Into the Data Vault via Data Warehouse Automation

April 2017

A ThoughtPoint by
Dr. Barry Devlin, 9sight Consulting
barry@9sight.com

The data warehouse is increasingly at the heart of business operations and management. Automating its design and delivery is mandatory to succeed with ever-changing, real-time business needs as data size and speed grow exponentially. Together, Data Vault and WhereScape Data Vault Express offer a firm design foundation and a powerful system for agile delivery.

Taking the business in hand

Data warehouse—love or hate it, you can't avoid it. For the business, whether you needed a single version of the truth or simply wanted reliable numbers for the CEO, it was at the heart of management and control. For IT, whether you were building your first BI system or filling a data lake, a data warehouse was the basis for delivering data to the business.

But today, the data warehouse has evolved from "just" a massive collection of all analytic and BI data—that's now the data lake—to the repository of core business information: the complete, historically correct record of all the legally binding activities that comprise a business.

The data warehouse is becoming the ultimate enterprise-wide system of record at the heart of every business decision. It is at the core of data quality and consistency—the only place where all the keys and relationships are collected and maintained together. You can play all you like with data in the lake, but the warehouse means business—serious business.

> *Today's data warehouse is rapidly becoming the core repository of the complete, historically correct record of all the business' legally binding activities.*

And that means getting the data right. An internal model that creates consistency but maintains agility as business continually changes. An infrastructure that gathers and reconciles data from diverse, ever-shifting sources with ease-of-use, reliability, and ongoing dexterity.

Driving good design and agile delivery

After thirty years a-growing, the data warehouse needs a facelift—a more powerful internal design and a better engineered delivery system. Modern data warehouse *design* demands a careful trade off between ease of use, consistency, and continuing maintenance. Data *delivery* today requires a comprehensive, well-managed process, from the first mapping of legacy or other systems to keeping abreast of evolving business needs and data sources.

Data warehouse design

The first data warehouse designs by myself and later Bill Inmon emphasized reconciliation and consistency of data from multiple sources, using a loosely normalized or subject-oriented approach. This requires defining a substantial part of an enterprise data model up front, leading to slow delivery and painful maintenance. Ralph Kimball's star schema / dimensional model design focused first on specific departmental query needs, and later extended to enterprise level. This approach offers faster roll-out, but has been criticized for a lack of generality, maintenance issues, and its handling of temporal data.

The *Data Vault* approach, since the early 2000s, promises a much-improved balance, with a hybrid of the normalized and star schema forms above. Version 2.0 introduced in 2013, consisting of a data model, methodology, and systems architecture, provides a design basis for data warehouses that emphasizes core data quality, consistency, and agility.

Data delivery

The earliest warehouses were manually populated, but data delivery strives to increase the level of automation and reliability by using off-the-shelf tools. Initially, this was through standalone ETL (extract, transform, and load) tools, which quickly evolved into comprehensive data preparation and delivery systems, covering every aspect of the process. Over time, many of these systems came to be seen as overly complex and expensive.

In the early 2000s, a new class of delivery and preparation tools—Data Warehouse Automation (DWA)—emerged. Their goal was to simplify, automate and accelerate the development and maintenance of data warehouses, while assuring data quality and consistency. In contrast to ETL, DWA operates within the data warehouse environment, and is thus described as some variation of extract, load, and transform (ELT). *WhereScape* was a pioneer of DWA and is the foundation for delivery of data with ease of definition and ongoing agility in maintenance.

BUILDING TOMORROW'S DATA WAREHOUSE—TODAY

The focus here is on WhereScape's support for the Data Vault design approach. This combination of automated preparation and structured design offers new and previously unseen levels of speed, quality, and maintainability of data warehouses.

Automating the Data Vault with WhereScape

In July 2016, WhereScape announced a strategic partnership with Dan Linstedt based on the Data Vault 2.0 model, methodology and best delivery practices. Since then, WhereScape has rolled out functionality in its products—3D and RED—offering easy-to-use automated design, code generation, documentation and management of data warehouses based on Data Vault 2.0. The latest iteration of this support is called WhereScape Data Vault Express.

Linstedt defines the Data Vault Model as a detail-oriented, historical tracking, and specially linked set of normalized tables that support one or more functional business areas. This design offers a flexible, scalable, consistent, and adaptable way to address enterprise-wide needs. The model consists of three specialized types of entities/tables: *hubs* based on rarely changed business keys, *links* that describe associations or transactions between business keys, and *satellites* that hold all temporal and descriptive attributes of business keys and their associations. A (Raw) Data Vault is not used directly by business users, but is the agreed source for a wide range of business-facing, user-specified data marts in any required format on any platform. Power users may access Business Data Vault objects such as Point-in-Time or Bridge tables that contain common derived data.

> *Data warehouse automation is a game changer for easier, faster, and more consistent and reliable implementation of a Data Vault.*

The Data Vault associated methodology and best delivery practices provide the basis for avoiding design errors and supporting proper auditing. Of course, this process could be undertaken manually. However, applying a powerful automation tool is a game changer: easier, faster and more consistent and reliable progress from first thoughts to final decommissioning.

A warehouse based on Data Vault begins with a model—conceptual, logical and physical—taking us from business needs and terminology through a uniquely structured data model to a design optimized for a specific database platform. With WhereScape 3D, business users and designers together explore business needs and discover applicable data sources to define a logical and generic physical model using a predefined, expert template of Data Vault hubs, links and satellites.

From detailed design, through build and operation, to ongoing maintenance, WhereScape RED provides the tools, platform, and infrastructure for rapid, automated delivery. While RED supports a range of architectural approaches, support for a Data Vault implementation has been rapidly extended over the past nine months.

In collaboration with business users, RED collects and manages the complete set of metadata required to build and operate the Data Vault, based on the discovery and design work done in 3D. It uses this metadata to automatically generate the specific SQL optimized for the deployment platform (including SQL Server, Teradata and DB2 among others) to build, populate and maintain hubs, links, and satellites, and most recently, to automate the latest development in Data Vault, including an insert-only architecture.

In essence, 3D generates the logical and physical models and metadata required to structure and populate the Data Vault, models and metadata that RED uses in the context of building, operating, and maintaining the specific technology platform on which the data warehouse and marts will run. The approach eliminates the need for a specialized ETL platform and minimizes the amount of manual effort required.

> *"Using WhereScape RED ... Vodafone now presents new sources to business users in days and then reverts back to build / generate the best possible Data Vault to support this data mart."*
>
> *Jos Driessen, Manager, BICC Strategy and Architecture, Vodafone NL*

Conclusions

As the data warehouse evolves into a complete, enterprise-wide, historically correct record of all the business' legally binding activities, design and delivery approaches that sufficed in the past will no longer work. We need a robust yet agile design and highly automated, repeatable delivery. The combination of Data Vault and WhereScape Data Vault Express offers a best-of-breed and cost effective approach to meeting these needs.

Dr. Barry Devlin is among the foremost authorities on business insight and one of the founders of data warehousing, having published the first architectural paper on the topic in 1988. With over 30 years of IT experience, including 20 years as an IBM Distinguished Engineer, he is a widely respected analyst, consultant, lecturer and author of the seminal book, "Data Warehouse—from Architecture to Implementation" and numerous White Papers. His latest book "Business unIntelligence—Insight and Innovation Beyond Analytics and Big Data" was published in October 2013.

Barry is founder and principal of 9sight Consulting. He specializes in the human, organizational and IT implications of deep business insight solutions that combine operational, informational and collaborative environments. A regular blogger, writer and commentator on information and its use, Barry is based in Cape Town, South Africa and operates worldwide.