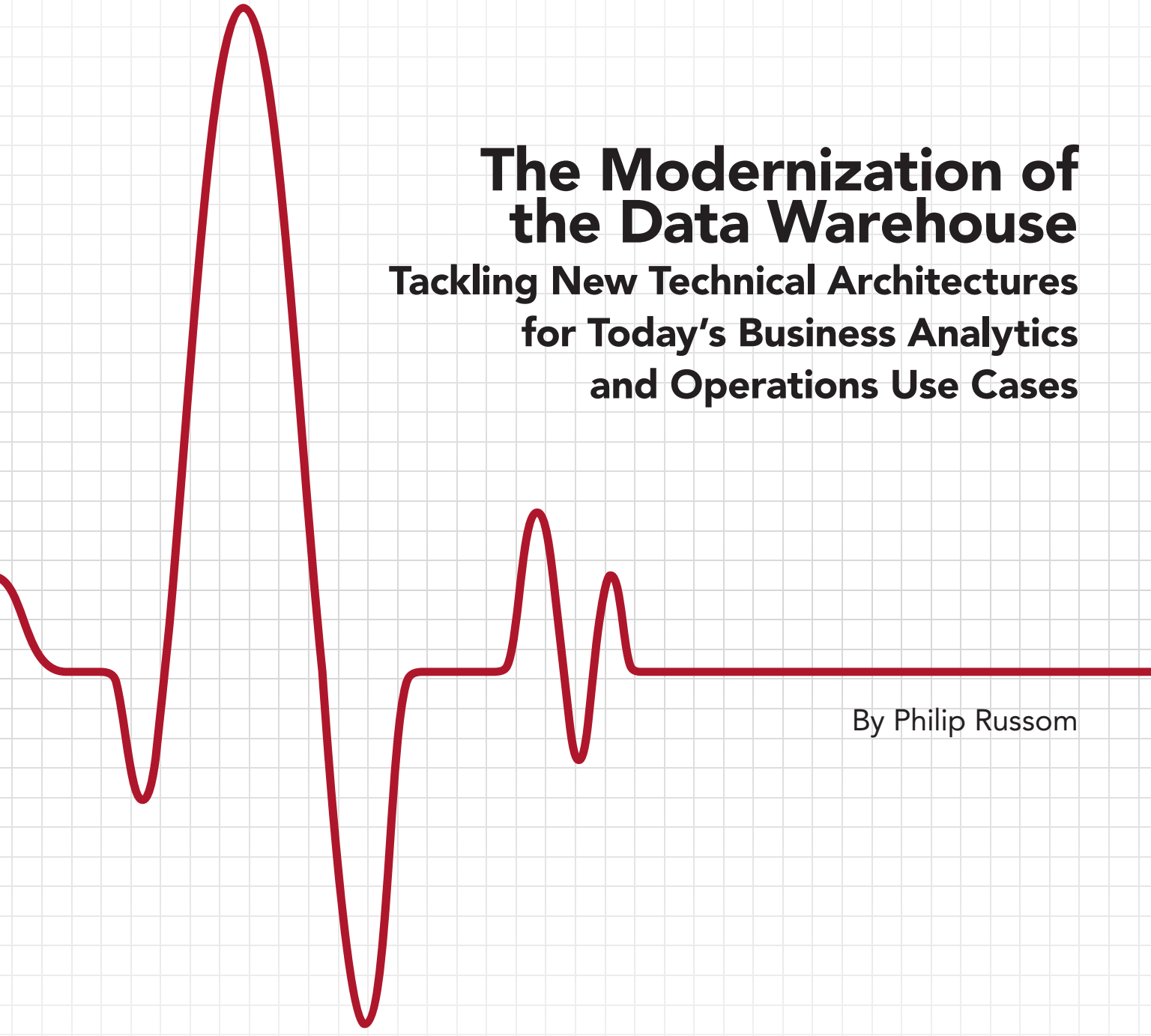
research

PULSE

# The Modernization of the Data Warehouse

## Tackling New Technical Architectures for Today's Business Analytics and Operations Use Cases



By Philip Russom

Sponsored by:

cloudera®

tdwi
Transforming Data With Intelligence™

research

TDWI PULSE REPORT

Q4 2019

# The Modernization of the Data Warehouse

## Tackling New Technical Architectures for Today's Business Analytics and Operations Use Cases

By Philip Russom



# Table of Contents

## About the Author

**PHILIP RUSSOM, Ph.D.,** is senior director of TDWI Research for data management and is a wellknown figure in data warehousing, integration, and quality. He has published more than 600 research reports, magazine articles, opinion columns, and speeches over a 20-year period. Before joining TDWI in 2005, Russom was an industry analyst covering data management at Forrester Research and Giga Information Group. He also ran his own business as an independent industry analyst and consultant, was a contributing editor with leading IT magazines, and a product manager at database vendors. His Ph.D. is from Yale. You can reach him at prussom@tdwi.org, @prussom on Twitter, and on LinkedIn at linkedin.com/in/philiprussom.

## About TDWI Research

TDWI Research provides research and advice for data professionals worldwide. TDWI Research focuses exclusively on data management 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 data management 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.

## About TDWI Pulse Reports

This series offers focused research and analysis of trending analytics, business intelligence, and data management issues facing organizations. The reports are designed to educate technical and business professionals and aid them in developing strategies for improvement. Research for the reports is conducted through surveys of professionals. To suggest a topic, please contact TDWI senior research directors Fern Halper (fhalper@tdwi.org), Philip Russom (prussom@tdwi.org), and David Stodder (dstodder@tdwi.org).

## Acknowledgments

TDWI would like to thank many people who contributed to this report. First, we appreciate the many professionals who respond to our surveys. Second, our report sponsor, who diligently reviewed outlines, survey questions, and report drafts. Finally, we would like to recognize TDWI's production team: James Powell, Peter Considine, Lindsay Stares, and Michael Boyda.

## Sponsors

Cloudera sponsored the writing of this report.

# The Pulse: A Modern Data Warehouse Is Required for Modern Business Practices in Analytics and Operations

## What makes the modern data warehouse *modern*?

Three characteristics of modernity stand out:

1. A modern data warehouse must enable new, data-driven business practices, especially those for advanced analytics, self-service, and data sharing across functional business areas
2. To achieve the first requirement, a modern data warehouse must incorporate new data platforms (e.g., cloud-based databases, Hadoop, or NoSQL), new computing platforms (clouds and clusters), and new data structures from new sources (the web, social media, and the Internet of Things (IoT))
3. Tools that accompany the warehouse—for analytics, reporting, and integration—must be equally modern, in that tools (from both vendor and open source communities) have been created or updated to deeply support all the new data-driven business practices, data platforms, and data types just mentioned

## Why is the modernization of the data warehouse such a pressing issue?

**The end game is the modernization of business.** Many enterprises are motivated to become more data driven so they have quantified facts and can be guided by analytics. In turn, these capabilities make substantial contributions to achieving innovative business goals, such as the digital enterprise and business transformation. Making all that work requires a fully modernized data warehouse. In many cases, similar modernizations should be made in other enterprise data environments, particularly those for marketing, finance, supply chain, and IoT operations.

**Business innovation requires bigger and better data.** This is especially apparent in emerging practices for advanced analytics (based on mining, clustering, statistics, machine learning, etc.) and self-service data access (for data discovery, prep, visualization, etc.). Data warehouses are under pressure to provision the voluminous and structurally diverse data required of these innovative practices. In fact, TDWI believes that the business need for new analytics is the strongest driver for the modernization of the data warehouse.

**Legacy warehouse designs need serious updating.** The average data warehouse today was designed by technical users to provision data for reporting, dashboards, and online analytic processing (OLAP). This kind of warehouse design is still relevant because those use cases are still needed by most enterprises. However, the traditional report-oriented warehouse is ill-suited to the advanced analytics, self-service, and new data sources that current businesses practices demand. Hence, many existing warehouse designs need to be modernized, augmented, and optimized to address the requirements of both yesterday and today. Furthermore, users designing a new data warehouse should keep old and new sets of requirements in mind.

**Technical users need to rethink their platform choices.** There is a long tradition of users choosing an on-premises relational database management system (RDBMS) as the primary or only data platform for a warehouse. The catch is that the on-premises RDBMS has limitations in speed, scale, agility, handling of unstructured data, and analytics workload support, plus it carries a high cost when deployed as a large massively parallel processing (MPP) configuration.

To achieve the technical goals of the modern data warehouse, many users have decided to "replatform" by migrating warehouse data (wholly or in part) to new data platforms (such as those based on Hadoop), columns, clouds, and new relational databases, with cloud-based data platforms emerging as a preference. Replatforming typically swaps an older platform for a newer one. However, a related modernization strategy is to keep the old warehouse platform (optimized mostly for reports) while augmenting it with additional new ones (optimized mostly for analytics) in an architecture where all platforms are tightly integrated.

This report will now drill into four critical success factors for the modernization of the data warehouse:

- Compelling business and technology drivers for the modernized data warehouse
- New platforms and strategies for data management
- How data modernization is aided by cloud and hybrid cloud environments
- Hybrid data architectures specifically for the modernized data warehouse

We will include examples of technical practices, platforms, and tool types, as well as how the modernization of the data warehouse supports data-driven business goals.

# Drivers for Data Warehouse Modernization

A recent TDWI survey asked: What are the leading drivers for the modernization of your DW? (See Figure 1).[1] The question generated 5.7 responses per respondent, on average, which indicates that the average user organization is working hard to satisfy several drivers simultaneously. The drivers identified in the survey fall into a few broad areas:

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

**Most technical modernizations of a DW enable greater scale and speed.** The second most common driver for modernization is to increase capacity for growing data, users, reports, analyses, and so on (37%). 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%).

[1] This figure and most of its discussion originally appeared in the 2016 *TDWI Best Practices Report: Data Warehouse Modernization in the Age of Big Data Analytics*. This and other TDWI Best Practices Reports are available at www.tdwi.org/bpreports.

## The leading drivers of modernization are DW-to-business alignment, speed and scale, new analytics, and new platforms and tools.

**One-third of DW professionals modernize for better and newer analytics.** Near the top of the 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 online analytic processing (OLAP) (12%). Note that new analytics complement—but don't replace—standard reports and OLAP; each delivers unique insights and guidance, so all are required by the modern business.

**Many users modernize to embrace new best practices and tool types.** Vendor, open source, and consulting communities have recently brought us new tools and methods for leveraging data for business 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 data virtualization (12%).

**Life cycle issues lead to redesigning or replacing some DWs.** 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 DW modernizations are driven by problems with the existing design or architecture (24%), or problems with the existing, underlying DW platform (16%).

**New types of big data and the data platforms built for them are emerging drivers.** Skills and tool portfolios tuned to SQL and the relational paradigm 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%). Organizations capturing new data assets should modernize their skills and portfolios of tools and data platforms.

**Replatforming has become a common manifestation of modernization.** Because older data platforms are not always suited to new big data volumes or formats, some users are turning to Hadoop (18%) and NoSQL (7%). However, TDWI believes that the trend is toward cloud or SaaS adoption (11%), which 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.

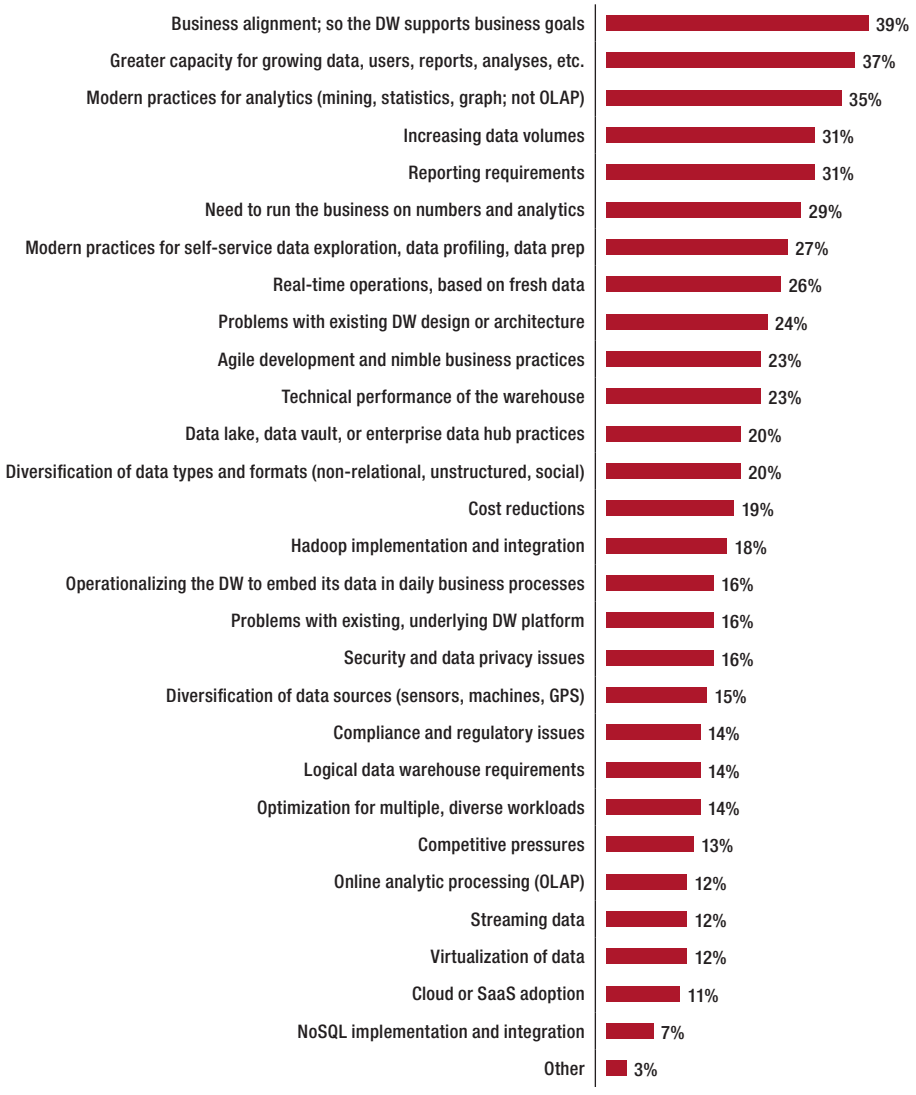



***Figure 1.*** *Based on 2684 responses from 473 respondents (5.7 answers per respondent on average).*

# New Platforms and Strategies for Modernization

It is good to have options and alternatives. With the modernization of the data warehouse, business and technology users are blessed with many of these so they can choose the platforms and strategies that best serve today's new data and uses cases.

## A modern DW may include multiple data platforms, both old and new.

That's because it is difficult (and sometimes impossible) to optimize one data platform for the highly diverse data types, use cases, and analytics workloads that users need today. In a related trend, users need a broad range of modern tools for analytics and integration. Having a diversified portfolio of data platforms and tools enables users to match data and use cases to the right storage and processing solutions. Figure 2 quantifies the types of platforms and tools found in and around the modern data warehouse and other complex data environments today.[2]

**Database management systems (DBMSs).** Given their history, maturity, and installed base, it is no surprise that RDBMSs are still the most common data platform in data warehouses and other data environments (56% in Figure 2). Also available are new analytics database management systems (ADBMSs, 31%), which are based on columns, graph, appliances, or clouds. Users with highly diverse business and data requirements regularly deploy old and new DBMSs of various brands and types.

**Hadoop.** Although it is not a DBMS per se, Hadoop (43%) has come on strong in recent years, especially as an extension or complement of RDBMSs (though rarely a replacement). For example, other TDWI surveys have shown that around 60% of data warehouse environments now include Hadoop, usually integrated tightly with an RDBMS.[3]

Modern data warehouse platforms include clouds, Hadoop, open source, and many types of DBMSs.

**On-premises data platforms and tools.** As a general-purpose compute platform, the cloud is clearly the strongest gainer in recent years, with compelling use cases in warehousing and analytics. However, on-premises systems (47%) continue to linger, especially those for data integration platforms on premises (50%), data warehouses on premises (50%), and analytics tools on premises (46%). Note that each of these was reported by roughly half of respondents, which is surprisingly low, meaning that migrations from on-premises to cloud-based data management tools and platforms must have already occurred at many organizations.

**Cloud-based data platforms and tools.** The cloud is progressively becoming a preferred platform for a variety of data-driven use cases. We see solid cloud adoption in survey responses, where roughly a quarter of respondents are already using data warehouses in the cloud (29%), data integration platforms in the cloud (26%), and analytics tools in the cloud (25%). In addition, TDWI is starting to see strong adoption of object storage on cloud platforms (21%) as a highly scalable and affordable alternative to on-premises storage. Other TDWI surveys show that users appreciate the cloud's elastic scalability for burgeoning big data and unpredictable analytics workloads.[4]

[2] This figure and most of its discussion originally appeared in the 2018 *TDWI Best Practices Report: Multiplatform Data Architectures*, online at tdwi.org/bpreports.

[3] See the 2015 *TDWI Best Practices Report: Hadoop for the Enterprise*, online at tdwi.org/bpreports.

[4] See the 2016 *TDWI Best Practices Report: BI, Analytics, and the Cloud*, online at tdwi.org/bpreports.

**Open source software (OSS).** As with the rest of the enterprise, modern data warehouses involve various types of open source software (24%), especially Hadoop (43%). However, Spark is coming on strong (31%) because data professionals want Spark's data-driven libraries for SQL, machine learning, and microbatching big data at speed and scale.

**For the complex data environment that you use most, what types of data and compute platforms are involved? Select all that apply.**

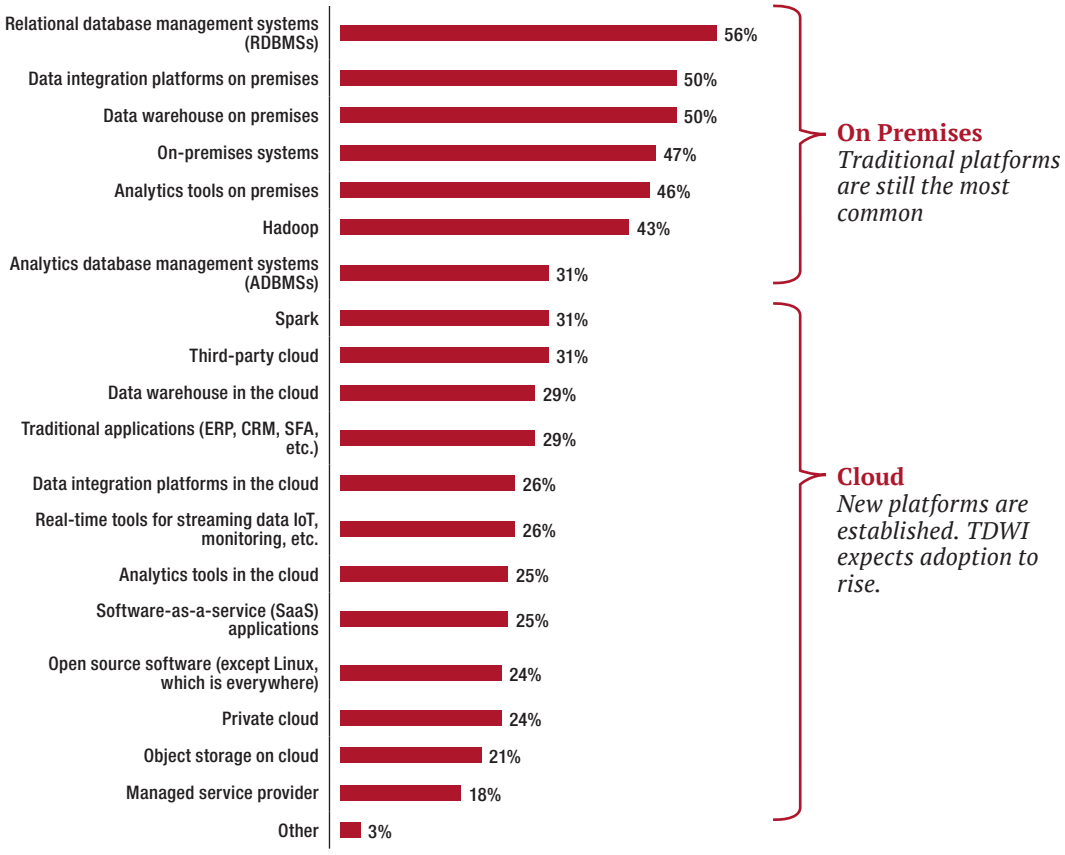



***Figure 2.*** *Based on 431 responses from 68 respondents (6.3 responses on average).*

## There are multiple strategies for the modernization of the data warehouse.

Figure 3 illustrates user preferences for various data warehouse modernization strategies.[5]

**Augment (but don't replace) existing a data warehouse's primary platform by adding additional data platforms and tools (42%).** In other words, most users will leave their primary DW platform in place but complement it with other systems, typically new data platform types. From a business viewpoint, this is a non-disruptive strategy in that it preserves existing investments in data warehousing and (when done well) extends the life of an expensive and useful system. For years, TDWI has seen users deploy on-premises data warehouse appliances and columnar databases as part of their warehouse augmentation and modernization strategy. More recently, Hadoop and cloud-based DBMSs have become prominent additions.

**Strategy determined on a case-by-case basis (24%).** Modernization hits many different layers of the overall technology stack, plus has diverse business drivers, so it is inevitable that a certain amount of case-by-case examination is appropriate.

**Replace existing data warehouse's primary platform (15%).** This is also called *replatforming* and *rip-and-replace*. For users with a deficient or outmoded DW platform, this approach is fully appropriate despite the disruption and expense. Here, 15% is rather conservative; the responses to a different question in the same survey show that 57% of respondents have already replaced their primary DW platform or will do so in a few years.[6]

**No strategy, because we do not need one (1%).** Almost no one surveyed thinks that a strategy for data warehouse modernization is not necessary. Even people who don't have a strategy (14%) know that they should.

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

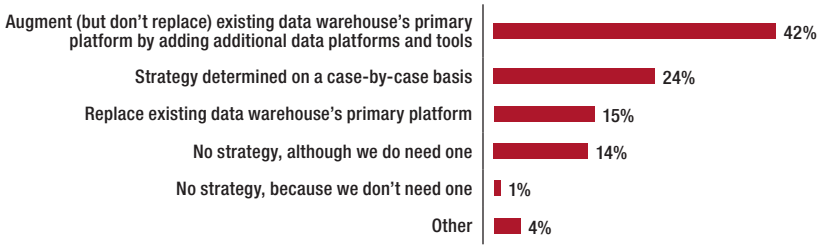



***Figure 3.*** *Based on 473 respondents.*

[5] Figure 3 in this report originated as Figure 11 in the 2016 *TDWI Best Practices Report: Data Warehouse Modernization*, online at tdwi.org/bpreports.

[6] Ibid, Figure 14

# Cloud and Hybrid Environments for the Modern DW

As a general computing platform, the cloud has been adopted by organizations of many sizes in all industries for many technical systems and business use cases. Thanks to the broad proliferation of new operational applications in recent years, deployed on clouds in the software-as-a-service model, cloud computing is now firmly established as the preferred, future-facing, modern platform for applications.

We are now well into a cloud maturity stage where users are building on their cloud success by deploying cloud-based systems for data warehousing, analytics, reporting, and data integration. As is often the case, operational applications adopted the new platform first and proved its business and technology value; cloud data management solutions and analytics applications are now accelerating to catch up.

TDWI defines *cloud data management* (CDM) simply as "data management that involves clouds." As with most forms of data management, CDM involves tasks and technologies for data persistence, data integration, and data semantics, as well as applications for analytics, reporting, and self-service. The data platforms and tools involved in CDM may run on premises, in the cloud, or in a combination of these.

Cloud is quickly becoming the preferred platform for the modern data warehouse and related data management solutions.

## Cloud data management (CDM) has compelling benefits.

The technical point of CDM is to extend data management technologies and practices to deeply support multiple cloud types, data originating in clouds, and data integration across on-premises and cloud systems. The business point is to achieve the greatest organizational value from cloud data and systems, typically via analytics and other data-driven applications. CDM enables several business and technology benefits, as shown in Figure 4.

According to survey responses, the leading benefit of CDM is that it boosts scalability for data storage and integration workloads (51% in Figure 4) and does so via automatic and elastic resource management (44%). In turn, CDM enables advanced analytics at scale but inexpensively (35%). With the right data integration capabilities, CDM enhances real-time access to all data, whether on premises or on cloud platforms (35%). From a business viewpoint, CDM assures that data and other assets or resources are more fully leveraged (32%). It also makes data easier to share with external suppliers, partners, and customers (30%). Finally, note that CDM modernizes mature data management infrastructure (30%), which includes the modernization of the data warehouse.[7]

[7] For more details, see the discussion about Figure 6 in the 2019 *TDWI Best Practices Report: Cloud Data Management*, online at tdwi.org/bpreports.

**If your organization were to implement cloud data management, what would its leading benefits be? Select all that apply.**

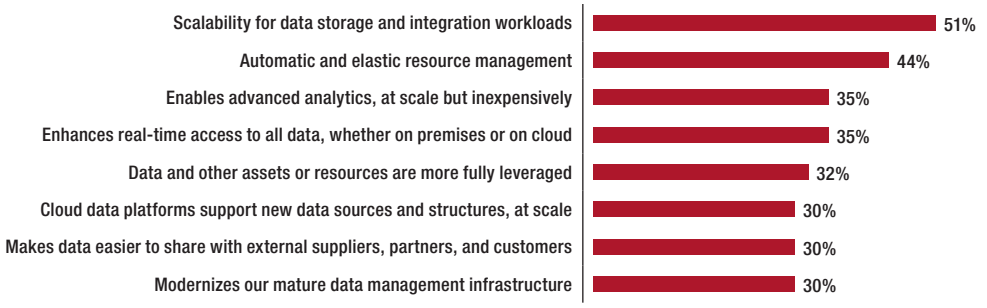



***Figure 4.*** *Based on 605 responses from 98 respondents (6.2 responses on average).*

## Cloud data management touches almost all business processes today.

CDM practices are in place today, supporting numerous production use cases in both analytics and operations (see Figure 5).[8] For example, the top four bars in Figure 5 are all for use cases in analytics, whereas the other bars are for operational use cases. This illustrates that CDM is a real-world best practice, providing valuable support for a wide range of business functions and departments. In other words, cloud data management has penetrated enterprises broadly, touching almost all business processes today.

**For what enterprise functions or use cases is your organization applying cloud data management? Select all that apply.**

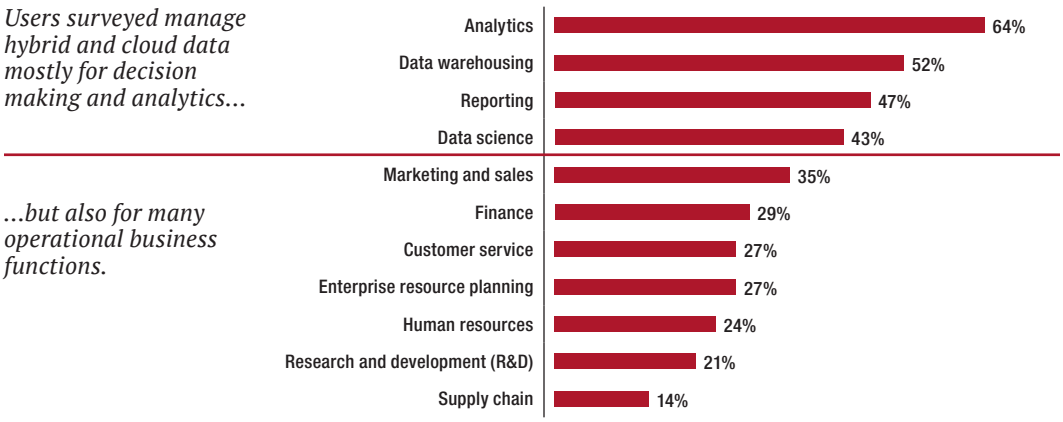



***Figure 5.*** *Based on 306 responses from 108 respondents (2.8 responses on average).*

[8] Figure 5 in this report originated as Figure 16 in the 2018 *TDWI Best Practice Report: Multiplatform Data Architectures*, online at tdwi.org/bpreports.

# Hybrid Data Architectures for the Modern DW

As we just saw, modern data warehouse architecture involves many data platform and tool types, which may be on premises, in the cloud, or in a combination of the two. Hence, data warehouse infrastructure is increasingly hybrid in the sense of on-premises and cloud-based systems that are integrated into a unified architecture. However, other combinations are also prominent in the modern data warehouse, such as when users integrate a traditional RDBMS and a modern data platform (e.g., Hadoop or NoSQL); integrate vendor-built, homegrown, and open source software; adopt a multi-cloud strategy; or deploy multiple brands of old and new DBMSs.

## A hybrid data architecture has challenges and successes.

**Hybrid data warehouse architecture presents challenges.** For example, users new to multiplatform environments struggle to wrap their heads around the extreme complexity. Data moves from platform to platform relentlessly—as users repurpose data for multiple use cases—making it difficult to govern data and track its lineage. One of the greatest challenges is to design, maintain, and optimize the performance of multiplatform processes. Finally, it is expensive to acquire and integrate multiple platforms as well as to hire and train personnel accordingly.

**Despite the challenges, TDWI sees users succeeding with hybrid data warehouse architectures.** Many are in production today by users in a wide range of industries and organizational sizes. Data warehouse professionals tell TDWI that the data warehouse has been multiplatform since circa 1990, typically including a platform for the core warehouse plus platforms for each data mart, operational data store, and data staging area. A fully modernized data warehouse does the same, but on new platforms and at a greater level of sophistication.

The modern data warehouse is increasingly organized as a hybrid data architecture that integrates many data platforms and tools, both on premises and in clouds.

Note that HDAs are not just for the data warehouse. Similar hybrid data architectures are seen in modern solutions for marketing, supply chain, financials, IoT, and multicloud SaaS apps. Across all these use cases, users tolerate the complexity and cost because HDAs offer business and technology benefits.

## A hybrid data architecture has compelling benefits.

A recent TDWI survey asked respondents to rank in priority order the benefits of HDA. (See Figure 6.)[9] Survey results show that there many compelling benefits:

1. **Get more business value from data, whether in operations or analytics.** HDAs support business value by repurposing and provisioning data for multiple business use cases from operations to analytics, as well as broad data sharing and cross-functional collaboration.
2. **Expand analytics into more advanced forms, such as machine learning and AI.** Some analytics forms work best with massive volumes of diverse data, which an HDA can provide. Likewise, an HDA enables broad self-service data exploration and discovery.
3. **Put the right data on the right platform for the right storage, processing, or provisioning.** Each form of analytics and each data-driven operational process has unique data needs. The numerous platform and tool options of the HDA ensure that all requirements are met.
4. **Capture and leverage emerging data types and sources, especially those from the Internet of Things (IoT), social media, customer channels, big data, and Web apps.** You cannot leverage data that you cannot capture, store, and process. The breadth of platform and tool choices of the HDA can handle the breadth of new data assets and use cases.
5. **Unify existing, siloed data environments without consolidating or restructuring them.** A well-designed HDA will be accompanied by substantial data integration, semantics, modeling, and virtualization. The integration infrastructure offers options ranging from physical data consolidation to virtual views of distributed data.
6. **Impose architecture on distributed data for complete and up-to-date views.** An HDA can retrofit just enough architecture, typically via virtualization and semantic techniques, to support priorities such as business value, expanding analytics, and integrating data silos.

**What is the point of hybrid data architectures? Rank the following in priority order.**

1. Get more business value from data, whether in operations or analytics
2. Expand analytics into more advanced forms, such as machine learning and AI
3. Put the right data on the right platform for the right storage, processing, or provisioning
4. Capture and leverage emerging data types and sources, especially those from the Internet of Things (IoT), social media, customer channels, big data, and Web apps
5. Unify existing, siloed data environments without consolidating or restructuring them
6. Impose architecture on distributed data for complete and up-to-date views

***Figure 6.*** *Based on 140 respondents.*

[9] Ibid.

# Final Thoughts

It is easy to get wrapped up in the technology side of the modernization of the data warehouse. After all, there are many exciting new data platforms and tools that any data warehouse professional would love to work with. That's fine, as long as you never forget that the modern data warehouse must continue to do what DWs have always done: demonstrate value by enabling operational and strategic analytics and reporting in ways that support the goals of the enterprise.

Business modernization is the true end game, and the modernization of a particular data warehouse should align with the unique goals of the evolving business that funds it.

With that in mind, note that enterprises—from for-profit corporations to nonprofit organizations, both big and small—are struggling to adapt to ever-changing markets, economies, politics, ecologies, governance requirements, and information technologies. *Business modernization* is the true end game, and the modernization of a particular data warehouse should align with the unique goals of the evolving business that funds it.

Every data warehouse modernization project is different, but most (especially in midsize to large or technically sophisticated organizations) are subject to the four critical success factors discussed in this TDWI Pulse Report. The factors include: drivers for change, new platforms, clouds, and hybrid architectures. These success factors should provide guidance for organizations planning to modernize their data warehouse, its portfolio of platforms and tools, and its increasingly hybrid architecture.

# About Our Sponsor

cloudera®

At Cloudera, we believe that data can make what is impossible today, possible tomorrow. We empower people to transform complex data into clear and actionable insights. Cloudera delivers an enterprise data cloud for any data, anywhere, from the edge to AI. Powered by the relentless innovation of the open-source community, Cloudera advances digital transformation for the world's largest enterprises. Store, analyze, and manage all your data in all its forms in a modern data warehouse wherever it works best for you. With Cloudera Data Warehouse you're in control. Run on premises, in the public cloud, or in any combination you'd like. With Cloudera the choice is yours. Visit the Cloudera Modern Data Warehouse Kit Hub to learn more.


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.


**Transforming Data With Intelligence™**

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