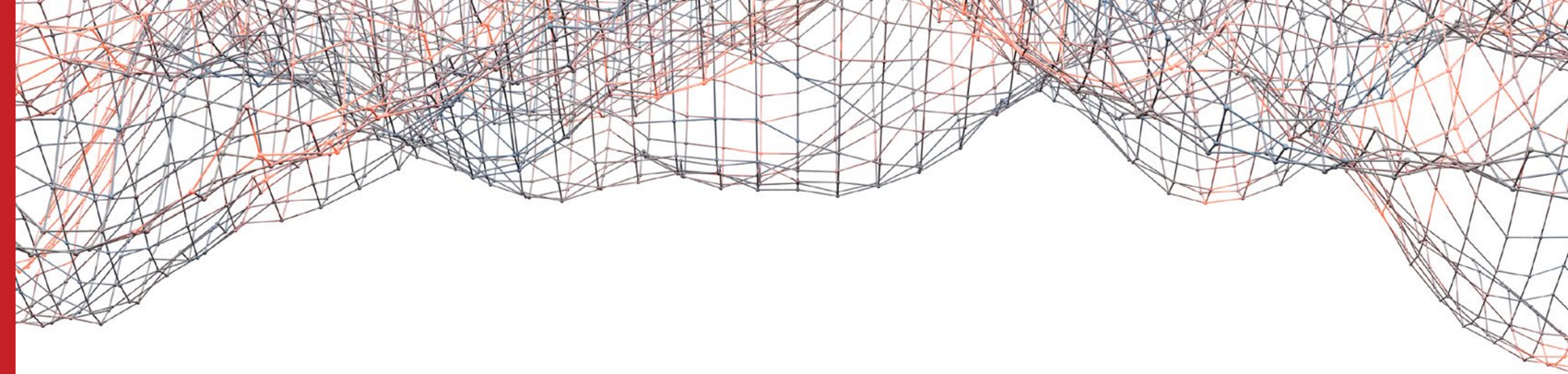
WhereScape®

Is Data Warehouse Automation Right for Your Organization?

A Primer and Readiness Guide

A White Paper by:
Dave Wells
Data Management Practice Director
Eckerson Group

Table of Contents



Is Data Warehouse Automation Right for Your Organization?

A Primer and Readiness Guide

Data Warehouse Automation is a mature technology that fills an important role in this age of complex data management and technological automation. Recent research shows that most organizations today are operating more than one data warehouse with the majority having two to five legacy data warehouses. Sustaining these warehouses and keeping them current and relevant in the face of changing data sources and business requirements is costly, labor intensive, and time consuming. Recent advances in data strategy, data architecture, and data management technologies compound the challenges of modern data warehousing. As organizations work to modernize data warehousing with cloud migration, real time and very low latency data, integration of unstructured data, ingestion of streaming data, and connecting to sensor data and IoT, the workload exceeds an organization's capacity to do the work manually. Today, data warehouse automation fills a critical and essential role in data warehouse modernization and agile data warehousing.

Automation is a must for those who want to gain efficiencies and improve effectiveness in data warehousing processes—development processes, operations processes, and modernization 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. It readily adapts to your data integration architecture with ability to automate traditional hub-and-spoke data warehouses, data marts, data vaults, and data lakes.

Automation changes the way that we think about building, maintaining, and evolving data repositories. 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 agile development practices and that enables requirements to be fluid.

Automation changes the way that business operates. It is one of the keys to digital transformation of business.

Automation also changes the way that business operates. Data is a critical component of digital transformation. An organization's digital capabilities increase substantially as they master data management agility. Automation accelerates data management processes—faster onboarding of new data, faster ingestion and processing of that data, faster delivery of business insights, and typically deeper insights as the scope of data grows. Fast data combined with big data is fundamental to cutting-edge technologies such as artificial intelligence and machine learning.

Ten Myths of Data Warehouse Automation

A surface-only look at data warehouse automation can be deceptive. Common misconceptions captured from an informal survey of people with limited understanding of data automation are shown in the table below.

Automating Only Data Warehouses

The term “data warehouse automation” comes from the early days of automation tools when data warehouses were the only widely adopted form of shared data repository. Modern data infrastructure supports data sharing and data integration with a variety of repositories including data warehouses, data vaults, data lakes, data marts, operational data stores, and master data and reference data stores. Full-featured automation tools are valuable across the infrastructure. It is time to update terminology and refer to this technology as Data Infrastructure Automation.

Automating ETL Only

Many assume that automation simply means generating ETL code. Robust automation tools do much more than simple ETL generation. They do generate code and scripts to move and process data (ETL code) but they also generate database objects (tables, indexes, and cubes) as well as comprehensive documentation that is always in sync with deployed data warehousing systems.

Automating Only the Development Processes

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, scheduling, 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, logging and auditing, validation, and change management, and they struggle to handle complex transformations without hand coding. It is highly unlikely that they can continue to adapt as data architectures and infrastructures evolve to embrace new kinds of data and new technologies. Investing time, talent, staff, and funding to custom build an automation tool is a poor investment of resources better used to respond to pressing business needs.

Only for Relational Data

The belief that warehouse automation is limited to data warehouses implemented using RDBMS is quite common. In fact, automation tools support many kinds of shared data repositories implemented using relational databases, multi-dimensional databases, columnar databases, in-memory databases, document stores, graph 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.

Not Suited for Streaming Data

The misconception that automation is focused solely on ETL leads to the belief that it works only with batch processing. Today’s robust automation tools are fully capable for low-latency, real-time, and streaming data processing. Many of today’s analytics use cases require a blend of batch processing and data streams. Batch and streaming must co-exist and not be treated as separate data silos. Automating the complex data pipelines that integrate streaming data is especially important for advanced analytics applications.

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 removes the burden of mundane and repetitive work and creates opportunities for people to do more interesting, creative, and value producing work. Companies using automation have consistently demonstrated 500% productivity increases for data warehousing teams. 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 management team doesn’t have more work than they have time and resources to complete the work?

Automation Limits Flexibility and Inhibits Creativity

Developers who are accustomed to hand crafting schema and processes may resist automation because they mistakenly believe that it limits the kinds of solutions that can be built. With experience, they’ll quickly discover the tremendous flexibility of automation tools and realize that creative problem solving is much more rewarding than creative coding.

Why Automate Data Warehousing?

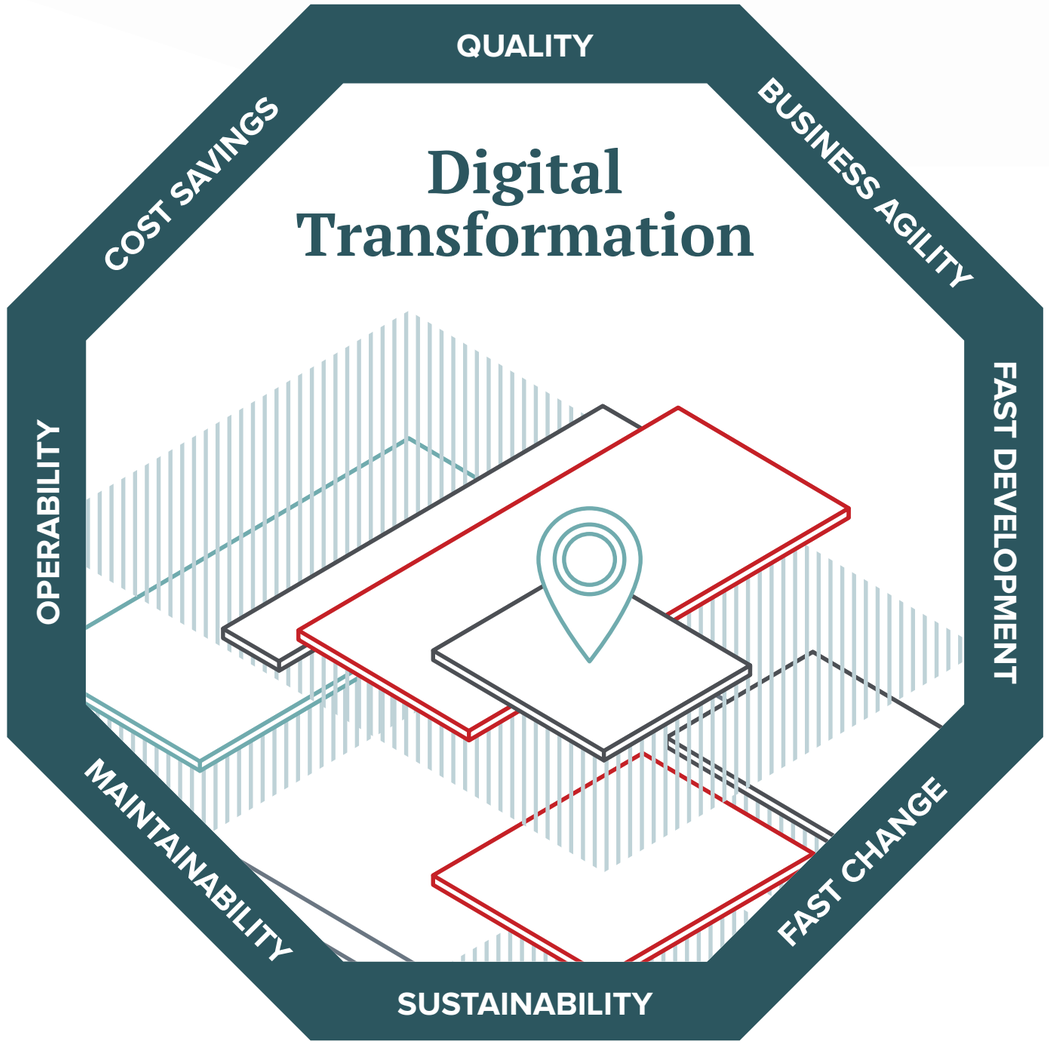
The corporate experience, almost without exception, is that data warehouses are necessary but painful. They take too long and cost too much to build, are out of sync with requirements by the time they are deployed, and are difficult to grow and change after deployment.

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.

Despite proclamations that the data warehouse is dead, we need data warehouses and will continue to need them.

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. The many benefits of data warehouse automation are illustrated here.

The Benefits of Automation



Five Reasons to Automate Data Warehousing

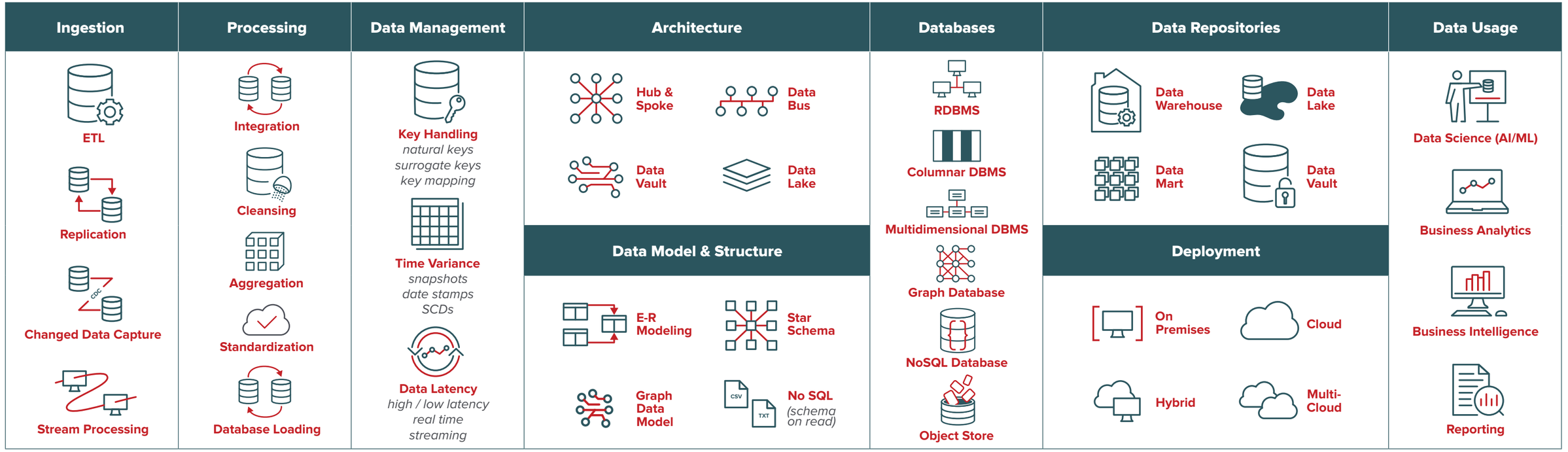
Data Warehouse Quality: Data warehouse automation delivers quality and effectiveness through the ability to build better solutions. Better solutions are those that best meet real business requirements, and it is especially difficult to get complete and correct requirements when limited to an early phase 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 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 the 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 Power of Patterns

Design patterns are fundamental to data infrastructure automation. Identifying and reusing patterns is at the core of the ability to achieve consistency, quality, speed, agility, and cost savings simultaneously. Design patterns encapsulate architectural standards as well as best practices for data ingestion, processing, management, architecture, design and modeling, databases, repositories, deployment, and usage. The diagram above illustrates common design patterns in all of these categories. Common data warehousing design patterns include:

Ingestion Patterns range from batch ETL to real-time stream processing. Data replication has become an increasingly popular way to push data to shared data repositories, often supporting near-real-time data with a combination of replication and micro-batches for data intake. Changed data capture (CDC) replicates data changes in real time and works well when very low latency is desired and when individual changes need to be tracked.

Processing Patterns support reusable operations and workflows for data integration, cleansing, aggregation, standardization, and database loading. Data integration—the heart of data warehousing—has an abundance of patterns with data movement patterns ranging from ETL to data virtualization. For data vaults the patterns for hubs, links, and satellites are fundamental.

Data Management Patterns include those for management of 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 latency patterns support standards, consistency, and reusable practices for high-latency, low-latency, real-time, and streaming data.

Architecture Patterns include the hub-and-spoke architecture popularized by Bill Inmon, the bus architecture championed by Ralph Kimball, the data vault architecture created by Dan Linstedt, and data lake architectures defined by pragmatic and in-the-trenches data management professionals.

Data Design and Modeling Patterns include those for structured data organized using entity-relationship concepts and multi-dimensional techniques. They also include patterns for differently structured and unstructured data such as graphs, documents, files, and text.

Databases Patterns naturally correspond to the variety of database types that are available in the age of Big Data and NoSQL. Unique patterns exist for each of relational databases, columnar databases, multidimensional databases, graph databases, NoSQL databases (with varied patterns for document stores, key-value stores, wide column stores, etc.), and object stores. These patterns support automation across the many database choices that are used today.

Repository Patterns reflect the best practices and conventions commonly used when building and operating data warehouses, data lakes, data marts, and data vaults.

Deployment Patterns support reusable operations, workflows, and management practices for data deployed on premises, in a cloud environment, with on-premises/cloud hybrids, and for multi-cloud deployments.

Data Usage Patterns are important considerations for full-lifecycle data warehouse automation. Patterns ranging from simple reporting to business intelligence, business analytics, and extending to data science, artificial intelligence, and machine learning may all occur simultaneously.

Patterns are powerful. Applied patterns in automation maximize the benefits of reusability, and support the goals of accelerated design and development. Equally important, they drive compliance with standards and best practices to achieve consistency and maintainability of data infrastructure results.

Are You Ready for Automation?

The greatest success with data infrastructure automation technology occurs when you're well prepared to reap the benefits. The ideal way to be ready is to evaluate your readiness. The checklist below describes twelve criteria to consider when evaluating your organization's readiness for automation.

- [ ] **Data Management Architecture**
  Does your architecture use a combination of best practices and specialized elements unique to your organization's needs?

- [ ] **Requirements Gathering**
  Do you define requirements through user stories and discovery processes instead of 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?

- [ ] **Time to Delivery**
  Do your business stakeholders expect fast and frequent delivery of data access, analysis, and business capabilities?

- [ ] **Project Risk**
  Do your data infrastructure projects experience a high level of risk from poor data quality, lack of source data knowledge, insufficient budget, understaffing, scope creep, and other factors?

- [ ] **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?

- [ ] **Operations**
  Are the processes and procedures for operation of your data infrastructure complex, detailed, time consuming, labor intensive, or fragile when something doesn't work right the first time?

- [ ] **Data Infrastructure Maintenance**
  Is your data infrastructure maintenance difficult, challenging, and dependent upon the knowledge of a few key individuals?

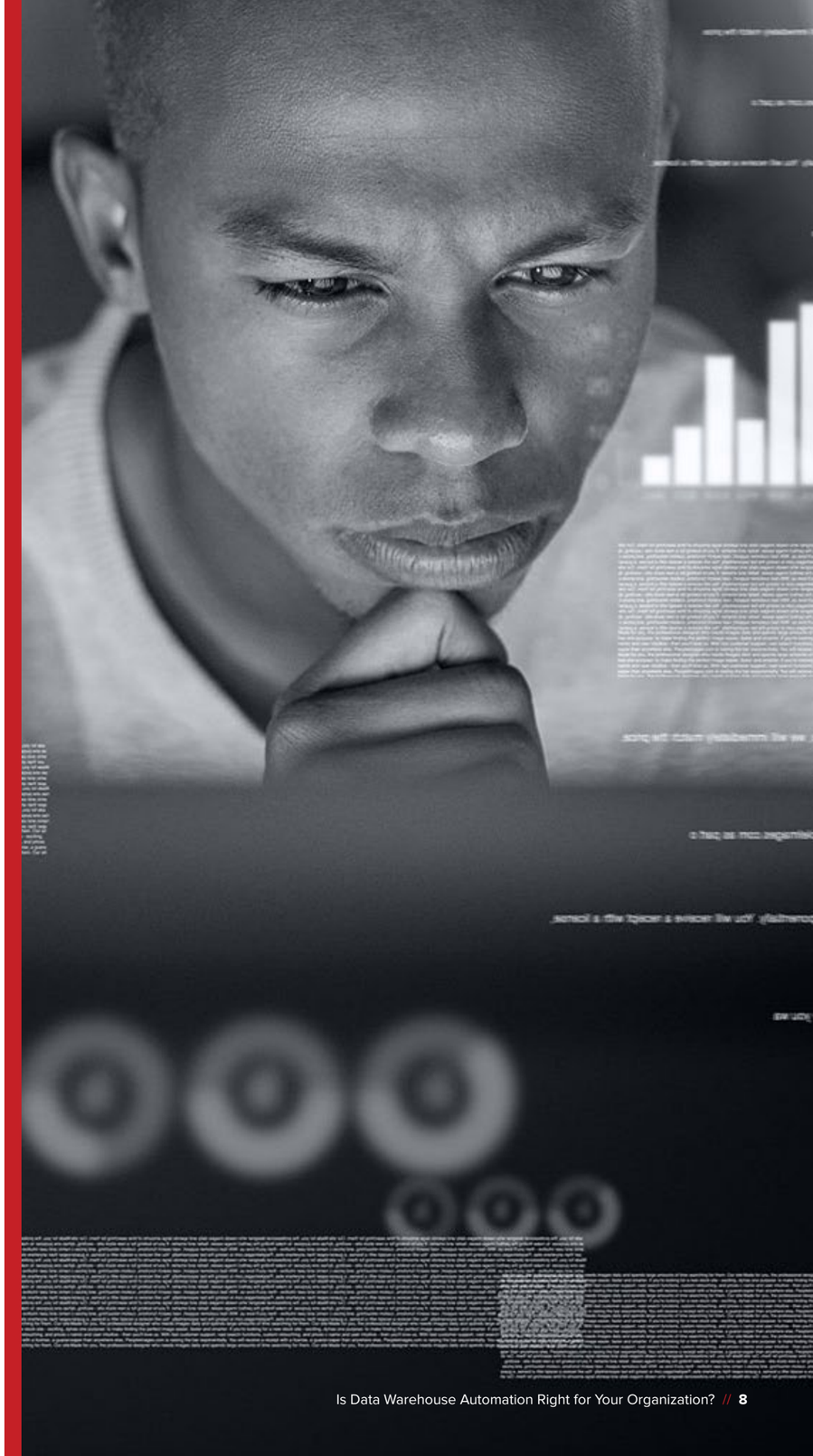
- [ ] **Documentation**
  Is the documentation for your data management processes and databases sparse, dated, and frequently out-of-sync with the implementation?

- [ ] **Testing**
  Do you lack consistent, reliable, and repeatable process for data warehouse testing including unit, stream, and integration testing during development and validation testing during operation?

- [ ] **Organization and Culture**
  Is your data warehousing team oriented to teamwork and collaboration? Is the IT relationship with business stakeholders collaborative?

- [ ] **Data Management Future**
  Are big data, cloud hosted data, cloud analytics, data science and artificial intelligence among current expectations of your business leaders and data consumers? Are they on the horizon in the foreseeable future?

Use the checklist as an aid to think through your organization's readiness for data infrastructure automation. If you answered "yes" to many of these questions then you have the need, motivation, and culture to successfully adopt and benefit from automation in your organization.



What Are Your Next Steps?

The due diligence to determine if data infrastructure automation is right for your organization begins by sharing your organization's needs and future goals with potential automation technology providers. Engage with providers at industry events geared to data management and data warehousing, or by attending automation software test drives. You can also start by contacting automation technology providers directly for a personal consultation or demonstration. Only by engaging and sharing your organization's aspirations, challenges, needs, goals and timelines will you gain the knowledge that you need to decide if and when data infrastructure automation is right for you.

Sponsored by

WhereScape®

WhereScape helps IT organizations of all sizes leverage automation to design, develop, deploy, and operate data infrastructure faster. More than 700 customers worldwide rely on WhereScape automation to eliminate hand-coding and other repetitive, time-intensive aspects of data infrastructure projects to deliver data warehouses, vaults, lakes and marts in days or weeks rather than in months or years. WhereScape has global operations in the USA, UK, Singapore and New Zealand. www.wherescape.com

Request a WhereScape automation demonstration today.

(04/19)

About the Author

Dave Wells

Dave Wells is the Data Management Practice Director at Eckerson Group, a business intelligence and analytics research and consulting organization. He brings a unique perspective to data management based on five decades of working with data in both technical and business roles. Dave works at the intersection of information management and business management, where real value is derived from data assets. He is an industry analyst, consultant, and educator dedicated to building meaningful and enduring connections throughout the path from data to business value. Knowledge sharing and skills development are Dave's passions, carried out through consulting, speaking, teaching, and writing. He is a continuous learner – fascinated with understanding how we think – and a student and practitioner of systems thinking, critical thinking, design thinking, divergent thinking, and innovation.