Best Practices in Data Warehousing

Les Pang
University of Maryland University College, USA

INTRODUCTION

Data warehousing has been a successful approach for supporting the important concept of knowledge management—one of the keys to organizational success at the enterprise level. Based on successful implementations of warehousing projects, a number of lessons learned and best practices were derived from these project experiences. The scope was limited to projects funded and implemented by federal agencies, military institutions and organizations directly supporting them.

Projects and organizations reviewed include the following:

- Census 2000 Cost and Progress System
- Defense Dental Standard System
- Defense Medical Logistics Support System Data Warehouse Program
- Department of Agriculture Rural Development Data Warehouse
- Department of Defense (DoD) Computerized Executive Information System
- Department of Energy, Lawrence Livermore National Laboratory, Enterprise Reporting Workbench
- Department of Health and Human Services, Health Care Financing Administration (HFCA) Teraplex Integration Center
- Environmental Protection Agency (EPA) Envirofacts Warehouse
- Federal Bureau of Investigation (FBI) Investigative Data Warehouse
- Federal Credit Union
- Internal Revenue Service (IRS) Compliance Data Warehouse
- Securities and Exchange Commission (SEC) Data Warehouse
- U.S. Army Operational Testing and Evaluation Command
- U.S. Coast Guard Executive Information System
- U.S. Navy Type Commander's Readiness Management System

BACKGROUND

Data warehousing involves the consolidation of data from various transactional data sources in order to support the strategic needs of an organization. This approach links the various silos of data that is distributed throughout an organization. By applying this approach, an organization can gain significant competitive advantages through the new level of corporate knowledge.

Various agencies in the Federal Government attempted to implement a data warehousing strategy in order to achieve data interoperability. Many of these agencies have achieved significant success in improving internal decision processes as well as enhancing the delivery of products and services to the citizen. This chapter aims to identify the best practices that were implemented as part of the successful data warehousing projects within the federal sector.

MAIN THRUST

Each best practice (indicated in **boldface**) and its rationale are listed below. Following each practice is a description of illustrative project or projects (indicated in *italics*), which support the practice.

Ensure the Accuracy of the Source Data to Maintain the User's Trust of the Information in a Warehouse

The user of a data warehouse needs to be confident that the data in a data warehouse is timely, precise, and complete. Otherwise, a user that discovers suspect data in warehouse will likely cease using it, thereby reduc-

ing the return on investment involved in building the warehouse. Within government circles, the appearance of suspect data takes on a new perspective.

HUD Enterprise Data Warehouse - Gloria Parker, HUD Chief Information Officer, spearheaded data warehousing projects at the Department of Education and at HUD. The HUD warehouse effort was used to profile performance, detect fraud, profile customers, and do "what if" analysis. Business areas served include Federal Housing Administration loans, subsidized properties, and grants. She emphasizes that the *public trust* of the information is critical. Government agencies do not want to jeopardize our public trust by putting out bad data. Bad data will result in major ramifications not only from citizens but also from the government auditing arm, the General Accounting Office, and from Congress (Parker, 1999).

EPA Envirofacts Warehouse - The Envirofacts data warehouse comprises of information from 12 different environmental databases for facility information, including toxic chemical releases, water discharge permit compliance, hazardous waste handling processes, Superfund status, and air emission estimates. Each program office provides its own data and is responsible for maintaining this data. Initially, the Envirofacts warehouse architects noted some data integrity problems, namely, issues with accurate data, understandable data, properly linked data and standardized data. The architects had to work hard to address these key data issues so that the public can trust that the quality of data in the warehouse (Garvey, 2003).

U.S. Navy Type Commander Readiness Management System – The Navy uses a data warehouse to support the decisions of its commanding officers. Data at the lower unit levels is aggregated to the higher levels and then interfaced with other military systems for a joint military assessment of readiness as required by the Joint Chiefs of Staff. The Navy found that it was spending too much time to determine its readiness and some of its reports contained incorrect data. The Navy developed a user friendly, Web-based system that provides quick and accurate assessment of readiness data at all levels within the Navy. "The system collects, stores, reports and analyzes mission readiness data from air, sub and surface forces" for the Atlantic and Pacific Fleets. Although this effort was successful, the Navy learned that data originating from the lower levels still needs to be accurate. The reason is that a number of legacy systems, which serves as the source data for the warehouse, lacked validation functions (Microsoft, 2000).

Standardize the Organization's Data Definitions

A key attribute of a data warehouse is that it serves as "a single version of the truth." This is a significant improvement over the different and often conflicting versions of the truth that come from an environment of disparate silos of data. To achieve this singular version of the truth, there needs to be consistent definitions of data elements to afford the consolidation of common information across different data sources. These consistent data definitions are captured in a data warehouse's metadata repository.

DoD Computerized Executive Information System (CEIS) is a 4-terabyte data warehouse holds the medical records of the 8.5 million active members of the U.S. military health care system who are treated at 115 hospitals and 461 clinics around the world. The Defense Department wanted to convert its fixed-cost health care system to a managed-care model to lower costs and increase patient care for the active military, retirees and their dependents. Over 12,000 doctors, nurses and administrators use it. Frank Gillett, an analyst at Forrester Research, Inc., stated that, "What kills these huge data warehouse projects is that the human beings don't agree on the definition of data. Without that . . . all that $450 million [cost of the warehouse project] could be thrown out the window" (Hamblen, 1998).

Be Selective on What Data Elements to Include in the Warehouse

Users are unsure of what they want so they place an excessive number of data elements in the warehouse. This results in an immense, unwieldy warehouse in which query performance is impaired.

Federal Credit Union - The data warehouse architect for this organization suggests that users know which data they use most, although they will not always admit to what they use least (Deitch, 2000).