Data Warehouse Automation

A Decision Guide

A White Paper by Dave Wells

Infocentric LLC

Table of Contents

Data Warehouse Automation: A Decision Guide

Data Warehouse Automation is an emerging class of data warehousing tools that use technology to gain efficiencies and improve effectiveness in data warehousing processes. Data warehouse automation is more than automation of warehouse design and build processes. It encompasses the entire data warehousing lifecycle from planning, analysis, and design through development and extending into operations, maintenance, change management and documentation.

Adoption of data warehouse automation changes the way that we think about building data warehouses. The widely accepted best practice of extensive up-front analysis, design, and modeling can be left behind as the mindset changes from "get it right the first time" to "develop fast and develop frequently" – an approach that is aligned with today's agile development practices and that enables requirements to be fluid.

Seven Myths of Data Warehouse Automation

A surface-only look at data warehouse automation can be deceptive. Some misconceptions captured from an informal survey of people who have little or no understanding of data warehouse automation include:

Automating ETL only: Many assume that automation simply means generating ETL code. Robust automation tools do much more than simple ETL generation. The functions span development processes from requirements gathering to deployment, and extend well beyond development.

Automating the warehouse build: Another common misconception is that automation is limited to building the data warehouse—generating ETL code and database schema. Data Warehouse Automation is much more than simply automating the development process. It encompasses all of the core processes of data warehousing including design, development, testing, deployment, operations, impact analysis, change management and documentation.

Build your own is practical: Some organizations have already built their own processes to generate ETL and database schema, believing that they have achieved the goals of data warehouse automation. Homegrown solutions typically lack important functions such as test automation, metadata management, and documentation generation. They rarely extend to include functions for scheduling, operations, validation, and change management, and they struggle to handle complex transformations without hand coding. Investing time, talent, staff, and funding to custom build a comprehensive data warehouse automation tool is a poor investment of resources better used to respond to pressing business needs.

Only for relational databases: The belief that warehouse automation is limited to data warehouses implemented using RDBMS is quite common. In fact, automation tools support relational databases, multi-dimensional databases, columnar databases, in-memory databases, and much more.

Not suited for big data: Some believe that automation will become a barrier when making the move to big data. This misconception simply is not true. Just as automation extends beyond RDBMS to other database types, full-featured automation tools are also able to work with Hadoop, NoSQL databases and cloud-hosted data. Furthermore, automation tools work quite well with data virtualization technology, making even the least traditional data sources accessible and practical.

High cost and late payback: Dismissing automation as out-of-reach due to cost is a mistake. Cost of entry for automation tools is actually quite low, especially when compared to cost of staffing. Automation also eliminates much of the cost of bringing new resources—internal or consulting—up to speed on tribal knowledge, business processes, data sources, business rules, etc. Automation enables your existing staff that is already armed with this knowledge to be more effective. Perhaps more importantly, time-to-payback is exceptionally quick for data warehouse automation. Some tool providers point to examples where the initial cost was entirely recovered with completion of the first project.

Automation replaces people and reduces headcount: This misconception is a common cause of resistance to automation whether for data warehousing or any other endeavor. Time and again, history has shown that automation remove the burden of mundane and repetitive work and creates opportunities for people to do more interesting, creative, and value producing work. Data warehouse automation offers opportunity to redirect talent to the kinds of work that truly make a difference for the business. This is a substantial benefit. What data warehousing team doesn't have more work than time and resources to complete the work?

Why Automate Data Warehousing?

The corporate experience, almost without exception, is that data warehouses are necessary but painful. They take too long and cost too much to build, are out of sync with requirements by the time they are deployed, and are difficult to grow and change after deployment.

Many technologists and thought leaders are ready to declare the data warehouse dead – no longer relevant in the age of big data. But these prognosticators are mistaken, perhaps misled by seeking an easy answer to the pain of data warehousing. Big data can extend and enrich a data warehouse, but cannot replace it. The data warehouse integrates critical and valuable enterprise data—data that is not found in big data sources and that continues to be the primary data resource for descriptive, prescriptive, and decision analytics. It serves as corporate memory, collecting the body of history that makes time-series and trend analysis possible. Equally important, the data warehouse organizes and structures data to make it understandable and useful for consumption by many different business stakeholders.

We do need data warehouses, and will continue to need them for the foreseeable future. Big data, discovery tools, and advanced visualization tools do not eliminate the need to integrate enterprise data and maintain enterprise history. But we need data warehouses that can be built quickly and at reasonable cost, that readily adapt to changing requirements, and that are responsive to business and technical change ... all without compromising solution quality.

Data Warehouse Quality: Data warehouse automation delivers quality and effectiveness through ability to build better solutions. Better solutions are those that best meet real business requirements, and it is especially difficult to get complete and correct requirements when limited to an early phase of a linear development process. With data warehouse automation the business can make changes much later in the development process and change can occur more frequently with less disruption, waste, and rework. Iterative requirements discovery, however, is only one aspect of data warehouse quality. Automation brings quality benefits through standards enforcement and standardizing the development processes.

Business Agility: Ability to change fast and frequently extends beyond the warehouse development process. Changes that occur in business requirements can be met with quick response. Responding to change in real time and without the delay of lengthy projects is the essence of business agility.

Fast Development and Fast Change: Speed is the critical factor that enables agility both for agile business and for agile development. Ability to generate quickly and to regenerate equally fast when change occurs are fundamental automation capabilities. The ability to fail fast is also important. Sometimes warehousing teams can't deliver what the business needs due to data unavailability, data quality issues, or elusive and difficult to define business rules. Discovering these issues as early as possible reduces waste of time and resources.

Cost Savings: Ultimately building better, building faster, and changing quickly when needed bring substantial cost savings to data warehouse development, operation, maintenance, and evolution.

Sustainability, Maintainability, and Operability: Beyond developing and changing a data warehouse, automation offers many technical benefits that contribute to extended lifespan and ease of operations for the warehouse. Consistency of components in a data warehouse is improved through ability to build in standards and conventions. Automated documentation capabilities ensure comprehensive documentation that stays in sync with the implementation. Impact analysis for planned changes is supported with extensive metadata capabilities. Testing is simplified with test automation both during development and as a validation capability of operations processes. Maintenance becomes easier with improved consistency, better documentation, simplified testing, version control, automated implementation of changes and standardized deployment methodology.

The Basis of Data Warehouse Automation

Design patterns are fundamental to data warehouse automation. Identifying and reusing patterns is central to the capabilities to achieve consistency, quality, speed, agility, and cost savings simultaneously. Design patterns encapsulate architectural standards as well as best practices for data design, data management, data integration, and data usage. Figure 1 illustrates common design patterns in all of these categories. Applied patterns in a data warehouse automation tool support the goals of accelerated design and development, but equally importantly they drive compliance with standards and consistency of data warehousing results. Common data warehousing design patterns include:

Architectural Patterns such as the hub-and-spoke architecture popularized by Bill Inmon, the bus architecture championed by Ralph Kimball, the data vault architecture, and hybrid architectures defined by many pragmatic and in-the-trenches data warehousing professionals.

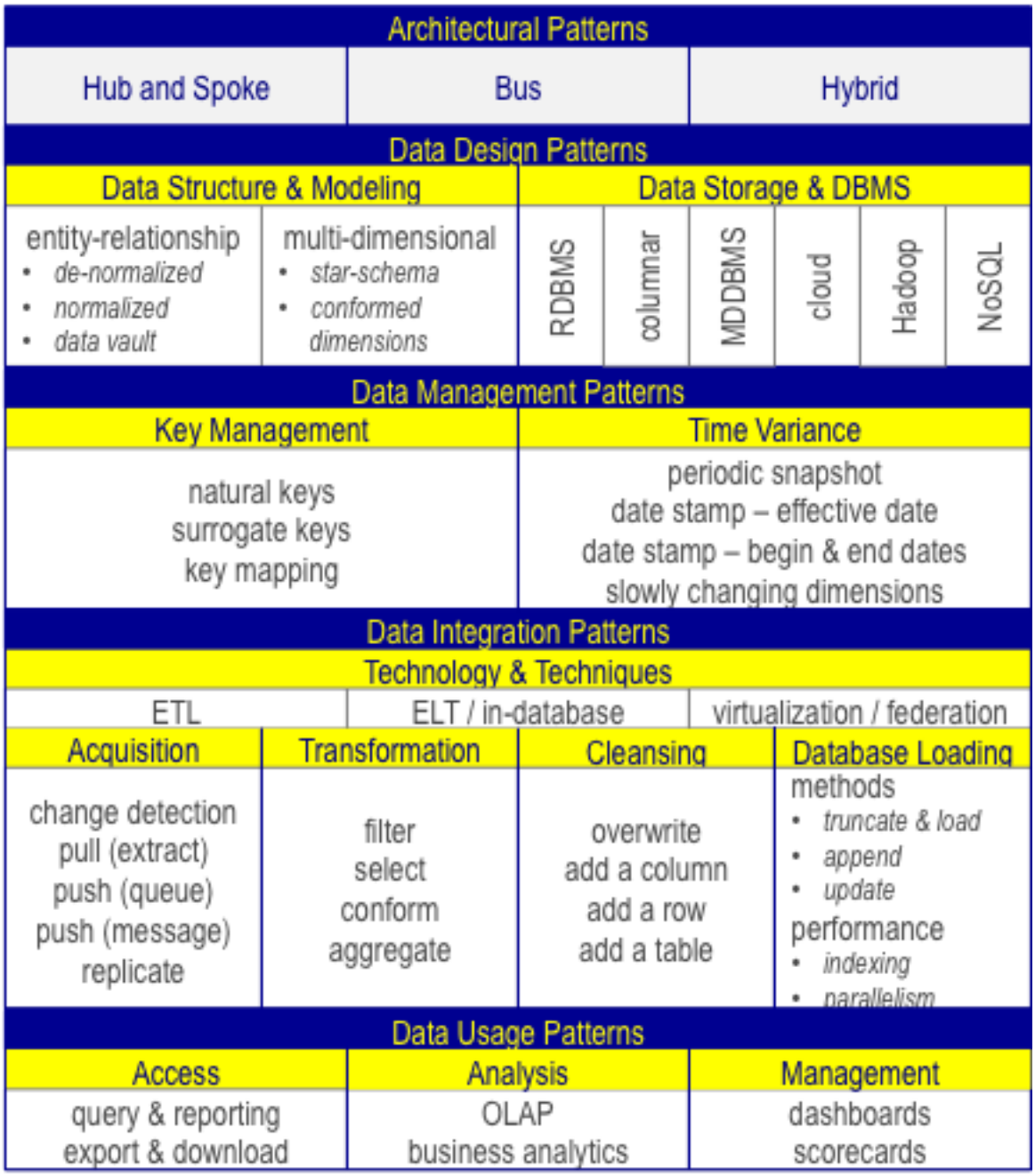


Figure 1: Data Warehouse Design Patterns

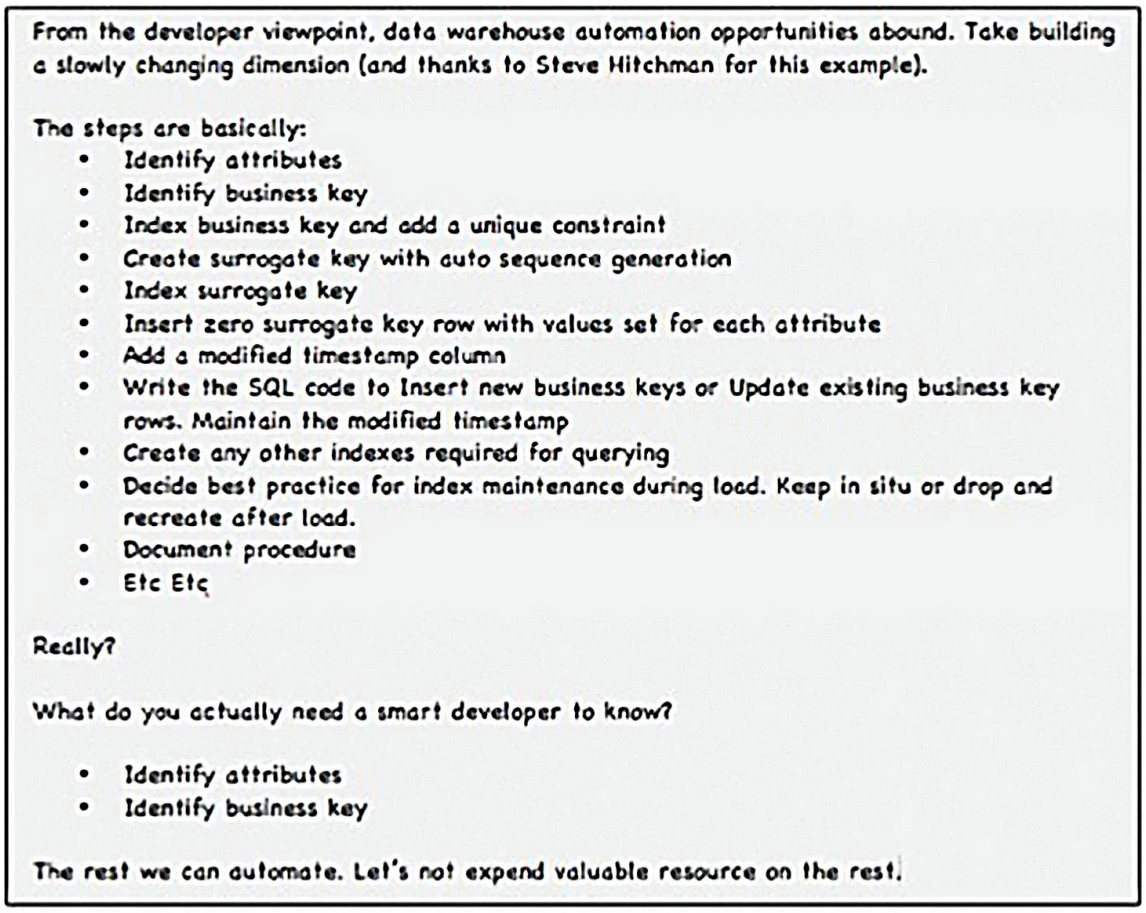
Data Design Patterns including those for data structures and modeling, both entity-relationship based and multi-dimensional. Data design patterns also include those for data storage – relational, columnar, multi-dimensional, cloud, Hadoop, and NoSQL.

Data Management Patterns such as key management patterns for natural keys, surrogate keys, and key mapping. Time variance is a substantial part of warehouse data management where patterns include snapshots, date stamps, and the various types of slowly changing dimensions.

Data Integration Patterns encompass those for technology and tools, for data acquisition transformation, and cleansing, and for database loading. Data integration – the heart of data warehousing – has an abundance of patterns with data movement patterns ranging from ETL to data virtualization, data acquisition patterns using push, pull, and messaging techniques, data transformation and cleansing patterns, database loading patterns, and much more.

Data Usage Patterns are also important considerations for full-lifecycle data warehouse automation. Patterns ranging from simple data access to performance management with dashboards and scorecards, and extending to analysis and advanced analytics may all occur simultaneously as essential considerations for data warehouse implementation, operation, and change.

The power of design patterns for data warehouse automation becomes clear with an example. The simple example shown in Figure 2 is taken from a blog posting by Michael Whitehead, CEO and Co-Founder of WhereScape.


From the developer viewpoint, data warehouse automation opportunities abound. Take building a slowly changing dimension (and thanks to Steve Hitchman for this example).

The steps are basically:

- Identify attributes
- Identify business key
- Index business key and add a unique constraint
- Create surrogate key with auto sequence generation
- Index surrogate key
- Insert zero surrogate key row with values set for each attribute
- Add a modified timestamp column
- Write the SQL code to Insert new business keys or Update existing business key rows. Maintain the modified timestamp
- Create any other indexes required for querying
- Decide best practice for index maintenance during load. Keep in situ or drop and recreate after load.
- Document procedure
- Etc Etc

Really?

What do you actually need a smart developer to know?

- Identify attributes
- Identify business key

The rest we can automate. Let's not expend valuable resource on the rest.

Figure 2: Design Pattern Example

To Automate or Not to Automate?

Although data warehouse automation is powerful, the decision to automate should not be taken lightly. Moving to automation brings change, and change inevitably brings resistance. A measured and thoughtful decision process helps to facilitate the necessary changes and minimize the risks of resistance to change. Give careful thought to the fit of data warehouse automation into your data warehousing program – architectural fit, methodological fit, and cultural fit. This guide offers twelve criteria (and a decision tool) to assist that process.

Data Warehouse Architecture: Is your architecture based in best practices or proprietary and specialized to your organization and technology platforms?

Requirements Gathering: Do you find requirements through user stories and discovery processes or using a waterfall approach of gathering business requirements, functional requirements, and technical requirements for stakeholder signoff?

Requirements Volatility: Do you experience frequent changes to requirements including regular change throughout the development process, or do you succeed in gathering requirements that are highly stable with change of requirements being a rare occurrence?

Time to Delivery: Do your business stakeholders expect fast and frequent delivery of data warehousing and business intelligence capabilities, or is a slow and deliberate approach to development projects the norm in your organization?

Project Risk: Do your data warehousing projects experience a high level of risk from poor data quality, lack of source data knowledge, insufficient budget, understaffing, scope creep, and other factors? Or are your projects relatively free of data, technology, funding, and staffing risks?

Project Backlog: Do you have an outstanding list of projects in waiting with a pattern of new projects being added to the backlog faster than older projects can be completed? Do you experience competing and conflicting priorities for project funding and staffing? Or is your data warehousing team able to keep pace with the demand for projects with little or no backlog?

Warehouse Operations: Are the processes and procedures for operation of your data warehouse detailed, time consuming, and labor intensive? Or are they relatively easy and painless even when errors occur and things don't go as planned?

Warehouse Documentation: Is the documentation for your data warehouse processes and databases sparse, dated, and out-of-sync with the implementation? Or do you have comprehensive and up-to-date documentation for all facets of the data warehouse?

Warehouse Testing: Do you have consistent, reliable, and repeatable process for data warehouse testing including unit, stream, and integration testing during development and validation testing during operation? Or are your testing plans inconsistent, uncertain, and routinely handcrafted?

Warehousing Organization Culture: Is your data warehousing team oriented to teamwork and collaboration, or is warehousing success driven primarily by individual talents and heroic efforts? Is the IT relationship with business stakeholders collaborative or contentious?

Warehousing Future: Are big data, cloud hosted data, and analytics in the cloud among current expectations for your data warehousing program? Are they on the horizon in the foreseeable future? Or is your program relatively stable and static with little disruption expected from emerging technologies?

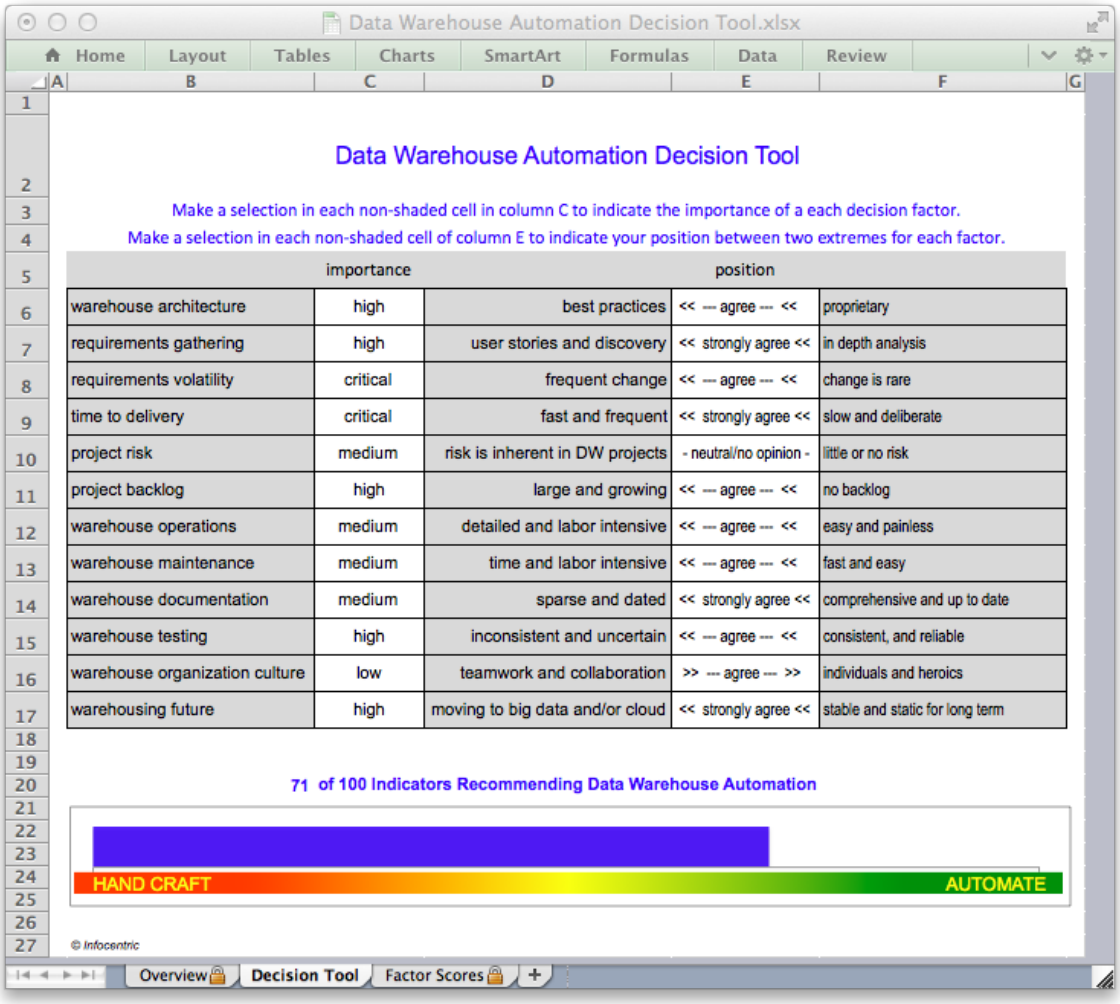


Figure 3: Data Warehouse Automation Decision Tool

(click the screenshot above to download the tool)

You can self-evaluate your position for each of the twelve decision factors using the Data Warehouse Automation Decision Tool. Figure 3 illustrates the main functions of the decision too. The tool evaluates all of your responses to develop a recommendation whether to automate or not to automate. The recommendations are not a simple, binary "yes" or "no" but a continuum that ranges from a score of zero to one hundred. A score

of one hundred is a strong case for automation - the benefits of automation are substantial and certain. A score of less than fifty raises questions because the benefits of automation are doubtful without changes to architecture, processes, or priorities.

In Conclusion

Data warehouse automation is here, it is real, and it is valuable. More importantly, it is not just a fad or a "technology thing" to help developers of data warehouses. It has real and substantial business benefits that are good reason for everyone to consider automation. Among those benefits are:

- Quality and Effectiveness – Automation enables warehousing teams to build solutions that best meet real business requirements. It is especially difficult to get complete and correct requirements when limited to an early phase of a linear development process. With data warehouse automation the business can make changes much later in the development process and change can occur more frequently with less disruption, waste, and rework.

- Business Agility - Ability to change fast and frequently extends beyond the warehouse development process. Changes that occur in business requirements can be met with quick response. Responding to change in real time and without the delay of lengthy projects is the essence of business agility.

- Speed - Speed is the critical factor that enables agility both for agile business and for agile development. Capabilities to generate quickly and to regenerate equally fast when change occurs are fundamental automation capabilities.

- Cost Savings - Ultimately building better, building faster, and changing quickly when needed bring substantial cost savings to data warehouse development, operation, maintenance, and evolution.

Perhaps automation isn't for everyone, but it certainly deserves consideration. Take a few minutes to download the decision tool and consider the degree to which it can bring value to your data warehousing efforts, your business intelligence program, and your business as a whole.

Sponsored by

WhereScape is the leading provider of data warehouse automation software. Over 650 customers use WhereScape to accelerate development while delivering a fully documented data warehouse or analytic solution in SQL Server, Teradata, IBM DB2 and Oracle targets, as well as Greenplum and Netezza analytic appliances. For more information, visit us.wherescape.com

About the Author

Dave Wells is actively involved in information management, business management, and the intersection of the two. As a consultant he provides strategic guidance and mentoring for Business Intelligence, Performance Management, and Business Analytics programs - the areas where business effectiveness, efficiency, and agility are driven. As an analyst he tracks trends in emerging and high-impact business intelligence and analytics technologies. As the founder of and Director of Community Development for the Business Analytics Collaborative, he is building a community of analytics professionals and practitioners that encompasses everyone from data scientists to data gurus and de facto analysts. On a personal level, Dave is a continuous learner, currently fascinated with understanding how we think, both individually and organizationally. He studies and practices systems thinking, critical thinking, lateral thinking, divergent thinking and the art and science of innovation.