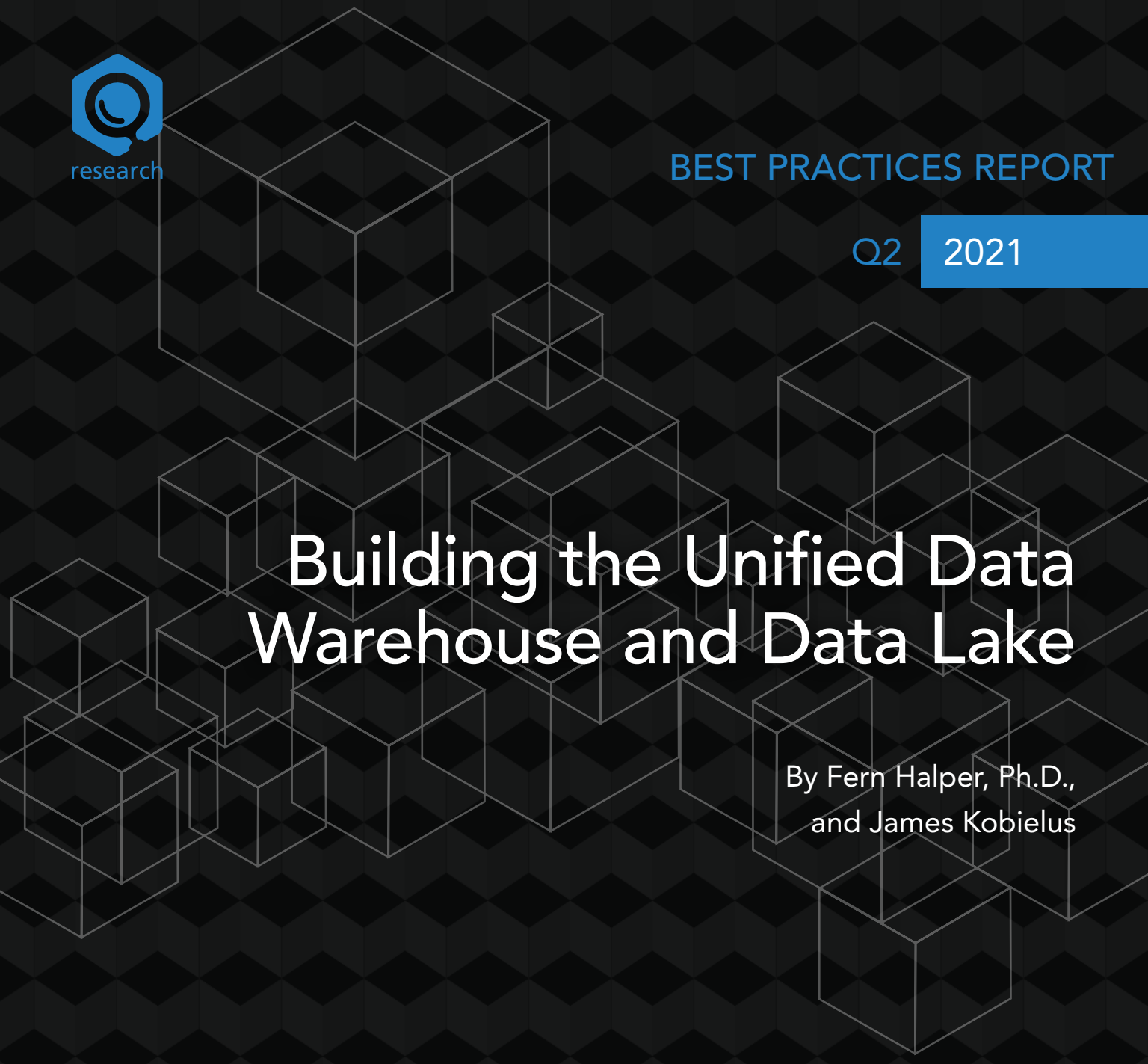


research

BEST PRACTICES REPORT

Q2 2021

# Building the Unified Data Warehouse and Data Lake

By Fern Halper, Ph.D., and James Kobielus

Co-sponsored by:

HITACHI
Inspire the Next

tdwi
Transforming Data With Intelligence™

# Building the Unified Data Warehouse and Data Lake

By Fern Halper, Ph.D., and James Kobielus

## Table of Contents



© 2021 by TDWI, a division of 1105 Media, Inc. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. Email requests or feedback to info@tdwi.org.

Product and company names mentioned herein may be trademarks and/or registered trademarks of their respective companies. Inclusion of a vendor, product, or service in TDWI research does not constitute an endorsement by TDWI or its management. Sponsorship of a publication should not be construed as an endorsement of the sponsor organization or validation of its claims.

This report is based on independent research and represents TDWI's findings; reader experience may differ. The information contained in this report was obtained from sources believed to be reliable at the time of publication. Features and specifications can and do change frequently; readers are encouraged to visit vendor websites for updated information. TDWI shall not be liable for any omissions or errors in the information in this report.

## About the Authors

**FERN HALPER, Ph.D.,** is VP and senior director of TDWI Research for advanced analytics, focusing on predictive analytics, social media analysis, text analytics, cloud computing, and other "big data" analytics approaches. She has more than 20 years of experience in data and business analysis, and has published numerous articles on data mining and information technology. Halper is co-author of "Dummies" books on cloud computing, hybrid cloud, service-oriented architecture, and service management, and *Big Data for Dummies*. She has been a partner at industry analyst firm Hurwitz & Associates and a lead analyst for Bell Labs. Her Ph.D. is from Texas A&M University. You can reach her at fhalper@tdwi.org, @fhalper on Twitter, and on LinkedIn at https://www.linkedin.com/in/fbhalper/.

**JAMES KOBIELUS** is senior director of TDWI Research for data management, focusing on data management. He is a veteran industry analyst, consultant, author, speaker, and blogger in analytics and data management. Kobielus focuses on advanced analytics, artificial intelligence, and cloud computing. Previously, he held positions at Futurum Research, SiliconANGLE Wikibon, Forrester Research, Current Analysis, and the Burton Group, and he has served as senior program director, product marketing for big data analytics, for IBM, where he was both a subject matter expert and a strategist on thought leadership and content marketing programs targeted at the data science community. You can reach him by email (jkobielus@tdwi.org), on Twitter (@jameskobielus), and on LinkedIn (https://www.linkedin.com/in/jameskobielus/).

## 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, educational conferences, topical educational seminars, role-based training, onsite courses, certification, solution provider partnerships, an awards program for best practices, live webinars, resource-filled 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 Senior Research Directors David Stodder (dstodder@tdwi.org), James Kobielus (jkobielus@tdwi.org), and Fern Halper (fhalper@tdwi.org).

## Sponsors

Denodo, Dremio, Hitachi, Matillion, SAP, Snowflake, Trifacta, and Vertica are Platinum Sponsors of the research and writing of this report. Qlik is a Gold Sponsor.

## 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 agreed 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: James Powell, Richard Seeley, Lindsay Stares, and Rod Gosser.

# Research Methodology and Demographics

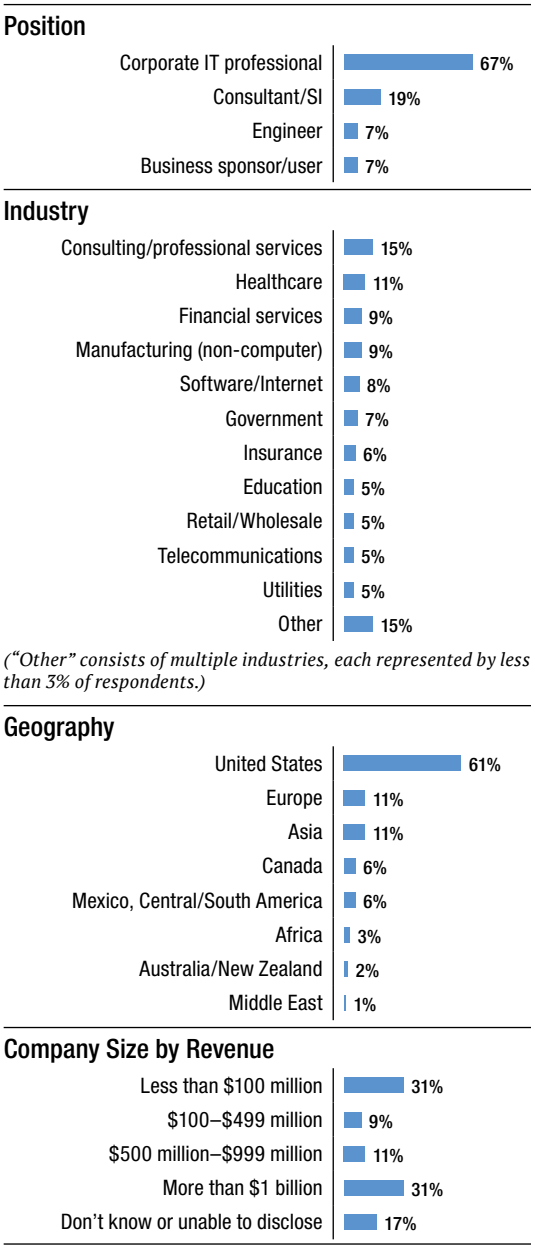
**Report purpose.** For years, TDWI research has tracked the modernization and evolution of data warehouse architectures as well as the emergence of the data lake design pattern for organizing massive volumes of analytics data. The two have recently converged to form a new and richer data architecture. Within this multiplatform environment, the warehouse and lake may each be discernable, and each may have its own internal microarchitecture. Yet, the two also integrate, interoperate, and share data standards to form a larger macroarchitecture, namely the unified data warehouse and data lake architecture. This report helps technical and business users understand new directions in data architecture, with a focus on the convergence of data warehouses (DWs) and data lakes (DLs).

**Survey methodology.** In February 2021, TDWI sent an invitation via email to the analytics and data professionals in our database, asking them to complete an online survey. The invitation was also posted online and in publications from TDWI and other firms. The survey collected responses from 220 respondents. One hundred and fifty of them completed the entire survey. This group was used for the analysis.

**Research methods.** In addition to the survey, TDWI conducted telephone interviews with technical users, business sponsors, and analytics experts. TDWI also received briefings from vendors that offer products and services related to these technologies.

**Survey demographics.** Respondents act in a variety of roles. The majority of survey respondents (67%) are directly involved in IT (including BI/DW), followed by consultants (19%) and engineers and business sponsors/users (both 7%).

The consulting (15%), healthcare (11%), and financial services and manufacturing (both 9%) industries dominate the respondent population, followed by software/internet (8%) and government (7%). Most survey respondents reside in the U.S. (61%), Europe (11%), or Asia (11%). Respondents come from enterprises of all sizes.



# Executive Summary

The unified DW/DL architecture is fairly new, and not many organizations have embraced it yet; the majority of respondents to this survey see it as an opportunity.

For years, TDWI research has tracked the modernization and evolution of data warehouse architectures as well as the emergence of the data lake design pattern for organizing massive volumes of analytics data. The two have recently converged to form a new and richer data architecture. The architecture is fairly new, and not many organizations have embraced it yet. The majority of respondents to this survey see it as an opportunity because it provides more options for managing an increasingly diverse range of data structures, end user types, and business use cases.

Within this evolved environment, data warehouses and data lakes can incorporate distinct but integrated, overlapping, and interoperable architectures that incorporate standard functional layers. These unifying layers include data storage, mixed workload management, data virtualization, content ETL, and data governance and protection. This unified DW/DL architecture continues to evolve, blurring the architectural distinctions between these formerly discrete approaches to deploying, processing, and managing analytics data.

In this study, 64% of respondents stated that the point of the unified data warehouse/data lake is to get more business value from data, whether in operations or analytics. Top value drivers include unifying silos (53%), providing a better foundation for analytics against new and traditional data types (49%), and storage and cost considerations (28%). Eighty-four percent of respondents to the survey stated that the unified DW/DL was either extremely important (48%) or moderately important (36%).

Organizations are accomplishing unification in different ways. This includes physical consolidation as well as using semantic layers and data virtualization. They are making use of tools such as modern data pipelines and data catalogs. They are utilizing disciplines such as data governance, master data management, and metadata management. Organizations attempting unification face challenges as well. Data governance ranks at the top of the list of challenges for the unified DW/DL environment.

This TDWI Best Practices Report examines the convergence of the data warehouse and data lake. It looks at how organizations are currently using their data warehouse and data lake environments and how they are bringing the two together. It examines the drivers, challenges, and opportunities for the unified DW/DL and provides best practices for moving forward.

# Introduction to the Unified Data Warehouse/ Data Lake

Data warehousing continues to evolve. As organizations collect and analyze large amounts of disparate and diverse data, they are often looking to modernize their data warehousing environments to support new use cases, such as powering the machine learning pipeline at the heart of enterprise AI.

TDWI research indicates that newer data types such as machine data, text data, image data, and other unstructured and semistructured data sources are gaining popularity for use in analytics. Different users—such as data scientists, business analysts, and business users—want to derive insights and take action on this data. Yet in many cases, the evolution of complex data has outstripped a company's ability to manage it for business value.

For years, TDWI research has tracked the modernization and evolution of data warehouse (DW) architectures, as well as the emergence of the data lake (DL) design pattern for organizing massive volumes of analytics data.[1] We have seen both the DW and the DL grow in popularity, especially in the cloud. The new generation of DWs are, in fact, DLs that are designed, first and foremost, to govern the cleansed, consolidated, and sanctioned data used to build and train machine learning models.

In recent years, enterprise data practitioners have seen DW and DL architectures converge into a powerful new type of platform. Within this evolved silo-busting environment, DWs and DLs incorporate distinct but integrated, overlapping, and interoperable architectures that include standard functional layers. This unified DW/DL architecture continues to evolve, blurring the architectural distinctions between these formerly discrete approaches to deploying, processing, and managing analytics data.

**In recent years, enterprise data practitioners have brought DW and DL architectures together into a powerful new type of platform.**

One of the hallmarks of the unified DW/DL architecture is its ability to support a wider range of data structures, end user types, and business use cases than either of its constituent micro-architectures. This may account for the reason why 89% of respondents to this survey view the unified DW/DL as an opportunity.

## The Current State of the Data Warehouse and Data Lake

DWs have their roots in business intelligence (BI). Most DWs—whether legacy or modern—were designed primarily for business reporting and related practices in performance management, dashboards, self-service, and OLAP, enabled by squeaky-clean, aggregated, and transformed data.

BI remains a core use case of the unified DW/DL. As organizations strive to derive value from their data, they are often modernizing their DW environments to support self-service, advanced analytics, and data sharing.

Nevertheless, artificial intelligence's many use cases are the principal driver behind the evolution of DWs into unified DW/DLs.

Initially built on the Apache Hadoop open-source data analytics platform, DLs have evolved over the past decade to include object stores and run on public, private, hybrid, and other cloud architectures. DLs primarily support artificial intelligence (AI), machine learning (ML), and other advanced analytics that may require a wider range of unstructured and semistructured data types, may scale to much larger volumes of stored data, and often handle more complex and dynamic analytics workloads than the traditional DW.

[1] See, for instance, the 2018 *TDWI Best Practices Report: Multiplatform Data Architectures*, available at tdwi.org/bpreports.

DLs can function as a single store of all enterprise data, including raw copies of source system data and transformed data used for tasks such as reporting, visualization, analytics, and machine learning. They incorporate a distributed file or object store, machine-learning model library, and highly parallelized clusters of processing and storage resources. Rather than enforce a common schema and semantics on the objects they store, data lakes generally do schema-on-read and use statistical models to extract meaningful correlations and other patterns from it all.

In our research into the trends and requirements catalyzing deployment of unified DW/DL platforms, we asked survey respondents about their current analytics ecosystems. Here are TDWI's principal findings.

DWs and DLs in the cloud are already mainstream.

**On-premises data warehouses still rule.** As illustrated in Figure 1, the majority of respondents (53%) have a data warehouse on premises. The on-premises data warehouse is a staple for many organizations, especially in large enterprises. We do not expect that to change any time soon. Many fewer enterprises (23%) have a data lake on premises. This may be because the first generation of data lakes (often on Hadoop) turned into data swamps because they lacked strong data governance and information life cycle management practices. In fact, in this survey, 53% (not shown) believed that data lakes need more robust data curation, data and model governance, and query optimization capabilities.

**Data warehouses and data lakes in the cloud are already mainstream.** Figure 1 also illustrates that the data warehouse and the data lake in the cloud are already mainstream, with 36% of respondents reporting that they had one or the other. Interestingly, about half of those with data warehouses in the cloud did not yet have a data lake in the cloud and vice versa (not shown). This supports the fact that today, many organizations use either the data lake or the data warehouse in the cloud, and a growing number use both. The cloud provides elasticity, scalability, and flexibility. The provider often deals with software and infrastructure management and updates so the IT team does not need to.

In your analytics data ecosystem, which of the following are in production today?

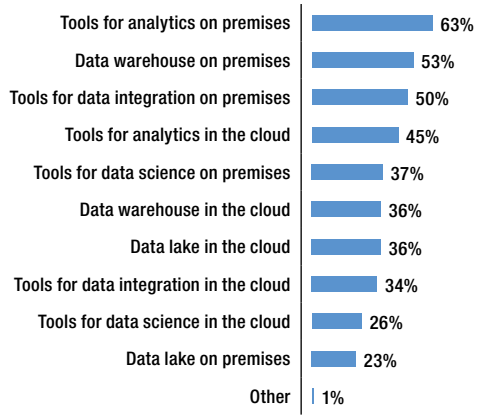



***Figure 1.*** *Based on 150 respondents. Multiple responses allowed.*

At TDWI, we've seen the enterprise data warehouse environment evolve to include cloud-based platforms, NoSQL databases, and Hadoop. Drivers of DW evolution include the need to support modern analytics, to manage new data types, to replace legacy data platforms that have reached the ends of their useful lives, and to scale out compute and storage capacity in support of growing, shifting data analytics workloads. On this latter point, decoupling of compute from storage within a modern DW/DL architecture enables these hardware resources to be scaled out independently. In older data platform architectures, including Hadoop, it has not been possible to decouple compute from storage to cost-effectively provision sufficient capacity for processing compute-intensive versus storage-intensive workloads.

## Use Cases for the Data Warehouse and Data Lake

As we explained, the core function of the data warehouse is to drive queries, reporting, dashboarding, and other decision-support analytics on data that is structured, cleansed, and curated. The data lake, on the other hand, is meant to ingest raw data that could be used by data scientists and others who want to support and develop more advanced analytics. In this survey, we asked respondents what they are using their data lake for today (see Figure 2).

Two popular use cases for the data lake are source data staging and advanced analytics. The data lake is not being used as a replacement for the data warehouse.

**Source data staging.** Tied for the top use case was source data staging (37.3%). This has been a popular use case for the data lake and continues to be so. Here, data from many sources is sent to the data lake, which serves as a staging area for the data. The data is then cleansed, conformed, transformed, and sent to the data warehouse for reporting and analytics. In other words, the data is staged in the data lake but is analyzed for reports, dashboards, and visualizations in the warehouse.

**Advanced analytics and big data.** Using the data lake for advanced analytics is another popular use case (37.3%). Advanced analytics (such as machine learning) often uses large amounts of diverse data for model training. The data lake is a prime, low-cost area for storing large amounts of raw data. For instance, a machine learning model in healthcare may use images in addition to structured patient data and doctor's notes to make a diagnosis. This kind of unstructured data is a good candidate to be stored in the data lake for analysis. The data lake may be used as the big data repository (36.7%) for unstructured data (30%). A data scientist may then create a sandbox environment that marries data from the data warehouse and the data lake as a space in which to perform the analysis using structured data from DWs and multistructured data from DLs.

**Extending the data warehouse.** Also a top use case, some organizations use the data lake as a complementary extension to the data warehouse (36.7%). For example, you could use a data lake to process multistructured data or analyze IoT data and then feed the data or the results back into the data warehouse for use in reports or visualization tools. As DW and DL architectures converge, ETL's scope is broadening to handle both structured and unstructured sources and then load the transformed multistructured data into a variety of downstream DBMSs, including but not limited to relational databases.

What is clear from the responses to this question is that the data lake is not being used much for operational reporting or data (both 19.3%), nor is it currently being used as a replacement for the data warehouse (8.7%).

## CONVERGENCE OF THE DW/DL

The unified DW/DL is coalescing on several architectural levels, incorporating functional services that are common across all data types, workloads, applications, and use cases. These unifying architectural levels include:

**Data storage.** A unified DW/DL incorporates a centralized or distributed repository of multistructured data used in BI, AI, and other analytics applications. It supports schema-on-read operations on data stored in its natural format, usually as object blobs or files, as well as schema-on-write on data stored in relational, columnar, key-value, graph, and other structured formats.

**Mixed-workload management.** A unified DW/DL includes tools for allocating, monitoring, and controlling a wide range of batch, streaming, and low-latency application workloads that execute dynamically across servers, storage devices, and other infrastructure components.

**Data virtualization.** A unified DW/DL uses data virtualization to integrate the structured and unstructured data from disparate sources in a unified semantic data layer. This unification involves harmonizing the disparate information into common data formats, vocabularies, schemas, dimensions, and hierarchies.

**Online transaction processing.** A unified DW/DL supports ACID transactions on stored data. It provides granular operations on transactional CRUD tables. It supports robust transactional controls such as optimistic concurrency serializability, snapshot isolation, data versioning, rollback, and schema enforcement.

**Data and model governance.** A unified DW/DL includes tools and services for business and IT professionals to discover, profile, cleanse, enhance, and curate data sets that are persisted within the repository. It also includes tools for policy-driven automation of the processes that govern running machine learning and other data-driven statistical models that are built and trained on the unified DW/DL.

**Content ETL.** A unified DW/DL includes tools for unifying data from different sources and then formatting it all into a common model for storage and processing in the repository.

**Statistical modeling.** A unified DW/DL includes a library of machine learning, deep learning, natural language processing, and other statistical models and algorithms that work with data that is stored, processed, and managed in the repository.

**Data protection.** A unified DW/DL includes tools for security, backup/recovery, and life cycle management of all data that is stored, managed, and processed in the repository. These include tools for enforcing policies on the creation, updating, deleting, storage, retention, classification, tagging, distribution, discovery, utilization, preservation, purging, and archiving of information from cradle to grave.

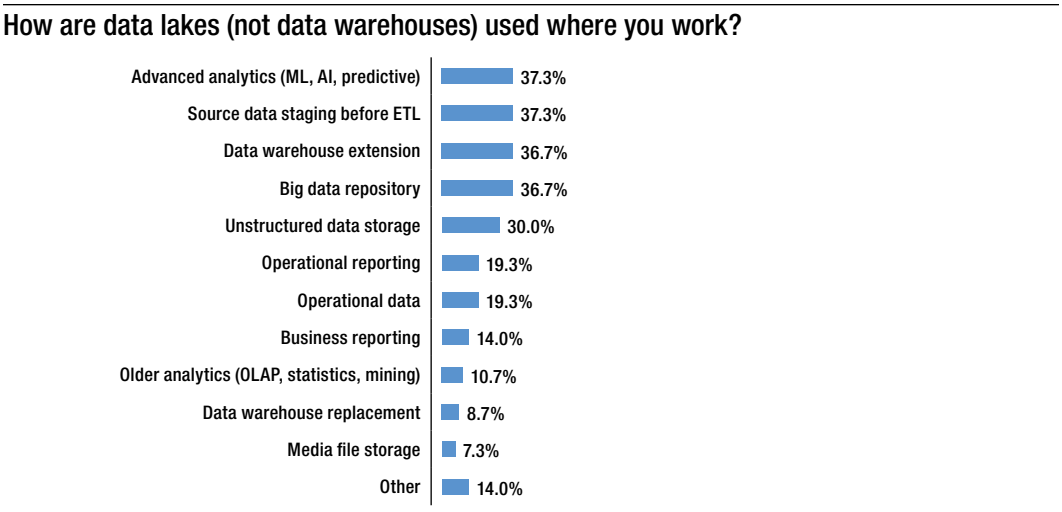



***Figure 2.*** *Based on 150 respondents. Multiple responses allowed.*

In a separate question, we asked respondents to select truth values (true, false, maybe, N/A) about data warehouses and data lakes and how they are used. As illustrated in Figure 3, the majority of respondents (51%) are in agreement that the data warehouse was not really meant to support all data types. They agree (46%) that the warehouse can be "stretched" by the data lake to support more advanced analytics. Respondents seemed to be a bit on the fence in terms of where data for self-service belongs. For instance, they were split in terms of it belonging in the data warehouse; yet the majority did not believe it belonged in the data lake. The majority of respondents did agree, however, that the data lake can support multiple use cases (58%) and that organizations need the functionality of both the data lake and the data warehouse (59%).

However, adoption of unifying DW/DL technologies is expanding and more vendors are offering solution portfolios that enable these capabilities—either in comprehensive packages or as separate services, platforms, and infrastructure that can be integrated and deployed incrementally as needed. In fact, some vendors no longer distinguish between the data warehouse and data lake (believing them to be arbitrary constructs that evolved over time) and instead enable features such as zones that perform traditional DW or DL functions. Vendors may also provide data lakes that can be queried and have DW properties (e.g., ACID compliance, data versioning, or concurrent transactions).

This is, in fact, the core trend in this convergence: DWs are evolving into DLs and vice versa, with common storage, backup/restore, ETL, workload management, semantic layer, data virtualization, query processing, transaction processing, data and model governance, continuous integration and delivery, modeling/visualization infrastructure, and tools. Some data lakes go beyond separating compute and storage to separate compute and data. Decoupling the architecture this way can make it easier to scale the system up or down based on the data volumes and workloads while preserving the open file formats from traditional data lake architecture.

**The core trend in convergence: DWs are evolving into DLs and DLs are evolving into DWs.**

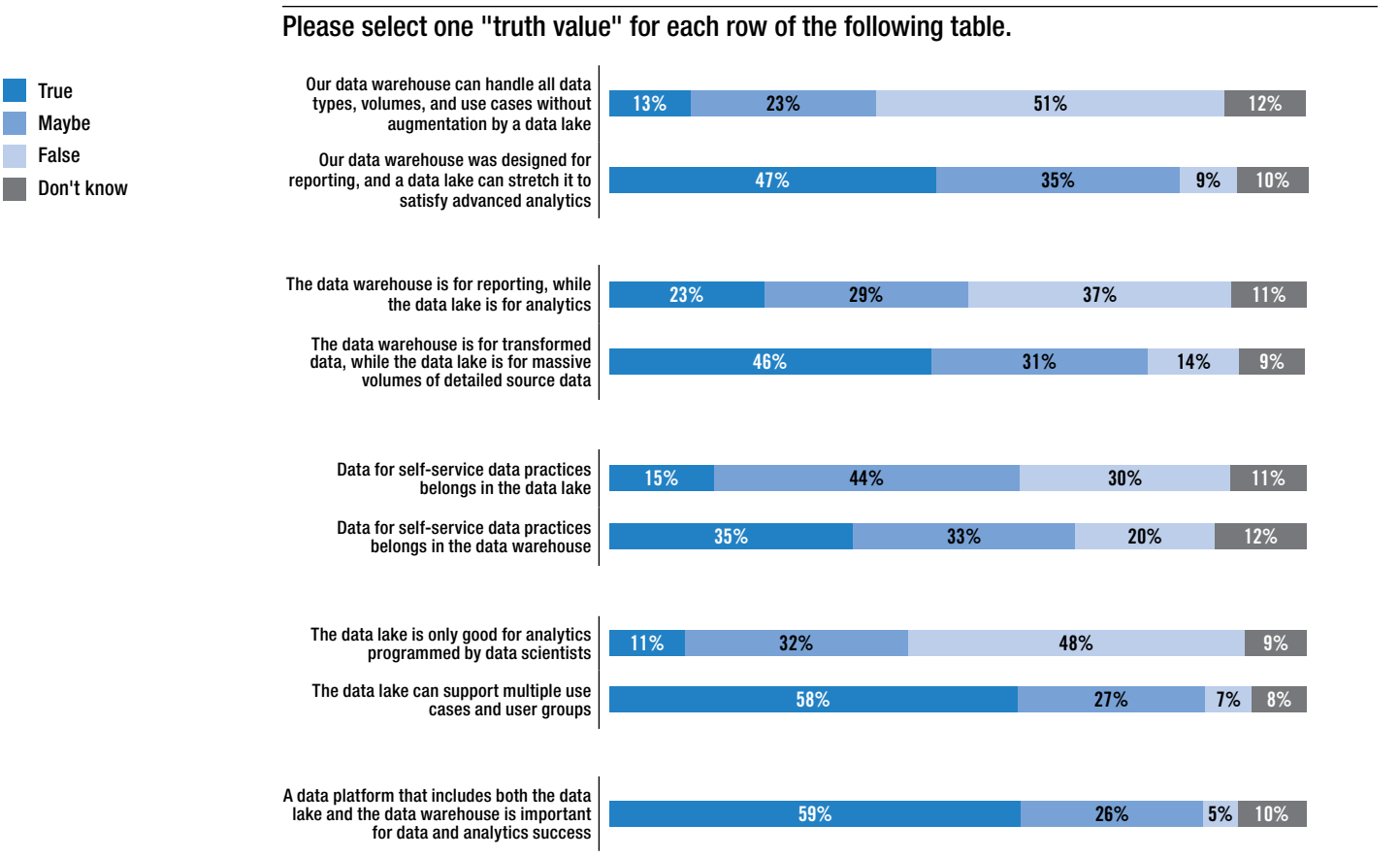



***Figure 3.*** *Based on 150 respondents.*

## Terminology Used for the Unified Data Warehouse/Data Lake

Although unifying the data lake and the data warehouse is relatively new, organizations understand the motivation behind it and many are moving in that direction. To get a sense of which terms users apply to the general practice of unified DW/DL, the survey asked, "What term(s) do you or your team use for complex environments such as the unified DW/DL?" (see Figure 4). Each of the terms adds nuance to our understanding of the term *unified DW/DL*.

The top term used to describe a unified DW/DL is enterprise data architecture.

**Architecture-related terms.** The top terms used to describe the unified DW/DL are the phrases *enterprise data architecture* (43%) and *hybrid data architecture* (36%). Terms such as *modern data warehouse architecture* (35% ) and *multiplatform data architecture* (23%) are also used. These terms make sense because the unified DW/DL is an architecture. Some organizations will develop it as part of their enterprise architecture. It can be a hybrid architecture and often one that is built as organizations modernize their data warehouse environment.

TDWI coined the term *multiplatform data architecture* (MDA) several years ago to describe an environment that contains data distributed across multiple databases, open source or big data platforms, file systems, clouds, and other data platforms. An MDA is characterized by its large number and diversity of data persistence platforms, as well as its broad range of data structures, types, and containers. Equally important, however, is the MDA's substantial data management infrastructure, which unifies the MDA's architecture by integrating, synchronizing, cleansing, mastering, and documenting data across the MDA's many platforms and beyond.

**Lakehouses, data fabrics, and bimodal IT.** The term *lakehouse* may have been coined as early as 2016 by Pedro Javier Gonzales Alonso to describe the convergence of the data warehouse and data lake approaches in his master's thesis.[2] A lakehouse is a combination of a data lake and a data warehouse that utilizes warehouse data structures and data management functions on low-cost platforms, such as those used for data lakes. In this survey, 14% of respondents use this term.

The term *data fabric* was coined at about the same time by NetApp. A data fabric unifies data management across distributed resources and provides control, choice, integration, access, and consistency.[3] Twelve percent of respondents to this survey use the term when talking about a unified environment.

*Bimodal IT* was coined by industry analyst firm Gartner in 2014 as the practice of managing two separate but coherent styles of work: one focused on predictability, the other on exploration. About 7% of respondents use this term when describing the unified DW/DL. Another term not included here is *data cloud*, which refers to a big network of data with secure access and governance across organizations.

**What term(s) do you or your team use for complex data environments, such as the unified data warehouse and data lake?**

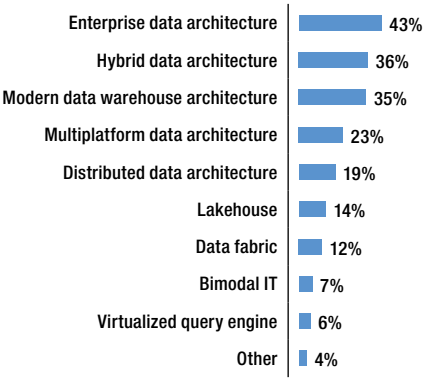



***Figure 4.*** *Based on 150 respondents. Multiple responses allowed.*

# The Importance of the Unified DW/DL

Whatever you call it, 84% of respondents to the survey stated that the unified DW/DL was either extremely important (48%) or moderately important (36%, not shown). To better understand users' views of the unified DW/DL, our survey asked respondents why it is important. Representative responses include:

**Most survey respondents (84%) stated that the unified DW/DL was either extremely important or important.**

"It would serve as a single source of truth, enabling high-impact BI reporting and analytics solutions." Corporate IT professional, professional services

"Because it provides more options for managing an increasingly diverse range of data structures, end user types, and business use cases." Corporate IT professional, healthcare

"Data is being siloed far too often, and there's no visibility into that data, its costs, and its usefulness." Independent consultant

"DW has its limitations and long tail for enhancements, and while defined process [leads] to better, refined data and quality of data, it does not lend to flexibility that DL can afford with quicker turnaround times." Corporate IT professional, healthcare

[2] Alonso, Pedro Javier Gonzales, "SETA, a suite-independent agile analytical framework," Universitat Politècnica de Catalunya, BarcelonaTech, 2016.
[3] https://cdn2.hubspot.net/hubfs/525875/Data-Fabric/Data_Fabric_Architecture_Fundamentals.pdf

"It is beneficial to provide business insights, data-driven decisions, and analytics." Business sponsor, healthcare

"Performance at scale. Easy integration with IoT, ML, and AI supporting advanced analytics. Flexibility. Improved quality of data." Corporate IT professional, transportation

"More than ever, we are working with data from multiplatforms and there is an urgency to have the data ready for ingestion in meaningful ways. A unified DW-DL architecture is necessary to meet the demands." Corporate IT professional, manufacturing

"It provides seamless access to the reporting my business requires. It also allows one to leave the data where it resides, without porting data into another store for analysis. Instead of wasting time to unify the data, you unify the analytics instead, and get the results you need much faster." Corporate IT professional, telecommunications

"Modern data is both counting/reporting and using data as an input into predictive models. The structure and rigor necessary for full DW may not be the best format for a model needing real-world data in low latency; a data lake can meet that need. An architecture allowing both would be a good thing." Corporate IT professional, software/internet

"A unified DW-DL provides the users the flexibility of doing data exploration and OLAP reporting from a single solution. One of the issues for data scientists and data analysts is access to consistent data while solving business problems. If your DW and DL are different stores, the reliability and consistency of data could be compromised." Corporate IT professional, retail/wholesale/distribution

**"We can tackle more use cases with a unified architecture…"**

"We can tackle more use cases with a unified architecture that were either difficult or not possible on DW or DL individually." Consulting/Professional services

As illustrated by these comments, there are numerous reasons for unification. These include organizations wanting to replace data silos with a single trusted source of data for reporting and analytics, supporting more advanced analytics utilizing diverse data types at scale, and leaving the data where it resides in order to meet increasing demands for data, analytics, and better data governance.

### USER STORY ELIMINATING THE DATA WAREHOUSE AND MOVING TO A CLOUD ODS

According to one senior director of DataOps at a healthcare company, about eight years ago the company wanted to perform remote diagnostic testing of equipment for maintenance purposes—an early IoT deployment. This required lots of data, and their on-premises data warehouse was not up for the task. Some of this data was structured data, some was unstructured. The company implemented an open source Hadoop data lake that they used for several years, but a few years ago they decided to merge the data lake and the data warehouse into one unified platform.

Drivers for the unification included the need to reduce operational and maintenance complexity. The company also wanted to "go full force into the cloud." As part of this, the company plans to ultimately get rid of the data warehouse and move all of its data to the Google Cloud Platform ODS. In fact, the DataOps team has written code to organize the data to look like a data warehouse so users are comfortable and see what they are used to seeing when they query a data warehouse.

An important part of the process is to make sure that data governance is in place. It started with data quality initiatives but grew from there. This includes knowing where data resides, utilizing data catalogs, and making use of data lineage. Currently, the company has a data governance team in place, which has executive-level support.

## Deriving Business Value from Data with the Unified DW/DL

Two-thirds of survey respondents (64%) stated that the point of a unified DW/DL was to get more business value from data.

Of course, all of this leads to getting more business value out of data, which was the top reason given by respondents when asked about the point of the unified architecture. As illustrated in Figure 5, 64% of respondents stated that the point of the unified data warehouse/data lake is to get more business value from data, whether in operations or analytics. Other top reasons include:

**Unifying silos.** Previous TDWI research has pointed out the growth of data silos. Silos arise when one business unit assembles its own data, which is separated from other data sources in other business units. For instance, business users often become impatient with the rollout of enterprise BI systems. They take matters into their own hands, spinning up shadow IT systems, often in the cloud. They stand up data warehouses and marts in the cloud. Business units might develop their own data lakes. In this research, for example, we saw that although IT and analytics teams typically own data lakes, business units across the enterprise also own them. For example, 17% of respondents reported that Operations owned a data lake. Ten percent reported that Finance owned one (not shown). This arbitrary expansion can lead to problems that make growth in functionality, scalability, and governance difficult, expensive, and complicated. In this survey, unifying existing silos was a top reason for the unified architecture, with 53% of respondents citing this as the purpose of unification.

**Expanding data types and analytics.** The data warehouse was designed to support analyses that use structured data such as reporting and dashboards. It is not always the best place to perform more compute-intensive and iterative kinds of analytics such as machine learning. In previous TDWI research, we have seen that organizations want to digitally transform and COVID-19 has compressed that timeline. More advanced analytics (such as predictive analytics and machine learning) are at the heart of this recent wave of digital transformation. These techniques are used against all data types, including structured and unstructured data. TDWI has seen a rise in adoption of data such as text data, machine-generated data, image data, and other data types. This data is often put into the data lake but with mixed results. If the data across the lake and warehouse is unified and data in the lake is more structured and able to be queried, the unified DW/DL can provide a better foundation for analytics against new and traditional data types (49%).

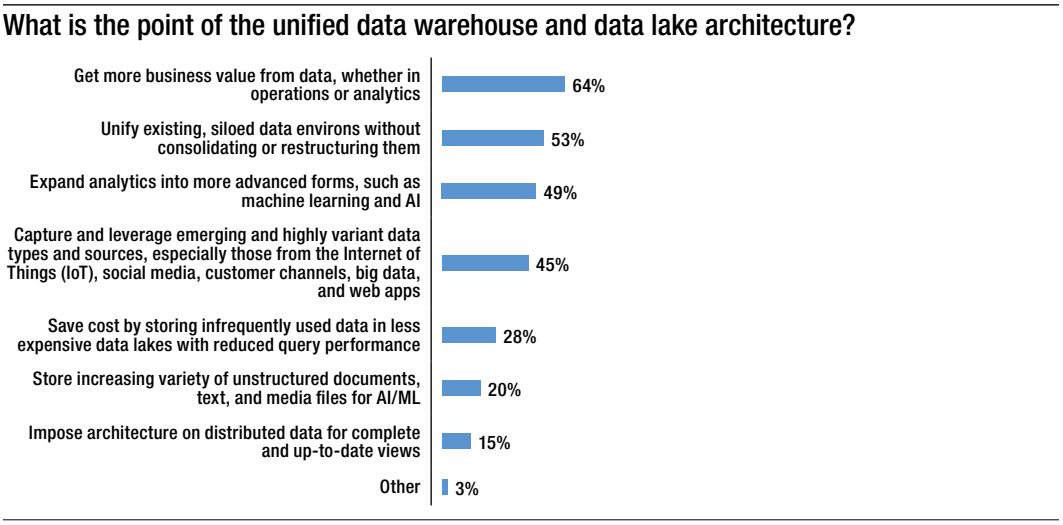



***Figure 5.*** *Based on 150 respondents. A maximum of three responses allowed.*

**Storage and cost.** Adopting data lakes can save money because data lakes are low-cost platforms that can store infrequently used data (28%). Some data lakes use an object data store, which can be inexpensive. As previously mentioned, the data lake is also used to store and (hopefully) analyze unstructured data such as documents, text, and media files (20%) for more advanced analytics such as machine learning. In other words, respondents still want a place to store and potentially analyze newer data types.

### USER STORY FROM SCRATCH TO IOT USING A CLOUD DATA PLATFORM

RD Offutt may be best known as a grower of potatoes across the Midwest and Northwest, but it is also a reseller of equipment for companies such as John Deere. When Howard Fulks, director of analytics, joined RDO three years ago, the vision was to develop a team and expand the company's data and analytics capacity. Fulks developed an IT team consisting of SQL developers, database administrators, data scientists, and an integrations specialist to develop the platform and perform analytics. They started to develop a warehouse as an operational data store using commercial tools on Microsoft Azure. The data is landed from various legacy and cloud applications and is then warehoused. The ODS does minimal transformation of the source data.

The team partitions the operational data store by schema, which partitions source data logically and enables schema-based security roles. The developers deploy SQL views as an abstraction layer to add naming conventions to the data layer prior to modeling the data in visualization tools such as Power BI.

Every one of RD Offutt's locations utilizes a slice of that data. The stores are run separately, like their own businesses, and report to regional managers by division for agriculture and construction equipment. The data is essentially the same and is partitioned like a data mart. The legacy systems let the stores run independently with economies of scale and administration to benefit the bottom line.

In terms of the data platform, Fulks says, "We started from scratch. Now we are leveraging capabilities of [Power BI] premium capacity and doing automated machine learning on data flows to deliver data science capabilities quickly. For instance, we are putting IoT sensors on potato pilers so the farms know the conditions going into storage. The sensors measure temperature and moisture. The solution automatically sends a message to the farm manager saying it is time to pull the plug. The goal is to maximize the potato output. We are testing potatoes all the time. This all would have been much harder without the data platform."

## Accomplishing Unification

Organizations are using a range of approaches to accomplish unification.

How can organizations accomplish the unification of the DW/DL? Do they physically move one into the other? Is it a logical unification? Survey responses indicate a mixed bag (Figure 6) and about a quarter either don't know or state that it isn't yet applicable.

**Physical consolidation.** Some respondents stated they are physically consolidating the data warehouse/data lake by either moving part of the data warehouse into the data lake or vice versa (17%). Others are physically moving data into another repository, sort of an "über data lake" that includes data from the data warehouse and the data lake (12%). Some organizations might use object storage—high-capacity, low-cost storage. The data in the object store might be registered with a data warehouse or ingested into a data warehouse but lives in the object store in native format. A number of data warehouses will allow the user to access the data in the object store using SQL. The data is made to look like a DW to the end user even though a DW is not managing it. That means data can be stored in a data lake in Parquet or JSON and other formats.

How is your company accomplishing the unification of the data warehouse and the data lake?

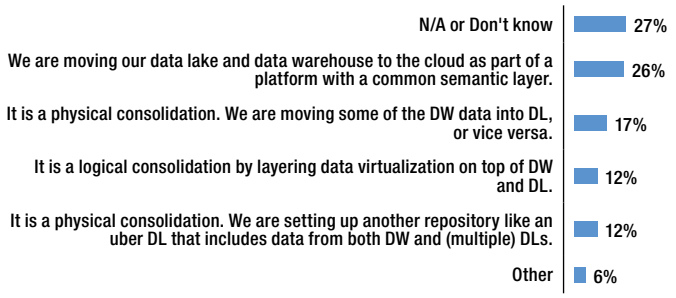



***Figure 6.*** *Based on 150 respondents.*

Respondents had different ways to accomplish DW/DL unification including physical consolidation, semantic layers, and data virtualization.

**Semantic layers.** A semantic layer is a layer that provides a consistent way of interpreting data. It helps users understand the business meaning of data (e.g., customer, product) that may be stored in the underlying data warehouse or lake. In this survey, 26% of respondents stated they were utilizing a common semantic layer across the cloud data lake and data warehouse as a way of unifying the data in the DW/DL. Here, the set up might be that the organization is using a semantic layer provided by their BI vendor. In some cases, BI vendors may provide their own semantic layer to help map complex terms and dimensions into something more easily understood. That means that they can operate on the underlying data warehouse or lake to help users find and access data. The DW/DL can coexist side by side and the data from them is "unified" in the semantic layer. However, these layers typically work only with one vendor's tools. That means that your organization would have to create multiple semantic layers for each tool or object data store.

**Data virtualization.** Data virtualization is a semantic layer that integrates heterogeneous and distributed data across multiple platforms without replicating it. It creates a single "virtual" data layer that unifies data and supports multiple applications and users. Data virtualization can create logical views in which the data looks consolidated though the data has not been moved or physically altered. This layer then connects to multiple BI and analytics tools. In this survey, 12% of respondents were using data virtualization to logically unify the data warehouse and the data lake. A good data virtualization platform supports query planning, in-memory functions, a catalog, self-service, and strategies for optimizing cross-platform performance, even across multiple cloud providers.

### EXPERT OPINION

Richard Winter is CEO and principal architect at WinterCorp, a consulting firm that specializes in the architecture, performance, and scalability of data warehouses and data lakes. According to Winter, "Companies would like to have an architected data analytics ecosystem which encompasses the warehouse and the lake—all of the data that is of interest—in their data strategy. And, it may be that in 5 years that is where we are." He recommends practices that span the whole DW/DL ecosystem for such issues as finding, accessing, curating, securing, and managing data, as well as applying user identity and permission management.

However, according to Winter, "The data warehouse is really a different proposition than the data lake. What you want in the data warehouse is the most intensively used, highest value, most integrated data. For that data, it may make sense to make a substantial investment to model, cleanse, and put it into its most usable form. The data warehouse is the one place where you can easily support many different

> uses of the same data that cross many subjects. It is the one place where you can implement and maintain enterprise business processes. It is also a place where you can deliver on challenging service-level objectives, manage mixed workloads with different performance requirements, or guarantee certain levels of quality. It really does make sense to think about which data is core and worth that special investment."
>
> That said, Winter also sees some of the boundaries between the data warehouse and the data lake blurring. "Some of the reasons for the blurring include the blossoming of object storage and the fact that some data warehouses can access data (e.g., using SQL) living in object storage. Another reason is due to the products offering federated queries and engines that can access data in a variety of databases and things like Parquet and columnar formats supported in data lakes. You can have tables in a data lake that were created in Hive or Parquet. These tables can be accessed via data virtualization."
>
> In terms of best practices, Winter suggests that "If you believe that the data warehouse has an important role then your practices are different for data in the DW than external data. A best practice would be to think about your data environment as a whole and put in place governance and processes to address all data of interest to the enterprise. This is complicated because not all of the data of interest is even housed in the enterprise. It may be housed with partners, suppliers, customers, or external providers. When you think of that data being available to users, if the data has an important role regardless of where it lives, you need a way to look at it and make decisions about how to curate it."

## Data Tools and Disciplines in Unification

For unification to succeed, no matter what the approach, data management disciplines and tools are important. Integration, interoperability, and standards are critical in a unified architecture. We asked respondents, "What tools are important for unification?" Their responses are illustrated in Figure 7.

**Data catalogs, data dictionaries, and business glossaries.** Data dictionaries (46%) and business glossaries (40%) provide a place to store the definitions of technical and business terms in data warehouses and data lakes. In modern environments, these are often giving way to data catalogs (66%)—searchable inventories that describe the data—that help users identify and understand what data exists and is available for analysis across multiple environments (such as the data warehouse and data lake). Analysts can use the catalog to search for data sets relevant to their analysis rather than having to spend time looking across multiple siloed disparate data sets. The catalog provides visibility to data, even across siloed environments. Catalogs are sold as tools or as cloud services. They span multiple environments, which helps to unify data findability and build trust in data across the unified environment.

There are numerous features in these modern catalogs. Some tools parse and deduce credible metadata. Other tools scan each new data set for sensitive data and tag that data appropriately so that tag-based security can be applied. Some tools automate the cleansing of data and some automatically discover and suggest missing lineage between data sets. Some modern catalogs embed natural language processing (NLP) functionality in the catalog that helps users ask questions of the catalog in a natural way. They provide data lineage information that describes the origin of the data and how it has changed form. Other features include the ability to certify data sets as well as rate them. That means that data stewards can mark a data set as certified. Others may be able to rate and review the data in terms of usefulness.

**ETL and data pipelines.** A majority of respondents (57%) believe ETL tools will help them enable integration, interoperability, and cross-platform processes. ETL (extract, transform,

and load) is a mainstay of the data warehouse environment. This is a well-known process that includes extraction, transformation such as standardization, and loading the data into the data warehouse. With the data warehouse and the data lake on premises or in the cloud, the approach to data integration is no longer only ETL but includes different processes applied to different data sources. ETL is part of a broader, more modern category called data pipelines (43%) that also includes ELT (extract, load, transform) and complex orchestration (38%).

The modern data pipeline is a sequence of processes for retrieving data from sources and preparing it for delivery to downstream consumers (which may be individual users or other data-processing pipelines). It provides the pathway and processes from data ingestion through movement, cleansing, transformation, loading, integration, replication, preparation, and enriching data for analysis in the unified environment, often making use of automation. Often developed by data engineers, pipelines are critical for the unified DW/DL because they provide data to the unified platform and help keep it updated. The results of predictive models can go back into the pipeline. The characteristics of these pipelines are discussed in more detail later in this report.

Pipelines are critical for the unified DW/DL because they provide data and updates to the unified platform.

**Assuming the coexistence of a data warehouse and data lake in an analytics ecosystem, which of the following tool types can help unify the two by enabling integration, interoperability, data standards, and cross-platform processes?**

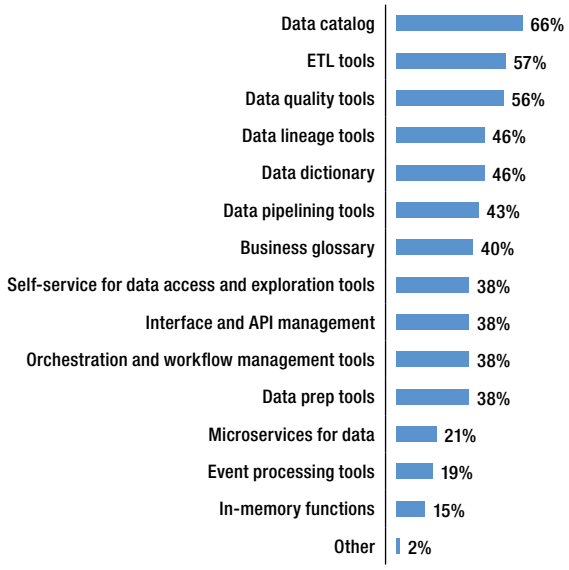



***Figure 7.*** *Based on 150 respondents. Multiple responses allowed.*

In addition to tools that help create the unified DW/DL environment, some data disciplines are also important for convergence. We asked about disciplines that can help to unify data warehouses and data lakes (see Figure 8).

**Data governance ranks at the top of the list.** Sixty-five percent of respondents cited governance as a key discipline for a unified DW/DL environment. Data governance involves the policies and processes organizations establish to ensure that the rules are followed when it comes to data as well as to build trust in their data. The core principles of data governance are important for the unified DW/DL to ensure that data is high quality and is trusted, compliant, and protected. Data integrity, of course, was a problem with the data lake, which became a data dumping ground for some organizations. The idea in the unified DW/DL is to provide a trusted and compliant source of data.

Two-thirds of respondents (65%) cited governance as a key discipline for the unified DW/DL environment.

For governance, data awareness is critical across the unified DW/DL to understand the data that might be available and whether it is sensitive, whether it complies with legal obligations, and who is using it. Accountability and ownership are critical—as are data quality and audit. If the unified environment is in the cloud, it will also be important to comply with any regional regulations.[4] Previous TDWI research indicates that organizations want to make use of a centralized data catalog, glossary, or metadata repository to address data governance challenges.[5]

Fifty-four percent of respondents cited MDM as an important discipline for the unified DW/DL.

**Master data management is also important.** Master data management (MDM) is the practice of defining and maintaining consistent definitions of business entities (e.g., customer or product) and data about them across multiple IT systems and possibly beyond the enterprise to partnering businesses. Fifty-four percent of respondents cited MDM as an important discipline for the unified DW/DL. The consensus-driven definitions of business entities and the consistent application of them across an enterprise are critical success factors for important cross-functional business activities, such as analytics. MDM is the reference data. For instance, many companies want a 360-degree view of each customer because it helps an organization retain and grow that customer.

**Assuming the coexistence of a data warehouse and data lake in an analytics ecosystem, which of the following data disciplines can help unify the two by enabling integration, interoperability, data standards, and cross-platform processes?**

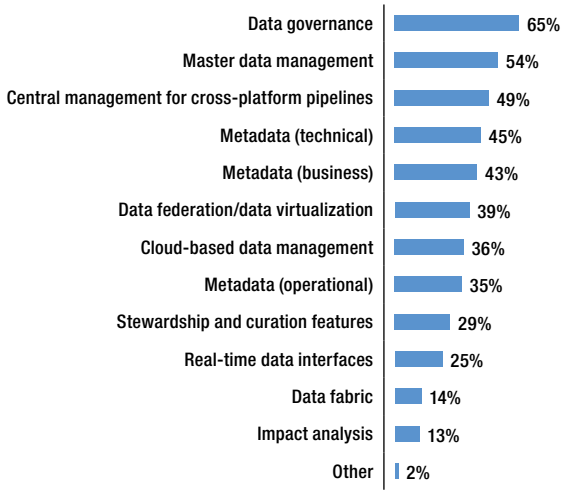



***Figure 8.*** *Based on 150 respondents. Multiple responses allowed.*

**Metadata is key.** Metadata was also cited as important for a unified DW/DL. In this survey, three types of metadata ranked highly as aids for unification. Technical metadata (45%) documents data's structures, components, and data types. This is a foundation for data extraction and load, other computerized processes, and highly technical interfaces. Business metadata (43%) describes data in user-friendly terms that people with basic tech skills can understand. It enables new practices, such as self-service data access, exploration, prep, and visualization. Operational (or usage) metadata (35%) records access to data by users and applications. These records can be analyzed to understand compliance, security, capacity, and chargeback accounting issues relative to data access and use.

**Central management for cross-platform pipelines.** As described above, data pipelines are important to both feed and update data in the unified DW/DL. The modern pipeline environment can become quite complex, with numerous pipelines that are often redundant. That is why it



[4] For more information about cloud data governance, please see the 2019 *TDWI Best Practices Report: Cloud Data Management* available at tdwi.org/bpreports.
[5] See, for instance, the 2020 *TDWI Best Practices Report: Evolving from Traditional Business Intelligence to Modern Business Analytics*, available at tdwi.org/bpreports.

makes sense to manage pipelines centrally, and 49% of respondents agreed central management for pipelines enabled unification. In this way, pipelines can be scheduled, tracked, and managed in one central place, avoiding redundancy and helping with reuse.

**Data virtualization for logical data integration.** Thirty-nine percent of respondents identified the importance of using data virtualization to enable a logical integration of the data within the DW/DL without having to physically replicate the DW/DL data into another repository. With data virtualization, the data can be delivered rapidly and in real time to business users via BI tools.

### USER STORY: MOVING TO A CLOUD DATA LAKE

Kent Maxwell, a data architect at property casualty insurance company SECURA, says his company "wanted to establish a single source of truth, but our organization had a diverse data landscape. We had systems that generated data that were specific to a business segment and systems that did similar functions. The result was that there were different versions of the same data. There was a lot of confusion and mistrust of data systems."

The company also had a problem with data access. "Not everyone had the capability to reach into a system to grab data or know what data was accessible to them. We also had data governance issues and issues with data retention for analysis. For instance, people wanted to see how quotes changed over the past ten years, but our systems couldn't support that."

To deal with these issues, the company moved to a cloud data lake using Apache Parquet. According to Maxwell, "Before the cloud data lake, we would put everything into an on-premises data warehouse. However, if we couldn't provide the IT resources, we would get rid of the data. The cloud data lake is a good approach for storage at the lowest IT overhead."

Now the company has a data warehouse and a data lake. The data warehouse is used for low-latency requirements. The data lake is used when speed isn't a significant factor. "We originally planned to integrate the two. However, we are learning that the only reason to put data into a data warehouse was if we needed low-latency data. We are also able to read Parquet files as if they were database tables using commercial tools that we have, so we determined that people could query the data lake. The group that experiences the latency is the data science group and they are okay with it." The company is now looking at open source options, too.

Maxwell recommends taking a phased approach. "We started about 10 months ago and right now about 40 percent of our data routinely accessed is in the data lake. We are still putting pipelines into place. As we encounter new requirements, we move to a process that brings data to the data lake. We didn't aggressively attack it; we are doing it in a progressive manner. That means that we could look at pipelines and see different and better ways of doing it rather than simply replicating the pipeline."

## Barriers to Unification

In the perceptions of survey respondents, there are a number of potential barriers (see Figure 9) for the unified DW/DL. A few areas stand out in their responses:

**Data governance.** Although data governance is one of the top disciplines organizations need for the unified DW/DL, it is also viewed as a top barrier. More than four in ten respondents (44%) cited this in their survey response. We've noticed in past research that many organizations feel good about the governance of the data in their warehouse. However, that isn't the case with the data lake, where data may be stored without regard to governance or compliance. That makes governance a barrier if in the unified DW/DL the data can't be trusted or isn't compliant and

**Data governance is also a barrier for unification. In fact, 44% of respondents cited data governance as a top barrier to the unified DW/DL.**

secure. In fact, in this survey, 21% cite data quality as a barrier. Data may have been landed into the data lake without thinking about quality. Sensitive data was also cited as a barrier by 23% of respondents. Here, the organization may not have put a process in place to deal with identifying and treating sensitive data. That means it can be stored in the data lake unprotected.

**In your organization, what are the most likely barriers to implementing a data lake that complements and integrates with an existing data warehouse?**

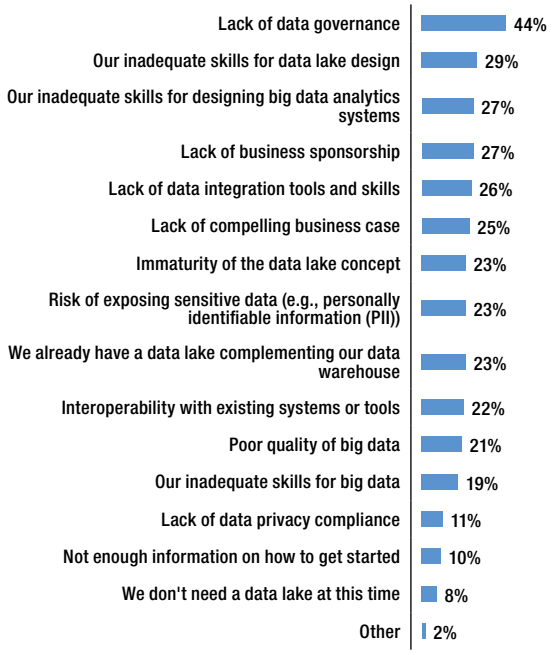



***Figure 9.*** *Based on 150 respondents. A maximum of three responses allowed.*

Organizations often view governance in the cloud environment as a challenge because it involves another platform as well as different cloud actors.

**Skills.** In addition to data governance, poor skills for data lake design (29%), designing big data analytics systems (27%), and managing big data in general (19%) were also cited as barriers for the unified DW/DL. Regardless of the solutions available today, organizations realize they will need to develop skills themselves to deal with more complex data in new cloud architectures. TDWI sees many teams successfully solving these problems with a combination of retraining existing employees, hiring more employees, and engaging consultants who have data management and advanced analytics experience. The growing popularity of visual development tools has opened up possibilities with low-code/no-code tools, which reduce design complexity without sacrificing capability. Data transformation designs managed through a graphical user interface can be understood by staff across a wider range of experience levels, thus facilitating knowledge transfer.

In this survey, 63% of respondents are training existing employees for new skills in architecture and integration.

In this survey, 63% are training existing employees for new skills in architecture and integration. Half are depending on consultants for new skills. Many fewer (39%) are hiring new employees with architectural and integration experience (all not shown). In addition to data architects, as organizations begin to utilize more complex data and build more complex pipelines for analytics, they will also need data engineers or staff in DataOps.

**Lack of business sponsorship or a compelling use case.** Some respondents stated that there simply isn't a business reason to move to a unified data warehouse or data lake. They may have one or the other and that is enough for now. There is nothing wrong with such a position.

Some organizations are still relatively early in their analytics journey; they may still be using structured data to build dashboards and reports, and that is fine. Other companies may move immediately to use a data lake and apply structure to it. They may have a cloud-first strategy and may not even use a data warehouse (on premises or otherwise), which can be the case with an internet company or even a midsized company. There is no one-size-fits-all approach. It will depend on your organization and the business problems your organization is trying to solve.

# Data Pipelines and the Unified DW/DL

If organizations are looking to expand their data and analytics footprint and use new and complex data in the unified data lake/data warehouse, they need to build pipelines for that data to either move it from one place to another or to update it (and potentially transform it). We saw the importance of the pipeline in the questions cited above.

The modern pipeline is an outgrowth of the traditional ETL approach. Modern pipelines associated with the unified DW/DL often have the following characteristics:

Modern pipelines are often automated and augmented. They support ETL/ELT, keep data fresh, and enable reuse and monitoring.

- **Comprehensive.** Whereas the ETL process covered extracting data from the source system, transforming it, and loading it into the target system (e.g., source to target), modern pipelines often cover the end-to-end process from data ingestion (source) all the way to analysis. These pipelines are often part of a bigger platform that includes a cloud repository as well as analytics functionality—all tightly integrated. They can be used by data engineers as well as other personas such as the data scientist or business analyst. Many come with a low-code/no-code or wizard-based environment.
- **Flexible.** The trend in many modern environments is to move from an ETL approach to ELT, which offers the flexibility of moving data transformations from intermediary nodes to downstream computing platforms. In ELT, data is extracted from the source system and then loaded into the target system. Typically, the tools used for this have numerous connectors that can extract data from multiple source systems, both on premises and in the cloud. Then, transformation occurs in the target system, such as the data lake. There are a number of benefits to this approach, especially in the cloud because ELT can use the processing power of the cloud platform itself. This is sometimes referred to as modern data loading. Modern pipelines have the ability to support both the data lake and the data warehouse. For instance, a pipeline can take data out of the source system and put it into a common landing zone to be provisioned by a data lake/data warehouse.

Modern pipelines are reusable, manageable, and can keep data fresh. They are often augmented and automated.

- **Reusable.** Organizations will reuse data pipelines to feed data for the data warehouse/data lake and for analytics. Sometimes one pipeline feeds another. Often the same pipeline is reused in another pipeline process or as part of a repeated process. To support this and to monitor pipelines in production, an enterprise must track its pipelines, what they are used for, who created them, and how they are performing. Some modern pipeline tools provide this functionality so users can manage and reuse pipelines (or even parts of pipelines). Some provide scheduling, monitoring, and alerting features.
- **Speedy.** Companies today want their data to be relevant and timely to support faster insights to changing conditions. Modern data pipelines are designed for low end-to-end latency, and some offer push-down instruction-set support, which leverages the native compute power of

DW/DLs. They can deliver data to a platform and then keep that data current. One way this is done is via change data capture (CDC). CDC is the process of identifying, capturing, and delivering changes made to a database, application, or mainframe system to a target database, data warehouse, or other type of data repository. Only changed records are copied to the target so as to minimize the need for bulk loading. This helps in real-time integration. Some tools let you specify how often you want to update your downstream systems.

- **Decoupled.** Some pipeline tools can decouple data from legacy systems and provide it to the pipeline environment. This approach enables organizations to resist the need to replatform their legacy systems, which are often reliable and performant. Without a decoupled data pipeline, some enterprise data may remain siloed in legacy systems and therefore be unavailable for new analytics applications.
- **Automated.** Modern pipelines typically have automation capabilities. Automation helps ensure consistency and enables organizations to scale. For instance, some pipelines will ingest metadata and write ETL code or allow users to visually transform data and address quality issues. Some validate the data as it is ingested into the pipeline. Others are connected to a data catalog so information about data is kept up to date.
- **Augmented.** Along with automation, many pipelines are also augmented—in other words, advanced analytics such as machine learning is embedded into the software to perform advanced functions. For instance, some tools may have augmented data quality tools to identify poor quality data. Others can augment transformations and recommend a ranked list of suggested transformations along with previews.

# Organizational Strategies for Unification

Unification isn't just about technology. Multiple teams contribute to the design of data structures and data integration solutions involved in successful unified DW/DL solutions. As we saw earlier in the report, skills are often considered a challenge in the convergence effort. We asked respondents who in their enterprise is responsible for designing the DW architecture and related data sets (see Table 1).

## Roles and Responsibilities

Architects own the overall design of the unified DW/DL.

**Architects own the overall design.** It is no surprise that architects are the top contributors to the design of the data warehousing environment. This includes data warehouse architects (49%), enterprise architects (43%), and IT architects (21%). This makes sense because many midsize and large organizations have an enterprise data architecture (EDA) team responsible for data architectures and data standards for enterprise environments that span multiple platforms and business units. The architects decide what route to take based on business needs, what already exists, and how well it is working. For instance, does a data-fabric approach using data virtualization make the most sense in a hybrid environment? What about federated queries? If the move is to the cloud, do some of the newer models (such as the data cloud or lakehouse) make sense for the organization? How do these fit with the existing architecture?

**Data integration specialists, ETL engineers, and data engineers integrate the data and the platforms.** The data integration specialists (33%), ETL engineers (31%), and data engineers (23%) are also critical to the overall architecture and data effort because they are the people integrating the environment and setting up the ETL and data pipelines to support analytics and other use cases. For example, when we ask data scientists what they need, they often say they

need more data engineers. Pipelines can be complex in a unified environment. As organizations conduct more advanced analytics, they often want more complex data from disparate data sources. Although modern pipeline tools often provide easy-to-use interfaces that make it possible for data scientists and others to construct pipelines, practicality can dictate who is building the pipeline and for what purpose.

**Others contribute components.** Data scientists (56%) are the top contributor of various components to the unified environment. Data scientists will often prepare the data for analysis and create features for machine learning models. In organizations where only a few models are in production, data scientists may also be responsible for building the pipeline. Other contributors include the data quality specialist (46%), the data modeler (45%), the database administrator (41%), and the metadata specialist (40%).

## Governing the Unified DW/DL

We already saw that data governance was considered a priority as well as a challenge for the unified DW/DL—especially in the cloud. New, complex data types for analytics in the cloud will require new policies and processes be put in place. New cloud actors will be involved, so governance plans will need to be updated. Organizations will want to trust the data in the cloud, just as they may trust the data stored in their on-premises data warehouse. Of course, the data will need to be protected and audit compliant, and governance plans for cloud data platforms will need to be revised.

We see organizations extending their governance processes to include the cloud environment. Data stewards are expanding their roles or new data stewards are put in place to help. They are

**Who designs or deploys the architecture for your data warehouse and related data sets? (Select one answer per row.)**

| | Owns overall design | Contributes components | Integrates data and platforms | N/A |
|---|---|---|---|---|
| Data engineers | 25% | 35% | 23% | 17% |
| Data integration specialists | 10% | 32% | 33% | 25% |
| Data management group or DataOps | 16% | 35% | 15% | 34% |
| Data modelers | 16% | 45% | 11% | 28% |
| Data quality specialists | 5% | 46% | 8% | 41% |
| Data scientists/ analysts | 11% | 52% | 6% | 31% |
| Data warehouse architects | 49% | 21% | 7% | 23% |
| Database administrators | 17% | 41% | 22% | 21% |
| Enterprise data architects | 43% | 21% | 8% | 28% |
| ETL engineers | 11% | 36% | 31% | 22% |
| IT architects | 21% | 39% | 11% | 29% |
| IT central services | 10% | 36% | 15% | 39% |
| Metadata specialists | 7% | 40% | 9% | 45% |
| Systems architects | 19% | 37% | 7% | 37% |

***Table 1.*** *Based on 150 responses.*

responsible for evaluating and monitoring data quality, integrity, accuracy, and consistency as well as identifying anomalies and discrepancies. They may be responsible for profiling the data to identify gaps and problems as well as for documenting metadata and ensuring compliance and security of data.

TDWI recommends a holistic approach to data governance in modern environments.

TDWI recommends a holistic approach to data governance in these modern environments. Holistic data governance seeks to create as few policies as possible but also make individual policies that apply broadly to many apps, data sets, and use cases. With fewer policies, data governance can scale to the complexity of hybrid data environments with fewer opportunities for confusion.

An emerging practice for the unified DW/DL will be model governance. In addition to data governance in this new environment, organizations will ultimately need to consider analytics governance, especially as analytics becomes more sophisticated in the unified DW/DL. This is a new and growing area, but it is important to put policies and procedures in place for analytics. For example, models will need to be registered to collect data about who built the model, when it was built, who has touched it, important attributes in the model, and so on. This will help to keep track of the models and information about the models. Models will need to be explainable to meet compliance requirements.

## COVID-19 and the Unified DW/DL

COVID has changed the way that many organizations work.

COVID-19 has changed the way many organizations work and has had an impact on how some organizations are approaching data and analytics. At TDWI, we see some organizations moving up their cloud migration timetable. Some have accelerated their digital transformation efforts as they see the competitive environment intensifying.

In this survey, we asked respondents how the current economic environment (caused by the COVID-19 pandemic) is affecting warehousing and analytics where they work. About 26% of the respondents stated that funding had been reduced or eliminated for DW work because of COVID-19. Another 28% responded that analytics work had ramped up to support new questions asked because of the pandemic. We have seen this in other research, as well. Data and analytics teams are being asked to answer new kinds of questions as a result of the pandemic and changes in the competitive environment. Forty-three percent said that there was no change (all not shown).

# Recommendations

This report has detailed many best practices for the unified DW/DL. In closing, we summarize the report by listing the top best practices for successful unification, along with a few comments about why each is important. Think of the best practices as recommendations that can guide your organization into successful model implementations.

**Know why you're unifying.** Not all companies need a unified DW/DL. Some organizations are still fine with their data warehouse on premises. Yet many organizations, as they mature, find that the data warehouse doesn't meet their needs. Perhaps they begin to collect unstructured data to answer business questions or they want a single source of the truth. Your modernization effort should tie to business needs.

**Plan the convergence strategy deliberately.** Aligned with the items above and as described in this report, the DW and DL environments need to converge. For some organizations, that convergence can be accomplished leveraging their current environment and virtualizing it. Others may be able to utilize their current DW and complement it with an object store they can query. Some organizations will decide to go with a converged DW/DL in the cloud because that makes the most sense for their needs. The method chosen will depend on your current environment and future needs.

**Architecture is key.** Build a high-level data architecture with the agility to support traditional BI/reporting/OLAP and emerging AI/ML requirements in a unified, flexible fashion. Determine whether you can cost-effectively repurpose/integrate existing DW and DL investments to support new AI/decision automation requirements alongside core BI, OLAP, dashboarding, reporting, and decision support requirements or if you need to replatform.

**Utilize a phased approach.** The key is to phase in the implementation of a DW/DL. If you're moving to a new cloud platform as part of the unification effort, don't try to do everything at once. Companies that quickly replatform often miss the opportunity to improve their processes and improve their data because they are trying to get everything done simultaneously.

**Plan for new skills.** Moving to a unified DW/DL environment will require new skills in emerging data disciplines and tooling. Although many organizations use third-party partners to help with the initial deployment, it will be important to have skills in-house. Train existing business, IT, and developers on the benefits, applications, infrastructure, and tools of unified cloud-based DW/DL platforms where possible. Where needed, hire externally. Modern advances in visual data transformation design utilize low-code/no-code design paradigms that can simplify the requirement to obtain complex SQL coding skills.

**Plan for modern pipelining and data engineering tools.** As part of the modern DW/DL environment, it will be important to plan for new pipeline tools. This may include tools that infuse machine learning into the pipeline to help automate some data integration and preparation steps. Some modern ETL/ELT tools leverage native push-down instruction sets to the DW/DL, thus improving performance, reducing cost, and simplifying the number of tools and skills required to load and prepare data. It may also include completely different kinds of tools to help extract useful data from text. That includes text mining or text analytics tools. Traditional vendors as well as newer entrants are offering these tools. Some vendors offer the tools as part of an analytics platform.

**Stay abreast of new technologies.** Many current products on the market use advanced technologies such as machine learning to help automate processes such as data profiling, data quality, and data mapping. Newer tools such as data catalogs can help build trust in data and

are an important component of a unified DW/DL. It is important to keep up with the changes in technology.

**Don't forget about data governance.** Data governance is going to be critical as you move into a new DW/DL environment. DL governance has historically been haphazard. To be successful in the converged environment, data governance will be key. Also, don't forget about governance of AI/ML models, which should be managed within the continuous integration and continuous deployment workflows central to the modern DevOps software development life cycle paradigm.

**Proactively nurture a better data culture.** As with any business transformation journey, it is the people who provide the impetus and vision for a better way of working in service of the wider business goal. Organizations that fail to invest in a deliberate effort to promote a better data culture in favor of a razor-sharp focus on just technology and processes risk putting their investment in peril. Look for opportunities to nurture closer cross-functional team cohesion and improved collaboration, and don't forget to recognize and celebrate your data achievements.

### EXPERT OPINION GOVERNING THE UNIFIED ENVIRONMENT

Evan Levy, a partner at Integral Data, sees many companies "forklift over the data warehouse or the data lake onto a single cloud platform, which is really colocation of systems, not really the integration of systems." In many cases, they have not done the data integration work because they don't have the organizational skill or time to do so.

However, Levy sees a growing awareness of some of the issues with this approach. For example, more companies are realizing that a "lift and shift" of data and jobs simply brings issues such as custom extracts and duplicate data from one platform to another. Levy points out that "Instead of a source system generating one master set of files with everyone pulling data out of that, you have systems generating hundreds or thousands of custom files every night. Over time, needs change and many of the files are no longer required. In many cases, these files continue to be loaded onto the CDW/DW, but many organizations don't keep track of all of them. In other words, they're loading data they no longer use into the cloud system."

Levy suggests that governance of data in data warehouses and data lakes should include tracking and cataloging the data coming out of the source systems. He says it is important to think about data going into the data lake as a controlled supply chain with the proper controls in place. This can include a data catalog to help with data governance. He sees that large enterprises have embraced the concept of the data catalog as a single repository of enterprise data information that helps to unify data and build trust in it.


Hitachi Vantara, a wholly-owned subsidiary of Hitachi, Ltd., guides our customers from what's now to what's next by solving their digital challenges. Working alongside each customer, we apply our unmatched industrial and digital capabilities to their data and applications to benefit both business and society. More than 80% of the *Fortune* 100 trust Hitachi Vantara to help them develop new revenue streams, unlock competitive advantages, lower costs, enhance customer experiences, and deliver social and environmental value.

Visit us at hitachivantara.com.


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.


**Transforming Data With Intelligence™**

A Division of 1105 Media
6300 Canoga Avenue, Suite 1150
Woodland Hills, CA 91367

E info@tdwi.org

tdwi.org