

Published: November 2006

Author: Infogile Inc.

A data warehouse is a valuable corporate asset to determine business strategies and make informed business decisions. With the enhanced access to information that a data warehouse provides, an organization can make the time-critical business decisions that are required to remain competitive. Simply put, data warehousing requires a comprehensive assessment of the impact on the entire organization and development of a plan for an organized, systematic solution. This includes implementation of tools for the entire data interaction, transformation, loading, and storage needs.

Data Warehousing - Providing Data Access to the Enterprise

A data warehouse is a consolidated view of your enterprise data, optimized for reporting and analysis. Basically it's an aggregated, sometimes summarized copy of transaction and non-transaction data specifically structured for dynamic queries and analytics. In data warehousing, data and information are extracted from heterogeneous production data sources as they are generated, or in periodic stages, making it simpler and more efficient to run queries over data that originally came from different sources. Data is turned into high-quality information to meet all enterprise reporting requirements for all levels of users. Interactive content can be delivered to anyone in the extended enterprise – customers, partners, employees, managers, and executives – anytime, anywhere. Data Warehousing is a field that has grown out of the integration of a number of different technologies and experiences over the last two decades. These experiences have allowed the IT industry to identify the key problems that have to be solved.

Data Warehouse Architecture

White Paper Overview

Abstract

This paper describes the basic concepts and methods for developing and documenting a data warehouse architecture for strategic information management. It attempts to answer the constituents of a data warehouse, how it should be architected, and what types of tools are needed. The key to success in scalable data warehouse development and the single factor that contributes most to data warehousing success is data warehouse architecture. An enterprise data warehouse provides management with information about what is happening in the business. It must present a single version of the truth from multiple perspectives—and quickly reflect organizational, regulatory, market and other business changes.

Document Audience

This document is primarily intended for Marketing, Sales, Product Support, Internet Services Group, Project Engineering and anyone who is interested in Data Warehouse Technology.

Linking an enterprise's strategic information requirements with its information architecture, application architecture, and technical architecture results in an "Enterprise Information Architecture". A subset of the Enterprise Architecture is the Data Warehouse Architecture.

A well-documented architecture (for the enterprise and its data warehouse) is a logical organization of information pertaining to the following corporate-level, enterprise-wide elements:

- Strategic goals, objectives, and strategies
- Business rules and metrics
- Information requirements
- Application systems
- Relationships between applications, activities, business requirements, and data elements
- Technology infrastructure

Data Warehouse Architecture also establishes guidelines, standards, and operational services that define the enterprise's operational system environment. When an enterprise's architecture is so documented, it can be used to accomplish the following:

- Facilitate change management by linking strategic requirements to systems (including the data warehouse and data marts) that support them and by linking the business model to application designs, including data warehouse designs
- Enable strategic information to be consistently and accurately derived from operational (and external) data
- Promote data sharing, thus reducing data redundancy and maintenance costs
- Improve productivity through component development, management, and reuse

The architecture and design of an enterprise's data warehouse should reflect the performance measurement and business requirements of the enterprise. Its **data model, structure, components, and metadata** should all be based upon internal information requirements -- not specific technologies.

A Data Warehouse Architecture (DWA) is a way of representing the overall structure of data, communication, processing and presentation that exists for end-user computing within the enterprise. The architecture is made up of a number of interconnected parts:

- Operational Database / External Database Layer
- Information Access Layer
- Data Access Layer
- Data Directory (Metadata) Layer



- Process Management Layer
- Application Messaging Layer
- Data Warehouse Layer
- Data Staging Layer

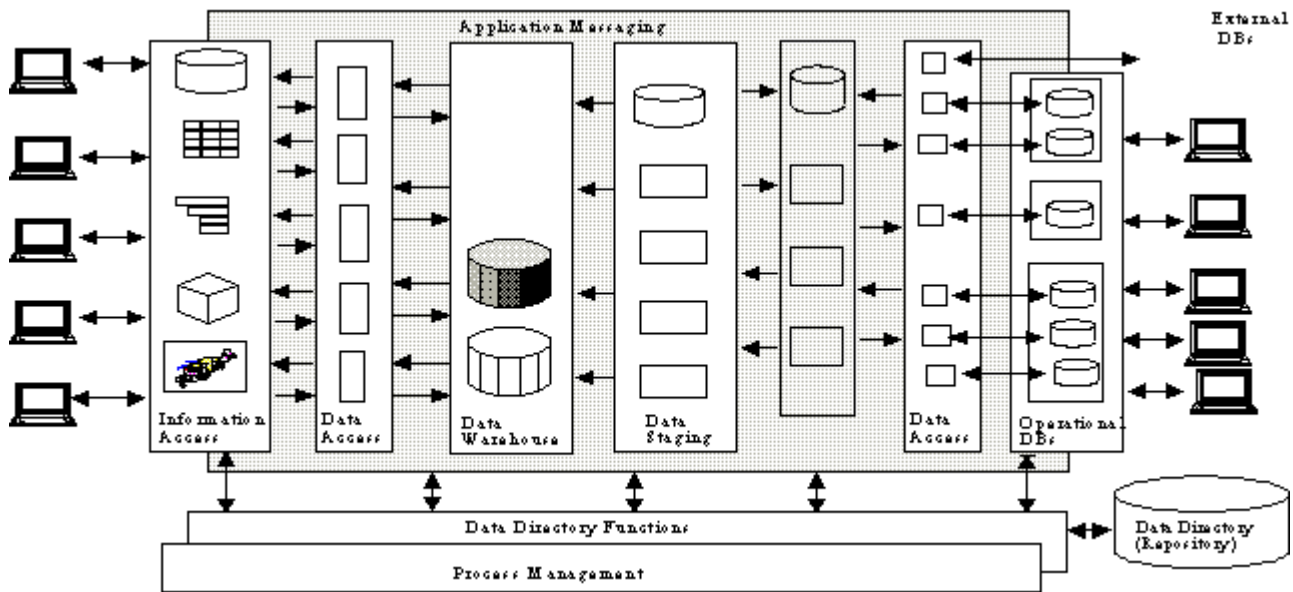


Figure 1 - Data Warehouse Architecture

Data Warehouse Data Model

A **data model** documents the data elements whose values at any point in time are necessary to tell data warehouse users how well their enterprise is performing. The data warehouse model provides a clear and unambiguous definition of every key data entity, describing the way each is used, as well as defining derivation formulas, aggregation categories, and refreshment time periods. The data warehouse model, linked with the enterprise information architecture, becomes both requirement documentation and a source for communicating the contents of the data warehouse to its users and developers. Issues that must be addressed in the data model include what legacy data will be used to populate the data warehouse, how data will be moved from legacy environments to the data warehouse, and how the legacy data will be integrated or transformed to ensure data quality and integrity in the data warehouse. *The two most important issues for any data warehouse are data quality and data access.*

Data Warehouse Metadata

Metadata, or data about data, is the nerve center of a data warehouse and is essential. Metadata is essential to all levels of the data warehouse, but exists and

functions in a different dimension from other warehouse data. Metadata used to manage and control data warehouse creation and maintenance resides outside the data warehouse, often in a digital repository. Metadata for data warehouse users is part of the data warehouse itself and is available to control access and analysis of the data warehouse. To a data warehouse user, metadata is like a "card catalog" to the subjects contained in the data warehouse. The two types of data warehouse metadata are called structural and access.

Structural metadata is used for creation and maintenance of the data warehouse. It fully describes data warehouse structure and content. The basic building block of structural metadata is the data warehouse model that describes its data entities, their characteristics, and how they are related to one another. The way potential data warehouse users currently use, or intend to use, enterprise measures provides insight into how to best serve them from the data warehouse; i.e., what data entities to include and how to aggregate detailed data entities. The data warehouse model provides a means of documenting and identifying structural metadata. This includes both strategic and operational uses of enterprise measures, as well as multi-dimensional summarization. Structural metadata also includes performance metrics for programs and queries so that users and developers know how long programs and queries should run. Data warehouse performance tuning also uses these metrics.

Access metadata is the dynamic link between the data warehouse and end-user applications. It generally contains the enterprise measures supported by the data warehouse and a dictionary of standard terms including user-defined custom names and aliases. Access metadata also includes the location and description of data warehouse servers, databases, tables, detailed data, and summaries along with descriptions of original data sources and transformations. Access metadata provides rules for drill up, drill down and views across enterprise dimensions and subject hierarchies like products, markets, and customers. Access metadata also allows rules for user-defined custom calculations and queries to be included. In addition, access metadata contains individual, work group, and enterprise security for viewing, changing, and distributing custom calculations, summaries, or other analyses.

Data Warehouse Components

The data warehouse architecture also contains descriptions data warehouse components: **current detail**, **summarized data**, and **archives** as well as **systems of record** and **integration/transformation programs**.

The heart of a data warehouse is its **current detail**. It is the place where the bulk of data resides. Current detail comes directly from operational systems and may be stored as raw data or as an aggregation of raw data. Current detail, organized by subject area, represents the entire enterprise, rather than a given application. Current detail is the lowest level of data granularity in the data warehouse. Every data entity in current detail is a snapshot, at a moment in time, representing the instance when the data are accurate. Current detail is typically maintained for two to five years, but some enterprises may require detail data for significantly longer periods. When initially implemented, a data warehouse may include current detail more than two years old, but the often questionable quality of older data must be considered and measures taken to ensure its validity. Current detail refreshment occurs as frequently as necessary to support enterprise requirements.

Lightly summarized data are the hallmark of a data warehouse. All enterprise elements (department, region, function, etc.) do not have the same information requirements, so effective data warehouse design provides for customized, lightly



summarized data for every enterprise element (see Data Mart, below). An enterprise element may have access to both detailed and summarized data, but typically much less than the total stored in current detail.

Highly summarized data are primarily for enterprise executives. Highly summarized data can come from either the lightly summarized data used by enterprise elements or from current detail. Data volume at this level is much less than other levels and represents an eclectic collection supporting a wide variety of needs and interests. In addition to access to highly summarized data, executives also should have the capability of accessing increasing levels of detail through a "drill down" process.

Data warehouse **archives** contain old data (normally over two years old) of significant, continuing interest and value to the enterprise. There is usually a massive amount of data stored in the data warehouse archives that has a low incidence of access. Archive data are most often used for forecasting and trend analysis. Although archive data may be stored with the same level of granularity as current detail, it is more likely that archive data are aggregated as they are archived. Archives include not only old data (in raw or summarized form); they also include the *metadata* that describes the old data's characteristics.

A **system of record** is the source of the best or "rightest" data that feed the data warehouse. The "rightest" data are those which are most timely, complete, accurate, and have the best structural conformance to the data warehouse. Often the "rightest" data are closest to the source of entry into the production environment. In other cases, a system of record may be one containing already summarized data. Often, "rightest" data is created from diverse sources through a reconciliation process.

The components that link operational systems with the data warehouse are the **integration/transformation programs**. Even the "rightest" operational data cannot usually be copied, as is, into a data warehouse. Raw operational data are virtually unintelligible to most end users. Additionally, operational data seldom conform to the logical, subject-oriented structure of a data warehouse. Further, different operational systems represent data differently, use different codes for the same thing, squeeze multiple pieces of information into one field, and more. Operational data can also come from many different physical sources: old mainframe files, non-relational databases, indexed flat files, even proprietary tape and card-based systems. Thus operational data must be cleaned up, edited, and reformatted before being loaded into a data warehouse.

As operational data items pass from their systems of record to a data warehouse, integration and transformation programs convert them from application-specific data into enterprise data. These integration and transformation programs perform functions such as:

- Reformatting, recalculating, or modifying key structures and other data elements.
- Adding time elements
- Identifying default values
- Supplying logic to choose between multiple data sources
- Summarizing, tallying, and merging data from multiple sources
- Reconciling data from multiple sources



When either operational or data warehouse environments change, integration and transformation programs must be modified to reflect that change.

Data Warehouse Structure

A data warehouse may have any of several structures. The structure that best meets the data warehouse needs of an enterprise is fully dependent upon the enterprise business, data, and access requirements. The basic data warehouse structures are:

Physical Data Warehouse - physical database in which all the data for the data warehouse are stored, along with metadata and processing logic for scrubbing, organizing, packaging and processing the detail data.

Logical Data Warehouse - also contains metadata including enterprise rules and processing logic for scrubbing, organizing, packaging and processing the data, but does not contain actual data. Instead it contains the information necessary to access the data wherever they reside. This structure is possible *only* when operational systems exactly reflect the enterprise data architecture and system capacities can support both operational and management functions.

Data Mart - subset of an enterprise-wide data warehouse. Typically it supports an enterprise element (department, region, function, etc.). The organization of data in a data mart reflects the needs of the enterprise element it supports, and may be different from the organization of the enterprise data warehouse. Specific data elements may be stored redundantly in both the data mart and the data warehouse. As part of an iterative data warehouse development process, an enterprise builds a series of physical data marts over time and links them via an enterprise-wide logical data warehouse or feeds them from a single physical warehouse.

Both within the Data Warehouse as a whole and within the individual Data Marts, different groups of users have needs for differing slices of data. For example, users at a branch generally need the "horizontal slice" of data that pertains to their branch (i.e. they need all the data elements - tables and columns - but only the rows pertaining to their branch). Other users need "vertical slices" or a combination of horizontal and vertical slices.

Framework and Methodology

Successfully implementing a data warehouse requires a proven framework, or blueprint. Just as you would not think of building a house without a blueprint, the data warehouse project manager should carefully consider what framework to use to build a warehouse (see Figure 3) in three basic steps.

Planning. Information discovery services, which identify business problems to be solved, provide a structured process that is the critical first step in building a data warehouse. These services can be independent of each other and can be done in any order or concurrently. Each planning area represents an entry point into the warehousing methodology.

Design and implementation. The data warehousing solution readiness process represents another entry point into the methodology and should take place when warehouse developers are ready to begin their first data warehousing project and each time additional warehouse projects are initiated as the warehouse grows. It provides a comprehensive analysis



of a company's current environment. The solution readiness process validates the effectiveness of the identified solution within the current environment. Solution readiness investigates the elements needed to support the implementation, including data readiness, technology readiness, functional readiness, support readiness, and infrastructure readiness.

This step is intended to protect the business from attempting to implement a solution for which it is not prepared or that might influence other functional areas within the company not included in the planning. Implementation project plans should be adjusted (as needed) based on the results of the assessments.

Support and enhancement. Data warehouse support and enhancement comprises a series of follow-up operational and value processes supporting the operations and maintenance of a data warehouse. These processes serve the following purposes:

- Supporting the day-to-day running of the warehouse solution, ensuring availability and ongoing performance.
- Assisting in expanding the use (and therefore the benefit) of the solution.
- Expanding the system, possibly to include new applications, users, or data or increased use of the solution through the education of end users.
- Helping relaunch the process at the business imperative step, when selling senior management on the project, or if additional needs or applications are discovered for the next project consulting cycle.
- Helping keep the system continually updated and growing, supporting better business decisions in a planned and controlled way to deliver business value.

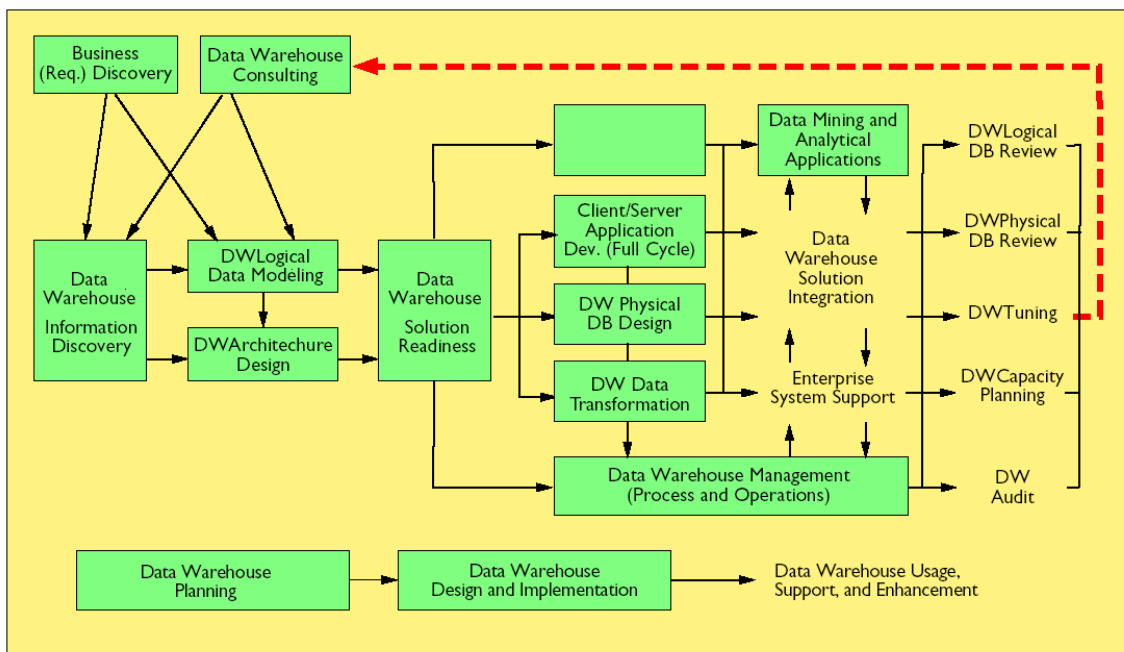


Fig. 2 A data warehouse methodology

A warehouse methodology has to address all of these steps. Building a data warehouse is iterative; therefore, it is critical there be multiple entry points into the chosen methodology. Use of a proven methodology, coupled with collaboration between the IT department and business users, will greatly enhance the chances of successfully building the system.

Figure 2 shows a proven data warehouse methodology representing an end-to-end solution

with multiple entry points. This methodology represents the steps through which service providers and a user company's staff can make decisions regarding the warehouse and then implement and maintain that warehouse.

Conclusions

There are five primary reasons because of which projects fail: lack of partnership between the IT department and business users; incorrect data warehouse architecture; not enough experienced people; improper planning, such as failure to use a proven methodology and a plan to ensure that no details are omitted; and depending on bleeding-edge technology.

On the other hand, experience shows the following precautions encourage successful warehouse implementations: win the highest possible level of executive support; identify a specific business problem to be solved; create a well-defined plan; use proven technology; and employ experienced people.

Cooperation and support at a business's executive level is important to success. Executive management can mediate political and funding issues while providing a foundation for collaboration between the IT department and business users. Choosing a specific business problem to solve and defining requirements and measurements for the solutions help focus the system's direction. Solving this problem by properly implementing a data warehouse

Ensures success and helps the warehouse grow to help solve even more business problems—resulting in even better business operations. Be sure a good plan is in place and that project

Management is top notch. A warehouse project is not a good place for a novice project manager to start. An experienced project manager helps create the plan and then keeps everyone on track.

Employ as many experienced people as possible from both inside and outside the business. They know the pitfalls and should be able to help mentor those who are less experienced.

Building a warehouse is a complex process requiring careful planning and alignment between the IT department and business users. Data warehouses are built to answer specific business problems, not to showcase the wonders of technology. Using the guidelines outlined here can significantly improve your chances of success.

For more information about Infogile products and services, contact the Infogile Sales Information Center at business@infogile.com. To access information on internet, go to www.Infogile.com.

© 2006 Infogile Corporation. All rights reserved.

This case study is for informational purposes only. INFOGILE MAKES NO WARRANTIES, EXPRESS OR IMPLIED, IN THIS SUMMARY. Infogile and the Infogile logo are either registered trademarks or trademarks of Infogile Corporation in the United States and/or other countries. The names of actual companies and products mentioned herein may be the trademarks of their respective owners.

