Key organizational factors in data warehouse architecture selection

Ravi Kumar Choudhary

ABSTRACT

Deciding the most suitable architecture is the most crucial activity in the Data warehouse life cycle. Architecture is the key factor in setting up the abilities and the limitations of a data warehouse.This articles was conducted to (1) better understand the factors that influence the selection of 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 many companies about the respondents, their companies, their data warehouses, the architectures they use, and the success of their architectures. The study findings provide interesting and useful thing about topics of long-standing importance to the data- warehousing field.

Keywords:*Datawarehouse, Architecture, Data Mining*

1. Introduction

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."[1] 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.[2] Each has strong proponents.

Considering the importance of the choice of architecture, there is surprisingly little articles 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.

2. 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 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[3].

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.

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.

3. The Five Architectures

The data warehousing literature provides discussions and examples of a variety of architectures[4]. 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.

3.1 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.

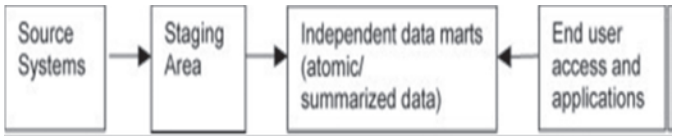

Figure 1. The Independent Data Marts Architecture

3.2 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.

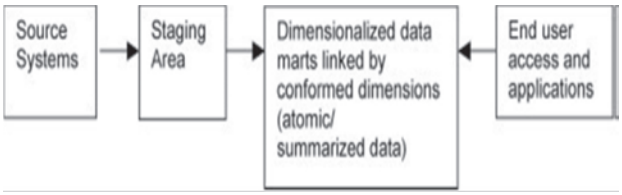

Figure 2. The Data Mart Bus Architecture with Linked Dimensional Data Marts

3.3 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, demoralized, or summarized/atomic dimensional data structures based on user needs. Most users query the dependent data marts. Figure 3 shows this architecture.

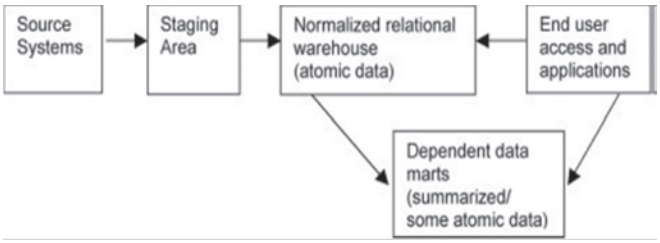

Figure 3. The Hub and Spoke Architecture

3.4 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.

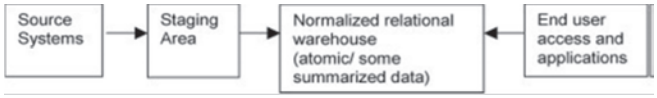

Figure 4. The Centralized Data Warehouse Architecture

3.5 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.

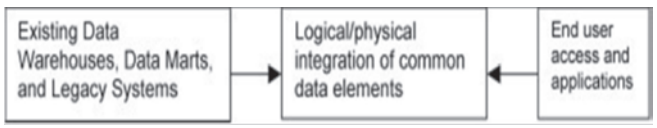

Figure 5. The Federated Architecture

4. 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[5].

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.

5. 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.[6]

5.1 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.

5.2 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.

5.3 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.

5.4 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.

5.5 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.

5.6 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.

5.7 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.

5.8 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.

5.9 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.

5.10 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 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.

5.11 Technical Issues

A variety of technical considerations can affect the choice of 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 architecture may be better than others.

Research Model for Research Question #1

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

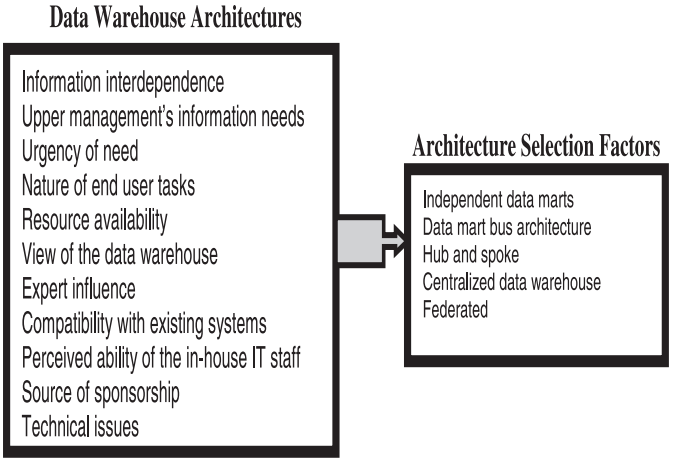

Figure 6. The Research Model that Relates the Selection Factors to the Architectures

6. 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.[7]

6.1 Information Quality

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

6.1.1 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.

6.1.2 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.

6.1.3 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.

6.2 System Quality

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

6.2.1 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.

6.2.2 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.

6.2.3 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.

6.3 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.

6.4 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.

6.5 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.

6.6 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.

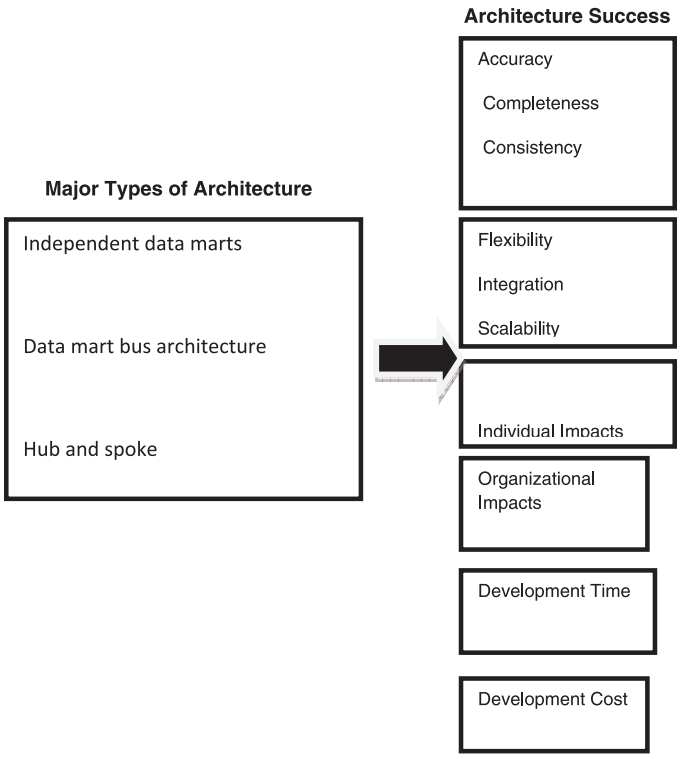

Figure 7. The Research Model that Relates the Success Metrics to the Architectures

7. 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 important* and 7 being *very important*. The

importance factors were described as:

7.1 Information interdependence between organizational units : The need to share information among organizational units.

7.2 Upper management's information needs: Upper management's needs for information from lower organizational levels.

7.3 Urgency of need for a data warehouse: The extent to which there was an urgent need to build the data warehouse.

7.4 Nature of end user tasks: The extent to which users' jobs required non-routine data analyses.

7.5 Constraints on resources: The availability of resources (IT personnel, business unit personnel, and monetary resources) for building the data warehouse.

7.6 Strategic view of the warehouse prior to implementation: The extent to which implementing a data warehouse was viewed as important to supporting strategic objectives.

7.7 Compatibility with existing systems: The extent to which the data warehouse architecture was compatible with existing systems.

7.8 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.

7.9 Technical issues: The extent to which technical issues affected the data warehouse architecture.

7.10 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.

8. 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 article, an overall selection model can be proposed that describes how companies choose architecture see Figure 8. 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.

Proposed architectural Selection model

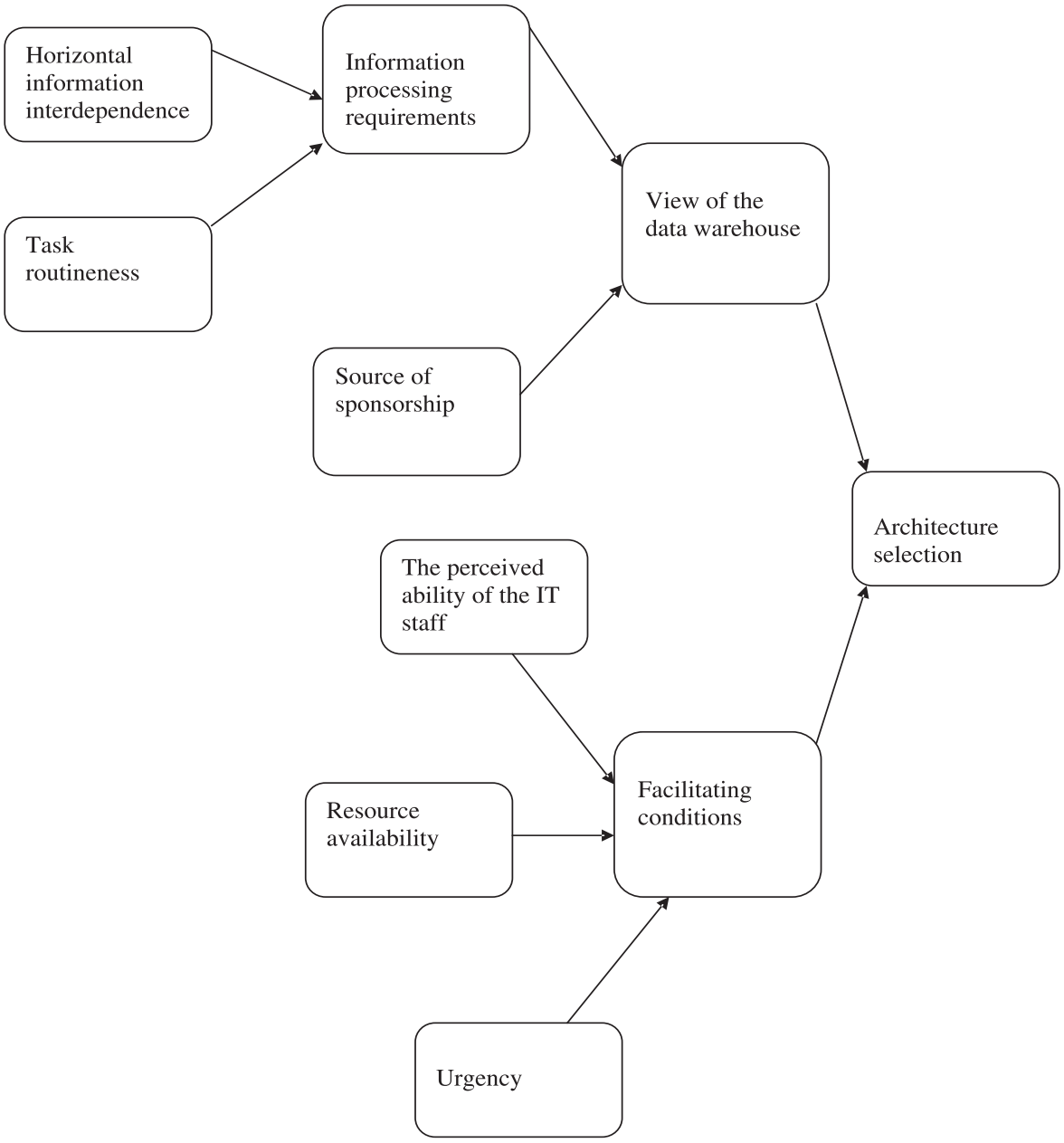

Figure 8. A Proposed Architecture Selection Model

References

[1] Santillo L., "Size & Estimation of data warehouse systems", *in Proceedings of the 4th European Conference on Software Measurement and ICT Control FESMA DASMA* ,Heidelberg (Germany), May 2001.

[2] Stankovski V., Swain M., Kravtsov V., Niessen T., Wegener D., Kindermann J., and Dubitzky W., "Grid- enabling data mining applications with DataMiningGrid: An architectural Perspective", *in proc. of Future Generation Computer Systems*, vol. 24, no. 4, pp. 259- 279, April 2008.

[3] Telbany M. Warda M., and Borahy M., " Mining the Classification Rules for Egyptian Rice Diseases", *The international Arab Journal of information Technology*, vol. 3, no. 4, pp. 303-307, 2006

[4] Vassiliadis P., Quix C., Vassiliou Y., and Jarke M., "Data Warehouse Process Management", *in proceedings of Information Systems*, vol. 26, no.3, pp. 205- 236, May 2001.

[5] Velpula V., and Gudipudi D., "Behavior-Anomaly-Based System for Detecting Insider Attacks and Data Mining", *in proc. of Intl. Journal on Recent Trends in Engineering*, vol. 1, no. 2, pp. 261- 266, May 2009.

[6] Wadhwa V., and Lin D., "Radio Frequency Identification: A new Opportunity for Data Science", *in proc. of Journal on Data Sciences*, vol. 6, no. 3, pp. 369- 388, July 2008.

[7] Wikramanayake G., and Goonetillake J., "Managing Very Large Databases and Data Warehousing", *in proc. of Sri Lankan Journal on Librarianship and Information Management*, vol. 2, no. 1, pp. 22- 29, 2006.

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

[9] 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.

[10] Inmon, W.H.: *Building the Data Warehouse*. John Wiley & Sons Inc., New York/ NY, 1992

[11] 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.

[12] Breslin, Mary. "Data Warehousing Battle of Giants: Comparing the Basics of the Kimball and Inmon Models," Business Intelligence