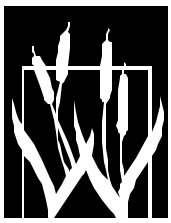

Whitemarsh
Information Systems Corporation

Guidelines for the Data Architecture Class: Data Warehouses

Whitemarsh Information Systems Corporation
2008 Althea Lane
Bowie, Maryland 20716
Tele: 301-249-1142
Email: mmgorman@wiscorp.com
Web: www.wiscorp.com

1
Introduction and Overview

A warehouse database is first and foremost a database. As such, it must represent coherent, well formed business policy execution results, that is, data. Data warehouse databases are not a fundamentally different kind of "beast." They are merely databases but of a different style. Their style is generally the inverse of "update at maximum efficiency." That is, they support reporting and analysis at maximum efficiency. This "battle," update versus reporting has been present and raging since the early 1960s. It is new only to those who have no memory. "Oldies" remember the wall-to-wall-to-wall extract-sort-and-print diagrams that transformed orginally collected data (i.e., punch cards) into many different reports. These print-formatted files were yesterday's data warehouses. So, what is new, is not the concept, but the source for the data. That is, multi-user, random-access databases and their access and reporting engines, database management systems (DBMS).

Because most warehouse databases are derivatives of "prior in the food chain databases," they rely on other databases as sources for data. A warehouse's database design cannot represent clearer or better semantics than those contained in the databases from which they are derived.

1.1 Data Architecture Classes

An enterprise-wide data architecture contains the following five distinct classes:

- Reference databases
- Original data capture databases
- Transaction data staging area databases
- Operational data store databases
- Warehouse databases (wholesale and retail (now called data marts))

Figure 1.1 depicts these database architecture classes within the context of an overall enterprise data architecture. At the bottom of each of the second to the fifth data class are its most significant characteristics. The table that follows Figure 1.1 presents a six column table that describes each data class in terms of: Data architecture class, persistent data characteristics, process characteristics, user considerations, technical considerations, and examples.

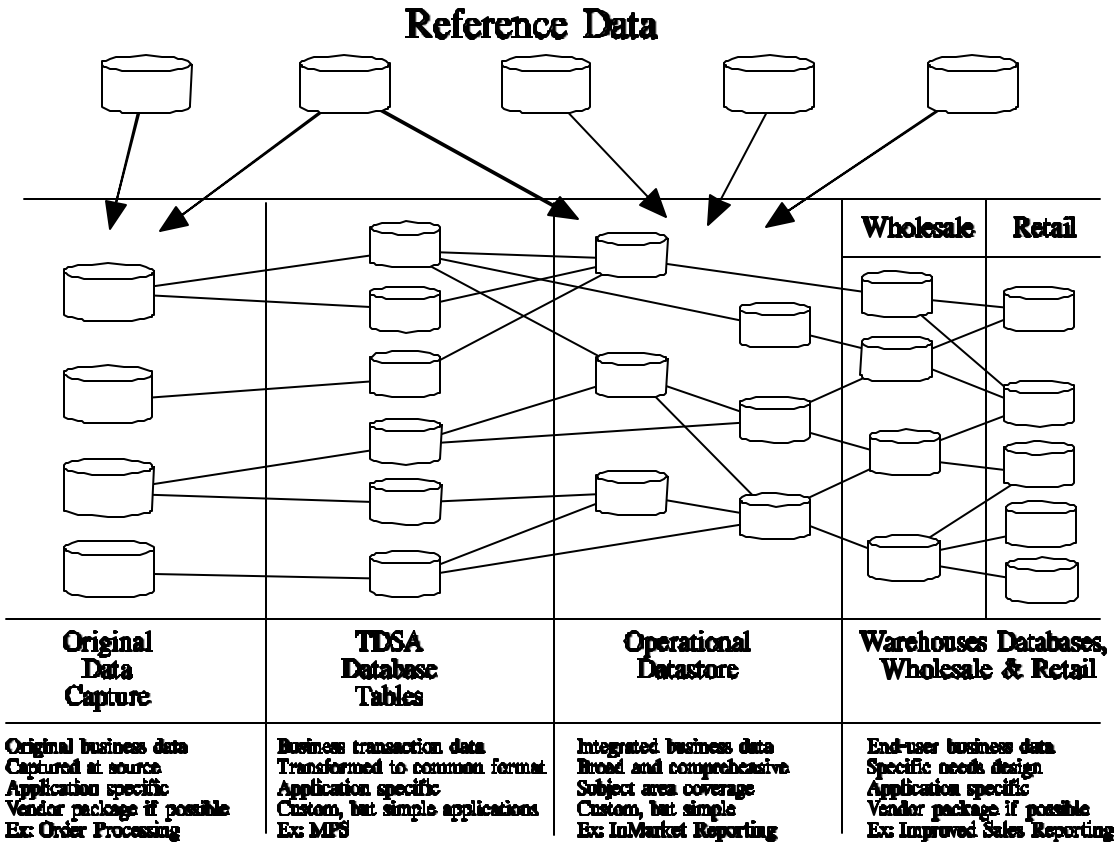

Figure 1. Data Architecture Classes

Data warehouses exist at the end of this file-class architecture sequence. Reasons to build warehouse databases include:

- To reduce or eliminate time consuming and resource intensive ad hoc queries on production databases.
- To bring into a single specialized report-only database data from multiple subject area databases that may have been operating on multiple platforms, or operating under different DBMSs
- Normalize granularity differences that may exist across subject area databases.

- Synchronize the as-of dates across data from multiple subject area databases so that specific and broad reports can be produced that are consistent for all requesters within the fixed time frame.

Regardless of the rationale for developing a data warehouse, the work plans for building them are similar. Differences center primarily on the amount of effort in each of the standard tasks. Broadly, the standard tasks are divided into the following three main phases:

- Requirements Analysis and Design
- Performance Modeling and Implementation
- Deployment and Evolution

The first phase, Requirements Analysis and Design, includes the following major steps

- Data Warehouse Concept Development
- Data Architecture Development
- Logical Data Model Development
- User Reference Data Development
- Data Mapping/Transformation Development
- Data Population Sequence Development

The second phase, Performance Modeling and Implementation, addresses:

- Data warehouse database design
- Data conversion and loading system development
- Data reporting and analysis system development
- Performance Testing
- Warehouse System Implementation

The final phase, Deployment and Evolution, addresses:

- Deployment
- Training
- Production Cut-over
- Interrogation Development
- Evolution and Maintenance

This paper addresses the activities contained in the first phase, Requirements Analysis and Design. The role of data management is restricted to the first phase. Other groups both define and perform the activities in the other two phases. Data management performs a review during the second phase to determine whether the physical database design is a faithful transformation of the first phase's

logical database design. The findings from the review are presented to the data warehouse project's manager.

Section 2 of this paper provides a high level review of the steps necessary for Requirements Analysis and Design. Section 3 presents the review guidelines that data management employs to accomplish quality control and configuration management on the data warehouse design produced during the Performance Modeling and Implementation phase. To illustrate the concepts contained in this paper, an adaptation of an actual case study is employed.