research

BEST PRACTICES REPORT

Q2 2016

Data Warehouse Modernization

In the Age of Big Data Analytics

By Philip Russom

Co-sponsored by

sas

tdwi

Advancing all things data.

Data Warehouse Modernization

In the Age of Big Data Analytics

By Philip Russom

Table of Contents

About the Author

PHILIP RUSSOM is a well-known figure in data warehousing and business intelligence (BI), having published over 500 research reports, magazine articles, opinion columns, speeches, webinars, and more. Today, he's the TDWI Research Director for Data Management at The Data Warehousing Institute (TDWI), where he oversees many of the company's research-oriented publications, services, and events. Before joining TDWI in 2005, Russom was an industry analyst covering BI at Forrester Research and Giga Information Group. He also ran his own business as an independent industry analyst and BI consultant and was contributing editor with leading IT magazines. Before that, Russom worked in technical and marketing positions for various database vendors. You can reach him at prussom@tdwi.org, @prussom on Twitter, and on LinkedIn at linkedin.com/in/philiprussom.

About TDWI

TDWI, a division of 1105 Media, Inc., is the premier provider of in-depth, high-quality education and research in the business intelligence and data warehousing industry. TDWI is dedicated to educating business and information technology professionals about the best practices, strategies, techniques, and tools required to successfully design, build, maintain, and enhance business intelligence and data warehousing solutions. TDWI also fosters the advancement of business intelligence and data warehousing research and contributes to knowledge transfer and the professional development of its members. TDWI offers a worldwide membership program, five major educational conferences, topical educational seminars, role-based training, onsite courses, certification, solution provider partnerships, an awards program for best practices, live webinars, resourceful publications, an in-depth research program, and a comprehensive website: tdwi.org.

About the TDWI Best Practices Reports Series

This series is designed to educate technical and business professionals about new business intelligence technologies, concepts, or approaches that address a significant problem or issue. Research for the reports is conducted via interviews with industry experts and leading-edge user companies and is supplemented by surveys of business intelligence professionals.

To support the program, TDWI seeks vendors that collectively wish to evangelize a new approach to solving business intelligence problems or an emerging technology discipline. By banding together, sponsors can validate a new market niche and educate organizations about alternative solutions to critical business intelligence issues. To suggest a topic that meets these requirements, please contact TDWI Research directors Philip Russom (prussom@tdwi.org), David Stodder (dstodder@tdwi.org), and Fern Halper (fhalper@tdwi.org).

Acknowledgments

TDWI would like to thank many people who contributed to this report. First, we appreciate the many users who responded to our survey, especially those who responded to our requests for phone interviews. Second, our report sponsors, who diligently reviewed outlines, survey questions, and report drafts. Finally, we would like to recognize TDWI's production team: Michael Boyda, Peter Considine, James Haley, and Denelle Hanlon.

Sponsors

IBM, Pentaho, SAP, SAS, and TimeXtender sponsored this report.

Research Methodology and Demographics

Report Scope. Technical users are scrambling to update, extend, and improve their data warehouse (DW) environments to satisfy their organizations' demands for new data types, subjects, sources, and targets for both operational and analytics use cases. The resulting practices and strategies for data warehouse modernization are documented here. This report also catalogs numerous tool types and features that are commonly applied to DW modernization, as well as drivers for modernization.

Audience. This report is geared to business and technical managers who are responsible for implementing and modernizing data warehouse environments that involve both traditional enterprise data and big data for analytics.

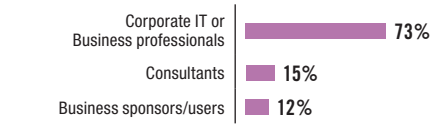
Survey Methodology. In November 2015, TDWI sent an invitation via email to the data management professionals in its database, asking them to complete an Internet-based survey. The invitation was also distributed via websites, newsletters, and publications from TDWI and other firms. The survey drew responses from 662 survey respondents. From these, we excluded incomplete responses and respondents who identified themselves as academics or vendor employees. The resulting complete responses of 473 respondents form the core data sample for this report.

Research Methods. In addition to the survey, TDWI Research conducted many telephone interviews with technical users, business sponsors, and recognized data management experts. TDWI also received product briefings from vendors that offer products and services related to the best practices under discussion.

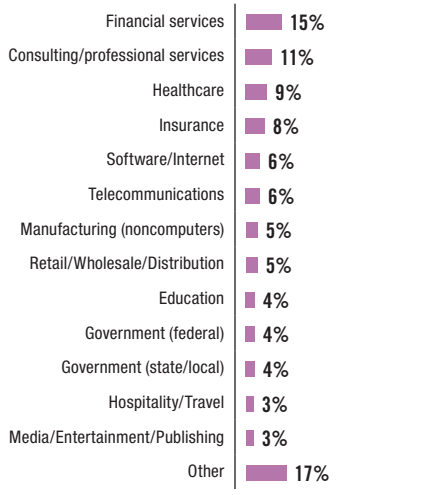
Survey Demographics. The majority of survey respondents are IT or BI/DW professionals (73%). Others are consultants (15%) and business sponsors or users (12%). We asked consultants to fill out the survey with a recent client in mind.

The financial services industry (15%) dominates the respondent population, followed by consulting (11%), healthcare (9%), insurance (8%), software/Internet (6%), telecommunications (6%), and other industries. Most survey respondents reside in the U.S. (60%), Europe (15%), or Canada (12%). Respondents are fairly evenly distributed across all sizes of companies and other organizations.

Position

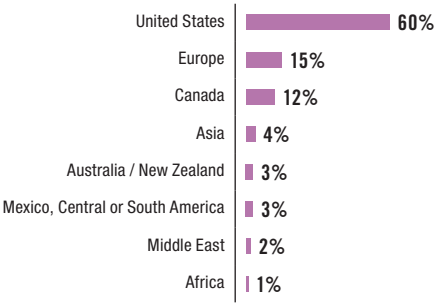


Industry

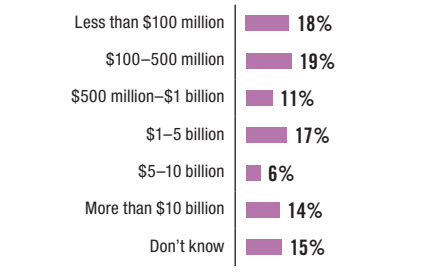


("Other" consists of multiple industries, each represented by less than 3% of respondents.)

Geography



Company Size by Revenue



Based on 473 survey respondents.

Executive Summary

No matter the vintage or sophistication of your organization's data warehouse (DW) and the environment around it, it probably needs to be modernized in one or more ways. That's because DWs and requirements for them continue to evolve. Many users need to get caught up by realigning the DW environment with new business requirements and technology challenges. Once caught up, they need a strategy for continuous modernization.

There are many manifestations of DW modernization.

DW modernization assumes many forms, from server upgrades and tweaks for data models, to adding new platforms into the extended data warehouse environment (DWE), to replacing the primary DW platform. Modernization may involve using features previously untapped, such as in-memory databases, in-database analytics, real-time functions, and data federation or virtualization. Systems integrated with the DW need attention, too. Analytics, reporting, and data integration are also modernizing, and the DW is under pressure to provision data in ways that enable modern end-user practices such as visualization, advanced analytics, data prep, and self-service data access. The arrival of big data has made such provisioning more business critical—and more difficult.

It's not just the DW. Other systems and practices are modernizing, too.

User best practices are also modernizing. For example, the move to agile development methods is one of the strongest trends in data warehousing. Similar trends involve lean, logical, and virtual methods. Modernization also affects users' skills, staffing, and team structure.

DW modernization is truly happening. This report's survey says that 76% of DWs are evolving briskly; 89% of respondents say modernization is an opportunity for innovation.

DW modernization can support business goals, DW scale, and analytics.

According to the survey, the leading drivers behind DW modernization include realigning the DW with new business goals, increasing DW scale for big data, enabling new analytics applications, and embracing new tools or data types and their attendant practices. The chief beneficiaries of modernization include analytics, business management, and real-time operations. The leading barriers involve problems with governance, staffing, funding, designs, and platforms.

Most modernization innovation is at the DW platform level.

The rise of the multiplatform DWE is an evolution of the DW system architecture. Hence, changes at the system architecture level are the most common form of DW modernization (53% of users surveyed). At one end, this involves simple upgrades and patches for hardware and software servers or tools. At the other end, many organizations are adding new data platforms and analytics tools to their extended DWEs to accommodate massive data volumes, new data types, and new analytics-processing workloads.

Modernization is more about integrating new platforms than replacing old ones.

Platform types being added to the DWE include those based on columns, appliances, event processing, advanced analytics, and Hadoop; these almost always complement the DW without replacing it. As an extreme measure, roughly half of organizations surveyed plan to rip out their current DW platform and replace it with a bigger and/or more modern one within three or four years. Compared to today, the average DW will be quite different in a few years—and hopefully more "modern" in the sense of bigger and better functionality, scope, speed, scale, user service, and business value. Yet the core platform (whether old or new) will continue to be relational (whether old or new), and new platforms (especially Hadoop) will improve substantially (especially with relational functions at scale).

As you can see, it's important to modernize a DWE to keep it competitive, relevant, growing, and aligned with new business and technology requirements. User organizations, however, struggle to understand the trends and take the right action. This report presents the many issues and categories of modernization, plus the strategies, methods, and enabling technologies that lead to success.

An Introduction to Data Warehouse Modernization

Manifestations of Data Warehouse Modernization

As any data warehouse (DW) professional can tell you, the DW is today evolving, extending, and modernizing to support new technology and business requirements, as well as to prove its continued relevance in the age of big data and analytics. This process has become known as *data warehouse modernization*; synonyms include *DW augmentation*, *automation*, and *optimization*. Every user organization and its DW is a unique scenario, so every modernization program is, too. Even so, a few common situations, drivers, and outcomes have arisen.

DW modernization takes many forms.

For example, common scenarios range from software and hardware server upgrades to the periodic addition of new data subjects, sources, tables, and dimensions. However, data types and data velocities are diversifying aggressively, so data modernization progressively involves users' diversifying their software portfolios to include tools and data platforms built for big data from new sources. As portfolios swell, most DWs are evolving—or modernizing—into complex and hybrid multiplatform data warehouse environments (DWEs). Though surrounded by complementary systems and tools, the traditional data warehouse is still the primary core of the modern DWE. Even so, a few organizations are decommissioning current DW platforms to replace them with modern ones optimized for today's requirements in big data, analytics, real-time operation, high-performance, and cost control. No matter what modernization strategy is in play, all require significant adjustments to the logical layers and systems architectures of the extended DWE.[1]

Looking inside the average data warehouse, we see many opportunities for DW professionals to initiate or expand the use of recent technology advancements, such as in-memory processing, in-database analytics, massively parallel processing (MPP), multiplatform federated queries, and Hadoop. Furthermore, there are many new database management systems purpose-built for analytics, based on columns, appliances, graph, MapReduce, NoSQL, and other innovations. Best practices can likewise be modernized by adapting agile, lean, logical, and virtual methods or by moving to modern team structures, such as the competency center or center of excellence.

Systems outside the DW need modernization, too.

Looking outside the warehouse, multiple disciplines have their own modern innovations that need support from a more modern DW. For example, new business practices need bigger, newer, and fresher data so the business can compete on analytics, get actionable business value from new big data, and monitor the business in real time. As another example, business intelligence (BI) is experiencing its own modernization right now, and BI needs the DW to provision data for modern BI practices, such as visualization, data exploration, and self-service. Likewise, many organizations are complementing their mature investments in online analytical processing (OLAP) with an exploding array of techniques for advanced analytics.

This report quantifies trends in data warehouse modernization and catalog technologies that are relevant. The report will also document strategies and user best practices for organizing modernization projects. The goal is to help DW professionals and their business counterparts plan the next generation of their data warehouse in alignment with business goals.

[1] For a more detailed definition of data warehouse environments, see pp.16–19 in the 2014 TDWI Best Practices Report *Evolving Data Warehouse Architectures in the Age of Big Data*, online at www.tdwi.org/bpreports.

Drivers for Data Warehouse Modernization

To get a sense of what issues are driving users toward modernizing their data warehouses nowadays—and with which priorities—this report's survey asked: What are the leading drivers for the modernization of your DW? (See Figure 1.) The question generated 5.7 responses per respondent, on average, which indicates that the average data warehouse professional is working hard to meet the requirements posed by several drivers simultaneously. The drivers identified in the survey group into eight broad areas, discussed below in rough survey priority order:

DW-to-business alignment is the leading driver for modernization.

Business concerns. Most of the drivers for modernization that data warehouse professionals are experiencing are technical in nature. Yet, the most pressing driver is the need to realign a DW so that it supports business goals (39% of respondents in Figure 1). Almost as pressing is the need to run the business on numbers and analytics (29%). Other business concerns reflected in the survey include cost reductions (19%), security and data privacy issues (16%), compliance and regulatory issues (14%), and competitive pressures (13%).

The second most common DW modernization is for greater scale and speed.

Technical scale and performance. The second most common driver for modernization is to increase capacity for growing data, users, reports, analyses, etc. (37%). This is no surprise because DW professionals have been improving their technology stack for decades to stay ahead of capacity. However, in recent years, the arrival of big data, the democratization of BI, and burgeoning programs for advanced analytics have greatly exacerbated this driver. Other scale and performance issues that need addressing include increasing data volumes (31%), the technical performance of the warehouse (23%), and optimization for multiple, diverse workloads (14%).

One-third of DW pros modernize for better and newer analytics.

Analytics. Near the top of the priority list (based on survey results) is the growing need for modern practices in analytics (mining, statistics, graph; not OLAP; 35%). Despite new implementations of advanced analytics, many organizations continue to modernize their mature investments in reporting (31%) and OLAP (12%). Note that new analytics complement—but don't replace—standard reports and OLAP; each delivers unique insights and guidance, and so each is required by the modern business.

Many users modernize to embrace new best practices and tool types.

New data-driven best practices. Vendor, open source, and consulting communities have recently brought us new tools and new ways of leveraging data for organizational advantage. Many users see business value in these and are eager to adopt modern practices for data exploration, data profiling, data prep (27%); data lake, data vault, or enterprise data hub practices (20%); the logical data warehouse (14%); and the virtualization of data (12%). Similarly, many data management professionals are adopting modern methodologies for agile development because they enable nimble business practices (23%).

Real-time operations. Already well established are data-driven methodologies that enable real-time operations for a business based on fresh data (26%). These methods include operational BI, performance management, and management dashboards. Most BI-driven organizations already have programs in place for these; however, the programs need modernization to gain faster performance for fetching and delivering real-time data and to give dashboards modern features, such as self-service data access, data prep, and visualization. In a related area, some users are operationalizing the DW to embed its data in daily business processes (16%), typically in near real time.

Life cycle issues lead to redesigning or replacing some DWs.

Problems that need fixing. DWs are like most other IT systems; as they age, their design and enabling technologies can become outmoded or simply no longer relevant to the evolving organization. Hence, some modernizations of DWs are driven by problems with the existing design or architecture (24%) or problems with the existing, underlying DW platform (16%). Although the last point ranked as a low priority in Figure 1, the rip and replace of DW platforms is shown to be common in Figure 14, later in this report.

New data types and platforms built for them are upcoming drivers.

New big data. Most DW professionals (and related personnel for data integration, business intelligence, and analytics) have worked mostly with data that is relational or otherwise structured. Their skills and tool portfolios—very much tuned to relational data and technologies (e.g., SQL)—are currently being challenged by the diversification of data types and formats (nonrelational, unstructured, social; 20%) and the diversification of data sources (sensors, machines, GPS; 15%). A special case that brings both of those together is the arrival of streaming data (12%). In organizations that are experiencing these forms of new big data, the data's unusual formats and sources are driving them to update both skills and portfolios of tools and data platforms.

New data platforms. Because older data platforms are not always well suited to new big data—as well as extreme volumes of traditional enterprise data—some users are turning to Hadoop implementation and integration (18%), as well as other manifestations of NoSQL implementation and integration (7%). For some organizations, cloud or SaaS adoption (11%) provides a data platform that is elastically scalable at a low cost.

What are the leading drivers for the modernization of your DW? Select one to seven answers.[2]

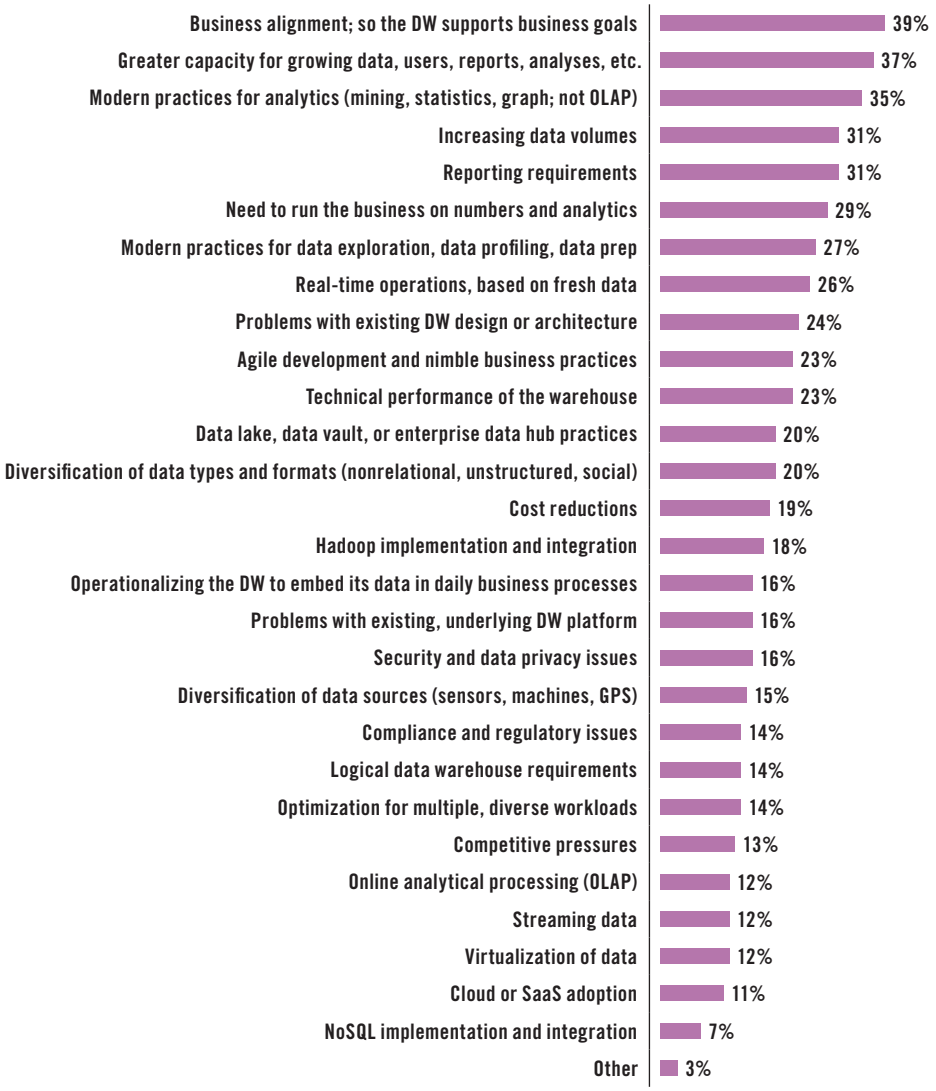


Figure 1. *Based on 2,684 responses from 473 respondents; 5.7 responses per respondent, on average.*

[2] For this and similar questions with numerous multiple-choice answers, the online survey randomly reordered answers to avoid respondents favoring the ones at the top of the list. The answer "Other" always remained at the bottom of the list.

USER STORY DW MODERNIZATION CAN HAVE MULTIPLE, DIVERSE DRIVERS

"The modernization efforts for my petroleum client's data warehouse are guided by multiple drivers," said Mervin van der Spuy, a senior management consultant at Integrationworx. "The leading driver comes from business people who want new solutions delivered quickly. Other business people are technical enough that they want to access data on their own. For both classes of business users, we've implemented tools and built data sets that enable self-service for data access, discovery, and visualization.

"Another driver is the low price of oil, which has narrowed our margins. In reaction, we have deployed new analytics that help people discover the new state of the business, bottom-line cost-cutting measures, and opportunities for top-line growth.

"On the horizon is the industrial Internet of Things. Our field operations have had sensors for years, but now they are modernizing by adding even more sensors to gear at the wellhead and in petroleum distillation facilities. With this new data, we can develop more in-depth analyses of geology, machinery, and processes—much of it in real time."

The State of Data Warehouse Modernization

The Importance of Data Warehouse Modernization

Very few respondents (8% in Figure 2) question the importance of DW modernization. Some data warehouses serve static businesses that are content with current conditions. For example, a warehouse professional at a U.S.-based transportation firm told TDWI: "Why would I need modernizing? My data warehouse meets my users' requirements." For others, the DW should mirror business needs, and some businesses don't want or need modernization. As a BI director at a retailer in the U.S. put it, "There has not yet been a heavy push from the business for DW modernization."

How important is modernization for the success of your data warehouse and related platforms today?

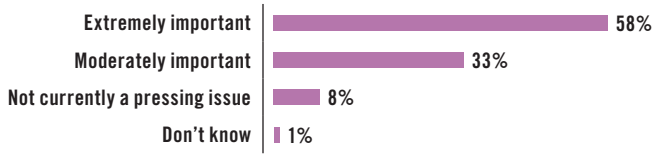


Figure 2. *Based on 473 respondents.*

Modernization is top of mind for data warehouse professionals worldwide.

The vast majority of respondents (91% in Figure 2) recognize the importance of modernizing a data warehouse. Over half feel that modernization is "extremely important" (58%), and an additional third see it as "moderately important" (33%).

Why are survey respondents unanimous in their zeal for data warehouse modernization? To get their unvarnished opinions, the survey asked the open-ended question: In your own words, why is DW modernization important (or not important)? The comments typed by respondents reveal a number of use cases, needs, and trends, as seen in the representative excerpts reproduced in Figure 3. Note that the users quoted work in a wide variety of industries and geographic regions. DW modernization is certainly top of mind for data professionals and their business sponsors in many contexts worldwide.

In your own words, why is modernizing a data warehouse or warehouse environment important (or not important)?

- "One has to keep up with the volumes and variety of data that can enhance your analytical results for better decision making and customer service." – Data architect, financial services, Africa
- "Need to adjust to business needs, as well as provide greater functionality and ease of use." – DW and BI architect, insurance, United States
- "The business landscape is constantly changing, and it's evolving the DW requirements. If you do not change with the times, you will become obsolete." – Enterprise architect, petroleum, Canada
- "In the past, week-old data might have sufficed, but today we need near-real-time data in most cases." – BI manager, state/local government, United States
- "To achieve low TCO [total cost of ownership], integrate with digital channels, support fast business decisions, allow complex analytics." – CTO team member, financial services, Asia
- "Current solution was built five years ago on 20-year-old technology and patterns. Latency, performance, and scope all lag far behind today's needs." – Data warehouse architect, insurance, United States
- "To allow the organization to increase its value by providing the decision makers with information targeted and aligned to the organization's short-term and long-term strategies." – Head of consulting, federal government, Mexico
- "Data is more vital to revenue and growth than ever before." – Chief digital officer, media, United States
- "Our data warehouse stores clinical and enterprise data. We are moving toward an information reservoir model with a near-real-time data warehouse at its heart. The business objective is to improve clinical outcomes." – BI technical lead, healthcare, Australia
- "Business processes and data have changed, but the warehouse and reporting systems have lagged behind. This causes difficulties in designing the outputs that my users really want." – BI architect, advertising, United States
- "Modernizing will enable the organization to lower the cost of historic data and benefit from new sources and types of data (e.g., social data, external information about prospects and customers, life events of customers, etc.)." – IT professional, financial services, Middle East
- "Modernization should be dictated by the business need, not the next shiny thing." – Senior BI analyst, transportation/logistics, Canada

Figure 3. *Drawn from the text responses of 413 respondents.*

Most Data Warehouses Are Changing Appreciably

Most DWs (76%) are evolving moderately or dramatically.

The average DW is definitely evolving but in varying degrees. Over half of respondents report moderate evolution (54% in Figure 4), and an additional fifth report dramatic evolution (22%).

A minority of DWs are evolving conservatively. Not all data warehouse professionals consider their work *modernization* per se. Some make continuous improvements to designs and upgrades to platforms regularly, and they consider these to be just the usual updates (20%).

Is your data warehouse evolving? Select only one.

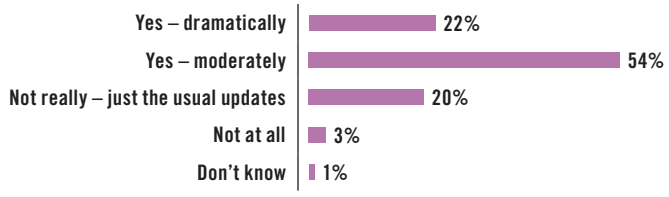


Figure 4. *Based on 473 respondents.*

Most Data Warehouses Have Room for Improvement

Most DWs (79%) are positioned for both recovery and growth.

The average DW is in pretty good shape. Right up the middle, most DWs are either mostly up to date (41% in Figure 5) or somewhat behind (38%). This positions DWs for both recovery and growth but also indicates ample room for improvement via various forms of modernization.

Few DWs are extremely modern or extremely outmoded. At one extreme, only 12% of respondents feel their DW is far behind, and the low percentage is good news. At the other extreme, only 7% consider their DW fully up to date. How can you achieve a state of perfection given the current rate of evolution in DWs and their organizations? As we'll see later in this report, some DW teams do it by rolling out brand-new platforms, architectures, or both. Others simply keep up with continuous, though less dramatic, improvements.

How modern is your data warehouse and its extended environment today?

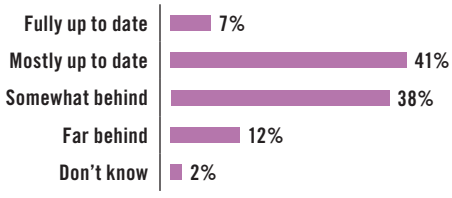


Figure 5. *Based on 473 respondents.*

Clearly, Data Warehouses Are Still Relevant

Most DWs (88%) are very or somewhat relevant to their enterprises.

The vast majority of DWs remain relevant to the current state of their enterprises. Half of respondents say their DW is very relevant (49% in Figure 6), and an additional third say it's somewhat relevant (39%). The survey results are quite clear. Data warehouses—regardless of design, architecture, or platform type—continue to ably provision data aligned with the way their organizations run their businesses. Therefore, it makes sense for organizations to deepen their investments in data warehousing and related data disciplines through modernization activities.

Very few users doubt the relevance of their DW. Only 8% of respondents consider their DW not very relevant. Even fewer see their DW as not relevant at all (3%).

Is your data warehouse relevant to the way your organization runs its business today?

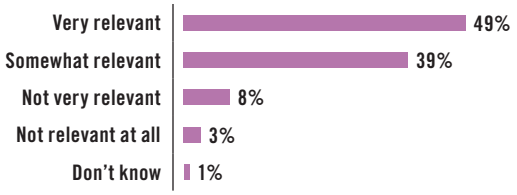


Figure 6. *Based on 473 respondents.*

USER STORY NEW CUSTOMER CHANNELS GENERATE BEHAVIOR DATA THAT SHOULD BE CAPTURED AND ANALYZED

"Our goals for data warehouse modernization are rather aggressive, so we've developed a plan for what we call 'the data warehouse of the future,'" said a data warehouse architect at a U.S.-based retailer of sporting equipment. "The first phase of the modernization plan is to build a data lake for capturing new types of data from new sources. For example, we recently launched a loyalty program which—as a new customer channel—generates valuable customer behavior data that will make our customer views more complete. Similarly, we will soon release a smartphone app that our customers can use to plot mountain bike excursions, and we're working on an app through which users can see consolidated exercise data gathered by multiple third-party devices. Customer data aside, the data lake will also enable analytics sandboxes for our logistics analysts, data scientists, and e-commerce team.

"Our data lake must support the eclectic data gathering and advanced analytics requirements of these new business programs—but scale in a cost-effective manner. Our tests have shown that Hadoop has the scale, analytics processing power, unstructured data handling, and low cost that we need. For even lower costs and more elastic scale, we will most likely go with a cloud-based implementation of Hadoop.

"Once that phase is in place, we'll move on to the second phase of our data warehouse modernization plan. At that point, we'll migrate terabytes of data and off-load several workloads from our traditional relational data warehouse to the new implementation of Hadoop."

Benefits and Barriers

Data Warehouse Modernization: Problem or Opportunity?

The vast majority (89%) feel DW modernization is an opportunity.

Depending on the kind of modernization activities with which you're involved, it can be a considerable amount of work and soak up a substantial amount of resources. Furthermore, modernization is risky when not planned or supported well enough (as we'll see in the later section on barriers to modernization), and a few data warehouse professionals consider most forms of modernization a distraction from the data-to-day, meat-and-potatoes work that has to be done. These issues lead us to wonder whether data warehouse modernization is truly the opportunity it's hyped up to be or is simply a problem to be endured. It's clearly the former—an opportunity, not a problem—according to most respondents taking this report's survey (see Figure 7).

An overwhelming majority (89%) consider data warehouse modernization an opportunity. As we'll see in the next section of this report, users feel that data warehouse modernization leads to improvements in analytics, decision making, near-time data usage, business operations, and development speed and productivity. Survey aside, users interviewed for this report talked about how modernization helps them roll out their innovations in data modeling and architecture (data lakes, hubs, analytics archives, and sandboxes) and new end-user practices (data exploration, discovery, and visualization).

A small minority (11%) consider modernization to be mostly a problem. As we'll see later in this report, users are aware of the common barriers to modernization, namely inadequate governance, staffing, funding, skills, and sponsorship—but the barriers don't stop most organizations.

Is data warehouse modernization mostly a problem or mostly an opportunity?

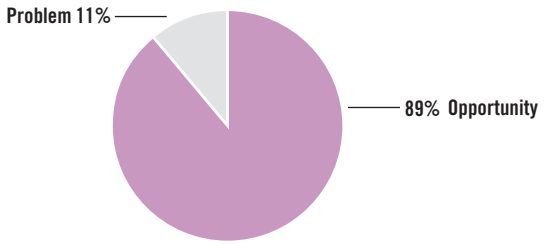


Figure 7. *Based on 473 respondents.*

Benefits of Modernizing a DW and Related Programs

The chief beneficiaries of modernization are analytics, business activities, and real-time data.

In the perceptions of survey respondents, data warehouse modernization offers several benefits (see Figure 8). Five areas stand out in their responses:

Analytics. At the top of the chart, the most common beneficial area concerns analytics in general, including visualization and exploration (53%). To a lesser degree, users also see benefits for specific analytics applications such as fraud detection (15%), customer base segmentation (12%), risk management and mitigation (quantification of risk; 11%), understanding business change (10%), and understanding consumer behavior as seen in clickstreams (10%).

Business. Several business activities ranked high among the potential benefits of modernization, ranging from decision making (52%) to operational efficiency (34%). Fewer respondents feel that modernization can address new business requirements (28%), enhance competitive advantage (28%), and reinvigorate both business and technology processes (10%).

Real time. A recurring theme throughout the survey is how modern tools, features, and platforms are key to enabling frequent report and analysis cycles, operating at near real time (37%).

Methods. There is also a need for modern methods and best practices, which can improve the agile delivery of solutions (33%), the management and maintenance of the DW environment (20%), and automation for the design, deployment, and operation of the DW (12%).

Finances and funding. A few respondents feel that modernization could help leverage big data with a return on the investment (16%), monetize data assets (12%), and contain costs for the DW environment (7%).[3]

What are the top business and technology tasks that would benefit if your organization implemented the forms of data warehouse modernization you are contemplating? Select one to five answers.

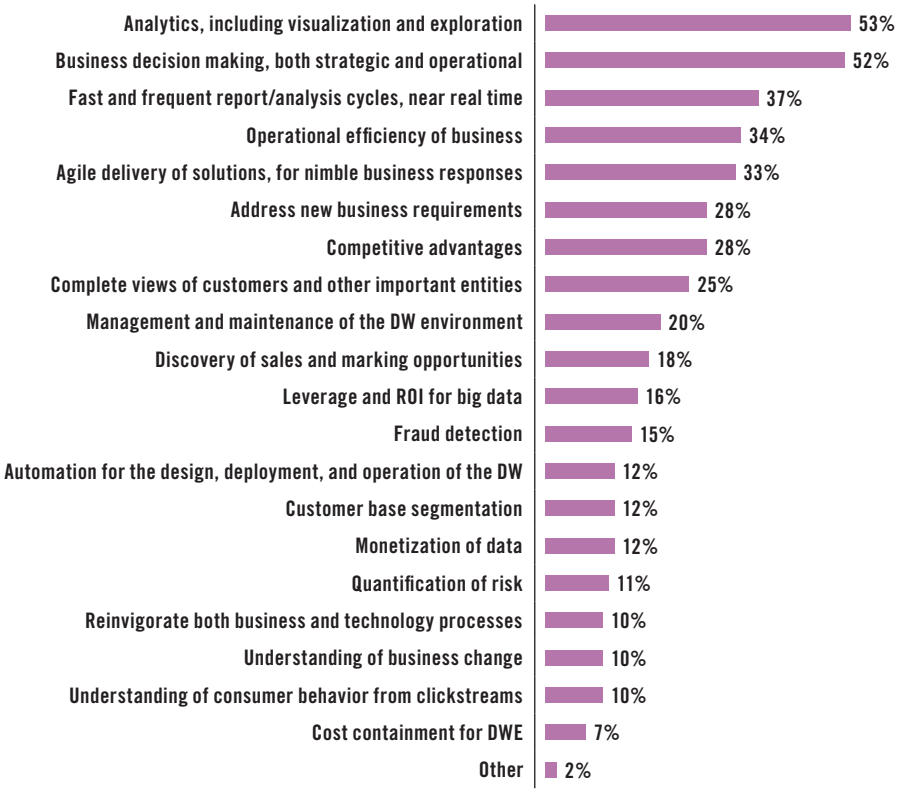


Figure 8. *Based on 2,102 responses from 473 respondents; 4.4 responses per respondent, on average.*

[3] For a discussion of how cost can guide multiplatform data warehouse architectures and their modernization, see "Number Eight" in the 2015 TDWI Checklist *Eight Tips for Modernizing a Data Warehouse*, online at www.tdwi.org/checklists.

Barriers to Making Modernization Happen

The leading barriers to modernization are inadequate governance, staffing, funding, design flaws, and platform weaknesses.

Data modernization has its benefits, as we just saw. However, it also has many potential barriers, according to survey results (see Figure 9). The issues span multiple areas.

Inadequate organizational support. As with most data-driven programs, DW modernization can be limited by poor stewardship or governance (40%) or a lack of a business case or sponsorship (30%).

Technical team deficiencies. Technical success depends on the team, which may suffer inadequate staffing for data warehousing and related disciplines (39%), inadequate skills for new technologies and practices (33%), or a lack of experience with new big data types and their analytics (28%).

Cost issues. Financing modernization can be inhibited by the cost of implementing new technologies (34%) and the cost of hardware and software upgrades (21%).

Data limitations. Whether focused on new big data, traditional enterprise data, or both, modernization can be threatened by the poor quality of data (27%) or metadata (18%).

Design challenges. Applying new architectures to an existing solution requires substantial retrofitting when the current DW was designed for standard reports and OLAP only (20%). Likewise, moving to the complex, multiplatform system architectures typical of modern DWs can be stymied by the difficulty of architecting a modern, complex environment (21%) and the difficulty of managing a multiplatform DW environment (14%).

DW platform limitations. The DBMS and hardware platform under an existing warehouse can be a substantial barrier when the current DW environment cannot scale up to big data (16%) or ingest data fast enough (14%) to leverage large volumes or streaming data.

Missing ancillary tools. Modernizing the ecosystem around a warehouse requires the acquisition or upgrade of many tool types. Otherwise, the results of modernization are limited by a lack of tools for analyzing new big data types (12%) or for integrating and managing new big data types (12%).

Stodgy mindsets. Twenty-seven respondents selected "Other" (6% in Figure 9) and entered additional barriers to modernization, most of them relating to mindset issues. Sometimes the problem stems from upper management mindsets, as when "management does not prioritize infrastructure investment"; "the business does not understand the potential of data"; or "top management is not committed to innovation." At other times, everyone suffers from an "inability to rethink the technological choices made earlier" or "the momentum of current or traditional thinking." When it comes to getting resources for modernization, few organizations are immune to "company politics."

In your organization, what are the top barriers to data warehouse modernization? Select one to five answers.

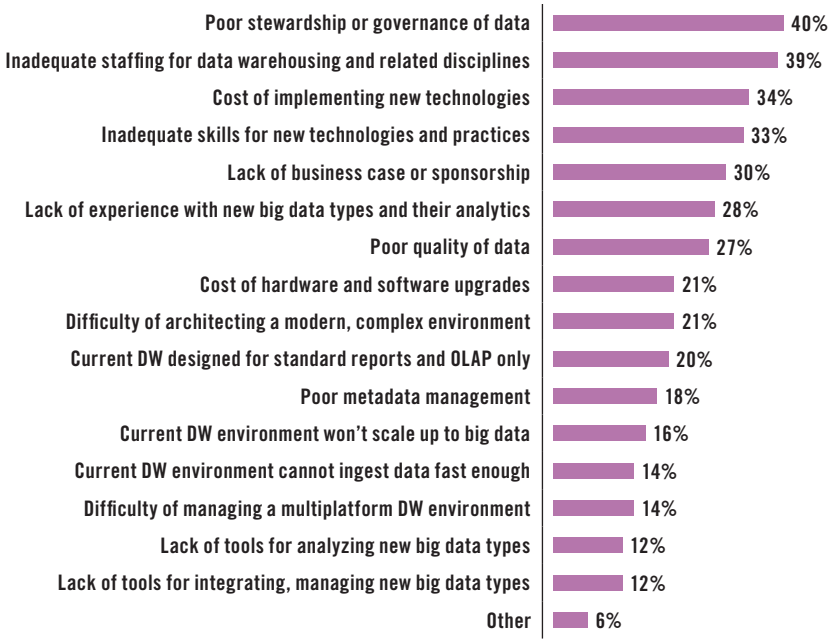


Figure 9. *Based on 1,816 responses from 473 respondents; 3.8 responses per respondent, on average.*

USER STORY IT'S NOT JUST THE DATA WAREHOUSE THAT NEEDS MODERNIZATION

"When I first came into my job, I found a large operational data store (or ODS) that consolidated data from multiple applications for integrated cross-department reporting," said a data architect at an educational institution. "That's not a warehouse, so we started from square one, building a true data warehouse. First, we revamped the ODS so it would serve as a scalable data staging area. Then we got to the meat of the matter—designing dimensional models, which are the sign of a true data warehouse. Subject areas for finance and human resources are already in place because those had the most burning needs. More dimensional models are coming for students, donors, and science research data.

"Warehouse modernization has gone well, but it's not all that needs modernizing. We are now committed to design and deploy a modern infrastructure for ETL, using a major data integration platform from a leading vendor. We are also committed to modernizing reporting and analytics by deploying a vendor platform known for data exploration and visualization. With those in place, we expect to roll out new dashboards by mid-2016."

Best Practices for DW Modernization

Categories of Modernization

As we have seen repeatedly in this report, data warehouse modernization can take many forms at varying scopes. Let's boil down the possibilities to the most common categories (see Figure 10).

The rise of the multiplatform DWE is forcing the modernization of system architectures.

System modernization (53%). At one end, this involves upgrades and patches for hardware and software servers or tools. At the other end, many organizations are adding new data platforms and analytics tools to their extended data warehouse environments (DWEs) to accommodate new data types, massive data volumes, and processing workloads.

Arbitrary modernization (47%). On the one hand, it's good that many modernization tasks are based on the business needs of a specific project, thereby attaining a level of DW-to-business alignment. On the other hand, data warehouse professionals interviewed observed that arbitrary tasks tend to pop up unpredictably; some demand an immediate response whereas other tasks can be folded into regular cycles for continuous modernization.

A DW integrates with many systems, so modernization must accommodate these.

Non-DW modernizations (44%). Users interviewed and surveyed for this report kept mentioning that the warehouse is rarely all that gets modernized. The DW is often modernized to better gather or provision data for new or evolving business processes, data integration solutions, reports (usually in dashboard style), and analytics (usually advanced forms such as data mining, statistics, and graphs).

Optimization modernization (42%). This is a significant productivity issue because the average data warehouse professional spends up to 30% of his or her time on performance tuning and similar optimizations.[4] In one direction, modern tools from both vendor and open source communities have become proficient at automated optimization, especially for SQL-based queries. In the opposite direction, however, the growing number of standalone platforms in users' extended DWEs has driven up the number and complexity of cross-platform queries, which are not so easily optimized.

Develop a recurring cycle for most DW modernizations.

Continuous modernization (37%). The issue here is to foster continuous improvement but in an organized fashion that assures proper standards and quality—and sanity. Many successful data warehouse professionals have spoken at TDWI events about how they stick to a regular, quarterly cycle for applying updates to the primary DW. This is similar to the controlled release cycles seen in software firms and open source incubation projects, but more frequently it is due to the pressing needs of the enterprise. The quarterly cycle works very well with small-to-midsize tasks, ranging from tweaking existing data models (for performance or to extend a customer view) to rolling out a new subject area (hot ones being about employees or locations). However, the quarterly cycle can also apply to grander modernization tasks, such as implementing a new tool or platform or rolling over groups of users from one tool or platform to another.

The more disruptive a modernization is, the more critical to success is the multiphase plan.

Disruptive modernization (21%). A modernization project can be rather dramatic and disruptive to business users when it involves the rip and replace of major data sets, platforms, or tools. As one respondent put it, "[We] have started a new BI program that will rebuild the entire DW on a new platform." Modernization on that scale must be planned carefully as a multiphase (and probably multiyear) project. At that level, you're not only modernizing the data warehouse, but you're also modernizing the whole technology stack (as mentioned above), and modernization of that magnitude will also be accompanied by changes to business processes, sometimes the entire business model. Hence, the multiphase plan is not just for the warehouse or technology; the plan must also lay out how business users and processes will migrate and roll over as the modernization progresses. Obviously, business managers will have to be deeply involved in drawing up and following the modernization plan.

[4] See the discussion around Figure 12 in the 2012 TDWI Best Practices Report *High-Performance Data Warehousing*, online at www.tdwi.org/bpreports.

Other (3%). A few respondents selected "Other" and entered additional modernization tasks. These involve extending and remodeling existing data sets, new DW architecture, projects to consolidate other data sets into the data warehouse, and the implementation of Hadoop.

What categories of DW modernization tasks in your organization have you been involved with or seen as an observer in the last three years or so? Select all that apply.

Figure 10. *Based on 1,167 responses from 473 respondents; 2.5 responses per respondent, on average.*

Modernization Strategies

The warehouse's data and its platform are two distinct layers.

Before continuing, let's recall that TDWI defines a data warehouse as a data architecture that's populated with data. In other words, the DW is *the data*. Furthermore, the data is captured and managed on the *data warehouse platform*, which consists of a database management system (DBMS; or an equivalent system, such as Hadoop), operating system, server hardware, networking, and so on. Hence, the warehouse and its platform are two different but related layers. The distinction is relevant to this discussion because DW modernization may hit only the data, only the platform, or both.[5]

Furthermore, modernization can involve disruptive rip-and-replace tasks (as described in the previous section of this report). Your data is not something you replace; instead, modernization may extend, remodel, consolidate, and improve data. However, replacing your DW platform can be a viable modernization approach when the platform is deficient or no longer a good fit for business and technology goals.

Replacing a DW platform is disruptive and expensive for a business.

The catch is that rip and replace is inherently expensive and disruptive. That's why many organizations prefer to update and improve the existing platform instead. An increasing number of organizations complement the existing DW platform by adding other standalone platforms to the DWE. Of course, update and complement strategies can be combined.

With these issues in mind, it's no surprise that many users are considering DW platform replacements as they plan their DW modernizations. To quantify the situation, this report's survey asked: Which of the following best describes your organization's strategy for data warehouse modernization? (See Figure 11.)

Most users will leave their DW platform in place but update it and complement it with other systems.

Augment (but don't replace) existing data warehouse's primary platform by adding additional data platforms and tools (42%). By far, more survey respondents selected this strategy. This is consistent with the movement toward multiplatform data warehouse environments, which is one of the strongest trends in data warehousing today. From a business viewpoint, this is a nondisruptive task, it preserves existing investments in data warehousing, and (when done well) it extends the life of an expensive and useful system. For years, TDWI has seen users deploy data warehouse appliances and columnar databases as part of their warehouse augmentation and modernization strategy. More recently, Hadoop has become a prominent data and analytics platform for such strategies.

[5] For more details about modern data warehouse platforms, see the 2009 TDWI Best Practices Report *Next Generation Data Warehouse Platforms*, online at www.tdwi.org/bpreports.

Replace existing data warehouse's primary platform (15%). For a minority of users (typically those with a deficient or outmoded DW platform), this approach is fully appropriate despite the disruption and expense. Here, 15% is rather conservative; the responses seen later in Figure 14 indicate that a higher percentage of organizations are committed to replacing their DW platform.

Strategy determined on a case-by-case basis (24%). As explained earlier, arbitrary modernization has its place but should be controlled to reduce its chaos. From a different viewpoint, consider that modernization hits many layers of the overall technology stack, plus has diverse business drivers; so it's inevitable that a certain amount of case-by-case examination is required.

No strategy, although we do need one (14%). A few users interviewed for this report complained that they have worked at many enterprises where the continuous modernization of the DW was one fire drill after the other due to business impatience. The result is too often a disorganized DW that suffers many architectural warts, has inconsistent data quality, and is difficult to optimize and maintain.[6]

Other (4%). A few respondents selected "Other" and entered additional modernization strategies, namely: upgrade existing platform (same vendor); upgrade current capabilities together with combined technologies for big data; and upgrade and redesign existing dimensional DW but add Hadoop data lake as a staging and advanced workflow environment.

Which of the following best describes your organization's strategy for data warehouse modernization?

Figure 11. *Based on 473 respondents.*

USER STORY SOMETIMES, DATA WAREHOUSE MODERNIZATION MEANS STARTING OVER

"I was hired three years ago to modernize business intelligence and build a modern data warehouse," said Ted Balzano, the head of global BI for the insurance division of Axis Capital. "Our new data warehouse follows a centralized architecture where data marts are generated from the DW for better control, quality, and governance. Our new ETL framework provides further standardization and control.

"We're modernizing BI by replacing old reports with dashboards built atop a modern platform for data visualization and analytics. The next step in modernization is to build new analytics for actuarial functions.

"In the near future, we'll modernize again by developing predictive analytics for risk management and underwriting. We'll upgrade the new warehouse and ETL so they refresh data more frequently and efficiently, which in turn will make the information in management dashboards more current."

[6] See the discussion around Figure 11 in the 2014 TDWI Best Practices Report *Evolving Data Warehouse Architectures in the Age of Big Data*, online at www.tdwi.org/bpreports. Comparing this report to the other, we see that the percentage of organizations without a strategy has dropped from 23% to 14%—luckily!

Ownership and Sponsorship

Chief officers and architects lead DW modernization efforts.

CIOs regularly fund and sponsor DW modernization (23% in Figure 12). Other chief officers involved in modernization include chief technology officers and chief operating officers. A few respondents pointed to their chief data officer.

Architects have an obvious stake in DW modernization. That explains why so many take control of its planning and execution (20%). These are mostly data warehouse architects, but other participating people are data architects, IT architects, and system architects.

Miscellaneous managers and directors are involved (47%). Their functions include BI, IT, DW, and line-of-business management. Data analysts and data scientists are also involved. Among the many job titles entered by respondents were vice president for analytics, BI, IT, data management, and information management.

In your organization, who is most responsible for planning and executing strategies for data warehouse modernization?

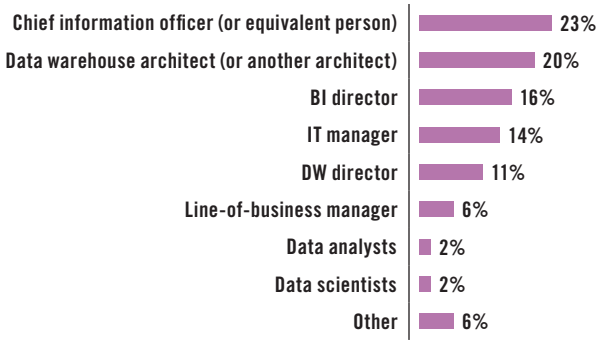


Figure 12. *Based on 473 respondents.*

Aligning Modernization with Business Goals

Most (72%) DW modernizations align well with business goals.

To be sure that data warehouse modernization delivers significant benefits to the organization, the people planning and executing it need to know the business's goals, how these relate to data-driven programs, and how to steer data-driven programs toward business goals. This is what most BI, analytics, and data warehouse programs do anyway, so it's not surprising that most organizations have aligned DW modernization efforts with business imperatives (see Figure 13).

For example, DW-to-business alignment is somewhat close in half the organizations surveyed (49%), with another quarter being very close (23%). Relatively few organizations are not very close (21%), although very few are completely unaligned (4%).

In your organization, how close is the alignment of DW modernization to business imperatives?

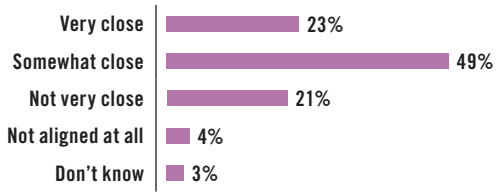


Figure 13. *Based on 473 respondents.*

USER STORY REORGANIZING A TEAM STRUCTURE CAN BE A FORM OF MODERNIZATION

"We recently reorganized a mature program for business intelligence to become an agile BI competency center," said Marty Afkhami, a BI director at Laureate Education. "The re-org really helped us modernize how we deliver BI products, with a focus on providing more self-service data access and visualization. As part of this change, we complemented our traditional BI tool with one known for high ease of use with data exploration, visualization, and analytics. To improve the ease of use even more (which we consider key to the success of self-service), we're evaluating the new generation of data prep tools. Our next step will be to socialize the BI competency center by training business users in how to get the most out of self-service data access and high ease of use for reporting, visualization, and analytics."

Data Warehouse Trends Relative to Modernization

The purpose of this section of the report is to predict possible directions users are going (or should go) with their DW modernization projects. A series of survey questions asked respondents about prominent trends in data warehousing, from platform replacements, architecture, and Hadoop usage to scalability and data types. Most questions ask what users are doing today versus what they think they'll do within three years. The comparison indicates directions that DWs and their users are going, and each trend is interpreted in the context of DW modernization. This provides the reader with tips and ideas for planning DW modernizations and other upcoming life cycle stages.

Ripping and Replacing DW Platforms

As we saw in the discussion of Figure 11, ripping out and replacing a data warehouse platform can be a viable modernization strategy—or part of a larger strategy combined with other approaches—when the platform is deficient or outmoded. However, rip and replace can be expensive and disruptive for both business and technical users. Despite those risks, some organizations are considering—or are already committed to—a DW platform replacement.

Yet how many organizations are involved and when might they execute the rip and replace? To quantify the situation, this report's survey asked: When do you anticipate replacing your current primary data warehouse platform? (See Figure 14.)

Rip and replace is real and will become more common.

A third of organizations surveyed will not replace their DW platform. It speaks well for the capability and value of the average data warehouse that so many users have no plans to replace the current primary DW platform (32%).

One-tenth of respondents have already made the replacement. When we consider that an appreciable number of organizations have recently replaced the primary DW platform (9%), we see that the rip-and-replace strategy is actually happening in the real world.

A quarter of respondents plan to replace the DW platform now. Many users surveyed report that they will make the replacement in 2016 (24%).

Another quarter anticipate DW platform replacements in coming years. Respondents report plans for replacement in 2017 (16%), 2018 (3%), and 2019 or later (5%).

When do you anticipate replacing your current primary data warehouse platform?

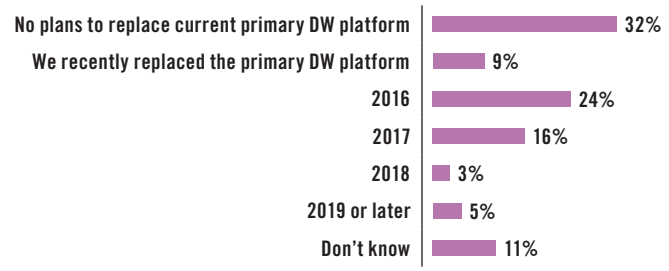


Figure 14. *Based on 473 respondents.*

If we pull together the above information, we see that roughly half of organizations surveyed will make platform replacements within three or four years. Hence, the systems architecture of the average data warehouse (where the platform resides) will be quite different in the future.

Half of organizations surveyed will replace their DW platforms.

Anecdotal evidence heard from users suggests that many are leaving older relational DBMSs for newer ones. Mature vendors now have newer platforms based on racks, grids, or appliances, so for some users the old platform and the new one they're migrating to are from the same vendor. Other organizations are migrating their DW data from mature relational DBMSs to newer brands based on columns, appliances, or open source. These tend to cost less and perform faster, despite having less functionality than mature brands. In related cases, users keep the mature relational DBMS brand in place and complement it with instances of newer relational DBMS brands; one trend is to optimize the former for enterprise reporting and the latter for exploration and analytics.

Despite replacements, relational DBMSs are still preferred.

TDWI has found a few rare cases where all the data of a warehouse is migrated to Hadoop as the primary DW platform. Far more common, however, is to migrate some of the DW's data (especially detailed source data and ODSs) to Hadoop (as a secondary platform on a peer level with the DW). In DWEs, Hadoop is now commonly used as a platform for data staging (for all data, both old and new), analytics processing, data lakes or hubs, and data archiving.

Hadoop is rising as an important but secondary DW platform.

Many other platform possibilities are also playing out in the DW community. Regardless of the direction taken by individual user organizations, vendors, or open source contributors, it's clear that in aggregate we're experiencing a dramatic and exciting evolution in the type and use of data warehouse platforms.[7]

Evolving Data Warehouse Platform Architectures

Once again, recall that TDWI defines a data warehouse as a data architecture populated with data, and the data is managed by a DW platform that's usually based on a relational DBMS and its hardware. Both are layers of a larger architecture. Unfortunately, most of us are sloppy with terminology, such that the term *data warehouse architecture* usually means one of those layers but rarely both. The situation is exacerbated by the fact that modern DWEs integrate multiple platforms.

For this discussion, let's ignore the data and its architecture so we can focus on the platform(s), which constitute a kind of systems architecture (or merely a portfolio of systems, depending on the level of integration). The systems architecture of DWs has become prominent as a source of innovation and modernization, because it includes the new platform types (appliances, columnar, Hadoop) and creative user practices (DWEs, data lakes and hubs, and architectures designed for extremely diverse data types and workloads) that have arisen in recent years.

The platform is but one part of a DW, yet it's where most innovation is occurring.

[7] Actually, DW platform replacement has been an ongoing modernization activity for several years now. See the discussion around Figure 2 in the 2009 TDWI Best Practices Report *Next Generation Data Warehouse Platforms*, online at www.tdwi.org/bpreports. In that study and this one, roughly half of respondents plan to replace the DW platform within three years. By now, we are well into that evolution and more is coming.

To get a sense of the current state and future direction of systems architectures for data warehouse platforms, this report's survey asked: Which of the following best describes the architecture of your extended DW environment today? What about in three years? (See Figure 15.)

Single-monolith DWs aren't as common as you might think. They're also getting rarer. This is how we were taught to build a warehouse back in the 1990s—one relational DBMS instance to assure "the single version of the truth." We put everything in that one instance, even structures that didn't necessarily belong there such as data marts, ODSs, detailed source data, and data staging. The single-monolith reigned for years until data and its requirements evolved to the point that one instance can't be optimized for it all. Today, only 19% of survey respondents report having a central, monolithic DW with no other data platforms. That number will dwindle to 10% as more organizations move to the multiplatform norm of the DWE.

The simple DWE is well established as the norm, and the complex DWE is coming.

The simple DWE has become the norm for DW systems architecture. All DWEs integrate multiple platforms, but the complexity is a matter of degree. Simple DWEs integrate a handful of platforms, whereas complex DWEs may integrate a dozen or more. The systems architecture norm has become the simple DWE, which includes a central DW with a few additional data platforms (34% today). It will continue to be the norm (35% in three years).[8]

Complex DWEs are coming on strong. That's where the DWE includes a central DW with *many* additional data platforms (15% today). The trend in the recent past has been toward the simple DWE; the prominent trend for the near future is toward the complex DWE (30% in three years).

The diversity of data types and workload processing is driving architecture. That's the point of the DWE. It gives users options so they can choose a platform with the storage, performance, and price characteristics that match a given data type or workload. Some users are pushing this practice to its extremes, toward many workload-specific data platforms, where the DW is present but not the center (6% today; 11% in three years).

Which of the following best describes the architecture of your extended DW environment today? What about in three years?

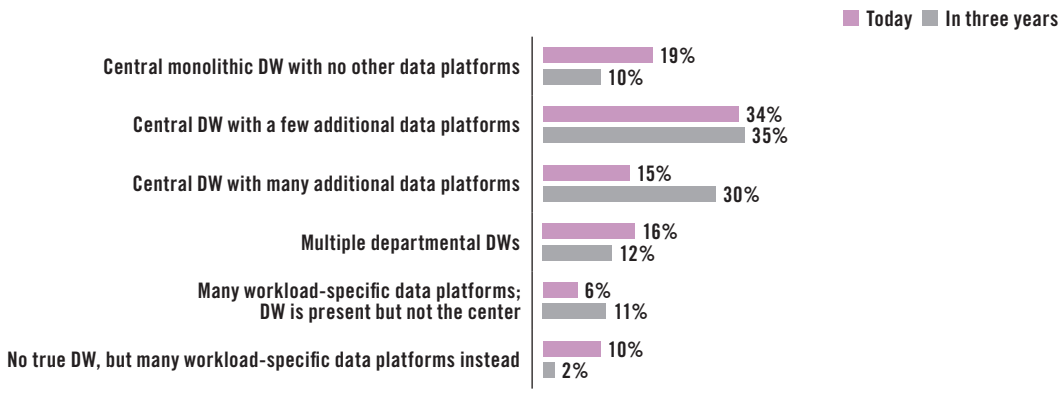


Figure 15. *Based on 473 respondents.*

[8] See the discussion around Figure 10 in the 2014 TDWI Best Practices Report *Evolving Data Warehouse Architectures in the Age of Big Data*, online at www.tdwi.org/bpreports. That earlier study also shows that single-monolith DWs are relatively rare and that central DWs with a few additional platforms have become the norm.

Hadoop's Role in DW Modernization

Hadoop plays many valuable roles in warehousing and analytics.

Recent studies by TDWI have shown that Hadoop is making steady progress as a platform well suited to many purposes in data warehousing and analytics. Many early adopters have already integrated Hadoop clusters and tools into the platform architectures of their data warehouse environments. Hadoop's massive and cheap storage off-loads older systems by taking responsibility for data staging, ELT pushdown, and archiving of detailed source data (retained for advanced analytics). Hadoop also serves as a massively parallel execution engine for a wide variety of set-based and algorithmic analytics methods.[9]

To further measure Hadoop's progress into warehouse architectures, this report's survey asked: What is the role that Hadoop plays in your extended data warehouse environment (DWE) today? What about in three years? (See Figure 16.)

Hadoop is poised for massive adoption by user organizations.

DW usage of Hadoop is still somewhat rare today, but this will change radically. This is clear from the high percentage of survey respondents who have no Hadoop in their DWE today (78%). However, this percentage will drop precipitously within three years (down to 15%) as a large number of organizations deploy Hadoop for DW, analytics, and data integration. With so many organizations committing to Hadoop for data warehousing, every organization should at least consider a role for Hadoop as they modernize data warehouse environments.

Conventional wisdom says Hadoop usually complements a DW without replacing it. That's what early adopters do with Hadoop in DWEs today (17% of respondents in Figure 16). In addition, the percentage of organizations integrating Hadoop with a DW will double within three years (36%).

Hadoop extends DWs and rarely replaces them.

In a few organizations, Hadoop may eclipse a DW without replacing it. For example, several Internet firms have multiple Hadoop clusters each managing petabytes of data in support of numerous analytics applications that operate on Internet-scale data sets. In these cases, the data volume in Hadoop clusters greatly outnumbers the data in their enterprise data warehouses. Many utilities, telcos, and federal intelligence agencies are heading in that direction, too, so it is possible to have Hadoop as the primary DW platform alongside a reduced traditional DW platform (2% today; 14% in three years).

Hadoop as a complete replacement for a DW platform is extremely rare but may get more common. TDWI has interviewed users in this situation, and the bespoke warehouse is usually just a very large operational data store set up like a data lake. In other words, these "warehouses" lack the dimensionality, aggregates, time series, and other advanced features we would associate with a true warehouse. Even so, we know that Hadoop functionality expands and improves almost daily, so it's just a matter of time before Hadoop can support the more advanced features of data warehousing.

Many people don't see a future for Hadoop in warehousing. A huge percentage of respondents (29%) said they don't know Hadoop's role in their DW's future. This is extremely large for a "don't know" response. Hadoop's future in warehousing may be a mystery to a large body of people due to it newness, its exotic nature (as nonrelational open source), or the fact that not everyone needs to handle the big data, massive volumes, advanced analytics processing, and unstructured data with which Hadoop excels. On the flip side, organizations with these requirements are numerous, and they are eagerly committing to Hadoop for modernization and other purposes, which means that Hadoop is here to stay.

[9] For in-depth discussions of Hadoop and its uses, see the TDWI Best Practices Reports *Integrating Hadoop into Business Intelligence and Data Warehousing* (2013) and *Hadoop for the Enterprise* (2015), online at www.tdwi.org/bpreports.

What is the role that Hadoop plays in your extended data warehouse environment (DWE) today? What about in three years?

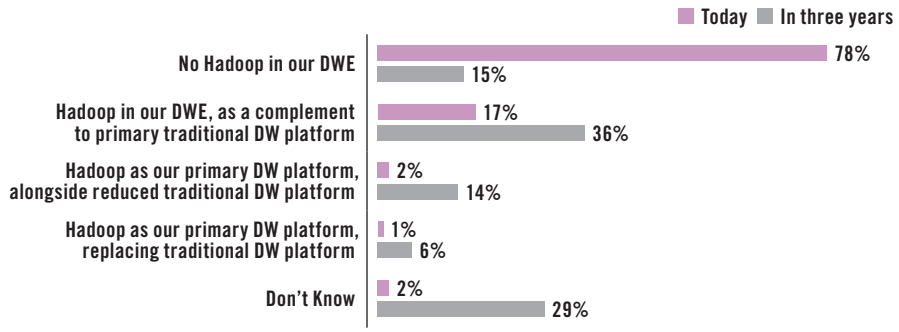


Figure 16. *Based on 473 respondents.*

USER STORY HADOOP FOR DATA STAGING IS A COMMON FIRST STEP FOR DATA WAREHOUSE MODERNIZATION

"We recently replaced our old data staging mechanisms with a data lake on Hadoop," said the assistant vice president of business analytics at an insurance firm. "Now that the Hadoop platform is up and running, we're building more solutions atop it, with analytics in R and Python rolling out in early 2016.

"We're new to open source software, but we're adapting to it quickly. The main thing is that our testing showed that Hadoop can satisfy our current data and analysis requirements but on a budget. At first, our biggest constraint was our lack of skills and experience with Hadoop. We solved that problem by acquiring a vendor distribution of Hadoop (instead of using pure open source from Apache) and by engaging intermediaries (consultants, that is) to set up Hadoop and get us started."

Exotic Data Types in the Modern Warehouse Environment

As we saw in Figure 1, some organizations are currently being challenged by the diversification of data types and formats (e.g., nonrelational, unstructured, and social data), plus the diversification of data sources (e.g., sensors, machines, and GPS). With these drivers in mind, the survey for this report asked: In your organization, which of the following data types are captured and managed in your extended DWE today? Which will be captured and managed within three years? For which do you have no plans?[10] (See Figure 17.)

Expect increased use of social, real-time, IoT, and unstructured data.

Exotic data types are poised for greater future usage. For many users, data is *exotic* when it's not structured, comes from unusual sources, or is generated and ingested in very short time frames. These are exotic when users have no prior experience with them. A number of exotic data types appear at the top of the chart in Figure 17, which is sorted by the data type users will embrace the most within three years. The exotic data types include social media data (51% in three years), real-time data (48%), data from the Internet of Things (IoT; 44%), unstructured data (43%), semistructured data (42%), and Web logs and clickstream data (42%).

Exotic data isn't used much, on average, today. For example, most of the exotic data types poised for future growth are also the least managed today. The issue relative to modernization is that skilled users and tools for exotic data types are missing from most organizations. These gaps are threats to success with modernization, and they must be confronted if organizations are to gain new and deeper insights from data types and sources that are new to them.

[10] For a similar discussion, see "Number One" in the 2015 TDWI Checklist *Eight Tips for Modernizing a Data Warehouse*, online at tdwi.org/checklists.

In your organization, which of the following data types are captured and managed in your extended DWE today? Which will be captured and managed within three years? For which do you have no plans?

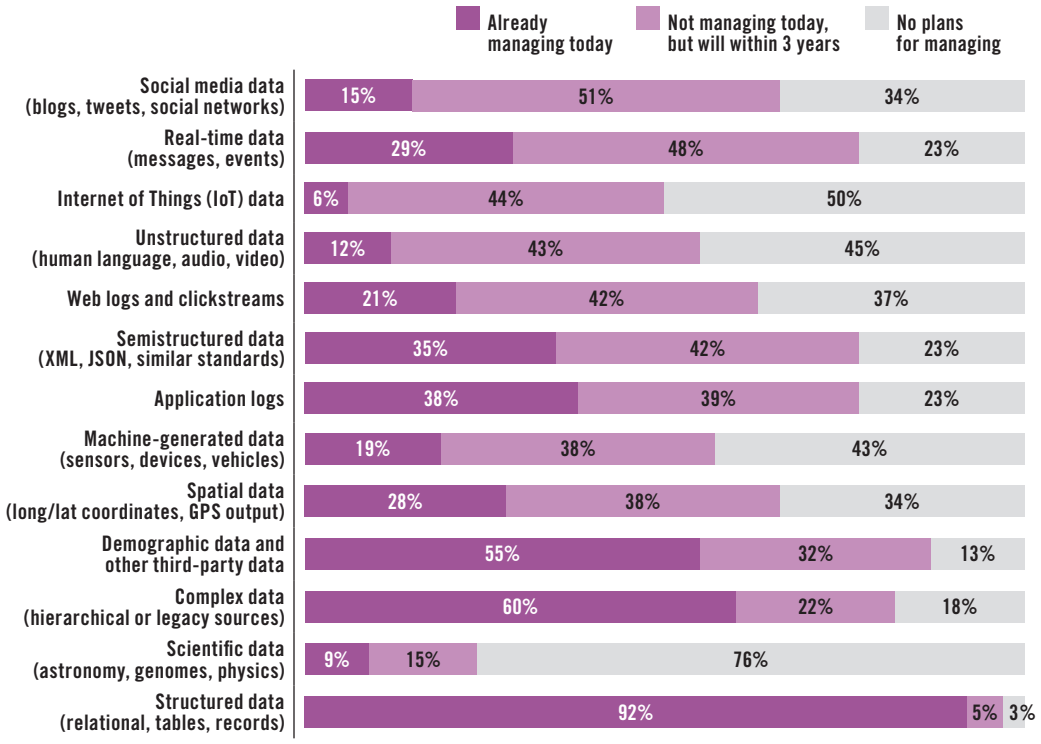


Figure 17. *Based on 473 respondents. Sorted by the column "Not managing today, but will within 3 years."*

Modernizing for exotic data usually requires new tools, platforms, skills, and personnel.

To bootstrap modernization with exotic data, TDWI sees organizations implementing platform types that are suited to the capture and management of nonrelational, nonstructured, and other exotic data types. This includes Hadoop, of course, but also systems for content management and enterprise search. Specialized tools are also needed to process unstructured data—especially human language text—which includes tools for text mining, text analytics, and other approaches to natural language processing. For data streams and other forms of real-time data, TDWI sees organizations implementing tools for complex event processing and other forms of stream processing and correlation.

The lack of skills can be addressed by cross-training existing personnel in the use of new tools and the management of new data. TDWI survey data shows this approach to staffing to be more likely to succeed compared to trying to hire very rare (and expensive) data scientists. Another approach is to hire consultants that specialize in exotic data and its processing, the same way that organizations regularly turn to consultants when attempting anything new and out of the ordinary.

Modernizing for Greater Capacity and Scale

In Figure 1, we saw that the second highest driver for DW modernization is to increase capacity for growing data, users, reports, and analyses. Similarly, number four on the driver list is to accommodate increasing data volumes. With these urgent drivers in mind, our survey asked respondents to estimate the data volume of their primary data warehouse platform for both today and in three years (see Figure 18).

The average DW manages 3–10 TB today, increasing to 10–100 TB in three years.

Today's norm for DW data volume is 3–10 TB. For the average-size organization, aggregated across industries and other characteristics, the data volume norm for data warehouses falls into the range of 3–10 TB today (23%). However, users will abandon this range in the next few years (down to 14% in three years) as their data stores swell into ranges that are 10 TB or greater.

In the near future, the norm will be 10–100 TB. The 10–100 TB range is already prominent (20% today), but it will grow even larger (27% in three years) as DW programs graduate from lesser data volumes to greater ones.

Small DWs will become less common, large ones more so. This is natural because, in almost all situations, the amount of data generated and captured for use with DWs, analytics, and other business intelligence continues to increase at a rapid pace.

These numbers are conservative. For some reason, this survey did not draw many users from utilities, telcos, Internet firms, huge global firms, and other industry types that are known for outrageously massive data volumes reaching into petabytes. If it had, the chart in Figure 18 would have longer bars in the "greater than 500 TB" range. Users with petabyte-scale data warehouses certainly exist, and a few have spoken at TDWI events. They are a small minority (3% today), but more will soon join "the petabyte club" (9% in three years).

Indicate the approximate total data volume that your organization manages in your primary data warehouse platform, both today and in three years.

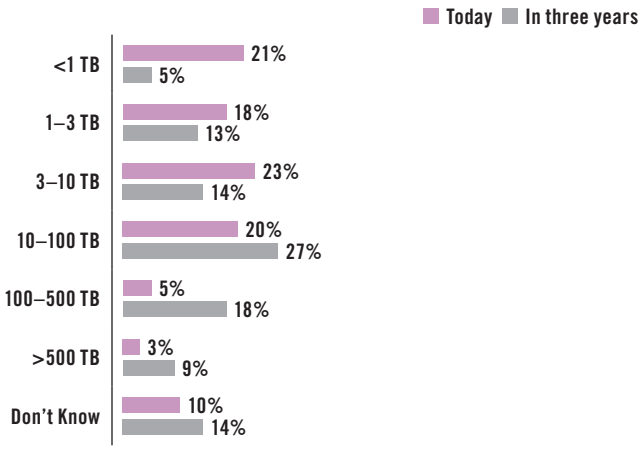


Figure 18. *Based on 473 respondents.*

10, 100, and 500 TB are known capacity goals for successful DWs.

What do these data volumes mean to DW professionals? The ranges included as multiple-choice answers in Figure 18 are based on interviews with real-world DW professionals over the years. Users with small or new DWs say that 3 TB is a comfortable beginner range, but they expect 10 TB soon if the DW and its organization succeed. Users with mature and successful DWs feel they need a system that can start near 10 TB and eventually expand to 100 TB.

In extreme, data-rich organizations (such as Internet firms and utilities), the DW team already needs 100 TB (sometimes 500 TB) and expects to expand into multiple petabytes in a few years. Admittedly, DW professionals tend to overbuild (as most IT disciplines do); yet 10, 100, and 500 TB are known capacity goals that can guide DW modernization.

USER STORY FOR MANY ORGANIZATIONS, OPEN SOURCE SOFTWARE IS KEY TO DATA WAREHOUSE MODERNIZATION

"We have 5 to 10 billion data points coming into our data warehouse environment every day," said Michael Liebman, a BI director at Bloomberg LP, "so not just any database management system can handle our data volumes and intense processing loads. We started taking open source software very seriously a few years ago when we started to realize that some of the best solutions in the market were open source technologies. After much testing and evaluation, we committed to a vendor distribution of Hadoop. The Hadoop cluster in our data warehouse environment is currently managing 200 terabytes today, and we feel that our implementation will scale into petabytes in the near future.

"We use Hadoop as a platform for both data staging and analytics processing, plus a bit of reporting. It integrates with our data warehouse platform, which is a large MPP configuration of a relational database managing around 400 terabytes of compressed data. Our MPP database is now open source as well.

"We feel that our technology stack is already very modern, but we know that we have to keep modernizing to stay ahead of our exploding data volumes and processing loads. For example, we have started evaluations of open source databases on top of Hadoop, such as Apache HAWQ. HAWQ is a Hadoop-native SQL query engine that combines the key technological advantages of relational MPP databases with the scalability and low cost of Hadoop.

"A catch to deepening our commitment to Hadoop is the need to continue to hire the best and brightest in the field of data science. Recently, we appointed a head of data science, Gideon Mann, who comes to us from Google. There are other recent hires that are moving us down the path, but there is also a growing need for more."

Vendors' Platforms and Tools for DW Modernization

The firms that sponsored this report are all good examples of software vendors that offer tools, platforms, and professional services that are regularly involved in data warehouse modernization. The sponsors form a representative sample of the vendor community. Yet their offerings illustrate different approaches for growing and enhancing the modern data warehouse and related systems.[11]

IBM

IBM sees the need for greater speed and scale; the need for IT to respond more quickly to the business for faster and new types of analytics; and the need to exploit new technologies as driving factors for organizations modernizing their data warehouse environments, moving toward a logical data warehouse and adopting hybrid architectures. For these use cases, IBM offers a suite of products for different deployment options. Because Hadoop is now an established platform for managing big data, IBM InfoSphere BigInsights for Apache Hadoop enhances open source Apache Hadoop to make it enterprise grade and to support text analytics, data exploration and discovery, visualization, developer tools, and connectivity to enterprise sources. For other approaches to data warehouse modernization, IBM offers PureData System for Analytics (an integrated and optimized appliance for analytics), IBM dashDB (for a managed data warehouse service on the cloud), and DB2 with BLU Acceleration (for building a software-defined data warehouse). Streaming data and other forms of real-time data are pressing modernization issues for many users, so IBM offers InfoSphere Streams, a complex event processing system that supports the continuous analysis of massive volumes of streaming data and large data bursts for real-time insights with subsecond responses.

Pentaho, A Hitachi Group Company

Pentaho, a Hitachi Group company, provides a commercial open source platform that tightly couples data integration and business analytics. In response to customers' needs for data warehouse modernization and optimization, Pentaho has developed deep support for Hadoop and other big data stores, facilitating the entire life cycle of big data integration and analytics. At the heart of the Pentaho platform is Pentaho Data Integration (PDI), providing an intuitive visual experience for preparing, blending, and processing data at scale. Pentaho's Adaptive Big Data Layer insulates users from a rapidly changing big data ecosystem and promotes maximum portability. Through it, Pentaho supports the latest Hadoop distributions from Cloudera, Hortonworks, Amazon Web Services, and MapR. Pentaho also integrates with NoSQL databases (Cassandra and MongoDB) and analytics databases (HP Vertica and Amazon Redshift). Pentaho's "Optimize the Data Warehouse" solution approach can help reduce the strain on existing data warehouses by off-loading less frequently used data and corresponding transformation workloads to Hadoop without the need for manual coding.

SAP

SAP provides a comprehensive set of solutions for big data, including analytics applications; rapid deployment solutions; BI; and advanced analytics tools, analytics databases, data warehousing solutions, and information management tools. All these and other SAP products or services can be applied to data warehouse modernization. Because big data and analytics are major drivers for modernization, SAP enables its customers to integrate Hadoop into their existing SAP HANA, BI, advanced analytics, and data warehousing environments in multiple ways so customers are able

[11] The vendors and products mentioned here are representative, and the list is not intended to be comprehensive.

to tailor Hadoop to their needs. Customers can use SAP Data Services to search and load data from HDFS or Hive into SAP HANA or SAP Sybase IQ. Furthermore, SAP BusinessObjects BI, SAP Visual Intelligence, and SAP Predictive Analysis users can query Hive environments, helping business analysts explore Hadoop data directly. Finally, customers can federate queries across SAP IQ, SAP HANA, and Hadoop environments or, alternatively, run MapReduce jobs across an SAP IQ cluster. To accelerate Spark and Hadoop-based queries, SAP offers SAP HANA Vora, an in-memory compute engine for Hadoop.

SAS

SAS's big data management and advanced analytics solutions help customers make better decisions faster. The following capabilities—all supported by SAS—can enable the modernization of your data warehouse:

- **Data integration:** Data movement, in-database processing, and native data access to traditional and emerging data sources such as Hadoop
- **Data quality:** Cleanse, standardize, and enrich data in real time and batch with prebuilt rules
- **Self-service big data preparation:** Business users profile, cleanse, and transform data on Hadoop without writing code
- **Business glossary and metadata management:** Track lineage, business rules, descriptive details, and workflow for improved governance of data assets
- **Event stream processing:** Analyze real-time streams of data in motion for better decisions
- **Data virtualization:** Provide blended, secure views of data without moving it
- **Hadoop support:** Access, deliver, and process data inside Hadoop across both the data management and analytics life cycle
- **Visualization and advanced analytics:** Deliver cutting-edge visualization and analysis capabilities without requiring analytical skills

TimeXtender

TimeXtender (TX) provides data warehouse automation (DWA) solutions built atop Microsoft SQL Server technologies. The TimeXtender software—TX DWA—modernizes the way a data warehouse is developed and maintained by automating all manual data warehouse processes—from design and development to operations and maintenance to change management. DWA tools blend user requirements and repeatable processes to automatically generate the necessary components of a modern data warehouse environment in a fully documented solution. By automating these tasks, TX DWA gives its users substantial speed and agility. This shortens development and delivery time, which in turn shortens the time to use and value for the business. Other benefits of TX DWA include reduced start-up and overhead costs, simplified access for business users, and enhanced IT-to-business collaboration. TimeXtender collaborates with VAR and OEM partners across six continents, providing more than 2,600 customers in 61 countries with its modern data warehouse automation software.

Top 12 Priorities for Data Warehouse Modernization

In closing, let's summarize the findings of this report by listing the top 12 priorities for data warehouse modernization, including a few comments about why these priorities are important. Think of the priorities as recommendations, requirements, or rules that can guide user organizations into successful strategies for implementing a modernization project.

1. **Embrace change.** Data warehouse modernization is real; our survey says that 76% of DWs are evolving moderately or dramatically. Given the rampant amount of change in markets and individual businesses, it's unlikely the status quo will serve you and your organization for much longer. Besides, change is an opportunity for improvement as long as you manage it with specific directions in mind.

2. **Make realignment with business goals your top priority.** This is the leading driver (39% of respondents in Figure 1). Learn the goals of the business and collaborate with business and technical staff to determine how business goals map to technology and data. Base your modernizations on the requirements you've defined. If your project aligns with your business goals, your entire business will modernize, not just your warehouse. After all, that's the real point.

3. **Make DW capacity a high priority on the technology side.** The second most pressing driver is greater capacity for growing data, users, reports, and analyses (37% in Figure 1). This is no surprise given the explosive growth of traditional enterprise data and new big data. Today's norm for DW data volume is 3–10 TB in the average-size organization; however, the norm will soon become 10–100 TB as DW programs graduate from lesser data volumes to greater ones. These and other volume ranges described in this report are known capacity goals for successful DWs, so keep them in mind when planning capacity modernization.

4. **Make analytics a priority, too.** One-third of DW professionals modernize for better and newer analytics. That's a technology challenge for the warehouse because diverse analytics techniques have diverse data preparation requirements, and they don't all fit the traditional warehouse. Therefore, additional data platforms and tools that complement older ones may be in order. Keep in mind that analytics is what business users want; your pristine data and elegant architecture won't mean much if modernization fails to deliver relevant analytics.

5. **Don't forget the related systems and disciplines that also need modernization.** Top priorities are analytics, reporting, and data integration followed by development methods and team characteristics. Align the modernization of the DW so it can ably provision the data in a manner that these other entities require for their success.

6. **Don't be seduced by new, shiny objects.** To paraphrase Duke Ellington: IT don't mean a thing if it only got that bling. In other words, there are lots of new and cool technologies and tools available today, and many get evaluated for DW modernization. Before adopting one, be sure it goes beyond the bling to satisfy real-world requirements in a performant and cost-effective manner.

7. **Assume that you'll need multiple manifestations of modernization.** This report describes common categories and strategies of DW modernization, namely: system, arbitrary, non-DW, optimization, continuous, and disruptive modernizations. To get the desired results, you should consider multiple strategies but try not to execute them all at once, in a big bang.

8. **Know the tools and techniques of the modern data warehouse environment (DWE).** These are commonly applied as the main modernization strategy or as support for others, because the DWE is one of the strongest trends in data warehousing.
9. **Adjust the large-scale architecture of your DWE.** The rise of the multiplatform DWE is forcing the modernization of system architectures. For most situations, you will keep and improve your centralized, relational DW. You should, however, expect to complement it with other platforms, and then migrate data and balance workloads among platforms. This requires you to rework the large-scale architecture, which determines how diverse platforms integrate and interoperate, plus which data goes where and how data flows among platforms.
10. **Reevaluate your DW platform.** The condition of your data is important, but it's all for naught if the platform can't capture, manage, and deliver data with speed, scale, and broad functionality at a reasonable cost. Replacing a DW platform is disruptive and expensive for a business. Therefore, consider leaving your DW platform in place, but update it and complement it with other systems. Even so, grossly deficient or outmoded platforms should be replaced.
11. **Consider Hadoop for various roles in the DWE.** Hadoop's massive and cheap storage off-loads older systems by taking responsibility for data staging, ELT pushdown, and the archiving of detailed source data (retained for advanced analytics). Hadoop also serves as a massively parallel execution engine for a wide variety of set-based and algorithmic analytics methods. Conventional wisdom says Hadoop usually complements a DW without replacing it. That's what early adopters do with Hadoop in DWEs today (17% of respondents in Figure 16), and the percentage of organizations integrating Hadoop with a DW will double within three years (36%).
12. **Develop plans and recurring cycles for DW modernization.** Most DW teams have settled on a quarterly schedule for updating DWs. This applies to tasks of many sizes; well-contained phases of some modernization projects may fit this scheme, as well. However, large-scale modernizations typically need their own plan. The more disruptive a modernization (such as rip and replace), the more critical to success is the multiphase (sometimes multiyear) plan. Modernization affects business users and their processes; for minimal disruption, business managers should be involved in developing and executing modernization plans.

sas.com/data

From Data to Decision: How SAS Can Help Modernize Your Data Warehouse

SAS is pleased to sponsor this TDWI Best Practices Report. SAS can help you meet the modernization challenges inherent in creating better-aligned, governed, and trusted data. Improve the productivity of data professionals by building an innovative analytics culture with an agile data architecture to support it.

SAS software provides the following capabilities:

- **Data integration:** Data movement, in-database processing, and native data access to traditional and emerging data sources such as Hadoop.
- **Data quality:** Cleanse, standardize, and enrich data in real time and batch with prebuilt rules.
- **Self-service big data preparation:** Business users can profile, cleanse, and transform data on Hadoop without writing code.
- **Data governance:** Business glossary and metadata management help track lineage, business rules, descriptive details, and workflow for improved governance of data assets and alignment of business and IT.
- **Event stream processing:** Analyze real-time streams of data in motion for better decisions.
- **Data virtualization:** Provide blended, cleansed, and secure views of data without moving it.
- **Hadoop support:** Access, deliver, and process data inside Hadoop across both the data management and analytics life cycles.
- **Visualization and advanced analytics:** Deliver cutting-edge visualization and analysis capabilities without requiring analytical skills.

How SAS Is Different

Trusted, decision-ready data is in our DNA. As analyst-validated leaders in data integration and data quality for many years, providing cleansed, reliable data is in our DNA. We're experts at empowering data professionals, modernizing data processes, and fueling confident decisions.

Manage data where it lives. Whether it's in stream, in database, in-memory, or inside Hadoop, SAS reduces data movement and improves performance by pushing the processing to the data.

Enterprise-grade stability. With 40 years of analytics and data management expertise, SAS is uniquely qualified to help you capitalize on the quickly evolving Hadoop ecosystem. Users can explore, manage, and analyze Hadoop data on their own using interactive and familiar SAS products—a critical feature given the skills shortage and complexity involved with Hadoop.

Superior analytics for Hadoop. Big data opportunities hidden in Hadoop are best exploited with the latest analytics techniques. We combine proven analytical algorithms—predictive analytics, machine learning, optimization, and text mining—with a unique, in-memory engine to gain unprecedented speed and accuracy.

Learn more: **sas.com/data**

TDWI Research provides research and advice for data professionals worldwide. TDWI Research focuses exclusively on business intelligence, data warehousing, and analytics issues and teams up with industry thought leaders and practitioners to deliver both broad and deep understanding of the business and technical challenges surrounding the deployment and use of business intelligence, data warehousing, and analytics solutions. TDWI Research offers in-depth research reports, commentary, inquiry services, and topical conferences as well as strategic planning services to user and vendor organizations.

tdwi
Advancing all things data.
555 S Renton Village Place, Ste. 700
Renton, WA 98057-3295

T 425.277.9126
F 425.687.2842
E info@tdwi.org

tdwi.org