Leveraging the Cloud for Data Warehouse Modernization

By David Loshin, President of Knowledge Integrity, Inc.

MOTIVATIONS FOR DATA WAREHOUSE MODERNIZATION

For nearly three decades, the organizational data warehouse has powered corporate reporting, analysis, and decision making. The data warehouse was designed as a segregated environment into which data sets extracted from transactional and operational systems could be organized and loaded for analytical purposes. Though there have been many optimizations and improvements made over the years, the fundamental architecture of the data warehouse has not significantly changed.

Yet there are signs that the conventional on-premises data warehouse is starting to show its age. Although that platform continues to support ongoing operational reporting and analysis, there are a number of factors that are inspiring initiatives to modernize the ways that data warehouses are designed and put into production.

- **Increased demand for analysts and data scientists.** According to recent analysis of the job market, there is both a rapid increase in the demand for data scientists and a growing gap in the number of skilled individuals to meet that demand. According to the January 2019 report from Indeed—one of the top job posting sites—there was a "29% increase in demand for data scientists year over year and a 344% increase since 2013." In contrast, "searches by job seekers skilled in data science grew at a slower pace (14%), suggesting a gap between supply and demand." An August 2018 report from LinkedIn reported that there was a "shortage of 151,717 people with data science skills in the U.S."[1]
- **Greater breadth of stakeholder communities.** There is an increasing demand from a variety of data consumer communities such as corporate sponsors and senior business stakeholders looking for faster, more reliable insight; a broader array of downstream data consumers who want to leverage a wide swath of available data assets; and the engineering facilitators looking to simplify the ways these two communities are satisfied.
- **Increased data democratization and more "aware" data consumers.** Informed user communities are begging for access to data assets in both their original and their "warehoused" formats.
- **Higher data velocity.** Increasing volumes of streaming data must be properly handled in real time to inform faster decisions.
- **Need for more flexibility.** Growing consumer communities with different demands require greater computational flexibility.
- **More complex workload management.** At the same time, those growing consumer communities require increasingly complex computational workloads.

Together, these factors lead the data management professional to rethink the organization's data warehousing, reporting, and analytics strategy. Existing on-premises data warehouses are reaching their performance and maintainability limits. Therefore, it may be time to assess the technical requirements for the future reporting and analytics

[1] Brian Holak, "Demand for data scientists is booming and will only increase," *SearchBusinessAnalytics*, January 2019, accessed June 28, 2019. https://searchbusinessanalytics.techtarget.com/feature/Demand-for-data-scientists-is-booming-and-will-increase

environment and to consider different platform alternatives for data warehouse modernization.

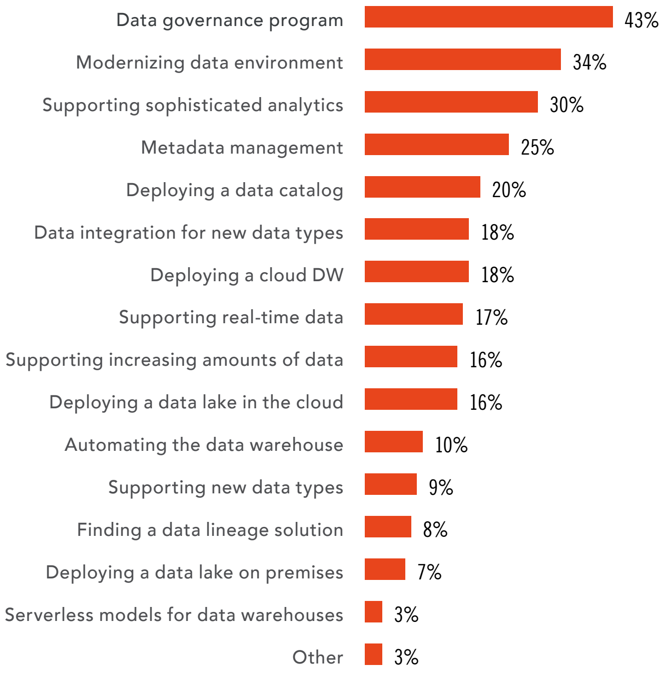
A DIVERSE DATA ENVIRONMENT

TDWI research points out that modernizing the data environment and supporting sophisticated analytics are two of the top three priorities for data management in 2019 (see Figure 1). The modernization process involves refactoring the data warehouse environment to meet business needs; therefore the functional and operational design of a modernized data warehouse environment is driven by information requirements solicited from the communities of downstream data consumers.

At the same time, there are fundamental architectural requirements for a high-performance, end-to-end data environment that incorporates a diverse set of static, dynamic, and streaming data sources to support a variety of reporting and analytics needs. This suggests that any platform for data warehouse modernization must satisfy five key technical architectural requirements.

1. **Data ingestion.** Every data warehouse involves some process for preparing data as a prelude to populating the target environment. However, because many data warehouses are populated with data from existing operational systems, the focus has been on extracting data from the source and then applying standardizations and transformations.

   A modernized data warehouse would not be limited to data sets extracted from on-premises systems. Instead, it should be able to ingest a range of data sources at scale, including static data sources originating within the corporate firewall and dynamic data sets furnished by external providers, as well as continuously streaming (i.e., high-velocity and high-volume)



What are your organization's biggest priorities for data management in 2019?

__Figure 1:__ Based on responses from 178 respondents. Source: 2019 TDWI Strategy Summit Guide.

data from both human-generated and machine-generated sources.

Additionally, although traditional data warehouses are typically limited to ingesting structured data, the modernized data warehouse should be capable of ingesting and processing structured, semistructured, and unstructured data.

2. **Data integration.** As opposed to conventional data warehouses that apply standardizations and transformations in bulk to data sets staged prior to loading, the modernized platform must also apply data preparation and integration procedures to data in transit, especially as more continuous data streams are incorporated.

3. **Data delivery.** The "data democratization" movement seeks to increase data availability and accessibility directly to data consumers through self-service mechanisms. The modernized data warehouse must provide the ability to serve data to all data consumers on demand.

4. **Storage scalability.** Data sources for analysis have become more diverse and are growing in number, volume, and velocity. The modernized environment must accommodate persistence of massive amounts of ingested and processed data.

5. **Computational elasticity and scalability.** A broader community of users will have different access patterns, leading to a need to support a range of analytical computational workloads at different times. The modernized environment should automatically scale its computational resources (both up and down) in relation to those workload demands.

Fortunately, the cloud is an optimal platform to satisfy these requirements.

CLOUD AWARENESS

From the perspective of enterprise data warehousing, we are still in the relatively early phases of cloud adoption. Though some early adopters have either lifted and shifted their existing on-premises data warehouse or have created cloud-based data lakes in anticipation of augmenting or replacing their enterprise data warehouse, more organizations are still looking to understand the value proposition of the cloud as the future platform for data warehousing and analytics.

This is borne out by TDWI research, which indicates that organizations at least recognize the cloud as a component of the future "extended information enterprise" (even though the terminology around cloud data management is still

What term(s) do you or your team use for data management that involves clouds? Select all that apply.

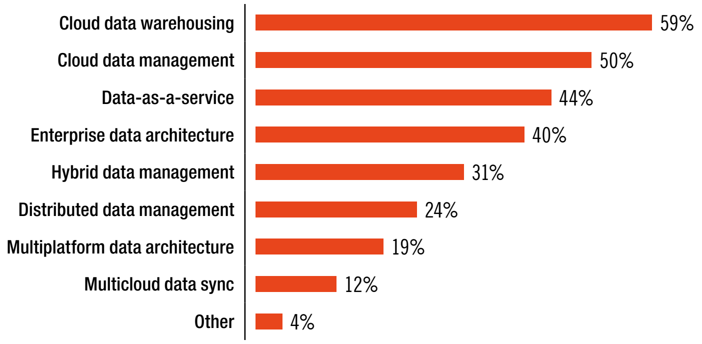


Figure 2: *Based on responses from 108 respondents. Source: TDWI Best Practices Report: Cloud Data Management, online at http://tdwi.org/bpreports.*

If your organization were to implement cloud data management, what would its leading benefits be? Select seven or fewer.

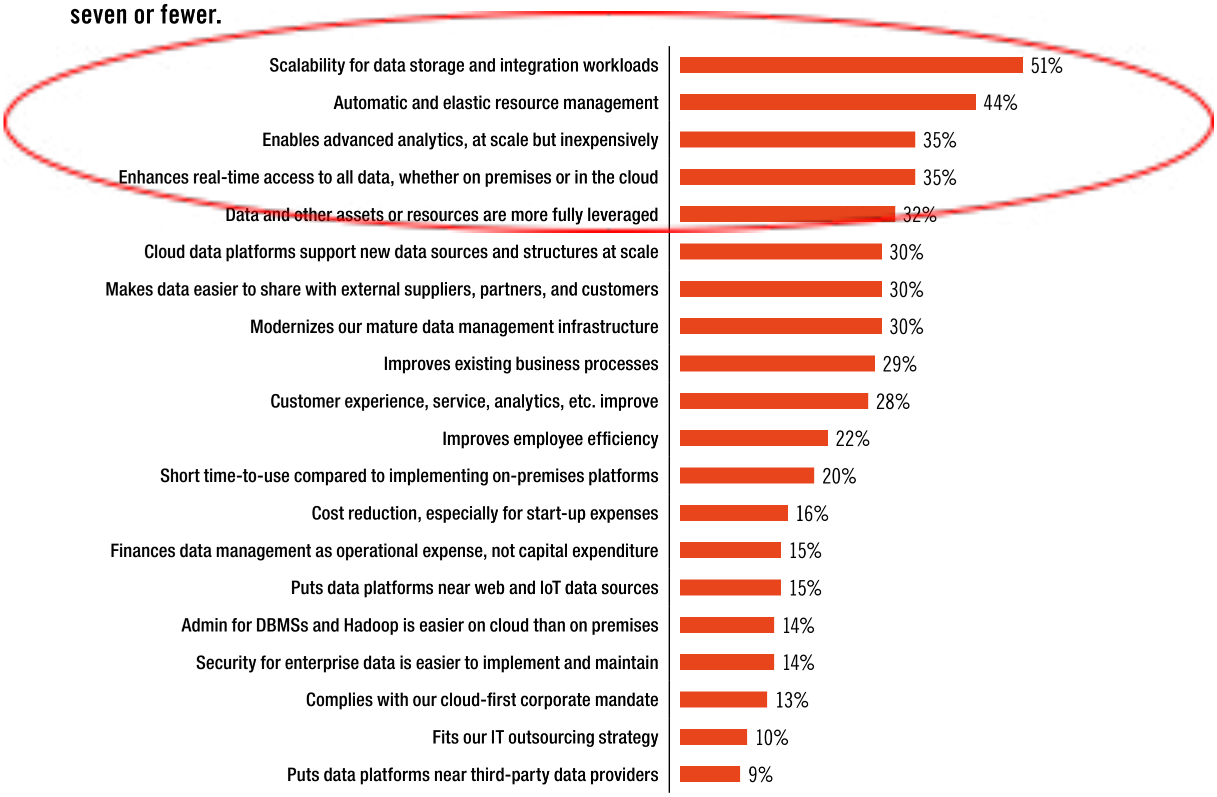


Figure 3: *Based on responses from 98 respondents. Source: TDWI Best Practices Report: Cloud Data Management, online at http://tdwi.org/bpreports.*

being refined). Figure 2 provides a list of terms that survey respondents use to refer to data management activities that involve clouds.

Although there are a number of different terms, it is interesting, and somewhat telling, to note that the most popular term used is "cloud data warehousing." Clearly the concept of deploying and running a data warehouse on a cloud platform has had some resonance.

The same respondents were asked about what they thought were the leading benefits of using the cloud (see Figure 3). The top four responses ("scalability for data storage and integration workloads," "automatic and elastic resource management," "advanced analytics, at scale, inexpensively," "enhances real-time access to all data") all touch upon some of our key architectural requirements, namely storage scalability, computation scalability and elasticity, rapid data delivery and accessibility, and a lowered cost of operations.

Cloud modernization—which includes both migrating existing data warehouse applications and new data warehouse development—relies on understanding the types of services (such as the different types of computing instances and the different types of storage services) that are available on the cloud. Modernization, however, does not happen overnight. There will be a period during which some applications will remain on premises, others will be migrated to the cloud, and still others are replaced by software-as-a-service (SaaS) or platform-as-a-service (PaaS) offerings. Modernization accounts for much of the ways that cloud services are adapted to a more flexible extended (and thereby hybridized) environment. That being said, cloud-based data warehouse architectures provide a suitable path for modernization for some key reasons.

- **Simplicity.** Data warehouses that are designed for cloud environments are easy to launch and easy to use. Because the host provider is responsible for installing and managing the data warehouse platform applicationware, there is a "short runway" for getting up and running.
- **Scalability.** On-premises data warehouses are limited in their performance by the hardware configuration that has been acquired. Cloud data warehouses rely on cloud services that can automatically make use of additional computing and storage resources without requiring the system to be taken down, updated, and restarted.
- **Maintainability.** On-premises systems require associated support: space, HVAC, power, staff for managing the space, staff for managing the system, performing system updates and upgrades, and so on. For a cloud-based data warehouse, all of these needs are handled by the host or cloud provider, simplifying maintenance of the environment.
- **Economics.** With a cloud-based system, you pay for what your system uses; you do not have to acquire hardware that sits idle when it is underused, nor do you have to purchase additional hardware as performance begins to lag. Cloud-based systems can automatically scale up or down according to user demand.

Once you have recognized the value of migrating your reporting, business intelligence, and analytics platform to the cloud, it is an easy jump to consider modernization. Simple migration of a system (also referred to as "lift-and-shift") replicates your capabilities using a cloud platform, but the process of modernization enables you to reconsider the needs of the growing communities of users and rearchitect the approaches to delivering actionable information to streamline corporate decision making.

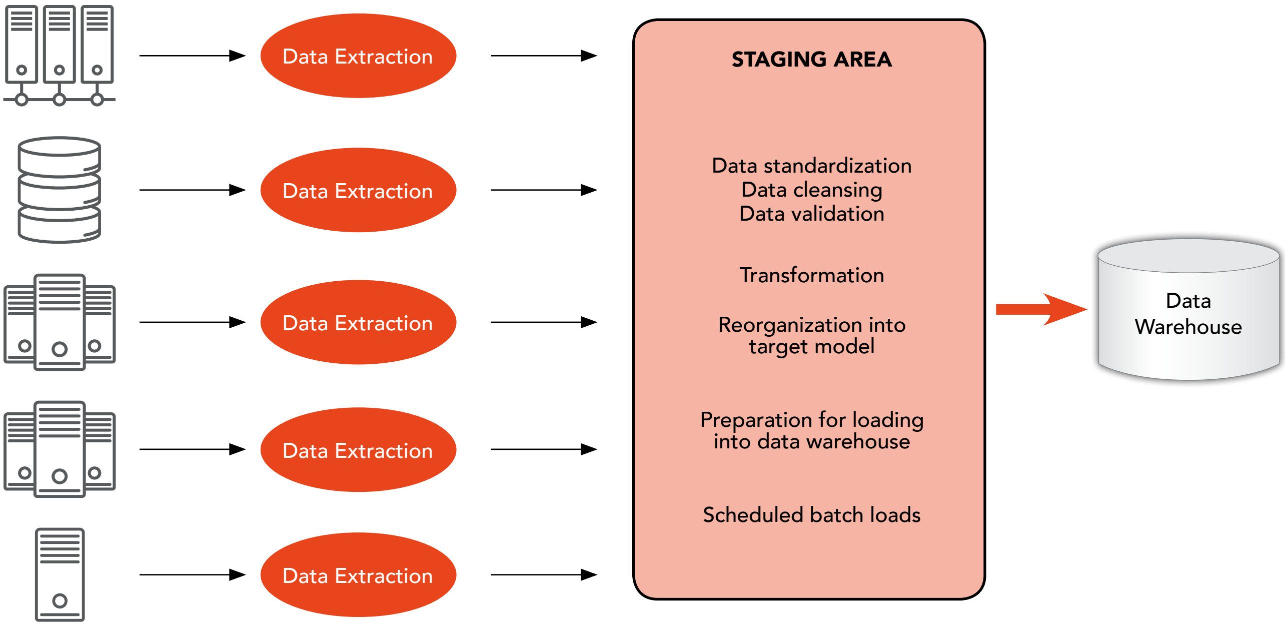


Figure 4: *The conventional data warehouse architecture.*

SHIFTING DATA WAREHOUSING PERSPECTIVES: TRADITIONAL ON PREMISES VS. MODERNIZED CLOUD

To understand how the services and capabilities provided by a cloud platform can free a data warehouse's design from the limitations of the typical on-premises data warehouse architecture, it is valuable to review the conventional architecture typically employed in on-premises implementations. Figure 4 shows the information from left to right: data sets are extracted from existing (and usually also on-premises) transaction processing or operational systems. Those data extracts are moved to a separate staging area, where the data sets are standardized, cleansed, and validated. Once these transformations have been executed, the table structures are reorganized to align with the dimensional model favored by data warehouse designers. Finally, the transformed data sets are loaded (generally in a synchronized manner) to the target data warehouse.

This approach is designed to support the collection of information from internal systems and enables operational reporting on a well-defined periodic basis according to the loading cadence. Yet this lock-step process creates the limitations and challenges of the conventional architecture, including (but not limited to) difficulties in:

- Providing real-time data integration and delivery
- Filtering data for persistence in limited storage resources
- Complexity in ingesting new data sources
- Creating new data pipelines for a new crop of citizen data analysts
- Balancing computational workloads

These (and other) limitations are attributable to technology choices driven by strict conformance to the conventional architecture.

Leveraging the cloud for data warehouse modernization allows the data warehouse engineer to overcome these limitations by adjusting the perception of how the data warehouse architecture maps to cloud resources. Figure 5 shows a more

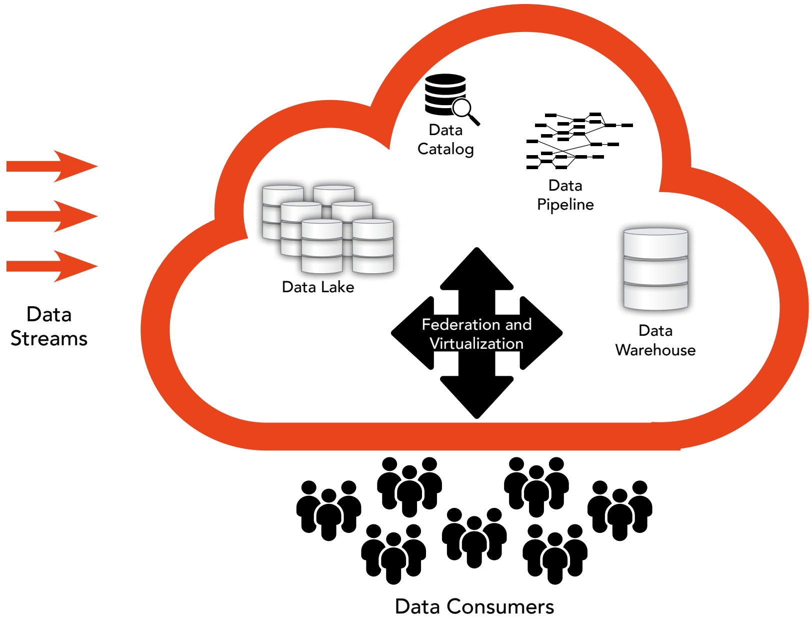


Figure 5: *Cloud-based data environment for reporting and analytics.*

flexible approach that layers both conventional and emerging techniques on top of cloud computing, storage, and stream management resources and services. Streaming data sources can be flowed directly into persistent object storage that scales automatically in relation to data volumes. Numerous data pipelines can execute in parallel, automatically launching computing instances to ensure conformance with defined service levels. Structured data can be pushed in real time to the data warehouse, with standardizations and transformations applied after loading. Data federation and virtualization techniques combine with more flexible data warehouse methods to layer structure on top of semistructured data assets in the warehouse and in the data lake. Finally, the needs of a variety of data consumers and citizen analysts accessing the reporting and analytics environment can be satisfied in an elastic manner, scaling up and down in relation to concurrent demand.

Using this high-level approach as the technology "palette," what is necessary to simplify the tasks of the data engineers becomes much clearer, especially in supporting data operations patterns that are not typically supported on premises (such as real-time analytics of continuously streaming data). The immediate and near-term challenge, though, is understanding that cloud migration is an incremental process—the journey to the cloud won't be made in a single jump.

That means that for the medium-term future, organizations will be operating within a hybrid environment that incorporates on-premises systems, SaaS/PaaS applications, and applications and systems deployed across a variety of cloud host platforms. To minimize risk, organizations are likely to choose a number of standard cloud host vendors to avoid vendor lock-in and provide the greatest degree of flexibility.

Partner with host data warehouse providers that natively take advantage of cloud resources (e.g., computing instances, object storage, metadata services and data catalogs, data virtualization, serverless computing, automated scale-up, and automated suspend, etc.). These providers can also guide your choices of cloud host and ensure portability of your data warehouse across different host platforms.

ADDRESSING CONCERNS ABOUT THE CLOUD

Of course, there have been and will continue to be concerns about migrating data and applications to the cloud, as noted in Figure 6. Many TDWI survey respondents continue to express apprehension about three key issues:

- **Data privacy and security:** (data privacy issues, data security threats, risk of exposing sensitive data)
- **Data governance across a hybrid environment:** (data governance, maintaining a single version of the truth, difficulty of sharing data)
- **Navigating the complexity of replatforming:** (replatforming to cloud-based systems, migrating data and apps to the cloud")

To address the issues around data privacy and security, cloud platform providers and product/service providers have been increasingly diligent in toughening their environments. An increasing cohort of cloud software application providers leverage host security and data protection protocols such as identity access management (IAM), but then augment that protection with additional levels of security, such as the use of network policies for granting access to users depending on the originating IP address, validation of individual user identity via a multifactor authentication, and data encryption and masking.

Cloud hosts are also supporting the ways that organizations can expand their corporate data governance initiatives to embrace data management on the cloud. One primary concern is data protection and prevention of unauthorized access. Look for vendors that can leverage existing

If your organization were to attempt cloud data management, what would its leading BARRIERS be? Select seven or fewer.

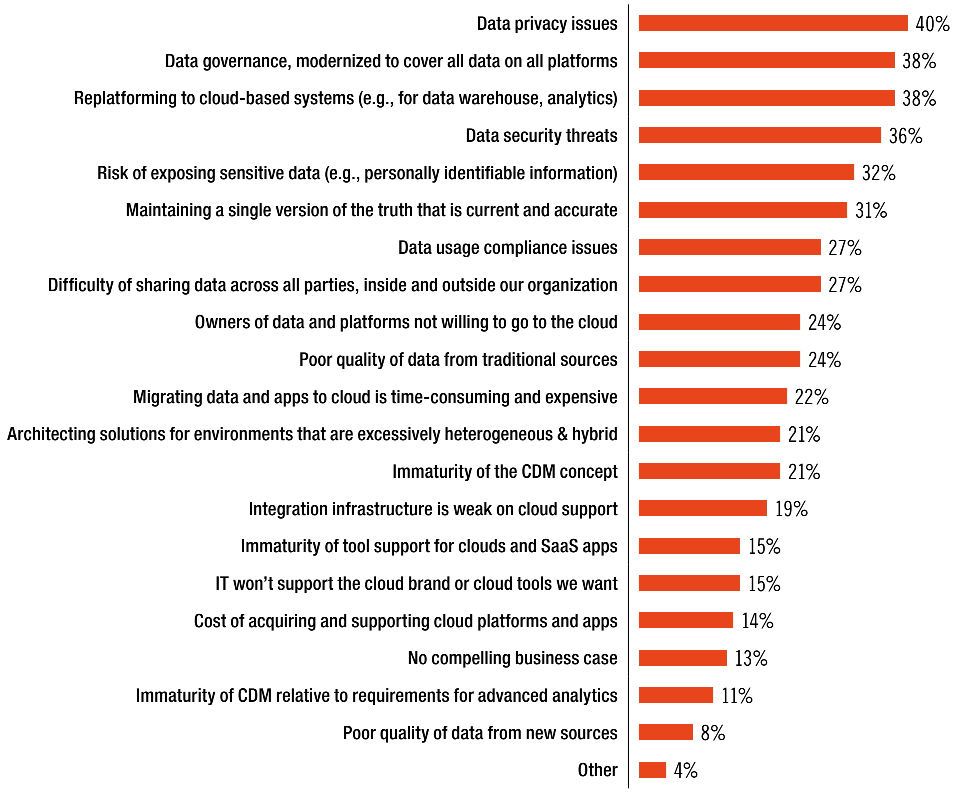


Figure 6: *Based on responses from 107 respondents. Source: TDWI Best Practices Report: Cloud Data Management, online at http://tdwi.org/bpreports.*

on-premises data protection practices and extend and deploy them in the cloud. An example would be a cloud data warehouse environment that can reuse, replicate, or extend existing security and protection within the cloud implementation (as opposed to trying to recreate the data protection profile inside the cloud separately). Aside from instituting policies for data access control and protection, the catalog of cloud services now includes tools for metadata analysis and management, data catalogs, data lineage mapping, data validation and quality control, and data integration. In essence, the cloud services are well-suited for implementing and enforcing data governance policies.

It is true: migrating any type of application to the cloud does involve a degree of complexity, especially when it comes to practical aspects such as determining which applications to migrate and in what order, determining the optimal architectures for migrated applications, and managing conformance to SLAs. However, applying some due diligence prior to migration can help reduce the complexity. By assessing and understanding the existing on-premises operational environment and assessing the corresponding downstream user workloads, you can work with your cloud/platform provider to determine how the computing requirements map to cloud resources and services, which will enable you to develop a core architecture for a cloud-based data warehouse. In addition, host providers should provide orchestration services for data pipelines to help manage and automate interoperability and integration of multiple processes to ensure all data consumer needs are met.

CONSIDERATIONS

For the near future, the lure of the simplicity, low cost, and flexibility of cloud environments suggests that the future of data warehouse modernization will involve migrating to a hybrid multicloud environment. Yet because there are risks in cloud migration, it is in your best interests to minimize them by partnering with technology vendors that provide both the platform and the expertise to help design and implement a data warehouse in the cloud.

At the same time, remember that modernization is more than just lifting and shifting a relational database from an on-premises system to a similar RDBMS implemented on top of cloud computing resources. Modernization involves understanding your current and future business process requirements, identifying where prior technology choices impede your ability to satisfy those requirements, and where you can leverage new technologies to develop a robust, flexible, and extensible framework for the future.

Data warehouse modernization in the cloud will require working with vendors that not only understand the end-to-end data life cycle for reporting, business intelligence, and analytics, but also have experience in cloud data management across the data flow spectrum: data streaming, ingestion, integration, persistence, computation, monitoring, security, and delivery. When considering a data warehouse modernization plan, make sure that your selected technology stack includes core capabilities that address ongoing concerns about data protection, governance, and complexity, such as:

- Integrated services supporting data protection
- Controls and protection for sensitive data
- Seamless data management that embraces streaming, storage, and computation
- Semantic consistency using shared metadata
- Intelligent resource management
- Workload analysis for scalability and elasticity
- Autoscaling and auto-suspend

These are all capabilities and services that are increasingly offered by most, if not all, cloud hosts as well as the vendors developing application services deployed on the cloud. Selecting these technologies will inform the modernization initiatives while simplifying the migration process.

ABOUT OUR SPONSOR

cloudera

www.cloudera.com

At Cloudera, we believe that data can make what is impossible today, possible tomorrow. We empower people to transform complex data into clear and actionable insights. Cloudera delivers an enterprise data cloud for any data, anywhere, from the Edge to AI. Powered by the relentless innovation of the open source community, Cloudera advances digital transformation for the world's largest enterprises. Learn more at Cloudera.com.

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.

ABOUT THE AUTHOR

DAVID LOSHIN, president of Knowledge Integrity, Inc, (www.knowledge-integrity.com), is a recognized thought leader, TDWI affiliate analyst, and expert consultant in the areas of data management and business intelligence. David is a prolific author on topics related to business intelligence best practices. He has written numerous books and papers on data management, including *Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph* and *The Practitioner's Guide to Data Quality Improvement*, with additional content provided at www.dataqualitybook.com. David is a frequent invited speaker at conferences, online seminars, and sponsored websites and channels including TechTarget and The Bloor Group. His best-selling book, *Master Data Management*, has been endorsed by many data management industry leaders.

David can be reached at loshin@knowledge-integrity.com.

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.