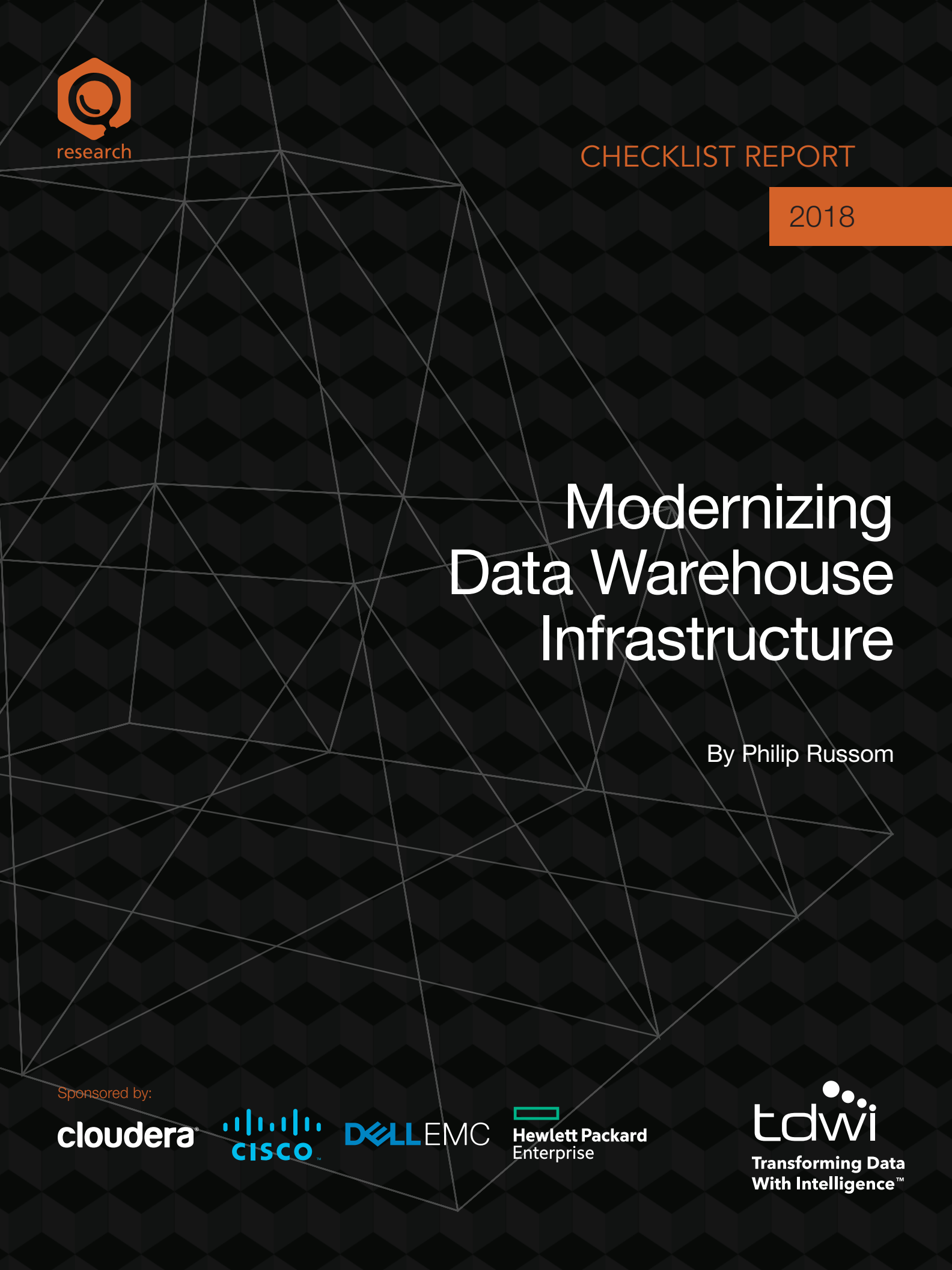

research

CHECKLIST REPORT

2018

Modernizing Data Warehouse Infrastructure

By Philip Russom

Sponsored by:

cloudera

cisco

DELL EMC

Hewlett Packard Enterprise

tdwi

Transforming Data With Intelligence™

MARCH 2018

TDWI CHECKLIST REPORT

Modernizing Data Warehouse Infrastructure

By Philip Russom

tdwi

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

TABLE OF CONTENTS

FOREWORD

The data warehouse continues to evolve and modernize so it can adapt to new data technologies (e.g., platforms, open source, and clouds) and remain relevant for new data-driven business requirements, such as advanced analytics, multichannel marketing, and other digital enterprise programs. To keep pace with these changes—and to support ever-increasing data volumes and user constituencies—user organizations across all industries need to modernize their data warehouses. The challenge has two aspects:

- Data warehouse modernization can take many forms.
- The overall infrastructure both inside and surrounding the data warehouse amounts to many components of many types.

It is good to have options, but the large number of options for data warehouse modernization can be overwhelming. To give business and technology people the information they need prior to deciding where to focus modernization, this report discusses data warehouse infrastructure, highlighting the components that are currently high priorities for data warehouse modernization. Let us begin by defining terms.

The data warehouse. TDWI defines the *data warehouse* as a data architecture that is populated with data, models, relationships among data structures, and data semantics. Note that the actual warehouse is largely data and should not be confused with the data platforms and their enterprise servers, which are key components of the data warehouse infrastructure.[1]

Data warehouse infrastructure. The virtual data architecture of a warehouse is physically managed atop one or more data platforms, including traditional relational database management systems (RDBMSs), newer DBMSs (for columns, NoSQL, and graph), file systems (of which Hadoop is becoming increasingly common), and object stores (typically cloud based). The data platforms are important, yet there is so much more to the total data warehouse infrastructure. It also includes a highly diverse ecosystem of tools, such as those for reporting, analytics, integration, quality, metadata, development, and administration.

Finally, for speed, scale, interoperability, and high availability all the above infrastructure components rely heavily on underlying enterprise hardware—namely CPUs, server memory, networks, storage, and clouds.

As you can see, a data warehouse of any maturity can involve large volumes of data in complex structures, compounded by the long list of platforms and tools in the extended warehouse infrastructure. Any and all of these can be subject to modernization.

Data warehouse modernization. This assumes many forms. Some are straightforward database upgrades, tweaks to data models, or new subject areas. Others are more dramatic, as when replacing the primary data platform or adding a new platform into the extended data warehouse infrastructure. Modernization may involve previously untapped features or platforms, such as in-memory functions, in-database analytics, clouds, and data types or analytics that are new to your organization. Besides the core warehouse, the systems and tools integrated with it need modernization too, especially those for analytics, reporting, and data integration. Many modernizations focus on hardware upgrades to give the warehouse greater speed and scale.[2]

Multiplatform data environments. Data warehouse infrastructure is inherently a multiplatform environment with technologies provided by multiple vendors and the open source community. A complex hybrid environment such as this will inevitably present challenges for system integration, interoperability among multiple systems, and data pipelines, flows, and queries that span multiple systems. Whether modernizing an existing multiplatform environment or designing a new one, users should seek out tools, platforms, and professional services that will enable cross-platform communication, integration, and analytics—at modern speed and scale—across data warehouse infrastructure and beyond.

Drivers for data warehouse modernization. Users ignore the modernization of deep warehouse infrastructure at their peril. Without it, they may achieve complete, clean, and beautifully modeled data, but without the ability to scale to big data, iterate data models on the fly, enable flexible self-service access, operate continuously and in real-time (as warehouses must in global businesses), and handle new data types and workflows for advanced analytics. Hence, modernizing data warehouse infrastructure is worth the effort and initial costs because it typically supports a modernization of the overall business, as enterprises seek to operate and compete based on broadly shared information, business monitoring, and analytics insight.

Note that data warehouse modernization is seldom for its own sake. In a few cases, updating a warehouse may be done for technology reasons, as when a database or hardware upgrade increases warehouse capacity or query response speed. In the majority of cases, however, there is typically a business reason that "trickles down" to the warehouse. For example, many business people are demanding advanced analytics and self-service data practices, which in turn demand data provisioned in new and unique ways by a modernized data warehouse.

[1] For an in-depth discussion of data warehouse architecture, see the 2014 *TDWI Best Practices Report: Evolving Data Warehouse Architectures in the Age of Big Data*, online at tdwi.org/bpreports.

[2] For an in-depth discussion of data warehouse modernization, see the 2016 *TDWI Best Practices Report: Data Warehouse Modernization in the Age of Big Data Analytics*, online at tdwi.org/bpreports.

NUMBER ONE

DIVERSIFY YOUR PORTFOLIO OF DATA PLATFORMS TO SATISFY THE DATA REQUIREMENTS OF THE MODERN WAREHOUSE

The modern data warehouse includes multiple data platform types.

Research from TDWI shows that only 15 percent of data warehouses surveyed manage data on one instance of one brand of RDBMS.[3] Instead, warehouse data is typically managed on a few data platforms (37 percent) or many (31 percent). In fact, the trend toward what TDWI calls multiplatform data architectures is one of the strongest trends in data warehousing. The same trend is also seen in other complex data environments, such as those for multichannel marketing and the digital supply chain. The trend can be described, in terms of software portfolio management, by saying that users are diversifying their portfolios of data platforms to include a growing number of types and brands of DBMSs and file systems—whether vendor-built, homegrown, or open source, both on premises and on clouds.

Portfolio diversification sounds good but presents some problems. It increases the complexity of multiplatform processes and architectures. Training and maintenance requirements go up with each additional platform. In some cases, licensing costs increase too. You need diverse platforms to satisfy diverse data requirements, but beware of creating more data silos. Luckily, the introduction of a new platform can also be an opportunity to consolidate data sets, thereby reducing complexity. TDWI sees savvy data warehouse professionals consolidating data marts, sandboxes, and other rogue data stores as new platforms are populated with data.

Multiplatform data warehouse infrastructure gives users options.

Given the challenges of complexity and cost, why are so many data warehouse teams diversifying their data platform portfolios? It is because more platform types mean more options from which technical users can pick and choose per use case or data type.

Relational data and relational processing. The relational database continues to be the most prominent data platform for data warehousing, and TDWI anticipates this will remain true for many years. This is due to the vast amounts of legacy relational data that existing warehouses have collected and developed. Equally important is the need to access and process data using relational technologies such as structured query language (SQL), SQL analytics, ad hoc queries for data exploration, online analytical processing (OLAP), materialized views, and star or snowflake schema for dimensional data. Even when a data warehouse infrastructure includes multiple data platform types, there is almost always one or more RDBMS instance present for the warehouse's relational requirements.

Just enough relational functionality. Mature brands of RDBMSs are known for their rich collections of relational functions. However, a growing number of users do not need much relational functionality or they prefer to keep it simple. Some of these users are turning to more modern data platforms, such as those based on Hadoop, clouds, or younger brands of DBMSs (for columns or graph analytics), which provide "just enough relational functionality" combined with linear scalability.

For example, data warehouse architectures often include data marts and operational data stores (ODSs); these can barely be considered relational as they consist of a few simple tables and keys. Instead of managing marts and ODSs on expensive RDBMSs, many data warehouse professionals are migrating these to cost-effective, scalable platforms where they get straightforward relational functionality from tools such as Apache Impala, Apache Drill, Apache Hive, and others—as well as shared metadata management from HCatalog and Navigator. These platforms can act as a complement or extension of the relational warehouse's core data platform and, for some use cases, a replacement.

Unstructured and multistructured data and documents. Storing these in a way that is conducive to query, indexing, and processing is a challenge for most traditional DBMSs. However, modern data platforms on Hadoop or on cloud-native storage have proved to be better homes for these nontraditional data types. Many of these data types are shared and integrated as files or documents (e.g., XML and JSON), which makes file systems a natural home for them. Some data warehouse teams are diversifying their data platform portfolios so they can finally get business value from unstructured and multistructured data—while also consolidating this data with structured data for more complete analysis.

Processing data inside the data platform for analytics, exploration, and data pipelining. Another strong trend in data warehousing is toward processing data *in situ* in the data platform where the data is stored rather than moving it elsewhere. There are good reasons for this trend. For example, the extreme size of big data prohibits the frequent movement of terabyte-scale data sets. Iterative practices such as data prep and exploration rely on the power of the data platform to develop a data set. ETL is increasingly deployed as ELT, where most or all of the data processing is pushed down into the target data platform. The maturity of modern data platforms allows this data to be processed where it is stored, supporting a wide range of exploratory analytics, BI, and reporting without the need to move or copy the data out. A diverse portfolio of data platforms should encompass a number of in-database processing techniques, whether for analytics, exploration, or data pipelining.

[3] See the discussion around Figure 10 in *TDWI Best Practices Report: Evolving Data Warehouse Architectures*, online at tdwi.org/bpreports.

New data store types. The average data warehouse was originally designed to be a clean and audited data store for reports and relational processing. These requirements are still with us but are being joined by data requirements for new practices. In particular, business and technical users need to explore large volumes of original source data to understand new customer channels, study data from new sources (such as sensors and other Internet of Things (IoT) devices), and make rich analytics correlations across data points from many diverse sources.

To enable these critical business tasks and allow for more immediate processing and analysis, new data store types have arisen that specialize in detailed source at scale. Data lakes, sandboxes, data labs, and self-service data stores have emerged to meet these needs. These design patterns may be deployed on premises, in the cloud, on RDBMSs, or on modern data platforms.

Hence, users need to diversify their arsenal of data store types as they diversify their portfolios of data platforms. To avoid creating silos, the new data store types should be tightly integrated with related warehouse platforms and analytics workflows. New data store types can also be a point of consolidation to reduce existing silos, especially data marts and ODSs.

NUMBER TWO

MODERNIZE YOUR DATA WAREHOUSE WITH CLOUD AND HYBRID PLATFORM STRATEGIES

The cloud is rising as a preferred data platform for data warehouses.

The cloud offers many technology and business advantages for a warehouse's data and analytics workloads:

- **Elasticity.** A cloud automatically allocates resources as analytics workloads ramp up, then reallocates resources as processing subsides. Cloud elasticity gives a data warehouse speed and scale with minimal capacity planning and administrative work, especially as compared to adding more nodes on premises or restructuring an MPP RDBMS configuration.
- **Low total cost of ownership (TCO).** Compared to on-premises data platforms, start-up costs and ongoing maintenance for cloud environments can be lower. Many analytics programs and data warehouses are owned and run by departments, for whom the low TCO of cloud implementations is not a barrier to entry. However, the cloud is not guaranteed to be lower cost. Ease of start-up can make it easier for costs to grow unexpectedly and more persistent workloads may be more cost-effective in on-premises environments.
- **Minimal system integration.** Intense system integration is often required when deploying an on-premises data platform. By comparison, acquiring a cloud license, uploading data into a cloud warehouse, and diving into analytics is fast and inexpensive. This also reduces time to business use, dependence on system integrators, and capital expenditures for hardware.
- **Business agility.** With the virtual resources of the cloud, onboarding new data sources and setting up analytics sandboxes or data labs in a data warehouse is fast, easy, and inexpensive—yet still effective.
- **Stability, reliability, and high availability.** The cloud provider handles all upgrades, patches, and environment tweaking while assuring high availability so your analysts, data scientists, and data warehouse professionals can focus on data warehouse development and business-driven analytics.

Migrate your data warehouse wholly or partially to the cloud.

At one end of the spectrum, TDWI has seen some organizations migrate all or most of their data warehouse infrastructure to the cloud. If the data warehouse team is allowed to select the cloud provider, they typically look for simple provisioning of the data warehouse and consistent SQL or access capabilities for their on-premises workloads. However, selecting the cloud is sometimes done by central IT as part of a "cloud first" enterprise strategy. In these cases, the data warehouse team should lobby for cloud-native object storage integration and vendors that meet security and governance requirements.

At the other end of the spectrum, organizations usually migrate one or more components of a data warehouse architecture to the cloud, leaving the rest of the warehouse on premises. For example, moving a data landing and staging area to the cloud provides the elasticity these disciplines demand and brings that area closer to external data from the Web, third parties (partners and clients), and IoT. As another example, capacity on a relational warehouse is very expensive, even though very large data sets of raw source data are often stored there. Migrating raw data for advanced analytics to a modern data platform in the cloud gives the data a platform that is cheaper and more conducive to its processing. In these cases, the cloud offloads a data warehouse platform, freeing up capacity on the RDBMS under the warehouse, so that capacity can be applied to the growth of data sets and use cases that demand advanced relational management and processing.

Whether they migrate a data warehouse to the cloud *in toto* or piecemeal, users can choose a number of house offerings from cloud providers or independent cloud-based services, which offer short time to use, zero system integration, pay-per-use pricing, and elasticity for speed and scale. Some modern data platform providers offer on-premises and cloud-based versions of their platform to

enable a consistent analytics experience across environments, which is particularly effective for hybrid and multicloud architectures.

Develop strategies for hybrid data warehouse infrastructure.

Data warehouse infrastructure involves many data platform and tool types. These may be on premises, in the cloud, or in hybrid combinations of the two. Data warehouse infrastructure is increasingly hybrid, in the sense of on-premises and cloud-based systems integrated into a unified architecture. On-premises and cloud integration is the most common definition of hybrid. However, other hybrid combinations are also prominent in modern data warehouse infrastructure, such as when users integrate a traditional RDBMS and a modern data platform; vendor-built, homegrown, and open source software; a multicloud strategy; or use multiple brands of old and new RDBMSs.

Prepare for the challenges a hybrid data warehouse infrastructure presents. For example, users find it difficult to wrap their heads around the extreme complexity of the extended infrastructure. Data moves from platform to platform relentlessly as users repurpose it 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. Users succeed in these situations by selecting vendor platforms that are conducive to data consolidation and aggregation at scale, complemented by tools for analytics and integration that provide unified and friendly views of distributed data.

Despite the challenges, TDWI sees hybrid data warehouse infrastructures in production today by users in a wide range of industries and organizational sizes. Beyond the data warehouse, similar complex data environments are seen in multichannel marketing, the digital supply chain, multimodule ERP, and financials.

Design and unify hybrid environments using "big picture" tools and techniques. These use cases succeed by relying on practices and technologies that can provide visibility across multiple platforms. For example, TDWI finds data architects and data governors designing and policing the big picture. Unifying technologies include integration infrastructure and centralized, shared metadata. Logical and virtual technologies and best practices are increasingly applied by users to draw the big picture via simplified views of heavily distributed data. Some vendors even provide platforms that support hybrid deployments natively.

NUMBER THREE

MODERNIZE DATA WAREHOUSE HARDWARE FOR GREATER SPEED AND SCALE, LOWER COSTS, AND INNOVATIVE BEST PRACTICES

For years, we data professionals dreamed of innovative data management practices, such as processing large data volumes for real-time analytics, virtualizing data sets, synchronizing data sets globally several times a day, and exploring data via complex queries that respond in subsecond time frames. All these and more are now a practical reality thanks to the speed and scale of modern hardware and software. In other words, innovations in hardware performance and price have enabled new best practices in data warehousing, analytics, reporting, and data management. Other modern practices enabled wholly or in part by fast and scalable hardware include continuous ingestion, self-service data prep, complex event processing, intra-day operational reporting, in-memory databases, and high-speed data pipelining.

With so many innovative and desirable practices within reach, it behooves data warehouse professionals and their IT colleagues to modernize hardware across the extended data warehouse infrastructure and beyond. Speed and scale are high priorities for such modernizations, but other notable goals include favorable economics, lower electricity consumption, greater storage capacity, and in-place processing capabilities.

Let's drill into specific hardware components commonly found in data warehouse infrastructure and how they contribute to modernization. According to a TDWI survey, the hardware components that data warehouse professionals value most are (in priority order) server memory, CPUs, storage, and networks.

Server memory. When 64-bit computing arrived over 10 years ago, data warehouses migrated immediately away from 32-bit servers to capitalize on the massive addressable memory spaces of 64-bit systems. Since then, the price of server memory has steadily dropped (though not as dramatically as the price of CPUs and storage), making server memory upgrades economically feasible. Multiterabyte memory is now common in servers for RDBMSs and modern analytics platforms (to be used for in-memory databases and frequently accessed tables) and data integration tools (used for in-memory data transformations and I/O-free data pipelining). "Big memory" alleviates the need to write data to disk, which speeds up complex SQL, multiway joins, and analytics model rescores.

Central processing units (CPUs). Server CPUs are obvious contributors to high performance. Moore's Law repeatedly takes us to a higher level of performance, as seen in the recent wave of reasonably priced multicore CPUs and commodity-priced server blades and racks.

Storage. For almost 20 years, the price of storage has been dropping steadily, even as storage capacity, bandwidth, and reliability improved. More recently, seek speeds improved. Today we are blessed with ample low-cost storage that enables desirable practices. For example, from an economic viewpoint, petabyte-scale data warehouses and the entire big data phenomenon wouldn't be possible without low-cost storage. Likewise for active archives that keep vast amounts of data on spinning disk, ready for access. For the ultimate in speed, solid state drives (SSDs) are now common in enterprise use for hot, frequently accessed data; SSDs are a common upgrade applied during a data warehouse modernization. Finally, these economic and technical innovations in storage systems have driven maturity in the new generation of data platforms that are very competitive due to their low-cost scalability.

Networks. Corporate networks are fundamental to how business people connect, communicate, and collaborate. We can say the same about the Internet. Too often, we forget that very few of our computer-driven practices would be possible without networks. Luckily, networks and the Internet have seen improvements in bandwidth, speed, and reliability. Without those advances, demanding multiplatform data architectures—such as the modern data warehouse—would be severely hamstrung. The recent success of cloud-based applications and data management solutions would likewise be limited. To assure support for innovative practices and infrastructures, we must all do what we can to encourage modernizations of network infrastructure.

NUMBER FOUR

COORDINATE DATA WAREHOUSE MODERNIZATION WITH BUSINESS MODERNIZATION AND ANALYTICS MODERNIZATION

We talk about and even perform data warehouse modernization as if it is an independent project with isolated goals. The reality is just the opposite. Data warehouse modernization is, in fact, usually one of many concurrent efforts for modernization that have project interdependencies. Here are some examples of dependent modernizations that need to be coordinated with data warehouse modernization.

Business modernization. In an ideal world, upper management leads the way by deciding how it must modernize the business to keep pace and stay relevant with evolving customers, partners, marketplaces, economies, and so on. The business modernization and its goals are in turn articulated "down the org chart." At some point in that process, people in IT and similar groups (such as a data warehouse group) should collaborate with business managers to determine how data, applications, and technology can support the stated business modernization goals. Even if you do not work in an ideal world, some semblance of that process should still be present to guide your alignment of warehouse modernization with business modernization.

For example, a common change of direction in business management is to modernize the business to run and compete on analytics. The analytics team determines what data the business needs for the new analytics, and a data management team then modernizes the warehouse accordingly. All data warehouse modernizations should begin with a business decision so that improvements to the warehouse and its infrastructure align with and support business goals. As a bonus, this level of alignment proves that the technical expense of modernization yields a return on the investment (ROI) for the business.

Analytics modernization. One way to describe analytics modernization is that it tends to introduce analytics methods that an organization has not deployed before. These are typically forms of *advanced analytics*, which are based on technologies for mining, clustering, graph, statistics, and natural language processing. In addition, we are currently experiencing an upsurge in the use of machine learning and artificial intelligence in support of predictive analytics; analytics modernization efforts typically incorporate these within end-to-end analytics workflows. Furthermore, the ad hoc query—with us since the dawn of databases—has evolved into real-time, iterative SQL analytics, which in turn enables innovative end-user practices such as self-service data access, exploration, visualization, and data prep.

Today's discovery-oriented advanced analytics and iterative self-service analysis demand massive volumes of detailed source data. This is a challenge for the majority of data warehouses, which were originally designed and optimized to provide aggregated, calculated, cleansed, and auditable data for standard reports, metrics-driven performance management, and OLAP. Such data warehouses must be true to their founding mandate while also extending and modernizing to fulfill the new data requirements of advanced analytics and self-service practices. Given the broad diversity of data types and processing required for the full range of both reporting and analytics, it has become impossible to satisfy all use cases with one data warehouse platform. For this reason, a common outcome of analytics-driven data warehouse modernization is a multiplatform data warehouse infrastructure that includes multiple types of data platforms.

Report modernization. The style of reports has evolved dramatically since the 1990s. Back then, reports were only on paper and consisted of giant tables of numbers. Because a single report

served dozens or hundreds of report consumers, the content of each report was mostly irrelevant to individual report consumers.

Luckily, waves of modernization have greatly improved reports by:

- Bringing reports online for greater distribution and ease of use, as well as drill-through and interaction
- Organizing reports around metrics and key performance indicators (KPIs) in support of performance management and Balanced Scorecard™ business methods
- Personalizing reports so users go straight to what they need for productivity and relevance
- Giving reports a visual dashboard presentation for interpretation at a glance and a more pleasant user experience

As the style of reporting has evolved, warehouse data structures have had little trouble modernizing to keep pace. However, more dramatic change is seen in users' portfolios of tools for reporting. These portfolios still include older enterprise reporting platforms, but they are now augmented with newer tools for dashboarding, data visualization, and data exploration. Additionally, new data platforms contribute to many reporting use cases, such as operational reports based on massive data volumes and reports that do not require complex data modeling or governance curation.

NUMBER FIVE

ADJUST DATA MANAGEMENT BEST PRACTICES TO FIT MODERN DATA WAREHOUSING AND ANALYTICS

Data itself is evolving, driven by the arrival of new data types, new data sources, big data, advanced analytics, and new data platforms. To address data's evolving state and its new requirements, data management best practices and related tool types need to modernize, including those for data integration, ingestion, quality, modeling, profiling, master data management, metadata management, and other semantics.

Data management best practices are critical success factors because they stitch together today's multiplatform data environments—within which clouds and on-premises systems can coexist, even when multiple premises and multiple clouds are involved. Despite recent evolutions in data structures, sources, and platforms, data management is more relevant than ever because it provides what users really want—regardless of platform—trusted data that is high quality, auditable, secure, governed, open for self-service, and fit for many purposes from operations to analytics.

The good news is that existing best practices work well in new big data and cloud environments. Even so, for users to succeed in new scenarios, they need to make a number of adjustments and upgrades to existing skills and tool portfolios. Here follow a few examples of those adjustments.

Fast and furious data ingestion.

One of the most prominent adjustments is to design new data management solutions (or adjust older ones) to capture and ingest new data faster and more frequently than in the past. This is sometimes called *early ingestion* or *continuous ingestion*, which is quite fast and frequent compared to overnight batch loads. As a trade-off, early ingestion does little or no transformation or aggregation of data prior to load because that would slow down ingestion. A benefit is that the data is captured in its original state, which means it can be repurposed repeatedly as new requirements for reporting and analytics arise. Furthermore, data is ready as soon as possible for reporting, analytics, and operational uses.

Note that early ingestion and traditional ETL for data warehousing are quite different, though regularly achieved with the same toolset. Before data is allowed to enter the average data warehouse, it is carefully aggregated, cleansed, and documented; reporting and performance management practices demand accurate and auditable data in support of recurring measurements. With early ingestion, data usually enters a raw data store (possibly a data lake on a modern Hadoop platform or cloud object store). That data store is designed for discovery analytics, which demands detailed source data but not the accuracy or auditability of standard reports. The data store may also serve as a data landing and staging area for the warehouse and other targets, though increasingly it also supports production-level BI, reporting, analytics applications, and many other use cases within the shared data store.

Increasingly, the modern data warehouse includes early ingestion that feeds the shared data lake or similar raw data store, traditional ETL and staging, ad hoc analytics, data science, data exploration, self-service BI, and curated or aggregate data, which then feeds traditional data warehouses for business reporting. For a modern data platform (acting as the ingestion layer) to integrate into this complex but increasingly common data warehouse architecture, it must support advanced metadata management and data governance activities.

Just enough data improvement, applied later and on the fly.

New data management best practices for early ingestion and raw data stores seem to preclude traditional data improvement practices, such as data quality, standardization, modeling, and metadata development. The traditional practices are still highly relevant, but instead of being applied prior to data loading (as with most data warehouses), they are applied later.

Up-to-date tools and fast modern hardware enable us to transform data considerably *on the fly at read time* as we explore, analyze, or otherwise process data. Even so, detailed source data captured

via early ingestion can also be improved later by established best practices. In fact, zero-latency and high-latency approaches can coexist. For example, the online processing of stream data from manufacturing robots can reveal bad lots or other problems that need immediate attention. The same data studied offline can reveal equally valuable long-term trends in supplier performance, supply quality, and yield on the production floor.

This is a major paradigm shift for data management best practices. Yet, TDWI sees data warehouse professionals successfully adjusting to early ingestion and on-the-fly improvements while continuing with older practices. After all, some data sets and use cases are conducive to improvements at read time (data exploration, real-time analytics) whereas others are not (batch processing for data deduplication).

Modern metadata management.

New data types (including unstructured and IoT data) are infamous for lacking metadata, even though we need metadata about all data—including the newest forms—for the sake of finding, querying, and using it for business operations and analytics. Luckily, modern metadata management systems support automation that can parse data and generate metadata from it based on machine learning, a rules engine, or both. Other automation tools can map sources to targets and set up new data sources without human intervention, which is critical to surviving the onslaught of new IoT sources.

Business users are increasingly demanding self-service access to data, which is not possible without business metadata from a modern metadata repository. Similarly, other new functionalities built atop the metadata repository are emerging, such those for master data management, business glossaries, data cataloging, and publish-and-subscribe mechanisms for sharing data sets.

Furthermore, the metadata repository still serves as a central hub for cross-functional collaboration and the sharing of common business rules, transformation logic, and data profiles. Hence, the modern metadata repository continues to be central to traditional data management best practices as well as to new ones. It's also important to understand how modern data platforms manage metadata and how they share it across these systems for consistent and trusted data access.

NUMBER SIX

LEVERAGE MULTIVENDOR PARTNERSHIPS FOR A UNIFIED, COMPREHENSIVE, HIGH-PERFORMANCE, AND TROUBLE-FREE DATA WAREHOUSE INFRASTRUCTURE

As discussed earlier, modern data warehouses—and other modern data environments—are inherently multiplatform and hybrid, which means that their infrastructures consist of many data platforms and tools, on premises and in clouds, acquired from multiple software vendors, hardware vendors, cloud providers, and the open source community.

Among the benefits, multivendor solutions can be powerful, especially when the assemblage of diverse products satisfies the long list of data requirements for modern analytics, reporting, warehousing, and data management. However, some multivendor solutions (especially when assembled by end users) suffer problems with cross-platform compatibility, interoperability, and optimization. No worries, though. These problems can be avoided by evaluating multivendor solutions produced by partnerships among multiple vendors. Such solutions also have a number of benefits to consider.

What to look for in multivendor partnerships that support modern data warehousing:

Reference architectures for multitool solutions. These validate that certain combinations of software and hardware products work well together, with high performance and reliability. User organizations can follow these road maps to implement their own complex data warehouse infrastructures. Look for reference architectures that include products from multiple vendors, as well as clouds, Hadoop and other open source Apache tools, and platforms for analytics and big data. Ideally, a reference architecture should address building a modern data warehouse, not just any data environment.

Unified infrastructure. A number of vendors are working toward this. Synonyms include *converged infrastructure, integrated infrastructure, common platform architecture,* and *unified data architecture*. Unified infrastructure is an important enabler of the cross-platform activities common in modern data architectures, namely data pipelines and flows, federated or cross-environment queries, materialized data views, and optimized performance of these. These systems also should provide unified security, governance, management, and data/metadata replication across all storage and access points. Unified infrastructure also simplifies the design and maintenance of today's complex architectures and assures trouble-free scaling over time. Ideally, unified infrastructure is not just for one vendor's platforms but for all or most components in a multiplatform data warehouse infrastructure.

Technical partnerships, not just sales and marketing partnerships. Look for partnerships with deep technical coordination, where technical personnel from multiple vendors have worked together in development, testing, and certification. Without this coordination, the value of reference architectures and unified infrastructure is limited.

Multivendor partnerships and solutions offer a number of benefits:

Kick-start your data warehouse modernization with professional services. A fully modern data warehouse includes technologies and best practices that you have not yet tapped. When embracing anything new to you and your organization, you should seek assistance from consultants and system integrators who have expertise in that practice area. Look for professional services that can handle complex and hybrid systems for data warehousing and analytics. As you evaluate multivendor solutions, take note of which have consulting specifically for that solution.

Centralized functions that reach multiple vendor products are an extra benefit. One of the many challenges of the complex and hybrid data warehouse infrastructure is that each platform may have its own siloed tools for administration and communication. Some multivendor solutions solve this problem by offering centralized tools for key cross-function tasks such as single-pane monitoring for cross-platform processes (or orchestration), multitenant management, authentication for single sign-on security, and metadata management that all platforms can share.

More partners mean more options. Don't forget: modern data warehouse infrastructure involves a long list of potential platforms and tools. It takes multiple products from many vendors to satisfy the diverse requirements generated by this list. An ecosystem of collaborating partners gives you options to get started with today and options you can grow into in the future as you complete your comprehensive data warehouse infrastructure.

ABOUT OUR SPONSORS

cloudera®

cloudera.com

Cloudera empowers people to transform complex data into clear and actionable insights. They deliver the modern platform for machine learning and analytics optimized for the cloud. The world's largest enterprises trust Cloudera to help solve their most challenging business problems.

cisco.com

Cisco is the worldwide technology leader that has been making the Internet work since 1984. Our people, products, and partners help society securely connect and seize tomorrow's digital opportunity today. Cloudera and Cisco enable you to gain value from data throughout your business. Cloudera Enterprise runs seamlessly on—and works intelligently with—Cisco UCS Integrated Infrastructure for Big Data. Together, we make it possible to store, protect, and access your data in an industry-compliant environment—wherever and whenever.

DELL EMC

dellemc.com

Dell EMC enables organizations to modernize, automate and transform their data center using industry-leading converged infrastructure, servers, storage and data protection technologies. Dell EMC services its customers with the industry's broadest, most innovative infrastructure portfolio from edge to core to cloud.

Hewlett Packard Enterprise

hpe.com

Hewlett Packard Enterprise is an industry leading technology company that enables customers to go further, faster. With the industry's most comprehensive portfolio, spanning the cloud to the data center to workplace applications, our technology and services help customers around the world make IT more efficient, more productive and more secure.

ABOUT THE AUTHOR

Philip Russom, Ph.D., is senior director of TDWI Research for data management and is a well-known figure in data warehousing, integration, and quality, having published over 550 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 was 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 BI professionals worldwide. TDWI Research focuses exclusively on analytics and data management issues and teams up with industry practitioners to deliver both broad and deep understanding of the business and technical issues surrounding the deployment of business intelligence and data management solutions. TDWI Research offers reports, commentary, and inquiry services via a worldwide membership program and provides custom research, benchmarking, and strategic planning services to user and vendor organizations.

ABOUT TDWI CHECKLIST REPORTS

TDWI Checklist Reports provide an overview of success factors for a specific project in business intelligence, data warehousing, or a related data management discipline. Companies may use this overview to get organized before beginning a project or to identify goals and areas of improvement for current projects.