Enterprise Data Warehouse (EDW)

UC Berkeley
Peter Cava
Manager Data Warehouse Services
October 5, 2006
What is a Data Warehouse?

- A data warehouse is a *subject-oriented, integrated, time-varying, non-volatile* collection of data in support of the management's *decision-making process.*

  • ---Bill Inmon
What is a Data Warehouse?

A Practitioners Viewpoint

“A data warehouse is simply a single, complete, and consistent store of data obtained from a variety of sources and made available to end users in a way they can understand and use it in a business context.”

-- Barry Devlin, IBM Consultant
<table>
<thead>
<tr>
<th>Enterprise Data Warehouse</th>
<th>Traditional Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integrated Data</td>
<td>Application-specific Data</td>
</tr>
<tr>
<td>Current/Historical Data</td>
<td>Current Data</td>
</tr>
<tr>
<td>Organized by Subject</td>
<td>Organized for Data Entry</td>
</tr>
<tr>
<td>Non-Volatile Data</td>
<td>Updated Data</td>
</tr>
<tr>
<td>Denormalized Data</td>
<td>Normalized Data</td>
</tr>
<tr>
<td>Descriptive Data</td>
<td>Encoded Data</td>
</tr>
<tr>
<td>Detailed/Summarized Data</td>
<td>Raw Data</td>
</tr>
<tr>
<td>Knowledge user (Manager)</td>
<td>Clerical User</td>
</tr>
</tbody>
</table>
What does a data warehouse do?

- Integrate divergent information from various systems which enable users to quickly produce powerful ad-hoc queries and perform complex analysis.
- Create an infrastructure for reusing the data in numerous ways.
- Create an open systems environment to make useful information easily accessible to authorized users.
- Help managers make informed decisions.
What are the users saying

- Data should be integrated across the enterprise
- Summary data had a real value to the organization
- Historical data held the key to understanding data over time
- What-if capabilities are required
Heterogeneous Information Sources

“Heterogeneities are everywhere”

- Different interfaces
- Different data representations
- Duplicate and inconsistent information
- Vertical fragmentation of informational systems (vertical stove pipes)
- Result of application (user)-driven development of operational systems
EDW UC Berkeley

Goal: Unified Access to Data

Integration System

- Collects and combines information
- Provides integrated view, uniform user interface
- Supports sharing

World Wide Web

Digital Libraries

Scientific Databases

Personal Databases
Key Concepts

- Data Warehouse vs. Data Mart
  - Data Mart – Single subject area
  - Data Warehouse – Integrated Data Marts
- Star Schema
  - Dimensions
    - Hierarchies
    - Descriptive Attributes
  - Fact Tables
- Metadata
Star Schema Architecture

- **Dimension tables**
  - Textual descriptions of the dimensions of data
  - Represents the data that you slice and dice

- **Fact tables**
  - Place where numerical measurements of business are stored
Star Schema Data Model

- Transaction Date Dim (View)
- Activity Date Dim (View)
- Organization Dim
- Organization History Dim
- Home Organization Dim (View)
- Fund Dim
- Account Dim
- Audit Dim
- Program Dim
- Activity Dim
- Transaction Type Dim
- Financial Transaction Dim
- Grant Dim

Finance Encumbrance Transaction Facts

Grain: Each Operating Ledger Encumbrance Transaction for the lowest level FOPAAL.
Reporting

- **Business intelligence tools**: These are software applications that simplify the process of development and production of business reports based on data warehouse data.

- **Executive information systems** (known more widely as **Dashboard (business)**): These are software applications that are used to display complex business metrics and information in a graphical way to allow rapid understanding.

- **OLAP Tools**: OLAP tools form data into logical multi-dimensional structures and allow users to select which dimensions to view data by.

- **Data Mining**: Data mining tools are software that allow users to perform detailed mathematical and statistical calculations on detailed data warehouse data to detect trends, identify patterns and analyze data.
Strengths of OLAP

- It is a powerful visualization paradigm.
- It provides fast, interactive response times.
- It is good for analyzing time series.
- It can be useful to find some clusters and outliers.
Typical OLAP Query

- Write a multi-table join to compare Budget for each Org YTD this year vs. last year.

- Repeat the above process to find the top 5 Donors to Funds.

- Repeat the above process to find the Funds of a Org to new vs. existing Donor.
Nature of OLAP Analysis

- Aggregation Comparison
- Budget vs. Expenses
- Ranking
- Access to detailed and aggregate data
- Complex criteria specification
- Visualization
ANALYTICS

Gain rich insight into your business.

EDW UC Berkeley
EDW UC Berkeley

Dashboard

Compliance Dashboard

<table>
<thead>
<tr>
<th>Compliance Metrics</th>
<th>Process Management</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Financial</strong></td>
<td><strong>Consolidations</strong></td>
<td><strong>Internal Controls</strong></td>
</tr>
<tr>
<td>Accounts</td>
<td>Entity</td>
<td>15-Oct-2003</td>
</tr>
<tr>
<td>Accounts Reconciled with Standards</td>
<td>East Region</td>
<td>Internal Avoid Assessment of Internal Controls</td>
</tr>
<tr>
<td># September 30, 2003 Quarterly</td>
<td>West Region</td>
<td>Date of Most Recent Controls Review</td>
</tr>
<tr>
<td>External Reports</td>
<td>Status</td>
<td>15-Oct-2003</td>
</tr>
<tr>
<td>with Errors</td>
<td>Notes to FS</td>
<td>Date of Most Recent Controls Review</td>
</tr>
<tr>
<td># December 30, 2003 Monthly</td>
<td>Italy</td>
<td>15-Oct-2003</td>
</tr>
<tr>
<td>Deadlines Met</td>
<td>Japan</td>
<td>On-time Tax Filings</td>
</tr>
<tr>
<td>Close and Reporting</td>
<td>Singapore</td>
<td>On-time SEC Filings</td>
</tr>
<tr>
<td># December 30, 2003 Monthly</td>
<td>UK</td>
<td>On-time Regulatory Filings</td>
</tr>
<tr>
<td>Tax, Legal and Regulatory</td>
<td>Account</td>
<td>Sign-Off</td>
</tr>
<tr>
<td>Tax Returns</td>
<td>Cash</td>
<td>382 Assertions</td>
</tr>
<tr>
<td>4% December 12, 2003 Weekly</td>
<td>Investments</td>
<td>404 Assertions</td>
</tr>
<tr>
<td>Negative Findings</td>
<td>AR</td>
<td>VP Planning</td>
</tr>
<tr>
<td>3% October 11, 2003 Northly</td>
<td>Inventories</td>
<td>VP Treasury</td>
</tr>
<tr>
<td># of Complaints</td>
<td>Long Term Investments</td>
<td>VP Accounting</td>
</tr>
<tr>
<td>SSNMAC</td>
<td>Long Term AR</td>
<td>VP Operations</td>
</tr>
<tr>
<td># October 20, 2003 Monthly</td>
<td>Long Term Assets</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accumulated Depreciation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AR</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Other Short Term Liabilities</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Equity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Revenue</td>
<td></td>
</tr>
<tr>
<td></td>
<td>COGS</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Direct Costs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Depreciation &amp; Amortization</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Taxes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cash Flow</td>
<td></td>
</tr>
</tbody>
</table>
EDW Today

- BAIRSII Financial (Operational reporting)
  - Actuals
  - Tempbudg
  - Permbudg
  - Payroll
- BIS Financial (Analytical reporting)
  - Actuals
  - Tempbudg
  - Permbudg
  - Payroll
- HRMS (Operational reporting)
  - Workforce Administration
- HRMS (Analytical reporting)
  - Workforce Summary
- Foundation
- Student Pilot (In development)
Technical Environment

- Oracle 9i
- ETL Tool – Informatica
- Reporting Tool – Hyperion Brio
- Data Modeling Tool - ERwin
Informatica PowerCenter 7 Architecture

Sources (heterogenous)

Native/ODBC

PowerCenter Server

Native/ODBC

Targets (heterogenous)

Network Protocol

PowerCenter Clients:
- Repository Manager
- Designer
- Workflow Manager
- Workflow Monitor
- Rep Server Administrative Console

Repository Server / Repository Agent

Native

PowerCenter Repository
ETL (Extract, Transform, and Load)

- **Data Extract**
  - Get Data from source

- **Data Transformation**
  - Data Cleansing - Data Quality Assurance
  - Data Scrubbing - Removing errors and inconsistencies
  - Processing Calculations
  - Applying Business Rules
  - Changing Data Types
  - Making the Data More Readable
  - Replacing Codes with Actual Values
  - Summarizing the Data

- **Data Load**
  - Load data into Warehouse
Data Quality – The reality

- Tempting to think that all that is there to creating a data warehouse is extracting operational data and entering into a data warehouse.
- Warehouse data comes from disparate questionable sources.
- Legacy systems no longer documented.
- Outside sources with questionable quality procedures.
- Production systems with no built in integrity checks and no integration.
EDW moving forward

- Hub and Spoke Architecture
- Streamline the existing process
- Standardize ETL processes
- Standardize Table/View design’s
- Standardize Naming conventions
- Address meta data requirements
Integrated Data Subject Area

- Workforce analysis
- Financial analysis of operations.
- Financial analysis – research and projects
- Curriculum and enrollment analysis
- Student pipeline analysis (pre-registration to alumni)
- Graduate support analysis
- Private support analysis (donors and gifts)
- Purchasing analysis
- Recharge analysis
- Classroom and facilities utilization analysis
- Absenteeism analysis
- Enterprise risk assessment
Warehouse Success Criteria

- Cross-function, campus-wide support
- Consistency and integration across subject areas.
  - Integration mechanism should be explicit.
  - Design should embody subject-matter expertise.
  - Standardized data definitions.
- Security; access control
  - Access limited by role and context.
  - Controlled by data owners.
- Support for operational users
  - Standard reports, customizable as needed.
  - Good documentation at the point of use.
- Support flexible access for decision-makers
  - Slice and dice many different ways. Tool support for this.
  - Documentation of available data.
  - Enough history for trend analysis.
  - Reconcile alternative views of same information; e.g., as-reported vs. final.
Metadata

Data about Data
- Business Meta Data
- Technical Meta Data
- Operational Meta Data
- Project Meta Data
Some of the challenges

- Governance and Prioritization
- Decentralized reporting systems
- Meta Data management
- Resource availability
- Central vs. local reporting needs
- Funding – BAI vs. Campus enterprise