some may perceive it as part of their strategic plans while other organizations may not. As a result, it may be developed to provide a "point solution" to a particular business unit's need, it may be a decision support infrastructure project to support a range of applications, or it may be a critical enabler to support a company's strategic business objectives. Depending on the view of the warehouse, some architectures are more appropriate than others.

7. Expert Influence

When building a data warehouse, there are many places to turn for help – consultants, the literature, conferences and seminars, internal experts, and end users. To varying degrees, these sources can influence the architecture that is selected. For example, a consultant may recommend an architecture that he or she has successfully implemented in the past.

8. Compatibility with Existing Systems

There are many benefits to implementing IT solutions that are compatible with the existing computing environment. Consequently, the selection of a data warehouse architecture is likely to be impacted by the systems and technologies that are already in place. This may include compatibility with source systems, metadata integration, data access tools, and technology vendors.

9. The Perceived Ability of the In-house IT Staff

The building of a data warehouse can be a daunting task and implementing some data warehouse architectures may be perceived as being more challenging than others, depending on the internal IT staff's technical skills, successful experiences with similar projects, and level of confidence. Consequently, the IT staff may chose an architecture that is compatible with what they think can be successfully built.

10. Source of Sponsorship

The source of sponsorship for a data warehouse may vary from a single department or business unit to the top management (i.e., CXO) within an organization. The sponsor can influence and may control many aspects of the data warehousing initiative, such as monetary resources and the architecture selected. For instance, sponsorship from a business unit may steer an organization to select a data warehouse architecture that provides more control to the business unit, such as a data mart.

11. Technical Issues

A variety of technical considerations can affect the choice of an architecture – the ability to integrate metadata; scalability in terms of the number of users, volume of data, and query performance; the ability to maintain historical data; and the ability to adapt to technical changes, such as in source systems. Depending on the importance of these technical issues, some architectures may be better than others.

Research Model for Research Question #1

Eleven factors were identified as potentially affecting the selection of an architecture. The research model that relates the factors to the architectures is shown in Figure 6.

Architecture Selection Factors

Information interdependence
Upper management's information needs
Urgency of need
Nature of end user tasks
Resource availability
View of the data warehouse
Expert influence
Compatibility with existing systems
Perceived ability of the in-house IT staff
Source of sponsorship
Technical issues

$\Rightarrow$

Data Warehouse Architectures

Independent data marts
Data mart bus architecture
Hub and spoke
Centralized data warehouse
Federated

Figure 6. The Research Model that Relates the Selection Factors to the Architectures

The Metrics for Assessing Architecture Success

Based on the literature and input from the experts, a variety of success metrics were identified. Some of them relate to information and system quality, such as data consistency and the ability to integrate data. Others relate to project management measures, such as whether the implementation was on budget and on schedule. Still others assess the impact on individuals and the organization, such as whether the warehouse is easy and intuitive to use and whether the warehouse has generated high, quantifiable ROI. The major success metric categories are identified and discussed below.

1. Information Quality

Information quality includes the following measures – information accuracy, information completeness, and information consistency.

Information Accuracy

Warehouse data should be as accurate as its intended use demands. Queries and reports should contain few errors because of data problems. Real-world objects and events should be correctly described.

Information Completeness

Over time, a warehouse should provide all (or nearly all) the decision support data that is needed. It should contain data for all of the required business processes and subject areas. It should provide the data that is needed by users and applications.

Information Consistency

A major reason for building a data warehouse is to create a "single version of the truth." It should eliminate the problem of having inconsistent data. The data warehouse should provide a single system of record for the organization.

2. System Quality

System quality includes three measures – system flexibility, system scalability, and system integration.

System Flexibility

Data warehouses should be flexible. It should be easy to add new business processes and subject areas. The warehouses should be able to adapt to new requirements quickly. They should be able to easily support future application needs.

System Scalability

The demands on data warehouses grow over time, so they must be scalable. They should be able to handle increases in the number of users, the complexity and number of queries, and the volume of data without negatively affecting system performance.

System Integration

Providing integrated data is an important requirement for a data warehouse. Using appropriate primary keys, a warehouse should integrate data from multiple sources, including both internal and external data.

3. Individual Impacts

By itself, a data warehouse does not create value. Value creation occurs when users employ the warehouse in their work. Users should be able to quickly and easily access data. They should be able to think about, ask questions, and explore issues in ways that were not previously possible. Overall, the warehouse should improve users' decision-making capabilities.

4. Organizational Impacts

Ultimately, the warehouse should have positive impacts on the organization. It should satisfy the business requirements for which it was built, facilitate the use of BI, support the accomplishment of strategic business objectives, enable improvements in business processes, lead to high, quantifiable ROI, and improve communications and cooperation across organizational units.

5. Development Time

A data warehouse should be developed in a timely manner to meet business needs. The time to rollout the first business process(es) or subject area(s) should be timely and on or ahead of schedule.

6. Development Cost

An organization's expenditure for the data warehouse should meet budgetary constraints for the project. The cost at key milestones during the development process, such as the cost to rollout the first business process(es) or subject area(s) and the annual cost to maintain the architecture, should be reasonable and at or below the budgeted amount.

The measures for development time and cost must be interpreted by considering the domain for which the data warehouse is implemented. An implementation in a large domain, such as the entire organization, typically requires more time and monetary resources than a warehouse implemented in a single business unit.

Research Model for Research Question #2

System quality, information quality, individual impacts, organizational impacts, development time, and development cost were used as metrics for assessing the success of the five architectures. The research model that relates these factors to the architectures is shown in Figure 7.

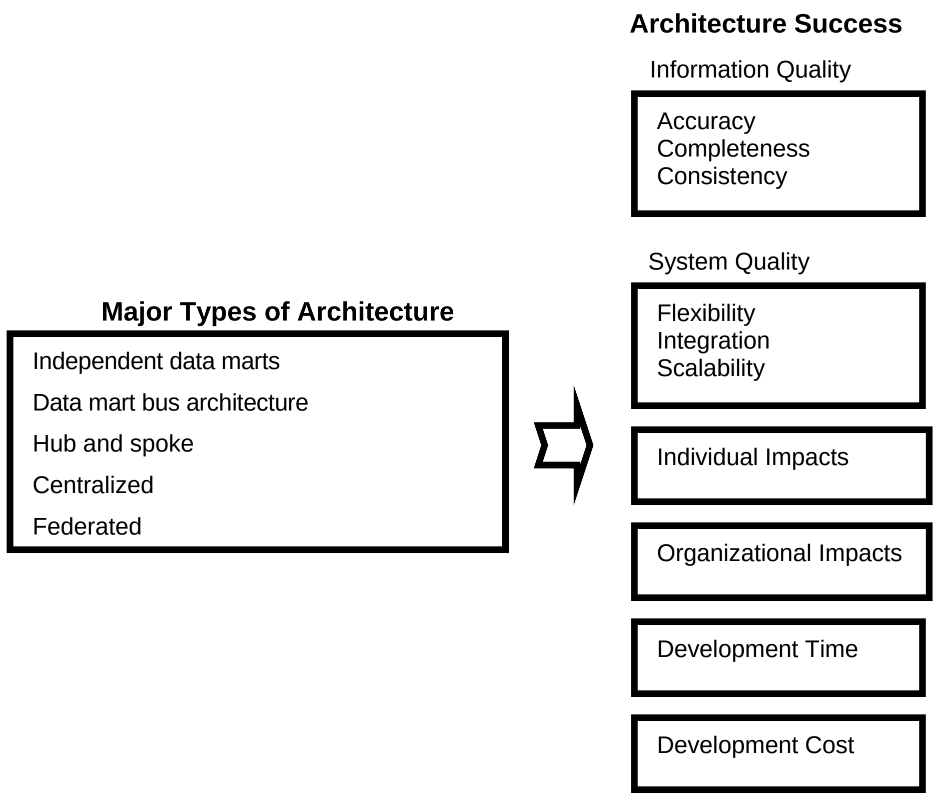

Figure 7. The Research Model that Relates the Success Metrics to the Architectures

The Study Findings

Data Collection

Four hundred and fifty four respondents provided usable survey data. Some respondents did not answer all of the questions; consequently, the sample size varied across the questions.

Some Caveats

We cannot claim that the data is random across the entire population of architectures. To assist in promoting the study, we turned to people and organizations with a strong interest in the topic (e.g., TDWI, *DM Review*) and some who advocate a particular architecture (e.g., Inmon Associates, Kimball Group). Because of the latter source of promotion, the percentages for each type of architecture may not represent the actual percentages for the population as a whole (some promotional efforts may have been more effective than others). However, the percentages for each kind of architecture are consistent with other studies, thus providing some evidence that the data collected is appropriate for the study's purposes.

The responses to the architecture selection and success measures are believed to be random within each kind of architecture. That is to say, there is no reason to think that only companies that have done especially well or poorly with their data warehousing initiatives completed the survey or that companies that used a particular architecture were more or less likely to participate in the study based on how well their data warehouse has done.

It should also be noted that there was the potential for bias in the responses to the study's success measures. In most cases, the survey respondents were associated with the development of the warehouses.   They may perceive the warehouse as being more successful than others in their organizations, such as users and management. To check on this potential source of bias, we asked the survey respondents to identify warehouse users and managers in their organizations. Some of them were contacted and asked to answer the success questions. Their responses were not significantly different from those of the warehouse developers. Consequently, we believe that the survey data for the success measures is unbiased.

The success metrics must be interpreted with caution because there are factors other than the architectures used that can affect success. For example, the quality of data in source systems can affect information quality. The data access tools and applications provided to users can affect the individual success factors. In addition, the degree of congruence between business and warehouse strategy can impact the organizational success metrics.

The study is *descriptive* rather than *normative*. The findings describe current practices rather than prescribing what practices should be. For example, the findings about the architecture selection factors do not provide a roadmap for selecting the best architecture.

They only provide insights on the importance of the factors and how the factors interact to help predict which architecture a company will choose.

Demographics

The Respondents

Figure 8 shows the position of the respondents. Almost 65 percent of the respondents were involved in the selection of the architecture. The respondents are relatively evenly distributed over data warehouse managers, data warehouse staff members, IS managers/professionals, and independent consultants/system integrators. Consultants and system integrators were asked to complete the survey with a particular client in mind.

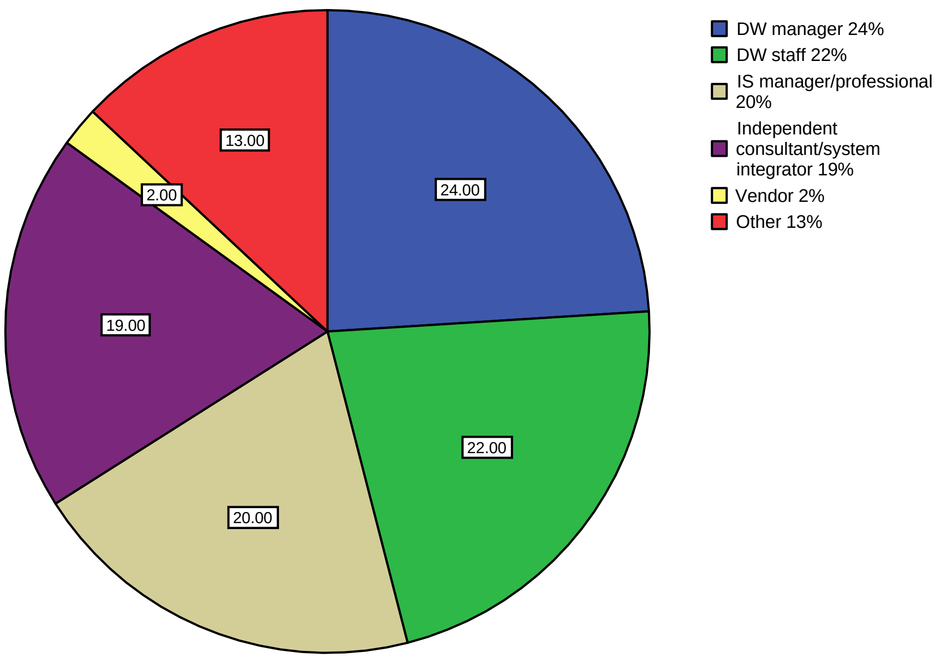

Figure 8.   Percentage of Respondents by Position (based on 454 responses)

The Companies

Close to 58 percent of the companies have revenues in excess of $750 million. As is seen in Figures 9 and 10, most of the companies are located in the United States (60%) and represent a variety of industries, with the financial services industry leading the way with 15 percent.

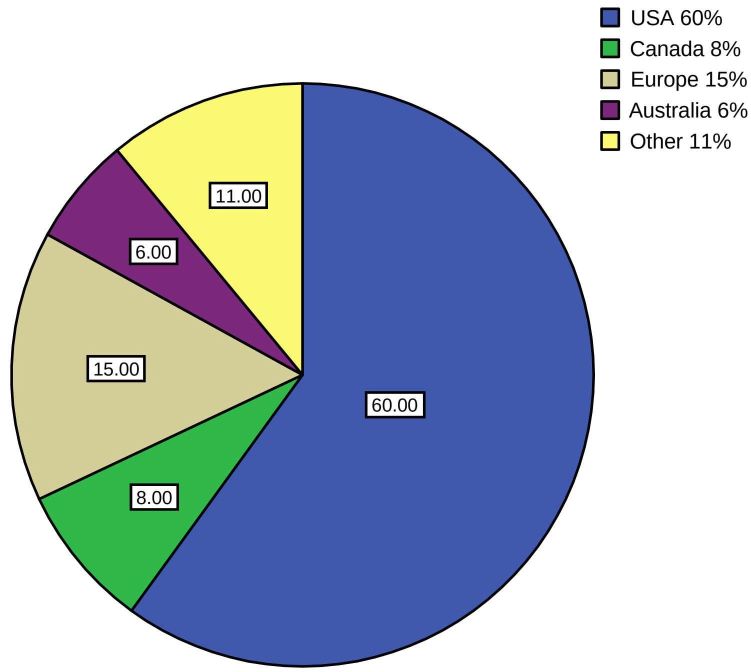

Figure 9. The Percentage of Companies by Country (based on 446 responses)

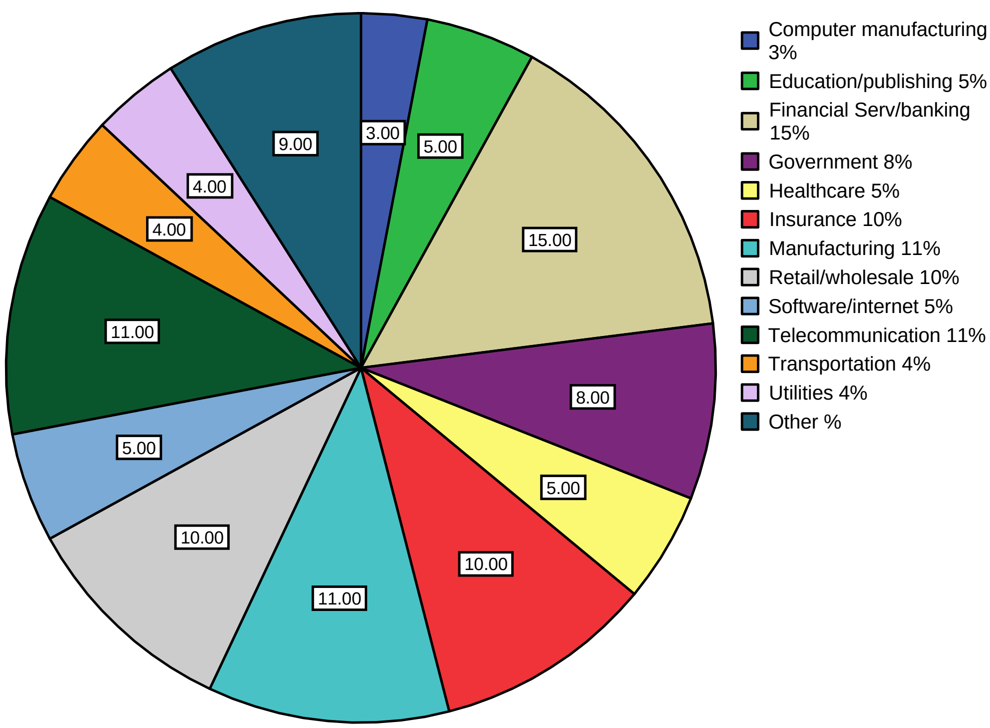

Figure 10. The Percentage of Companies by Industry (based on 393 responses)

The Architectures

Figure 11 shows the percentages of companies that are using the various architectures. The most predominant is the hub and spoke, with 39 percent, followed by the bus architecture with 26 percent. Slightly over 17 percent of the companies have implemented a centralized data warehouse. Only a little over 12 percent of the companies report having independent data marts as their architecture, but this may underestimate the actual percentage in the real world population as a whole since independent data marts were not the focus of the study. The number of responses for independent data marts, however, is sufficiently high to allow meaningful comparisons with the other architectures. Very few companies report having a federated architecture (4%), and because of the small number of respondents, any comparisons with the other architectures must be done carefully because of the small sample size.

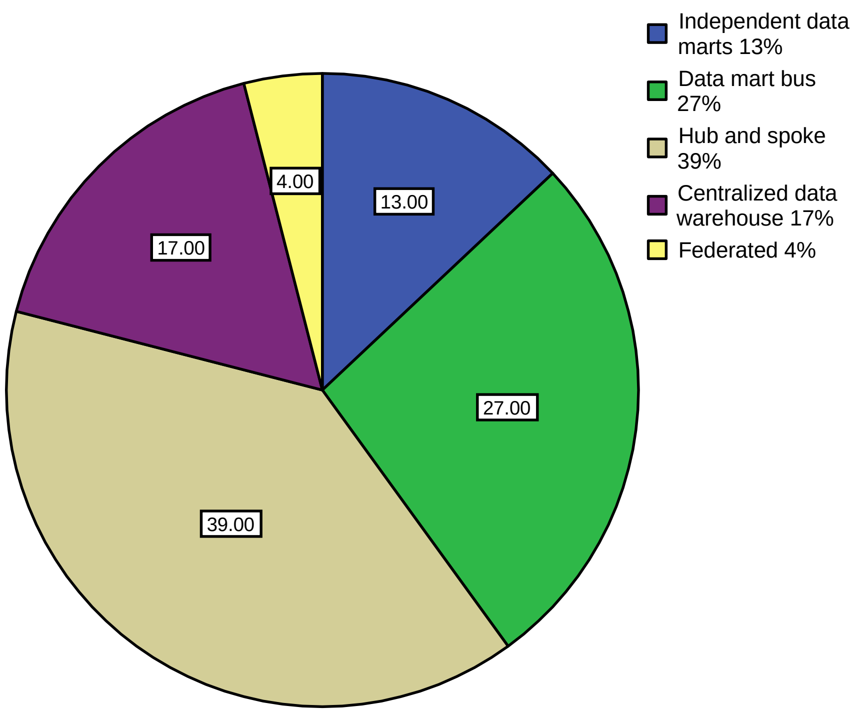

Figure 11. The Distribution of the Architectures (based on 454 responses)

The Match between the Actual and Reference Architectures

In addition to asking which of the reference architectures (e.g., bus, centralized) was most like the respondent's actual architecture, the respondents were asked to indicate on a 1 to 7 scale how closely the reference architecture matched their actual architecture. A 1 was *not a match* and a 7 was *an exact match*. Table 1 shows the averages for the five reference architectures.

Architecture	Average Match
Independent data marts	4.9
Bus architecture	5.6
Hub and spoke	5.6
Centralized	5.6
Federated	4.6

Table 1. How Closely the Actual Architectures Match the Reference Architectures (based on 454 responses)

Overall, the matches are close, especially for the bus, hub and spoke, and centralized architectures. This close match is important to the validity of the analyses that are based on the architecture used. The closeness of the matches is also somewhat surprising. As mentioned previously, there are many variations on the five architectures included in this study. Also, people often believe that "our architecture is different."

The federated and independent data marts architectures match the reference architectures the least. Why might this be the case? The bus, hub and spoke, and centralized architectures are widely written about, documented, and promoted by various vendors and consultants. Many companies have followed the architectural guidelines that are readily available, resulting in a close match. However, there are fewer models and advocates for independent data marts and the federated architecture. For various reasons companies may have them, at least as an interim solution; but they come in many "flavors," and do not match their reference architectures as closely as the others.

The Platforms

Figure 12 shows the platforms that host the data warehouses. Oracle is the clear leader with 41 percent, followed by Microsoft (19%) and IBM (18%).

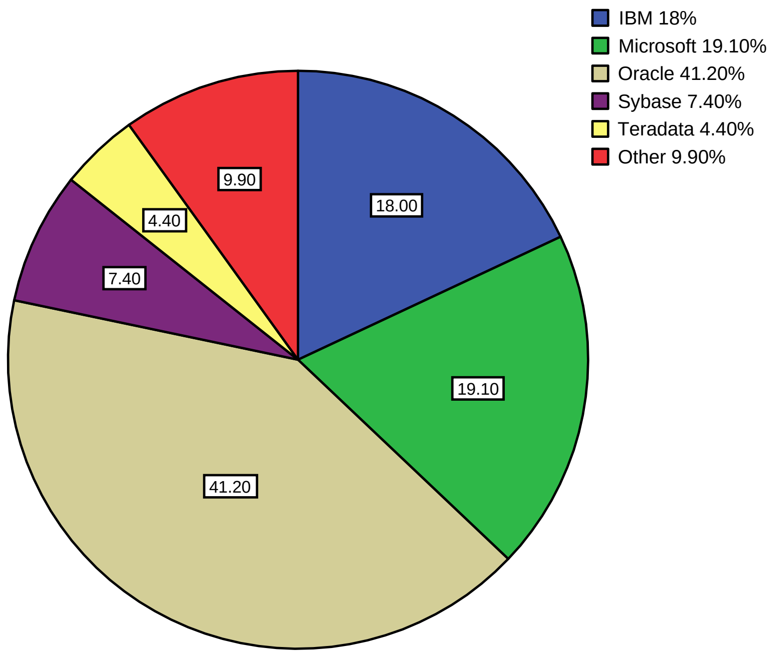

Figure 12. The Percentage of Platforms that Host the Data Warehouses (based on 454 responses)

Company Revenue and the Age of the Data Warehouse by Architecture

Table 2 shows how the various architectures vary in terms of average (i.e., mean) company revenue and age of the warehousing environment.

Architecture	**Average Gross Revenue**	**Average Age**
Independent data marts	3.7 billion	39 months
Bus architecture	4 billion	41 months
Hub and spoke	5 billion	56 months
Centralized	4.6 billion	49 months
Federated	6 billion	47 months

Table 2. How the Architectures Vary by Company Revenue and Age (based on 454 responses)

It is interesting that companies that have the highest average revenue use the federated architecture. A possible explanation is that these companies have gotten large through mergers and acquisitions and have inherited a disparate set of decision support platforms and use a federated architecture to tie the various platforms together.

Companies that are relatively large also use the hub and spoke architecture. The warehouses with this architecture also tend to be older. Because the hub and spoke architecture typically tends to be the most expensive and time consuming to build because of the many component parts, it is to be expected that the hub and spoke architecture tends to be used by companies with high revenues and a mature data warehousing initiative.

The Domain

Data warehouses vary in their domain (i.e., scope or coverage) ranging from departmental to enterprise-wide. As can be seen in Figure 13, most of the warehouses support either several business units (38%) or the entire company (35.5%).[1] Fewer than 12 percent of the data warehouses are for a single functional area or sub unit.

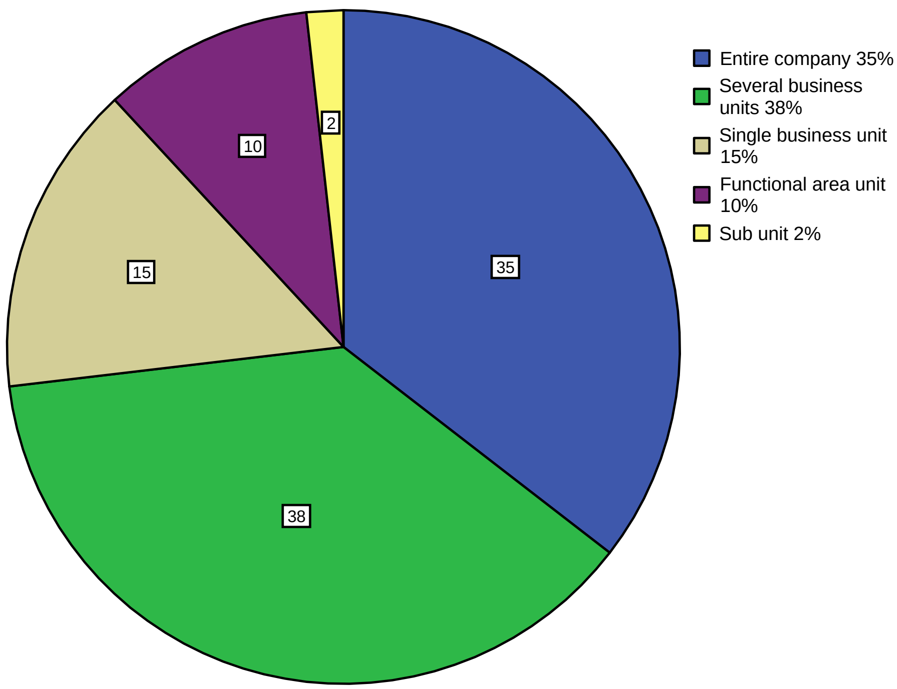

Figure 13. The Domain Percent of the Data Warehouses (based on 454 responses)

[1] The statistical package used to generate the pie chart rounds numbers down when 0.5.

The Domains Supported by the Various the Architectures

There are interesting differences in the scope or domain for the various architectures; see Table 3. When the entire company is the domain, the centralized (43.0%) and hub and spoke (41.8%) architectures are the most popular. When the focus turns to a single or several business units, the bus architecture is most common. This supports the common perception that the hub and spoke and centralized architectures tend to be used more with enterprise-wide implementations.

Domain	**Independent Data Marts**	**Bus Architecture**	**Hub and Spoke**	**Centralized (no dependent data marts)**	**Federated**
Entire Company	17.5	30.8	41.8	43.0	26.3
Several Business Units	40.4	40.0	37.3	32.9	36.8
Single Business Unit	17.5	20.0	13.6	8.9	15.8
Functional Area Unit	22.8	7.5	6.8	12.7	10.5
Sub Unit	1.8	1.7	0.6	2.5	10.5

Table 3. The Domain Percent of the Data Warehouses by Architecture (based on 454 responses)

Data Warehouse Size by Domain and Architecture

The size (in gigabytes) of the various warehouses is shown in Table 4. The largest warehouses are associated with the hub and spoke architecture and this is true across all domains.

Domain	Independent Data Marts	Bus Architecture	Hub and Spoke	Centralized (no dependent data marts)	Federated
Entire Company	738	2,571	3,430	2,178	830
Several Business Units	2,164	1,603	2,447	2,392	1,700
Single Business Unit	275	1,199	2,440	1,040	3,938
Functional Area Unit	450	136	1,723	531	113
Sub Unit	51	26	0	750	463

Table 4. The Size (in gigabytes) of the Data Warehouses by Domain and Architecture (based on 454 responses)

Getting It Right

Companies do not always stay with their initial architecture. It may prove to be inappropriate for the company's needs or the company's needs may change over time, resulting in an architecture change. One third of the companies in our study have switched architectures.

Of the companies that have switched, over 44 percent of them went to the hub and spoke architecture. Most of them switched from independent data marts (32%), the bus architecture (29%), or a centralized data warehouse (27%).

About a quarter of the companies switched to the bus architecture. Forty percent of them moved from independent data marts, 30 percent from a hub and spoke architecture, and 22 percent from a centralized architecture.

Fewer companies switched to the other architectures: 13 percent to a centralized data warehouse, 10 percent to independent data marts, and 7 percent to a federated architecture.

It is interesting that the percentage of companies that switched to a hub and spoke from the bus architecture (29%) is essentially the same that switched in the other direction (30%). This is an interesting finding because the supporters of the different architectures often talk about the failures of the other.

Factors that Affect the Selection of an Architecture

The survey respondents answered: *Please indicate the importance of each of the following factors on the selection of your data warehouse architecture*. A seven-point scale was used for the responses, with 1 being *not importan*t and 7 being *very important*. The importance factors were described as:

1. ***Information interdependence between organizational units*:** The need to share information among organizational units.
2. ***Upper management's information needs:*** Upper management's needs for information from lower organizational levels.
3. ***Urgency of need for a data warehouse:*** The extent to which there was an urgent need to build the data warehouse.
4. ***Nature of end user tasks***: The extent to which users' jobs required non-routine data analyses.
5. ***Constraints on resources:*** The availability of resources (IT personnel, business unit personnel, and monetary resources) for building the data warehouse.
6. ***Strategic view of the warehouse prior to implementation*:** The extent to which implementing a data warehouse was viewed as important to supporting strategic objectives.
7. ***Compatibility with existing systems:*** The extent to which the data warehouse architecture was compatible with existing systems.
8. ***Perceived ability of the in-house IT staff:*** The perceived ability of the in-house IT staff in terms technical skills, experiences, and confidence in developing a data warehouse.
9. ***Technical issues:*** The extent to which technical issues affected the data warehouse architecture.
10. ***Expert influence:*** The influence from sources of data warehouse expertise.

Source of sponsorship was not included in the list because it is a categorical rather than a continuous variable, and consequently, could not be assessed on a continuous scale. Figure 14 shows the average scores for the ranked selection factors across all the architectures.

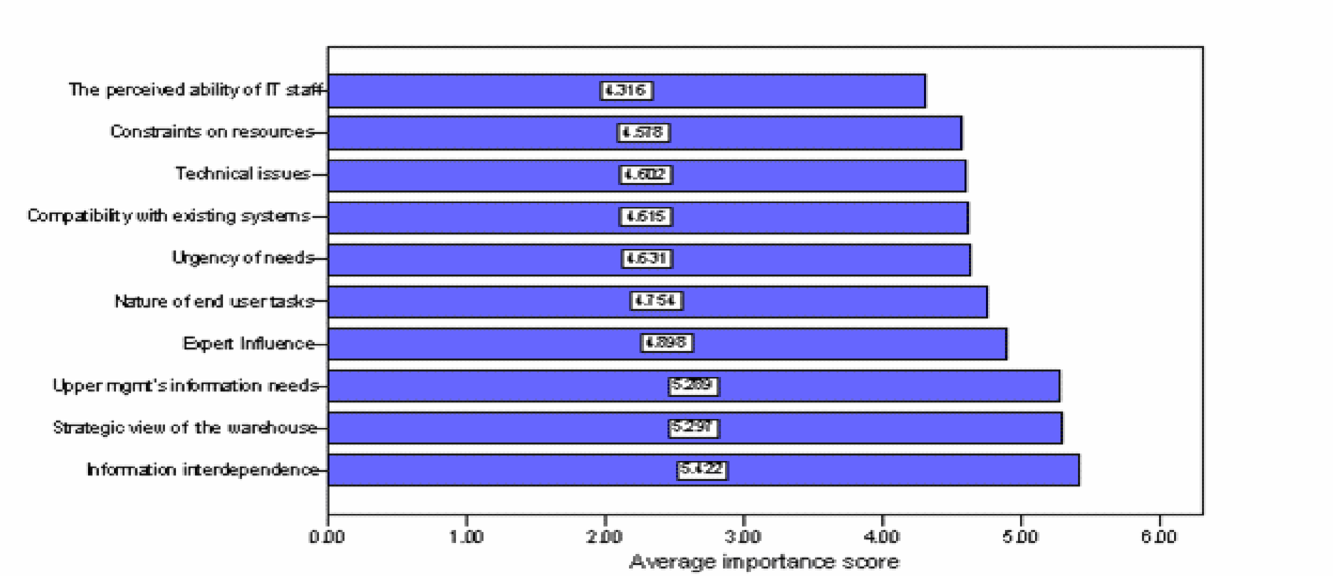

Figure 14. The Importance of the Selection Factors – All Architectures (based on 369 responses)

The data reveals that all of the selection factors have some influence. The lowest average score is over 4.3 (for the perceived ability of the in-house IT staff), thus indicating that even the lowest rated factor is important. The most important factors (with average scores over 5.0) are information interdependence between organizational units, the strategic view of the warehouse prior to implementation, and upper management's information needs. All of these are rational factors, suggesting that optimizing the architecture selection decision is of paramount importance.

Table 5 drills into the data further and provides the average score for every factor and architecture. While it is risky to speculate about the meanings based on just this data, none-the-less, there are some possible interpretations.

Selection Factor	Independent Data Marts	Bus Architecture	Hub and Spoke	Centralized (no dependent data marts)	Federated
Information Interdependence between Organizational Units	4.70	5.40	5.60	5.51	5.13
Upper Management's Information Needs	5.15	5.19	5.39	5.33	5.13
Urgency of Need for a Data Warehouse	4.50	4.85	4.47	4.82	4.73
Nature of End User Tasks	4.30	4.90	4.73	4.85	5.00
Constraints on Resources	4.97	4.30	4.50	5.06	4.46
Strategic View of the Data Warehouse Prior to Implementation	4.85	5.09	5.46	5.70	4.93
Compatibility with Existing Systems	4.55	4.35	4.75	4.75	4.53
Perceived Ability of the In-house IT Staff	4.25	4.37	4.18	4.59	4.20
Technical Issues	4.12	4.56	4.59	5.14	3.93
Expert Influence	4.40	5.05	5.06	4.75	4.46

Table 5. The Importance of the Selection Factors – For Every Architecture (based on 369 responses)

In general, the selection factors for the independent data marts receive lower average scores than the other architectures. This suggests that independent data marts are employed more by happenstance than the others. This architecture is often the consequence of a series of independent decisions rather than an overall plan.

The one notable exception (where the independent data marts architecture scored relatively high) is constraints on resources. A likely explanation is a lack of resources prevents some organizations from implementing a better architectural solution. Of course, the independent data mart architecture has its own costs – missed business opportunities, the need to support multiple decision-support platforms, etc.

Despite the arguments over the merits of the bus versus the hub and spoke and centralized architectures, the scores for the architecture selection factors are similar for most factors. Apparently, companies focus on many of the same factors, but depending on their company situation relative to the factors, arrive at different architecture decisions.

For three selection factors, the bus architecture averages at least 0.2 lower than either the hub and spoke or centralized architectures: constraints on resources, strategic view of the data warehouse prior to implementation, and compatibility with existing systems. A possible interpretation of these lower scores is that the bus architecture is sometimes selected because the availability of resources is less of an issue (perhaps being sufficient to meet the needs of the data warehouse initiative), the view of the warehouse is less strategic, and there are fewer concerns about being compatible with existing systems.

The centralized architecture scores higher than the hub and spoke architecture (by at least 0.2) on urgency of need (indicating a need for a relatively fast implementation), constraints on resources (the solution had to require fewer resources), the strategic view of the warehouse prior to implementation (the architecture had to support strategic company objectives), and the perceived ability of the in-house IT staff (suggesting confidence in being able to successfully implement the architecture). The hub and spoke architecture averages higher on technical issues (indicating that technical issues were a concern) and expert influence (suggesting that sources of expertise advocated the use of this architecture).

The relatively small number of companies with a federated architecture (n=15) makes it difficult to generalize. While recognizing this limitation, it is interesting to note that the score on technical issues is lower than for any of the other architectures. The highest scores are for information interdependence between organizational units and upper management's information needs. The IT staff may have been told to coble data together from various systems to meet senior management's information needs and not be concerned with a technically elegant solution.

Advanced Analyses

In addition to the use of descriptive statistics to analyze the selection factors, advanced, more powerful statistical methods were used. These methods helped to identify why particular architectures were selected.

Prior to performing the advanced analyses, the five architectures were reduced to three in number. This allowed us to focus attention on those that are most fundamentally different and are most often used. The hub and spoke and centralized architectures were combined into a single category. The rationale was that these architectures are similar, except for the inclusion of physical data marts in the hub and spoke architecture. Second, the federated architecture was dropped from the analysis because of the small number of firms using it (at least in our survey).

Stepwise binary logistic regression analysis was used to investigate why a particular architecture was selected. Logistic regression is similar to regression analysis, except that it is used when the dependent variable is categorical rather than continuous. In our study, the architecture selected (a categorical variable) was the dependent variable and the selection factors were the independent variables. The data for the selection factors came from the survey questions that assessed the organizational conditions in the firm at the time that the architecture selection decision was made.

With stepwise regression, independent variables are entered into the model only if they are statistically significant. In other words, only independent variables that help predict the dependent variable are included. In our study, the independent variables (the selection factors) help predict why a particular architecture was selected over the others. An alpha level of 0.20 was set for the independent variables to enter the model, which is the recommended significance level for stepwise binary logistic regression. Those variables that entered can be deemed to be statistically significant in predicting the particular architecture.

Independent Data Marts Versus the Others

In the first data analysis, those companies that selected independent data marts rather than any of the other architectures were analyzed. Three factors were found to influence the selection decision: constraints on resources, the view of the data warehouse prior to implementation, and the perceived ability of the in-house IT staff. When there are constraints on resources, the view of the warehouse is limited in scope (e.g., a subunit solution), and the perceived IT skills in-house are low, conditions exist for selecting the independent data mart architecture.

Bus Architecture Versus the Others

Next, the bus architecture was compared to all the others. Four factors were found to be statistically significant to the selection of the bus architecture: information interdependence between organizational units, urgency of need for a data warehouse, constraints on resources, and source of sponsorship. When there is a high need to share and integrate information across organizational units, an urgent need for the data warehouse, constraints on the availability of resources exist, and sponsorship at high organizational levels, the bus architecture is an attractive choice.

Hub and Spoke/Centralized Versus the Others

Another comparison was the hub and spoke/centralized against all the others. Information flow between organizational units, the view of the data warehouse prior to implementation, and the perceived ability of the in-house IT staff, all significantly influenced the selection of the hub and spoke/centralized architecture. When there is a high need for information integration among organizational units, the warehouse is viewed as being strategic, and the perceived ability of the in-house IT staff is high, the hub and spoke/centralized warehouse is a common choice. This is consistent with conventional wisdom that this architecture requires high level of skill to implement but provides the organization-wide infrastructure needed to support strategic objectives.

Bus Architecture Versus Hub and Spoke/Centralized

The final comparison was the bus architecture against the hub and spoke/centralized. In some ways, this is the most interesting comparison because of the on-going discussion as to which of the two is the best. Three factors were found to impact the selection decision: information flow between organizational units, the urgency of need for a data warehouse, and the view of the warehouse prior to implementation. When the need for information flow between organizational units is high, companies tend to select the bus architecture. This can perhaps be explained by the bus architecture's emphasis on using conformed dimensions, which allows data to be joined across different processes, subject areas, and consequently, organizational units. The hub and spoke/centralized architecture also supports information flow across organizational units, but is perhaps less in people's minds when the architecture selection decision is made. When the need for the data warehouse is more urgent, the bus architecture tends to be chosen. This finding is not surprising because of the shorter development time for the bus architecture. And finally, when the view of the warehouse is more strategic (more of an enterprise-wide implementation), the hub and spoke/centralized architecture is often selected. This finding is consistent with the finding that the hub and spoke/centralized architecture tends to have a wider domain of implementation in organizations.

The Impact of Domain

In order to explore the effect of domain on the architecture selection decision, the data was grouped and analyzed on the basis of domain. The first group included those respondents who indicated that the domain for their architecture was either the entire company or several business units. These are the largest domains. In the second group were those firms with either a functional area unit or a sub unit as the domain. These are the smallest domains. The middle domain, a single business unit, was not included in the analysis in order to better identify the differences between the largest and smallest domains.

Analysis of variance (ANOVA), a multivariate analysis technique that detects differences among groups, was used to examine which selection factors are statistically different for the large and small domains. Using the recommended alpha level of 0.05 as the significance level, one factor, information interdependence between organizational units, was found to differ between the large and small domains. More specifically, information interdependence is more important as a selection factor for large domains. This finding is consistent with the expectation that organizations that have a large domain need an architecture that supports information sharing across the organization.

Company Experiences and Expert Commentary

The aggregated data provides an overall understanding of the factors that affect the architecture selection decision. However, every company's story is different and interesting. Presented below are brief, high-level descriptions of how some of the companies arrived at their current architectures. Also provided are interesting and insightful comments by the survey respondents and experts.

A **data warehouse manager** describes his experience at the company where he used to work. "We went for a bus architecture for cost reasons. In the current economic environment, it is difficult to get funding for long-term, expensive projects. With the bus architecture, we were able to build the warehouse by budgeting for it in small chunks."

The **data warehouse manager** at one company described how organizational factors influenced their data warehouse implementation. According to the manager, "The business units needed to get something operational quickly but political issues involving some of the business units made the implementation more challenging. The bus architecture allowed us to have core dimensions that were conformed across the organization and some dimensions that were unique to each business unit."

Kathleen Spracklen, CIO at Petersen-Arne, was strongly influenced in college on the merits of E-R data modeling techniques and normalization principles. According to Kathleen, "In building the warehouse, I honored sound data design methodology. All of the data is maintained in a centralized architecture in 3rd normal form and users are given logical views of the data."

An **IT manager** at one company was deeply involved in his company's initial data warehousing activities in the 1980's, but then moved on to other assignments. About three years ago, he was asked to perform a comprehensive audit of their current data warehouse implementation. According to him, "Our initial architecture was a hybrid, but was most similar to what is now recognized as a Corporate Information Factory." As he investigated contemporary architecture alternatives, he was "heavily influenced by both Inmon's books and reports from the Gartner Group - which reinforce the Inmon-style architecture for a central data warehouse." His company now has a hub and spoke architecture, with some logical marts and one physical mart that supports applications with over 1,000 users and high computational requirements. Interestingly, late in the project to update their architecture, a junior member of the team reviewed Kimball's books and advocated the bus architecture, which led to considerable discussion about the alternatives. In the end, the team consensus was "an Inmon approach for the central data warehouse, with a dimensional approach endorsed for data marts."

According to data warehouse architect **Krishna Mohan** of the Commonwealth Office of Technology for the State of Kentucky, "We initially had a centralized architecture that stored all of the data. We had a difficult time with it for data separation, security, and performance reasons. With that experience and advances with the Corporate Information Factory, we chose the hub and spoke architecture for our next warehouse."

Julie Mae Longgood is Vice President, Manager of Internal Support, at Banner Bank. Julie Mae said, “We needed an architecture that supported a high level of data accuracy, consistency, and control. The new system had to provide a viable alternative to Excel spreadsheets. A high level of control on financial reports and compliance with Sarbanes-Oxley was critical. The hub and spoke architecture was judged to provide the best solution for our needs.”

Claudia Imhoff commented on independent data marts. “Independent data marts are not a long-term solution. Cost is one of their many problems. The costs of the marts are additive with no economies of scale. Interestingly, every company that I go into has one or more independent data marts, even if they have some other architecture for most of their data. Nothing is pristine in terms of architecture.”

Rob Armstrong discussed why some companies have independent data marts: “They may be deployed due to short sightedness or the ‘down scoping’ of deliverables. The IT teams are looking for explicit requirements in order to have a solid objective. This means that the requirements are ‘dumbed down’ to a point where simple reports and meeting departmental needs are the only true deliverables.” Rob also discussed how some companies move to the centralized architecture: “Some companies start with independent data marts and decide to move to a centralized architecture. This can be a long journey, so in the interim, you get a hub and spoke as the marts are absorbed. And some of the marts may stay in place for a long time for political reasons.”

Tim Feetham commented on the federated architecture: “When a company gets so large and is acquiring other companies at great speed, the only option they may have is a federated architecture, at least as an interim solution.” He also discussed the tradeoffs between the bus and hub and spoke architectures: "Users are likely to find that the star schema, on which the bus architecture is based, is easy to understand and provides relatively good performance. In addition, all of the data stored within a bus-based data warehouse is intended to be directly accessible by the user. However, this may lead the designer to make trade offs for performance sake between including everything in the source system that might be used in the future, and including only those items specifically identified by the user. The designer using a hub and spoke architecture can bring all relevant data that might be used in the future into the hub and then pipe those data that are immediately needed out to the spokes or dependent data marts. This gives the hub and spoke designer more latitude to meet potential future needs of the business, especially in large and complex environments."

Jim Thomann discussed switching architectures***:*** “It’s hard to pick the right architecture initially. What is intellectually appealing doesn’t always fit the realities of the company’s situation and this can lead to a switch to a new architecture.”

The Success of the Architectures

Multiple metrics were used to assess the success of the various architectures: system quality, information quality, individual impacts, organizational impacts, development time, and development cost.

Information Quality

Accuracy, completeness, and consistency were used to assess information quality. Table 6 shows the average scores for these measures across the various architectures.

Information Quality Metric	Independent Data Marts	Bus Architecture	Hub and Spoke	Centralized (no dependent data marts)	Federated
Accuracy	4.97	5.59	5.45	5.46	5.23
Completeness	3.78	4.29	4.92	4.80	3.74
Consistency	4.52	5.61	5.69	5.43	5.23

Table 6. Information Quality – For Every Architecture (based on 454 responses)

Independent data marts score relatively poorly in all aspects of information quality. With only a single exception, its scores are the lowest on all of the metrics. It rates very low on completeness, which indicates that the independent data mart architecture is especially weak in terms of providing all of the data needed for decision support.

For most metrics, the bus, hub and spoke, and centralized architectures have similar scores. The hub and spoke and centralized architectures do score higher on completeness, however, suggesting that they are used to provide a more comprehensive source of decision support data.

While not as low as for the independent data marts, the scores for the federated architecture are also relatively low. This is especially the case in terms of completeness. This is not surprising since a federated architecture is often used to quickly bring together some, but quite likely not all, of the data that might be useful.

It is interesting to note that the scores for completeness are lower than for accuracy and consistency for all of the architectures. Regardless of the architecture selected, companies are apparently doing a better job putting accurate, consistent data in their warehouses than including all of the possible relevant decision support data.

System Quality

Flexibility, scalability, and integration were used to measure system quality. The average scores for these measures across all architectures are shown in Table 7.

System Quality Metric	Independent Data Marts	Bus Architecture	Hub and Spoke	Centralized (no dependent data marts)	Federated
Flexibility	4.40	5.63	5.44	5.23	4.54
Scalability	4.70	5.42	5.38	5.50	4.60
Integration	4.67	5.77	5.87	5.50	4.93

Table 7. System Quality – For Every Architecture (based on 454 responses)

The averages for system quality follow the same overall pattern as for information quality. Independent data marts score lower on almost every metric. The scores for the federated architecture are also consistently lower than for the other three architectures.

The average scores for the bus architecture, hub and spoke architecture, and centralized warehouse are similar in most cases. The bus architecture scores slightly higher, however, on flexibility. The questions on flexibility included the ability to easily add new business processes and subject areas, satisfy new information requirements easily, and easily support new application needs.

In general, the scores are higher for system quality than for information quality. For all of the system quality dimensions, the average scores are over 5.0 for the bus, hub and spoke, and centralized architectures. The highest of all scores are for system integration. Companies appear to be doing an excellent job of being able to integrate a variety of data around common keys.

Individual Impacts

The impact that the data warehouse has on users is an important measure of its success. Five questions were asked about individual impacts and Table 8 shows the average scores across all architectures.

Individual Impacts Metrics	Independent Data Marts	Bus Architecture	Hub and Spoke	Centralized (no dependent data marts)	Federated
Users are using the data warehouse	5.39	5.77	5.67	5.71	4.95
Users can access data more quickly and easily	5.50	6.12	5.95	5.88	5.27
Data is easy and intuitive to understand and use	4.72	5.32	4.92	5.15	4.74
Users can think about and ask questions in new ways	4.72	5.84	5.69	5.69	5.11
The decision - making capabilities of users is improved	5.06	5.96	5.87	5.76	5.69

Table 8. Individual Impacts – For Every Architecture (based on 454 responses)

For all of the questions, across all of the architectures, the average scores are reasonably high, suggesting that all of the architectures are having positive individual impacts. Once again, however, the bus, hub and spoke, and centralized architecture consistently out perform the independent data marts and federated architectures.

Though the differences are small, it is interesting to note that the bus architecture scores the highest on all the questions. Its average score on “users can access data more quickly and easily” is a 6.12, which is the only score that exceeds 6.0 on any of the success measures in the study. Proponents of the bus architecture often state how easy and intuitive it is for users to access data.

Organizational Impacts

The ultimate impact of a data warehouse is on the organization as a whole. Six questions asked about organizational impacts and the average scores for the questions across all of the architectures and are shown in Table 9.

Organizational Impacts Metrics	**Independent Data Marts**	**Bus Architecture**	**Hub and Spoke**	**Centralized (no dependent data marts)**	**Federated**
Data warehouse has met the requirements for which it was built	4.93	5.61	5.63	5.54	5.00
Greatly facilitated the use of business intelligence	4.84	5.78	5.45	5.54	4.47
Enabled improvements in business processes	4.86	5.45	5.33	5.43	4.95
Supported the achievement of strategic business objectives	4.91	5.42	5.54	5.43	5.16
Led to high and measurable ROI	3.97	4.65	4.53	4.79	4.21
Improved communication and cooperation across organizational units	4.46	5.10	4.99	5.08	4.84

Table 9. Organizational Impacts – For Every Architecture (based on 454 responses)

The pattern seen with the other success measures holds for organizational impacts. The independent data mart architecture consistently ranks lowest on all metrics (with a single exception). Next lowest is the federated architecture. Overall, the data suggests that the architectures are having a very positive impact at the organizational level.

Respondents were also asked to provide an overall assessment of the success of their data warehouse, and were given three choices – *potentially in trouble, an up and coming system*, or *a runaway success*. Only about 10 percent of the warehouses with the bus, hub and spoke, or centralized architectures are *potentially in trouble*. This percentage rises to about 20 percent for the independent data marts and federated architectures. The centralized architecture has the highest percentage (35%) of companies reporting *a runaway success*, followed by the hub and spoke, federated, and bus architectures, each with about 26 percent.

Development Time

Depending on the criticality of the data warehouse, the time to rollout the first subject area(s) or business process(es) can be very important. As can be seen in Figure 15, on average, the hub and spoke architecture requires the most time, followed by the federated architecture. Given the large investment in infrastructure, processes, and the number of component parts, it is not surprising that the hub and spoke architecture takes the most time. It is interesting to note that the independent data marts, bus architecture, and centralized data warehouse all take about the same amount of time. Because the major difference between the hub and spoke and centralized architectures is the dependent data marts, the longer time for developing the hub and spoke architecture can probably be attributed to the dependent data marts. Independent data marts can be rolled out quickly because their development is independent of other decision support initiatives.

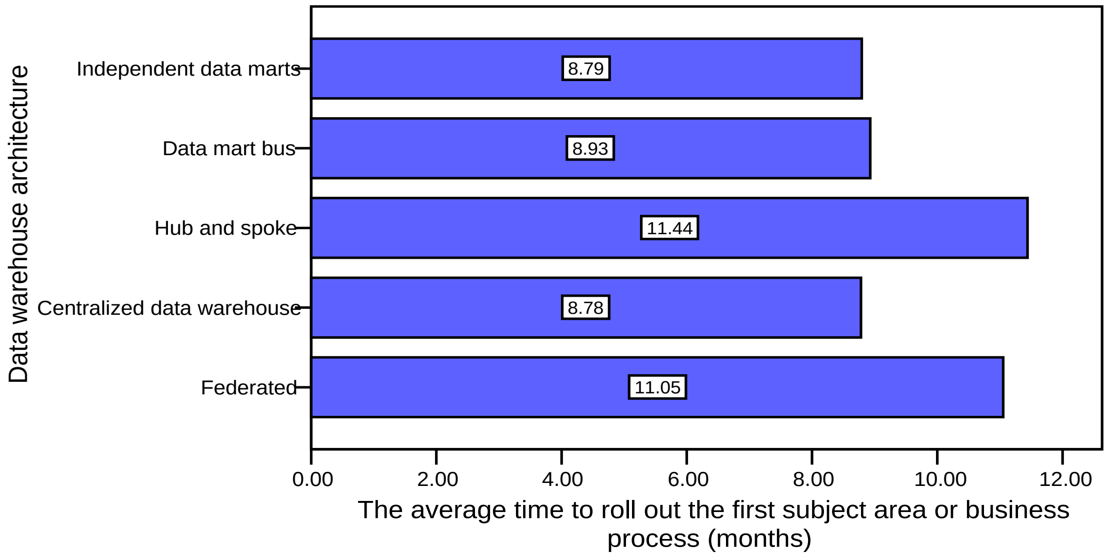

Figure 15. The Average Time to Rollout the First Subject Area(s) or Business Process(es) – For Every Architecture (based on 424 responses)

It is also important for a deployment to be on or ahead of schedule. Table 10 shows the percentages of the rollouts of the first subject area(s) or business process(es) that were *behind*, *on*, or *ahead of schedule* for the various architectures.

Development Time Metrics	Independent Data Marts	Bus Architecture	Hub and Spoke	Centralized (no dependent data marts)	Federated
Behind schedule	42.1	47.1	46.6	37.2	31.6
On schedule	49.1	47.9	50.0	48.7	63.2
Ahead of schedule	8.8	5.0	3.4	14.1	5.2
	100.0%	100.0%	100.0%	100.0%	100.0%

Table 10. Whether the Time to Rollout the First Subject Area(s) or Business Process(es) Was on Schedule – For Every Architecture (based on 454 responses)

Overall, the majority (i.e., over 50%) of implementations were on or ahead of schedule for all of the architectures. The centralized architecture was the most likely (14.1%) to be ahead of schedule, while the bus and the hub and spoke architectures were the most likely (47.1 and 46.6%) to be behind schedule.

Development Cost

Cost data was collected for rolling out the first subject area(s) or business process(es) and the annual maintenance costs. This data is shown in Figures 16 and 17 and includes hardware, software, and personnel costs.

The cost pattern for rolling out the initial version is similar to that for development time. The hub and spoke architecture is the most expensive, followed by the federated architecture. The centralized architecture is slightly more expensive on average than the bus architecture.

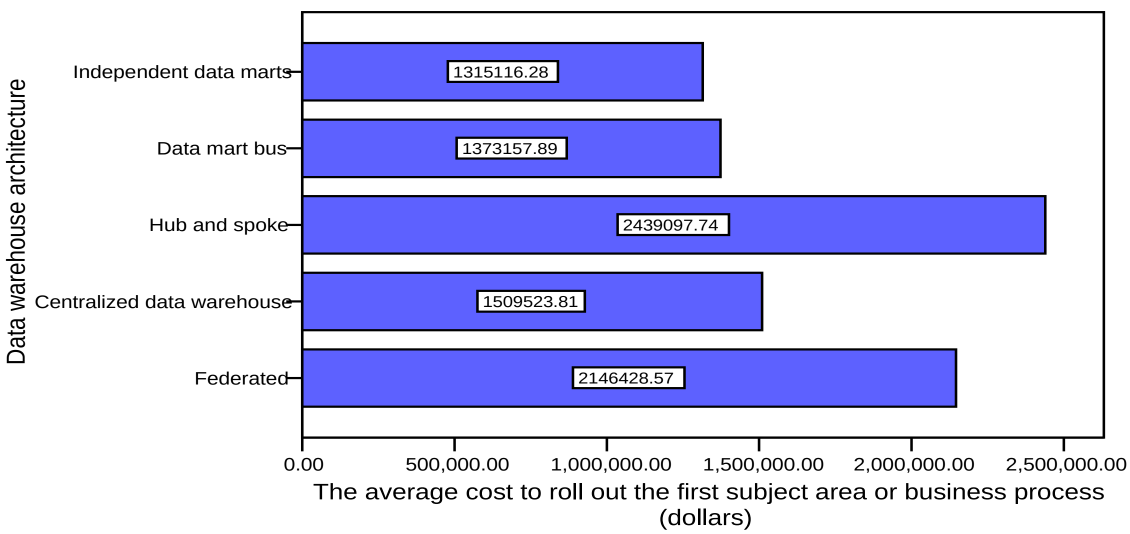

Figure 16. The Average Cost to Rollout the First Subject Area(s) or Business Process(es) – For Every Architecture (based on 348 responses)

The average annual maintenance costs show a slightly different pattern. The hub and spoke architecture is still the most costly, followed by the federated architecture. However, the bus costs slightly more on average than the centralized architecture. The independent data marts are the least expensive to maintain annually; however, intangible costs, such as missed business opportunities are not included.

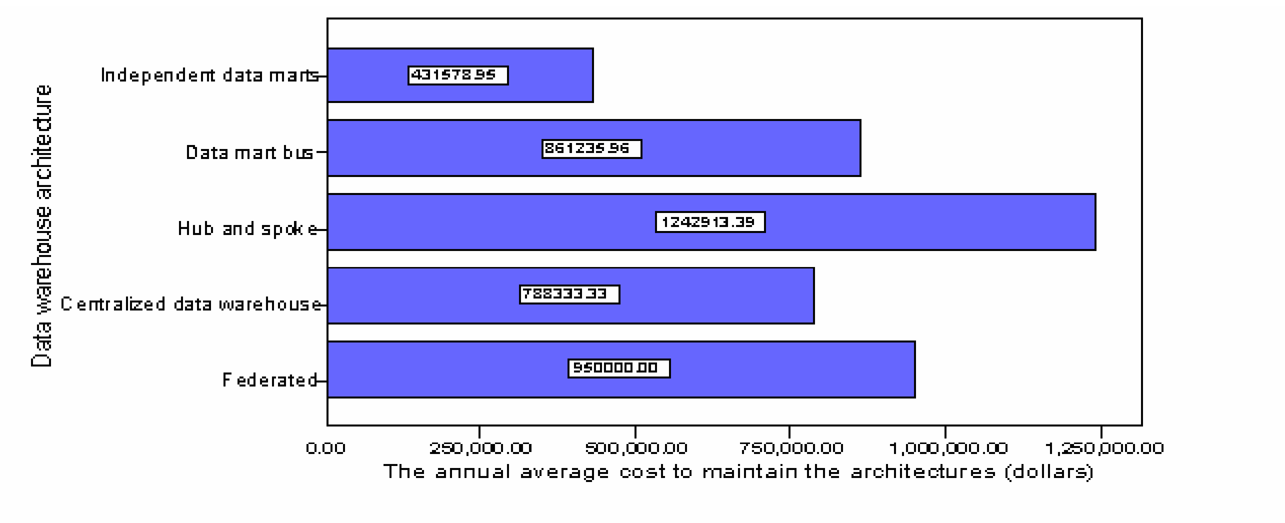

Figure 17. The Average Annual Cost to Maintain the Architectures (based on 329 responses)

Data was also collected on whether the cost of rolling out the first subject area(s) or business process(es) was *over*, *on*, or *under the budgeted amount*. Table 11 shows the percentages in each category for all of the architectures.

Development Costs Metrics	**Independent Data Marts**	**Bus Architecture**	**Hub and Spoke**	**Centralized (no dependent data marts)**	**Federated**
Over budget	33.3	42.9	35.6	29.5	42.1
On budget	59.6	51.2	59.8	61.5	52.6
Under budget	7.1	5.9	4.6	9.0	5.3
	100.0%	100.0%	100.0%	100.0%	100.0%

Table 11. Whether the Cost to Rollout the First Subject Area(s) or Business Process(es) Was on Budget – For Every Architecture (based on 348 responses)

Once again, the majority of the implementations met or exceeded expectations. Most were on or under budget. In fact, the initial rollouts did slightly better in terms of budget than time. The bus and federated architectures were the most likely (42.9 and 42.1%, respectively) to be over budget. The centralized was the least likely to be over budget (29.5%) and the most likely to be under budget (9.0%).

A Comment on Data Warehouse Failures

Some of the early literature on data warehousing mentions the high number of data warehouse failures. Obviously, this is of concern to everyone involved with data warehousing. It is important to recognize, however, that a "failure" can mean different things. It may be a project that is behind schedule, over budget, fails to meet requirements, is not fully used, or "bellies up."

While this study did not address data warehouse failures directly, it does provide interesting insights. For the bus, hub and spoke, and centralized architectures, the percentage of warehouses *potentially in trouble* is only about 10 percent. Most are doing well and many are *a runaway success*. The warehouses are less successful, however, when it comes to being on time and on budget for rolling out the first business process(es) or subject area(s). Depending on the architecture, 30 to 50 percent of the warehouses tend to be either behind schedule or over budget.

These numbers emphasize the importance of understanding how a data warehouse failure is defined. If the definition includes a measure of overall success, the percentage of failures is probably much lower than is often reported. However, if the definition includes *on time* and *on budget*, the failure rate is higher.

The Effect of Domain

Earlier in the report, the interaction of the domain of the warehouse and the warehouse architecture was considered. It was seen that independent data marts tend to be associated with smaller domains, while the bus architecture and especially the hub and spoke and centralized architectures have larger domains. This raises the question as to what effect domain has on the success measures.

To explore this issue, the data for only the highest domain category (i.e., the entire organization) was analyzed. In other words, only data for companies that indicated that their data warehouse was enterprise-wide was analyzed for this part of the report. This allowed an "apples to apples" comparison of the architectures to be made. It also resulted in smaller sample sizes for the architectures and the findings should be interpreted with this in mind. The federated architecture was not included in this analysis because of the small number of firms that reported both a federated architecture and a company-wide domain.

Overall, the average scores for information quality, system quality, individual impacts, and organizational impacts changed very little when adjusted for domain. That is, the success scores for enterprise-wide implementations were about the same as for the smaller domains.

When considering only company-wide implementations, independent data marts still scored the lowest on the information and system quality and individual and organizational impacts. The bus, hub and spoke, and centralized architectures continued to have the highest averages, with few differences among them.

Next, the development time and cost for enterprise-wide implementations were analyzed. Not surprisingly, these warehouses took longer and cost more to implement than for smaller domains. With the domain held constant for company-wide implementations, the average time to implement the first business process/subject area was about the same for the independent data marts, bus architecture, and centralized architecture. The hub and spoke architecture took the longest time to implement.

The cost of implementing the first business process/subject area and the annual cost was the lowest for the independent data marts. It should be kept in mind, however, that the average size of the independent marts architecture is much smaller than any of the others. The cost of implementing the first business process/subject area was essentially the same for the bus architecture and the centralized architecture. The annual cost was slightly higher for the bus architecture. The hub and spoke architecture had the highest initial development and annual costs. It also tends to be the largest in size.

The big picture view is that for enterprise-wide implementations, the development time and costs for independent data marts, the bus architecture, and the centralized architecture are similar. The hub and spoke architecture takes the longest time to initially develop and is the most costly, but is also associated with warehouses that are larger in size.

Advanced Analyses

Two powerful statistical analysis techniques – multivariate analysis of variance (MANOVA) and structural equation modeling (SEM) -- were used to further analyze the information quality, system quality, individual impacts, and organizational impacts measures. Once again, the hub and spoke and centralized architectures were combined for data analysis purposes and the federated architecture was dropped from the analysis because of the small sample size.

It was found that the independent data marts architecture scored statistically significantly lower than the bus and the hub and spoke/centralized architectures. When the bus architecture was compared to the hub and spoke/centralized architecture, no significant differences were found. Even though the average scores were slightly higher for the bus than the hub and spoke/centralized architecture on some success measures, the differences were not statistically significant.

The Impact of Domain

Analysis of variance (ANOVA) was used to investigate the potential effect that the domain of the data warehouse implementation has on the success metrics. Once again the data was placed into two groups – large (entire company and several business units) and small (functional area and sub unit) domains. Using an alpha of 0.05, the time and cost to implement the first subject area(s) or business process(es) were found to be significantly higher for the large domain. This finding reflects the greater time and cost associated with more comprehensive implementations. There were no statistically significant differences for any of the other success metrics.

Company Experiences and Expert Commentary

The experts and survey respondents commented on the success of the various architectures.

Barry McConnell is a data architect for the Florida Department of Education. For their architecture, they implemented Inmon's Government Information Factory. According to Barry, "Our data is highly flexible, scalable, integrated, and complete. It scores lower, however, on accuracy and consistency. This is not so much due to the architecture, but rather, to the various source systems that we have to use."

Eric Lofstrom is Manager of BI at Quicken Loans. He described his company's hub and spoke architecture and some of its greatest successes in the following way: "We source three major types of data – loans, call center, and website. The data is maintained in large tables in the hub and then loaded on servers in cubes for reporting and OLAP purposes. Users find the cubes to be easy and intuitive to use and query performance is great."

Tim Feetham discussed the high information and system quality scores for the bus architecture: "The bus architecture is easily accessible, understandable, and provides high performance. Everything is exposed to the user. Consequently, users perceive the data to be consistent and well integrated."

Margy Ross commented on the individual impacts of the bus architecture: "The success of any data warehouse is measured by the business' acceptance of the analytic environment and the benefits realized from it. You should use the data warehouse architecture that best supports these success criteria, regardless of the label. It's not surprising, however, that the bus architecture scored highest across the board on the five measures regarding the data warehouse's impact on users (e.g., data warehouse usage, quick and easy data accessibility, understandable data, ability to ask questions in new ways, and improved decision-making capability). The findings are consistent with our experiences."

Rob Armstrong commented on the similarity of the scores on the success metrics for the bus, hub and spoke, and centralized architectures: "I understand that the competing architectures got similar scores and the conclusion that they are equally successful. I think that they all meet the outset demands rather than all provide the same outputs. To that degree they are all successful, but do they deliver the same capacity?"

Karolyn Duncan Hepp commented on the importance of providing quick deliverables. "Any successful warehousing endeavor requires both architecture and methodology, which must be complimentary, yet are distinct components. Bottom-up methods are assumed to be faster, while top-down methods are assumed to be more enterprise oriented. The problem, however, comes when teams following top-down methods create sharable elements (e.g., enterprise requirements, data models) at the expense of implementation. It is bothersome that the research shows hub and spoke architected systems take more time than bus or centralized. Regardless of the underlying architecture, warehousing projects should always allow for something (e.g., new data, new functionality, or both) to go into production in 60 to 120 calendar days. Due to the managed data replication in a hub and spoke architecture, this usually means the results/deliverables are more narrow, addressing fewer requirements in each early project (than a bus or centralized environment)."

Claudia Imhoff made a recommendation for a future study. "It would be interesting to collect data for the analysis of the costs of the various architectures over time. An advantage of the bus, hub and spoke, and centralized architectures is that they have reusable components, which should help keep the long-term costs down."

Conclusion

This study sheds light on two questions of continuing interest to the data warehousing community: (1) what factors lead companies to select a particular architecture and (2) how successful are the various architectures? In this concluding section, we summarize and comment on the study findings.

Findings about Architecture Selection

Eleven factors affect the architecture selection decision. In general, the most important factors are information interdependence between organizational units, the strategic view of the warehouse prior to implementation, and upper management's information needs.

The factors that affect the selection of a particular architecture, however, depend on what the architecture is. In the case of *independent data marts*, when there are constraints on resources, the view of the warehouse is limited in scope (e.g., a subunit solution), and the perceived IT skills in-house are low, the independent data mart architecture is likely to be selected. When there is a high need to share and integrate information across organizational units, an urgent need for the data warehouse, low constraints on the availability of resources, and sponsorship at high organizational levels, the *bus architecture* is an attractive choice. When there is a high need for information integration among organizational units, the warehouse is viewed as being strategic, and the perceived ability of the in-house IT staff is high, the *hub and spoke/centralized* warehouse is a common choice.

Of particular interest to many people is why some companies select the bus over the hub and spoke/centralized architecture. The bus architecture may be the architecture of choice when there is a high need for information flow between organizational units, the urgency of need for a data warehouse is high, and the view of the warehouse prior to implementation is more limited in scope.

A Proposed Architecture Selection Model

Based on this research, an overall selection model can be proposed that describes how companies choose an architecture; see Figure 19. It takes the various selection factors and organizes them into a causal-flow model. The selection factors in the proposed model represent factors that emerged as having a significant influence on architecture selection based on advanced statistical analyses using multinomial logistics regression. In this model, the need for information interdependence between organizational units (i.e., horizontal information interdependence) and the nature of end user tasks (i.e., task routineness) combine to create the information requirements for the data warehouse. The information processing requirements and the source of sponsorship then combine to determine the view of the data warehouse; that is, whether perhaps the warehouse is a point solution for at particular department's needs or is an enabler for supporting strategic business objectives. The perceived ability of the IT staff, the availability of resources, and the urgency of need for the data warehouse combine as facilitating conditions for the selection of a particular architecture. And finally, the view of the warehouse and the facilitating conditions influence the architecture selection decision. This proposed model still needs to be tested, but it is consistent with this study's findings.

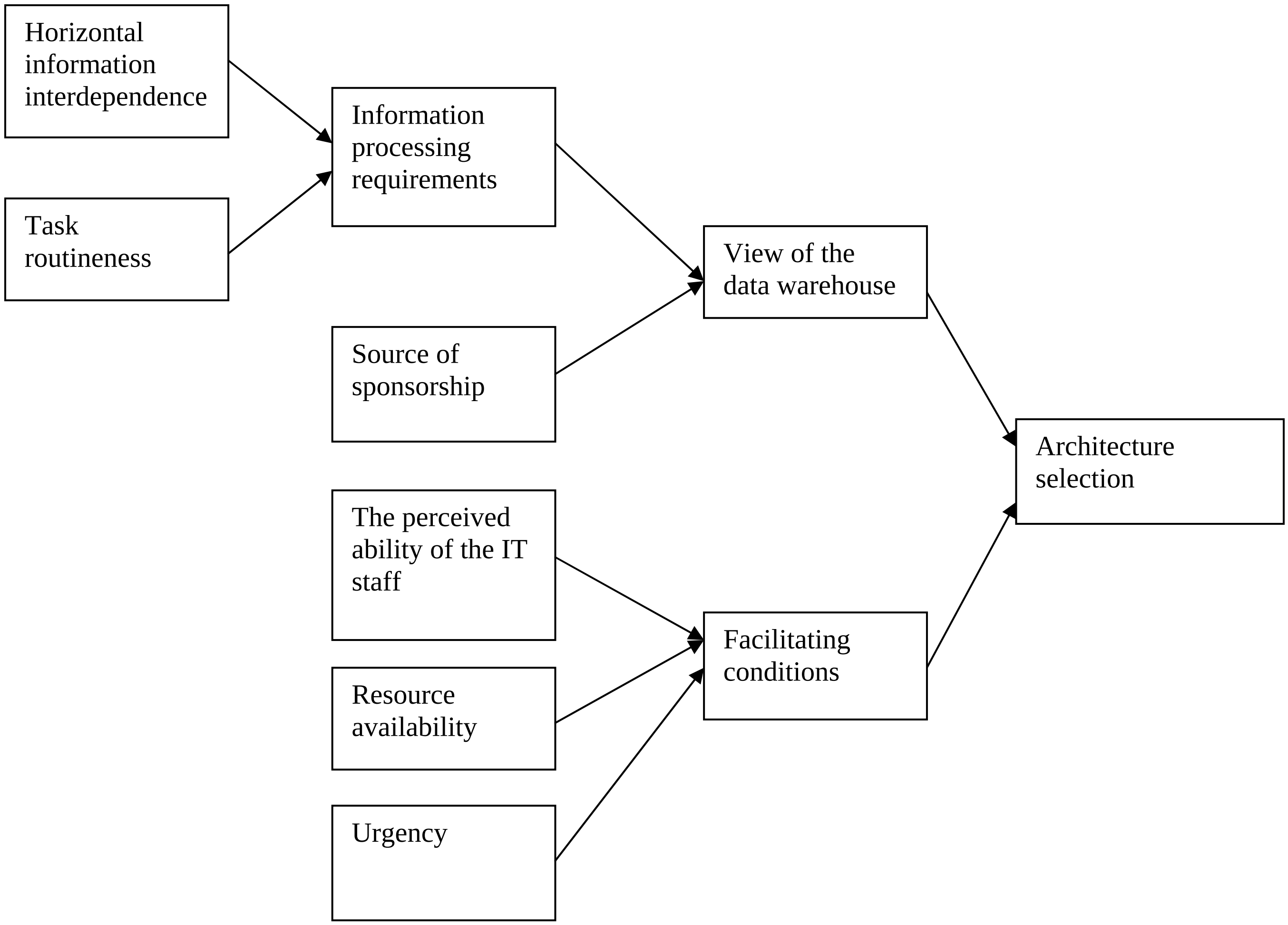

Figure 19. A Proposed Architecture Selection Model

Findings about the Success of the Architectures

This study suggests why there are agreements and disagreements over which architecture is best. The findings show conclusively that independent data marts are weaker than the alternatives in terms of information quality, system quality, individual impacts, and organizational impacts. This is consistent with the conventional wisdom. Though not as weak, the federated architecture tends to score relatively low on the success metrics. This is not surprising. A federated architecture must “make do” with an existing decision support infrastructure and to some extent has to live with its weaknesses.

Perhaps the most interesting study finding is how similar the bus, hub and spoke, and centralized architectures scored on the information and system quality and the individual and organizational metrics. It helps explain why these competing architectures have survived over time – they are equally successful for their intended purposes! Based on these metrics, no single architecture is dominant.

There are statistically significant differences, however, in terms of development time and cost. The hub and spoke takes the longest time to initially develop and is the most costly to initially develop and maintain. The other architectures tend to be similar in terms of development time and cost. Interestingly, when all of the success metrics are considered, the bus and centralized architectures tend to be the most alike.

The similarity of the success of the bus, hub and spoke, and centralized architectures is perhaps not all that surprising. Much like the development methodologies have converged, so too have the architectures. When the researchers were developing the descriptions of the architectures, the experts would sometimes question a description and point us to an early writing. When we compared that description to more recent writings, it became apparent that the proponents of the various architectures have often evolved their thinking over time, with the architectures becoming more alike. For example, the hub and spoke architecture may include dimensional data marts, which is at the heart of the bus architecture. This evolution is appropriate and good for the industry, but it is also a possible reason that the scores on the success metrics are similar.

Overall, this study found good news for the data warehousing community. The scores on the success metrics for the most common architectures – the bus, hub and spoke, and centralized -- are uniformly high. Few warehouses are potentially in trouble and companies are experiencing overall success with the architectures they have implemented. These findings bode well for the future of data warehousing.

Appendix A. Contributing Experts

Expert	Organization
Rob Armstrong	Teradata, a division of NCR
Wayne Eckerson	The Data Warehousing Institute
Vickie Farrell	Cerebra
Jonathan Geiger	Intelligent Solutions
Jane Griffin	KPMG
Doug Hackney	Enterprise Group Ltd
Karolyn Duncan Hepp	Toguchi, LLC.
Claudia Imhoff	Intelligent Solutions
Bill Inmon	Inmon Associates
Julie Kimball	Kimball Group
Ralph Kimball	Kimball Group
Pieter Mimno	Mimno, Myers & Holum
Jim Revak	Progressive
Margy Ross	Kimball Group
Don Stoller	Owens & Minor
Ron Swift	Teradata, a division of NCR
Jim Thomann	DecisionPath Consulting
Warren Thornthwaite	Kimball Group
David Wells	The Data Warehousing Institute
Todd Walters	Teradata, a division of NCR

Appendix B. Individuals and Organizations that Promoted the Study

B-EYE-Network
Business Objects
Claudia Imhoff
DAMA International Data Warehousing Professional Group
DM Review
Inmon Associates
Kimball Group
MicroStrategy
Teradata
The Data Warehousing Institute
William McKnight

Suggested Readings

Hackney, D., *Understanding and Implementing Successful Data Marts,* Addison-Wesley, 1997.

Inmon, W., Imhoff, C. and Sousa, *Corporate Information Factory,* 2nd edition, Wiley, 2001.

Inmon, W., *Building the Data Warehouse,* 3rd edition, Wiley, 2002.

Kimball, R., Reeves, L., Ross, M., and Thornthwaite, W., *The Data Warehouse Lifecycle Toolkit*, Wiley, 1998.

Kimball, R. and Ross, M., *The Data Warehouse Toolkit*, 2nd edition, Wiley, 2002.

Swift, R.S., *Accelerating Customer Relationships,* Prentice-Hall, 2001.

TDWI Data Warehousing Architectures, The Data Warehousing Institute, 2005.

Suggested Websites

www.baseline-consulting.com

www.datawarehouse.com

www.decisionpath.com

www.dmreview.com

www.dwinfocenter.com

www.informatica.com

www.intelSols.com

www.kimballgroup.com

www.mimno.com

www.tdwi.org

www.teradata.com