research

CHECKLIST REPORT 2021

Cloud Data Warehousing Straight Talk

By David Loshin

Sponsored by:

MATILLION

tdwi TRANSFORMING DATA WITH INTELLIGENCE™

MARCH 2021

TDWI CHECKLIST REPORT

Cloud Data Warehousing Straight Talk

By David Loshin

TABLE OF CONTENTS

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

FOREWORD

Admit it. Developing a data architecture used to be straightforward. Now lowered expenses and simplified management are inspiring senior managers to migrate many organizations' enterprise reporting and analytics infrastructures to the cloud.

Cloud service providers are rapidly producing a dizzying array of storage, computing, and other value-added services. Enterprises supplanted old-fashioned extract, transform, and load (ETL) processes with data onboarding and pipeline orchestration. With multicloud hybrid architecture only getting more complicated, even highly trained technical people have trouble keeping track of what's what.

Organizations increasingly need to support more sophisticated data scientists as well as a broader community of citizen data analysts—those business data consumers and business analysts who maintain a level of data awareness and are comfortable using end-user technologies coupled with self-service data access to produce their own reports and analytics. Both must be enabled without creating additional IT bottlenecks. We wrote this TDWI Checklist Report to raise awareness among various corporate personas about cloud data warehouse architectural paradigms.

We provide some straight talk about the current state of data warehousing, and answer questions such as:

- What is an enterprise data warehouse?
- What is the difference between a data warehouse and a data mart?
- Who is using a data lake and why?
- What is a data lakehouse?

Critically, more chief data officers and senior information management directors are compelling their teams to migrate their data warehouse environments to the cloud. Although there are many benefits to a modernized data warehouse environment in the cloud, the devil, as always, is in the details. This checklist discusses emerging platform paradigms that complement the traditional data warehouse and explains how enterprises can deploy these components across a hybrid cloud architecture.

1 UNDERSTANDING THE CHARACTERISTICS OF DATA WAREHOUSES

To set the stage for considering a cloud-based reporting and analytics environment, we need a working definition of a data warehouse. Conceptually, a data warehouse is a centralized repository of data organized to simplify data reporting, analytics, and production of actionable knowledge to drive advantageous and beneficial decision making.

Practically, though, a data warehouse is a database implemented as a segregated data architecture in which the data is organized using a schema that simplifies and accelerates reporting and analytics. The data schema often facilitates aggregation and summarization.

Data consumers can easily execute queries providing insight associated with aggregation along various dimensions of classification such as time and geography, as well as other defined categorizations. The data warehouse is populated on a periodic basis with data sets that are extracted from internal application systems as well as data sets acquired from other sources.

The data warehouse is often supported with additional capabilities and services, such as:

- A metadata repository documenting the data elements managed within the data warehouse
- Identified source systems from which data sets are extracted to load into the data warehouse
- Processes for cleaning and standardization to prepare data for loading into the data warehouse
- Data integration processes that execute the transformations and load the data
- Processes for filtering data by subject area and reducing data volumes into data marts
- End-user querying, reporting, and analysis tools enabling data consumers to gather actionable knowledge to motivate profitable actions

In essence, a data warehouse is not a *system* or even a *platform*. Rather, TDWI defines a data warehouse as a *data architecture* populated with data, metadata, and schema unified via an integration backbone.

This distinction is important because traditional data warehouses typically were implemented as monolithic systems, specialized "analytical appliance" hardware platforms that depreciate in both performance and value over time. Fortunately, the cloud provides an alternative architecture that eliminates limitations of rigid, traditional on-premises implementations.

2 THE RISE OF OPERATIONAL DATA STORES

Organizational platforms supporting business decision making have been part of the information management landscape for decades. The traditional data warehouse environment architecture that effectively jelled in the early 1990s embraces a set of staged processes to extract data from operational or transactional sources, land that data in a staging platform for cleansing and standardization, and perform bulk uploads to the data warehouse platform.

However, as data warehouses became established in the enterprise, the complexity of juggling the integration of numerous data sources and growing data volumes increased. The appetite for analytics grew in lockstep with increased data volumes, data variety, and the number of data sources. At the same time, downstream information consumers' expectations for data availability also grew—along with their desire to rapidly access data to support operational reporting demands, especially within independent lines of business.

The staged process for data warehouse population can be adapted to support both holistic organizational analyses and line-of-business reporting. A data warehouse ultimately supports strategic analysis and decision making and must provide visibility horizontally *across* the organization from the perspective of the lines of business and how they interoperate, as well as from a historical context over time. However, as data sets are extracted from individual application systems, they could be made available for operational reporting. This evolved into the concept of the operational data store (ODS)—an interim data platform intended to support the operational aspects of a line of business or a particular business process.

For example, a manufacturing company might use a data warehouse to analyze how item production correlates to sales over time across different geographies. However, the factory floor managers might want more up-to-date reporting about the status of the manufacturing and production processes—such as production speed, product quality, and injection of final product into the supply chain.

ODSs are valuable as the collection points for data relevant to lines of business, especially when they provide operational personas access to more real-time information for operational reporting and analysis tasks.

3 DATA WAREHOUSING MOVES TO THE CLOUD

The traditional data warehouse architecture is easily segregated. Extracted source data is easily shunted to data staging areas facilitating the integration and loading pipelines. Until now, our discussion presumed that the various components of the data warehouse architecture are generally managed within an on-premises platform environment within the organization's data center.

However, on-premises data warehouse platforms are reaching the limits of scalability without drastically increasing costs for hardware, software, and typical environment resources (e.g., space, power, cooling). As the capital acquisition costs and on-premises maintenance and management costs rise, organizations are considering migrating data warehouse environments to the cloud.

Conventional monolithic or appliance-based on-premises data warehousing is somewhat stodgy and rigid when compared to the increased flexibility of cloud-based implementations. Consider these factors when contemplating migrating to the cloud:

- **CLOUD ECONOMICS.** With the cloud, your organization transitions from capital acquisition of platforms to "renting" the required computational resources from the cloud service provider. This switch from capital expenses to operational expenses is economically desirable, especially when you only pay for what you use.
- **LIMITLESS RESOURCES.** Cloud service providers (CSPs) can deliver virtually unlimited storage and computing resources. As your organization's computing and storage needs change, it is easy to expand the cloud resource footprint.
- **ELASTICITY.** Indeed, computing and storage needs do change, and although system growth is easily accommodated, the benefit of only paying for what you use applies in both directions. The concept of elasticity means your organization can maintain on-demand access to scalable, high-performance computing, but as demand declines, you can dial back your use of cloud resources.
- **HYBRID INTEGRATION.** Once you are free from the on-premises environment, your organization is able to integrate across a hybrid environment connecting on-premises, multicloud, software-as-a-service (SaaS), and platform-as-a-service (PaaS).
- **CLOUD SERVICES.** CSPs have been diligent in developing and providing access to many value-added cloud-native services that simplify maintenance, management, and, especially, development.
- **ARTIFICIAL INTELLIGENCE/MACHINE LEARNING.** CSPs provide easy access to AI/ML services that are simple to blend with the cloud data warehouse.
- **DATA SECURITY.** Cloud service providers are extremely attentive to customer concerns about data privacy, protection of sensitive data, and data security and are very diligent in strengthening their data security services.

4 CLOUD SERVICES INCLUDE REPORTING AND ANALYTICS

Some organizations are comfortable with effectively lifting and shifting their existing vendor-based data warehouse platform from on premises to a similarly configured cloud-based platform. Other organizations use data warehouse migration as an opportunity to revisit the underlying architecture and consider the potential benefits of abandoning their existing vendor selections for cloud-based reporting and analytics services. These forward-thinking organizations can leverage the types of resources and services CSPs provide to radically improve their analytics environments.

For example, many organizations have a variety of reporting and business intelligence (BI) tools for building and producing reports. Although these tools can be implemented using a cloud data warehouse as their back end, the cloud offers alternatives. CSPs provide cloud-based services specifically engineered to replicate the same reporting and analysis techniques employed by users with traditional tools.

Because cloud hosts want to encourage their customers to increase their use of cloud resources, they have developed services supporting all aspects of the analytics/BI environment. For example, cloud hosts provide tools for data profiling, metadata management, and data management across a variety of levels of storage, including in-memory data provision, block storage using solid-state disk (SSD) storage, more traditional disk storage, and even archival storage. These different layers of a storage hierarchy can be leveraged to reduce costs by balancing accessibility and access latency against higher or lower costs of the different storage paradigms.

Importantly, the cloud hosts provide a variety of reporting/analysis services that can support an organization's operational and strategic reporting needs. This includes the ability to develop APIs for ad hoc data access, specialized services for developing reports, and integration with artificial intelligence and machine learning algorithms for more sophisticated advanced analytics.

5 LEVERAGING A DATA LAKE

OK, you are convinced that it makes sense to migrate your data warehouse to the cloud. Yet when you review the technology media, you find considerable buzz about data lakes.

What is a data lake? Can it replace the data warehouse? Why is so much written about failed data lake implementations?

The growing number of reporting/analytics consumers can be differentiated into distinct consumer communities, including traditional data analysts, citizen data analysts (those looking for transparent capabilities for producing simple reports), informed business analysts (who are a little more savvy), as well as more sophisticated data scientists desiring access to massive data volumes in their original forms.

For some of these analysts, the rigid structure of the processes that ingest, process, and transform data for loading into a data warehouse often "wash out" information with potential for analytics insights. Alternatively, data scientists may want greater control over their own data pipelines for data preparation.

This is where the data lake comes in—it provides a curated repository of source data sets in their original formats and makes those data sets available to a variety of consumers. TDWI defines a data lake as an unstructured data repository that contains information available for analysis. A data lake ingests data in its raw, original state, straight from data sources, without any cleansing, standardization, remodeling, or transformation.

In essence, a data lake is distinguished from the existing data warehouse by providing a more flexible platform for data availability and accessibility. In contrast with a data warehouse, a data lake imposes fewer limitations on what data elements are available for use, allows for data storage at scale, and can accommodate structured, semistructured, and unstructured data.

To address the risk of data lake failure, organizations are instituting processes for data curation and governance to assess data lake assets, document their structural and object metadata in a data catalog, and help data consumers find and use the optimal data assets for their specific needs.

6

WHAT IS A DATA LAKEHOUSE?

We have differentiated between a data warehouse and a data lake. Both have benefits and drawbacks that would inspire a data consumer to choose one paradigm over the other. Data in the data warehouse is cleansed and well-organized to simplify reporting and analysis, but it is limited because the included data sets are subject to filtering and transformation.

Data lakes allow for a much broader array of data options, but when the data sets remain in their original raw state, there are bound to be inconsistencies that potentially impact trust in the analytics insights derived from them. In a modern analytics environment, neither approach alone is likely to be satisfactory to meet every analyst's needs.

However, you do not have to choose one approach over the other. These two paradigms are not mutually exclusive, and your organization can benefit from the synergy that emerges from blending the approaches. This blended paradigm, known as a data lakehouse, looks at how the data warehouse and the data lake can complement each other and deliver the best of both worlds.

A data lakehouse is characterized as a reporting/analytics/BI environment that provides a semantic layer harmonizing accessibility to the structured, semistructured, and unstructured data assets managed in the combined data landscape afforded by the warehouse and the data lake.

A data lakehouse provides:

- Standardized data storage formats in object storage (such as using the columnar alignment provided by Apache Parquet)
- Separation of storage from compute, freeing data consumers from the limitations imposed by monolithic databases
- Virtual layering of structured schema over data managed in object storage
- Integrated governance over population, management, and access to data in object storage
- Support for query access to structured data in the data warehouse as well as both structured and semistructured data in the data lake using schema-on-read
- Capabilities for cross-platform integration (e.g., queries joining database tables with semistructured data sets)
- Practically unlimited reusability, in that you can load the data once into the data lakehouse and use that data resource as part of any number of advanced analytics workflows as well as attaching or loading the data directly into the cloud data warehouse's more structured reporting and analysis workflows

Essentially, the data lakehouse approach smooths integration and access across the data landscape and allows for even greater flexibility for the different consumer communities.

7 UNDERSTANDING HYBRID CLOUD

Most organizations are set on a trajectory to migrate their reporting and analytics to the cloud by employing a combination of a cloud-based data warehouse, data lake, and data lakehouse concepts. Yet why should an organization limit itself to a single virtual environment with a specific cloud service provider? We are already seeing organizations migrating to more than one cloud environment using the same CSP and even employing services from multiple cloud service, SaaS, and PaaS providers.

Different cloud service providers provide different services and capabilities, and the same economic and performance factors that are driving organizations to the cloud are often encouraging them to seek out the best of breed across a variety of CSPs and vendors for particular services and balance costs versus performance.

That being said, some challenges and drawbacks to this approach remain, such as:

- Increased **data chaos** due to unmanaged data distribution across a variety of CSP object stores
- Difficulty in managing **spiraling costs** for data egress and replicated invocation of the same or similar services
- Managing the **inventory of data assets** across on-premises and cloud storage resources
- Managing **data protection** across the different platforms
- **Performance bottlenecks** when a hybrid analytics application must move data from one CSP to another
- Increased complexity of **data integration**, especially as data sets transition from one platform to another

Hybrid cloud architectures hold promise when an organization has a clear understanding of the key performance criteria and how the architecture is designed to optimize for those criteria. Yet as the structure and offerings of hybrid cloud environments continue to grow in complexity, your data architects will require more scaffolding to maintain visibility across the landscape.

We suggest that hybrid cloud data warehouse architectures require more sophisticated supporting technologies. We recommend introducing platform management technologies that optimize management across different platforms and provide guidance for allocating data and resource usage for optimal performance and value. These tools provide the necessary visibility to ensure optimal use of hybrid cloud architectures.

AFTERWORD

Having read this overview of the increasingly complex cloud data warehousing alternatives, you may recognize the need for technical capabilities that can help in the architecture, design, implementation, and finally migration into your emerging hybrid cloud architecture.

We suggest that one key involves governed data management and integration, and that means looking at products that enable a variety of data consumers.

Necessary technical capabilities include:

- **Data onboarding** that balances data loading, ingestion, and integration with managed data pipelines that enable streamlined ingestion and integration of static data as well as continuous data in motion
- **Data preparation** tools that data scientists can use to produce (and publish) their specialized data pipelines
- **Data pipeline orchestration** to speed processing and data pipeline management
- Coordinated and integrated **data governance** to ensure a level of trust in the data
- Coordinated and integrated **data security** that applies data protection policies consistently across the hybrid environments
- **Data quality and data lineage** for validation and trustworthiness
- **Data catalogs** for enhanced visibility into curated data assets and their distribution across the data landscape
- Data management cost **analysis and optimization** that can help modulate budget decisions when implementing the data analytics architecture

ABOUT OUR SPONSOR

MATILLION

Matillion makes the world's data useful with an easy-to-use, cloud-native data integration platform. Matillion delivers rapid returns on cloud investments for global enterprise customers, helping them wield data as their most strategic asset. Optimized for modern enterprise data teams, Matillion is built on native integrations to cloud data platforms such as Snowflake, Delta Lake on Databricks, Amazon Redshift, Google BigQuery, and Microsoft Azure Synapse to enable new levels of efficiency and productivity in data programs.

Learn more at matillion.com.

ABOUT TDWI RESEARCH

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

ABOUT THE AUTHOR

David Loshin, president of Knowledge Integrity, Inc., is a recognized thought leader and expert consultant in the areas of data management and business intelligence. David is a prolific author regarding business intelligence best practices as the author of 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, web seminars, and sponsored websites and channels. David is also the Program Director for the Master of Information Management program at the University of Maryland's College of Information Studies.

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

ABOUT TDWI CHECKLIST REPORTS

TDWI Checklist Reports provide an overview of success factors for a specific project in business intelligence, data warehousing, analytics, 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.