



Meeting the Six Data Vault Challenges

Data warehouse automation to the rescue

AUGUST 2017

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

Implementing a data warehouse has long involved treading a narrow line between data quality and business agility. In the early years of the new millennium, two separate approaches emerged that promised to square this circle: The Data Vault offered a new data model optimized for data warehouse agility, while Data Warehouse Automation provided an environment for faster and more reliable development.

While Data Vault has continued to gain supporters and now offers a development methodology, implementation challenges remain. This paper outlines six common issues that businesses have encountered in the areas of initial acceptance, development, and ongoing evolution of a Data Vault-base data warehouse. We further show how these challenges are addressed by Data Warehouse Automation and, in particular, by the function offered by WhereScape® Data Vault Express.

Contents

- 2** Breaking into a Data Vault
- 4** Locking in a Data Vault
- 5** Living in a Data Vault
- 7** Conclusions

Data modelers love complexity! Or so you'd think when reading a full wall-sized enterprise data model, much of which is in 8-point print. Truth is: they don't. Business and its data is complex. It's a problem that demands solutions.

The complexity of commerce and its data has long proven troublesome to the IT people tasked with building systems to run and measure the business. Business people barely notice the intricacy of documenting what they do, leading to the fog of misunderstanding that seeps over every "requirements" discussion between business and IT. When the mist disperses, IT deploys its engineering—data models and software processes—to build solutions and, in time, maintain and upgrade them.

The latest data model engineering is the Data Vault, conceived by Dan Linstedt in the early 2000s and now at version 2. This hybrid of third normal (3NF) and star-schema forms (see my previous ThoughtPoint¹) offers significant benefits and interesting challenges. On the plus side, it promises agility in the face of the all-too-common changes in business needs, separating ingestion concerns from those of business use, and promotes data quality best practices. On the other hand, its structure is complex and involves often thought-provoking design choices. Data Vault offers a methodology to counter these difficulties but, like all such practices, it depends on implementers following it thoroughly and consistently. That may fly in the face of tight schedules and limited budgets.

To fully achieve the benefits promised and minimize the challenges encountered, engineering software is essential. Data warehouse automation (DWA) software, also dating from the early 2000s, automates the entire process of designing, building, operating and maintaining a data warehouse². WhereScape® Data Vault Express applies this engineering to the specific case of a Data Vault-based warehouse.

This paper explores six key challenges faced when using a largely manual approach to Data Vault and shows how WhereScape® Data Vault Express helps overcome them.

Breaking into a Data Vault

Data warehousing isn't new. Many business and IT people have long experience in how it's done. Data Vault requires new thinking and approaches from both parties. However, resetting old behavior patterns can be hard.

Much of the challenge of implementing a Data Vault falls to the data warehouse (DW) development team, but the well-known challenges of the business-IT relationship do need to be addressed up front. Of course, these issues are not unique to a Data Vault implementation. However, the very nature of the Data Vault makes it much more important to make sure that business and IT remain in the "romantic phase".

The rift between business and IT

The business-IT gap is as old as IT itself. You might say business is from Venus and IT from Mars. Business needs excitement and results; IT embraces structure and quality. This rift explicitly challenges Data Vault projects. Success depends on representing innovative, changeable business needs in a highly structured and technically complex environment.

CHALLENGE ONE

While the original Data Vault was “simply” a model structured with agility and flexibility in mind, version 2.0 adds a comprehensive methodology—consistent, repeatable and pattern-based. This methodology is built upon best practice software engineering, including a focus on CMMI (Capability Maturity Model Integration) and Agile Scrum techniques, to ensure that the agility offered by the model is delivered by its developers.

Admirable as this is, an overly engineering mindset in IT may alienate business people. As DW staff focus on implementing new enterprise-focused modeling techniques and best engineering practices in project management, they may have to reduce face-to-face time with the business, leading to poorer, less detailed, or delayed understanding of specific project needs. Creating the enterprise-level Data Vault may delay delivery of specific business solutions. As the rift between business and IT widens, the business looks elsewhere for quick-fix solutions.

Crossing the rift

Models and methods must fade into the background with a focus on seamless collaboration between business and IT to meet functional business needs and tight delivery timeframes. WhereScape Data Vault Express emphasizes joint design and development between business and IT, offering an integrated environment for collaboration and the ability to immediately prototype design decisions with real data.

Principles can be perilous

Data Vault proposes new principles for DW owners, users and designers from traditional business intelligence backgrounds. Challenging old beliefs and changing behaviors is vital for a successful Data Vault project.

Data Vault principles of agility and auditability start from a clean separation between ingestion of data from sources and creation/consumption of information by the business. Between these two lies a specialized data structure based on semantic meaning, data type, and rate of change. The goal of providing an auditable trail from data source to information consumer means that cleansing of data between sources and the warehouse is strongly discouraged. While accepted as the intrinsically correct approach since the earliest days of data warehousing, IT has long known that operational data is often dirty—incorrect in detail and inconsistent across sources.

However, business has conflicting needs and views. Achieving cross-functional consistency may indicate errors in existing operational reports which the business uses. Operational systems may be in error, although the business is unwilling to invest in fixing them. In an effort to accommodate these and other needs, such as the business demand for quick delivery, IT has often compromised: correcting and massaging data on its way into the warehouse, leading to complex and largely unmaintainable ingestion routines.

Dealing with principles first

The first step in adhering to Data Vault principles is to understand the source systems, their structures, relationships, and underlying data quality. This can be a time-consuming task, but it is necessary to validate the model design and implementation approach. Automated data and structure discovery, data quality profiling, and design, as provided by WhereScape Data Vault Express will reduce the time spent in design of the Data Vault and its population processes. Business and IT can collaborate in very compressed time-

CHALLENGE TWO

windows to iterate on model design and validate using live data—eliminating assumptions, enabling the model to be validated before deployment, and ensuring that the data warehouse can evolve at the pace needed by the business.

Locking in a Data Vault

Like most of us, data warehouse developers are creatures of habit. Data Vault introduces new ways of populating and structuring the warehouse. Designers and developers need new rules and new tools.

Of course, if your developers are from the A-team, anything is possible. But, for most IT shops, developers need help in moving to a new way of working. Training in a new methodology will certainly be needed, but in the modern data warehouse world, automated development tools can take most of the strain. And enforce good practice in the process.

A novel technical environment

Most DW developers come from substantial experience of implementing a 3NF or star-schema approach to data warehousing. While Data Vault is designed to offer the best of both these worlds, it also differs from them in significant ways.

A Data Vault does not enforce relationships in the database, although it contains many joins between tables. Using specialized keys during the population phase replaces this traditional model-level approach to increase agility in the warehouse and allow for future changes, as well as allowing the ingestion of data into multiple tables in parallel.

The approach is also predicated on the use of “deltas”—specialized record types containing effective date/time stamps as a way of tracking changes in the warehouse. Given that many operational sources do not provide such deltas, they must be computed at load time by comparing new and existing states.

Data Vault design patterns promote the concept of splitting source data (tables/files) into multiple target objects (usually satellite objects) within the Data Vault, for example to reduce the amount of data being stored based on rate-of-change, or to separate attributes based on business categorization. This generally results in a significant increase in the total number of target objects being constructed and associated processing code blocks. Manually building and managing such volumes of objects—although often individually simple—is very time-consuming, tedious, and error-prone.

Mastering a novel technical environment

All these technical aspects can and should be automated to lessen their impact on Data Vault development projects. WhereScape Data Vault Express provides specialized automation functions for this environment. For example, it auto-generates the correct syntax and associated processing logic for hash and change keys, speeding development and eliminating the many opportunities for human error in manual development.

While the Data Vault approach leads to a significant increase in the number of data and processing objects that must be created and managed, these objects are very compatible with pattern-based development. Software developers are certainly capable of devising

CHALLENGE THREE

and working with templates, but the WhereScape Data Vault Express approach of generating the schema, database objects, and population code in an integrated development environment based on metadata is better, faster, and more reliable, allowing developers to focus on iterative design and development with business users.

Memory loss and inconsistency in development

CHALLENGE FOUR

Give developers rule books and they'll soon put them to good use—raising their screens to the optimum height. The Data Vault model and methodology requires adherence to an extensive set of rules and even recommendations. In this respect, it differs little from any development methodology. And data warehousing teams differ little from most developers: standards are for others, habitual coding allegedly leads to fastest delivery, and, as for writing documentation—well—tomorrow is another day.

Data Vault's data objects—from common hubs, links and satellites to the lesser known PIT (point in time) and Bridge helper tables—must all adhere to very specific standards and definition rules to make the Vault work as intended and to ensure its agility and ongoing ease of maintenance. When developers “re-invent”, these structures, through ignorance, prior experience, or the best of intentions, problems will arise and rework will be demanded, both in the initial build and—often costlier—in ongoing maintenance.

Aligning the development team

Whether sharing tasks between diverse teams or onboarding new team members, strict observance of standards of design, documentation, error handling, and auditing is best practice. The built-in best practices and methodologies within DWA software enforce consistency across the entire data warehouse. WhereScape Data Vault Express provides built-in best practices and standards, while allowing template modification if required.

By eliminating the idiosyncrasies of each developer's coding style, the generated code looks and feels the same across the team and adheres to the same naming standards, easing maintenance and upgrades as well as on-boarding of new developers. With all the logic and associated documentation captured in metadata, the risk of having tribal knowledge leave when a developer changes jobs or leaves is reduced or eliminated. This provides the best and most cost-effective approach to achieving consistency and repeatability in Data Vault implementation.

Living in a Data Vault

Perhaps the most under-appreciated challenge for a data warehouse team—at least, until you have encountered it—is that of ongoing operation, maintenance and upgrade of the environment. Prepare for it now.

As in the case of the previous two topics, manual procedures are not going to win much favor with the business—they cannot keep up with the need for 24/7 operation and evolving business requirements. However, the automation work done in the preparation and build phases will stand you in good stead when the rubber hits the road.

Ongoing operational predicaments

CHALLENGE FIVE

Data Vaults, like data warehouses, need ongoing operations overhead to schedule, execute, and monitor the data feeds—including handling failed jobs, restarts, all the while ensuring everything is processed in the correct order. The challenge for Data Vaults is the added complexity of scheduling and management resulting from the numerous data and processing objects involved. Manual approaches—from design and development, through to operations itself—are inadequate to address this.

A particular challenge is that in manual development the necessary logging and auditing capabilities are often sidelined when projects fall behind schedule.

Automated operations

DWA software, driven by metadata, can easily handle the execution, dependencies, and related failure processing (including restarts) required by operations. WhereScape Data Vault Express automatically generates all necessary error handling, logging and auditing code, to ease ongoing operations and management tasks. Built-in scheduler and job management capabilities allow the definition of job execution, taking dependencies into account—ensuring that objects are processed in the correct order and allowing unaffected objects to continue processing, while halting dependent objects and picking up where they left off upon re-start.

Supporting non-stop, high-speed business change

CHALLENGE SIX

Business needs change—constantly, rapidly, and unpredictably—even before the first DW iteration has been rolled out to the business. A key driver of the Data Vault model and methodology is to ease the problems associated with such ongoing change.

At a practical, implementation level, managing change requires the ability to carry out extensive and effective impact analysis. What tables and columns will be affected by changing this code? What are the unintended, down-stream consequences? How can we reduce risk and, simultaneously, expedite necessary change? Documentation is supposed to provide answers, but the reality is that manual approaches seldom deliver the required documents and even if they do at first, they will never keep them updated as the environment evolves.

A metadata-driven development environment

The real answer to all these questions lies in the automation of standards, documentation, and so on. With these items based on and managed in metadata, change management can be automated and in large part reduced to review rather than decoding ancient programming. Such metadata-driven automation is key to keeping pace with the ever more rapidly changing needs of the business.

WhereScape Data Vault Express is a complete metadata-driven development environment. This metadata describes all the processing logic, ordering and potential impacts when changes are made. Documentation is automatically maintained via the metadata to keep both user and operations teams up-to-date with the current implementation, improving dependency management and operations. Risk is substantially reduced, as the downstream effects of any changes can be easily anticipated.

Conclusions

The real-world complexity of business data and the changeability of its usage demands adoption of the Data Vault model. This model requires another step forward in project management: embracing Data Warehouse Automation to navigate the challenges of successful and timely implementation.

Over more than a decade and a half, businesses have been gradually adopting the Data Vault model as a new foundation for their data warehouses. Its design and approach has been instrumental in successfully addressing the growing need for flexibility and agility in business analytics and decision-making support. However, many companies have found that the structural complexity of the model can be a challenge to the IT teams charged with its implementation.

We observe these challenges to span all aspects of implementation—from the relationship between business and IT, through the design and build process, to the longer-term operations and maintenance/upgrade phases.

The solution to the six challenges identified can be found in Data Warehouse Automation software tools that improve collaboration between business and IT, reduce manual labor, increase standardization, and produce lasting and reliable documentation across all aspects of implementation. WhereScape’s metadata repository-based Data Vault Express has been specifically designed to automate the process of implementing a Data Vault model at the scale and speed demanded by today’s data-driven business.

*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 with IBM as a 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 book, “**Business unIntelligence—Insight and Innovation Beyond Analytics and Big Data**” (<http://bit.ly/Bunl-TP2>) was published in October 2013.*



Barry is founder and principal of 9sight Consulting. He specializes in the human, organizational and technological implications of deep business insight solutions combining all aspects of internally and externally sourced information, analytics, and artificial intelligence. A regular contributor to Twitter (@BarryDevlin), TDWI, BACollaborative, and more, Barry is based in Cape Town, South Africa and operates worldwide.

Brand and product names mentioned in this paper are trademarks or registered trademarks of WhereScape and other companies.

¹ Devlin, B. “Into the Data Vault via Data Warehouse Automation”, April 2017, <http://bit.ly/2tPzYpl>

² Devlin, B. Four-part blog series, July 2017, <https://www.wherescape.com/blog/blog-posts/2017/july/>