

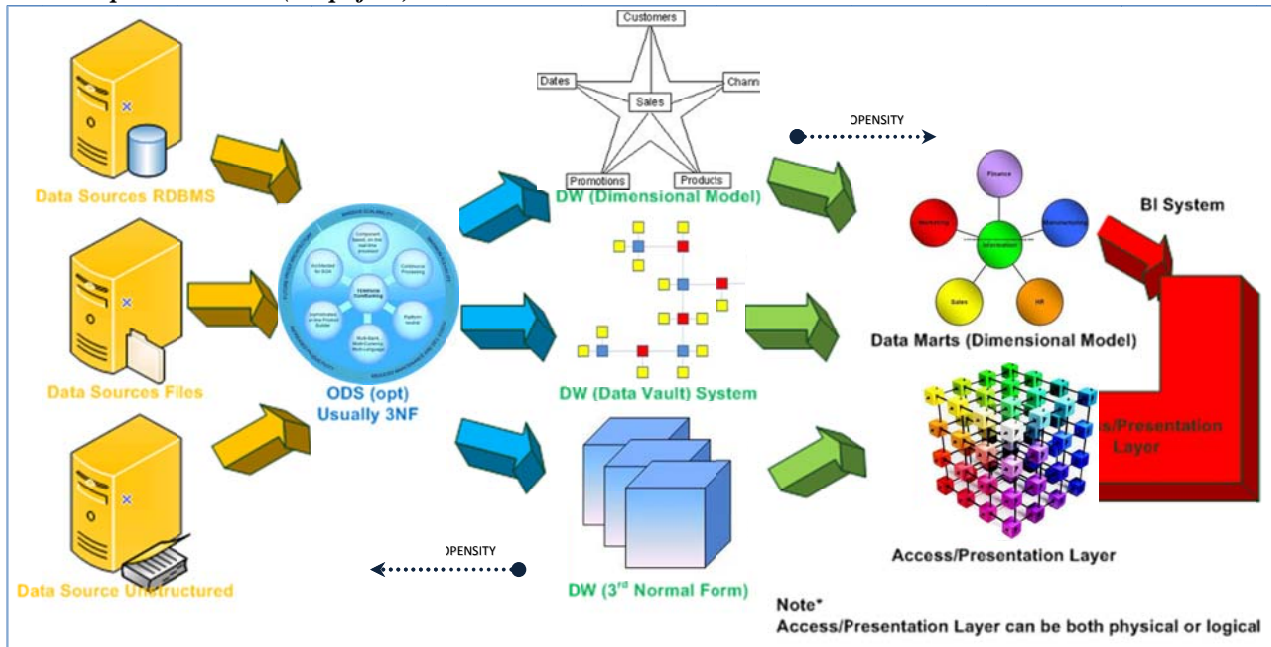
Data Vault case for EDW

By Raphael Klebanov, WhereScape, Inc.

Introduction

The purpose of this article is to review the traditional models of the Enterprise Data Warehouses in comparison to the Data Vault model as a method of developing an Enterprise Data Warehouse (EDW) project. Also discussed will be a real-world example where the Data Vault was chosen to replace a more “traditional” architecture.

Principal Data Flow (simplified)



Legend:

- ✓ **Data Source (DS)** is a model that represents one or many sets of data used by a DW application. DSs can be relational databases (DBs), a variety of file types (delimited, fixed-length, XML, etc), unstructured data (UD), and so on.
- ✓ **Operational Data Store (ODS)** is "...a subject-oriented, integrated, volatile, current-valued, detailed-only collection of data in support of an organization's need for up-to-the-second, operational, integrated, collective information." (*Bill Inmon*)
- ✓ **Data Warehouse (DW)** according to ...
 - Ralph Kimball : "...a copy of transaction data specifically structured for query and analysis" (*The Data Warehouse Toolkit*)
 - Bill Inmon: "...a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process" (*DW2.0*)
- ✓ **Data Mart** is a DB, or collection of DBs, designed to help business users make strategic decisions about their business. While a DW generally combines DBs across an entire enterprise, Data Marts are usually smaller and focus on a particular subject area or department. They are designed to answer specific business questions and are therefore limited in scope.
- ✓ **Access /Presentation Layer** is an optional physical or semantic structure that is situated between DW or Data Marts and BI system and maybe designed to accommodate specific needs for the BI tool such as snow-flaking for MicroStrategy.
- ✓ **Business Intelligence (BI)** is a system that refers to skills, processes, technologies, applications and practices used to support decision-making process in the enterprise.

Three Models for Designing an EDW

Since the pre-web times when informational trilobites roamed the cyberspace, there have been two famous competing options for designing data warehouse systems:

- one, according to Ralph Kimball, uses conformed dimensions and facts for composing the Enterprise Data Warehouse (EDW). This approach is generally known as the Dimensional Model (DM) or the Kimball Bus Architecture;
- and the other approach, proposed by W.H. (Bill) Inmon designs the data warehouse database in Third Normal Form (3NF) and uses data marts for end user access and reporting.

For an extended time, as architects we only had the option to pick between the two; let us look at them a bit closer.

1. Dimensional Model (DM)

Ralph Kimball's model: DW is the collection of all data marts within the enterprise. Information is always stored in the Dimensional Model.

In a DM, transaction data is separated into either "facts", which are generally numeric transaction data, or "dimensions", which are the reference information that gives meaning to the facts. E.g., sales transactions can be broken up into facts (measures) such as unit_sales_price, quantity, sales_value, etc. and into dimensions such as order_date, ship_date, customer, product, etc.

Key characteristics of a Dimensional Model approach are:

- ◇ clarity of the design to both developers and business folks;
- ◇ easy use for access/presentation layers due to design favorable to most BI tools;
- ◇ ability for business users to query the database directly.

The main disadvantages of the dimensional approach are:

- ◆ it is complicated to maintain the integrity of facts and dimensions, loading the DW with data from different source systems;
- ◆ expensive to modify the DW structure if the organization adopting the dimensional approach changes the way in which it does business;
- ◆ for conformed dimensions, you also have to scrub data - to conform it- and this is sometimes problematic due to requests of the DW to conform to the raw data to some standard that is not enforced in the various source systems.

All of these characteristics boil down to a main usage (propensity) of the DM in *data marts and access/presentation layers*. The DW can be created using this approach for small data volumes and stable business structures.

2. Third Normal Form (3NF)

Bill Inmon's model: a central DW is one component of the overall Business Intelligence system (BI), also referred to as the Corporate Information factory (CIF). An enterprise has one centralized EDW, and data marts obtain their information from the EDW. In the EDW, information is usually stored in 3NF (Codd's third normal form).

In the normalized approach, the data in the DW is stored by database normalization rules. Tables are grouped together in subject areas that reflect general data categories e.g., data on customers, products, etc.

The main characteristics of this approach are:

- ◇ It is more straightforward to add information into the database containing full historical data from the operational systems;
- ◇ the data structures are more resilient to change since data should only appear in one table (i.e., the data is normalized);
- ◇ due to optimized, normalized structure, NRT/RT- and VLDB-loading are supported in some cases.

Disadvantage of this approach is that, because of the number of tables involved

- ◆ it is difficult for user to join data from different sources into meaningful information and, subsequently,
- ◆ access the information without an exact understanding of the sources of data and of the data structure of the data warehouse (According to Bill Inmon, however, it is not recommended to directly access DW)

So all these characteristics lead up to realizing that main usage (propensity) if the 3NF model is *Operational Data Stores* rather than EDW.

Well, where is the "happy medium", a model that allows us to avoid pitfalls of the Dimensional and 3NF models? The answer is Dan Linstedt's evolutionary approach called Data Vault.

3. Data Vault (DV)

Data Vault is relatively young model (10 years or so) and is designed to avoid or minimize the impact of the issues related to DM and 3NF and to resolve drawbacks of both methods.

DV Modeling is a method of designing an EDW to provide historical storage of data coming in from many operational systems with complete tracing of the origin of all the data coming into the database. This method proved to be highly adaptable to change in the business environment. This is primarily achieved by taking the business organization structure and process flow as a starting point for the data model, since it is assumed that it will change less frequently than the operational systems used to support the business. (In some organizations, however, the business structure does change quite often)

Although DV is not always a “silver bullet”, it provides important advantages:

- ◇ Less complicated EDW loads resulting in greater stability and performance. The ETL/ELT processes can be effortlessly standardized.
- ◇ Improved flexibility allowing EDW to more easily adapt to changes in the business. Adding a new Unit of Work into existing EDW is rather trivial task.
- ◇ More suitability for incremental implementation (Agile DW) ensuing in quicker delivery of business value because the effort can be broken down to the smaller pieces for example, for Sprint model.
- ◇ Due to the highly granular nature of the DV model, it sustains Very Large Database (VLDB) capability resulting in no-need for redesign when EDW matures/grows.
- ◇ Design supports Real Time and/or Near Real Time loading making is suitable for Operational Data Warehouse(ODW) ;
- ◇ Great flexibility and adaptability of DV make is great spring board to future of the Data warehousing – ODW and Dynamic Data Warehouse (DDW)

The Stories

One of my customers, Independent Purchasing Cooperative (IPC), a procurement organization for the Subway restaurant chain, located in Miami, invited me into discussion about future of their EDW.

The topic of the discussion in the IPC was why the existing EDW does not work for decision-making folks. No, it was not busted; all the processes were intact populating series of Star Schemas on the daily basis. However, reports produced based on the data from this EDW were fragmented and inconclusive for the business. The IPC business evolved quite a bit since the EDW was complete and now requires the entirely new breadth of data for decision-making.

The IPC architecture of the EDW was “bottom-up” Dimensional Model, so common in the industry. It turned out to be rigid and inflexible to the evolving needs of the business. The scope of work adjusting it to the current needs of the company was becoming very expensive and time consuming... What to do? After some debates and commitments the decision has been made to re-build an EDW using the DV approach. Why Data Vault? The main considerations were:

- ✓ Existence of many, over a dozen, diverse DSs, both relational and flat files including a large segment of UD (unstructured data) in form of text blocks.
- ✓ Ability to capture all the data all the time, including Historical loads since 2003.
- ✓ Clear option of breaking down the scope of EDW by hierarchy of DS systems/Business Areas/Units of Work/individual hubs thus allowing creating truly Agile environment with daily stand-up meetings, burn down list of tasks, etc.
- ✓ Finite and rather short list of objects involved in EDW, such as Hubs/Satellites, SLinks and, in some cases, ancillary staging tables and PITs. This makes Modeling and, especially, ETL processes standardized and patterned.
- ✓ Ability to store all the changes on attributes in Satellites and associations in SLinks eradicates the need for SCDs (slowly changing dimensions) and SCF's (slowly changing facts) in the data marts.
- ✓ Full traceability of each bit of data back to the DSs establishes trust between the Business community and IT.
- ✓ Using DV, enables ability to take apart the data on basic “nuts and bolts”, put it in the right “compartments” and “assemble” data marts as needed cleanly and quickly according to the changing business rules.
- ✓ Apparent ability to accommodate the upcoming evolution of the EDW feeds, daily now, into hourly and even Near Real Time (NRT) frequencies.
- ✓ Ease to re-platforming the EDW to more volume-oriented engines, such as Teradata.

Briefly about the Environment:

DS Feeds: Variety of daily, weekly, monthly and ad-hoc from RDBMSs and flat files, some UD

EDW Platform: SQL Server 2005+. Projected volume of the EDW for 2010 is 4 to 5TB and growing 10-15% annually

Data Warehouse Builder: WhereScape RED 6

BI solution: Balanced Insight Consensus

BI Reporting: Microstrategy 9

The Phase I of the Data Vault EDW is completed (approx 500 objects) along with the Data Mart and BI reports. The Phase II has been developing now.

Conclusion:

In 2008 W.H. (Bill) Inmon stated that the "Data Vault is the optimal approach for modeling the EDW in the DW2.0 framework." (DW2.0).

As defined by Dan Linstedt, the creator of the method, the resulting database looks and feels like the following:
"The Data Vault is a detail oriented, historical tracking and uniquely linked set of normalized tables that support one or more functional areas of business. It is a hybrid approach encompassing the best of breed between 3rd normal form (3NF) and star schema. The design is flexible, scalable, consistent and adaptable to the needs of the enterprise"
(<http://www.tdan.com/view-articles/5054/>)

The number of Data Vault users surpassed 500 (<http://danlinstedt.com/about/dv-customers/>) and grows rapidly.