CHECKLIST 2021

The Wisdom of Transitioning to a Data Lakehouse Strategy

By David Loshin

databricks

tdwi TRANSFORMING DATA WITH INTELLIGENCE™

The Wisdom of Transitioning to a Data Lakehouse Strategy

By David Loshin

Organizations bent on deploying a strategy for digital transformation are rapidly transitioning their data and applications to the cloud. Cloud modernization has become the rallying cry for organizations looking to take advantage of advanced analytics using machine learning and artificial intelligence while continuing to support traditional data warehouse consumers.

An expanded array of data consumers with different needs and expectations has become more sophisticated in the use of analytics and data science tools, and their expectations for data democratization have created demand for a simplified way to access data assets shared across the enterprise without the constraints imposed by using the traditional data warehouse architecture.

One of the first innovative uses of cloud-based storage is the creation of the cloud data lake—a repository for capturing and storing data sets in their

Seven steps for transitioning to a data lakehouse:

1. Recognize the limitations of the traditional data warehouse model
2. Acknowledge the value of integration with semistructured and unstructured data assets
3. Understand the differences between a data lake and a data lakehouse
4. Leverage ACID transaction semantics to establish consistency and trust
5. Consider a decentralized architecture
6. Institute governance to centralize data awareness and protection
7. Seek opportunities for optimization

raw format to be made available for data analysts' and data scientists' analytics projects. However, the data lake model is flawed. Although the data lake approach does presumptively allow for sharing up-to-date data assets in their original formats, the absence of governance, the lack of data quality, and concerns about data awareness and data protection pose limitations.

This checklist examines the limitations of existing data architecture strategies when supporting emerging and future analytics needs and how a data lakehouse approach addresses those limitations. We suggest integrating structured data with unstructured data to enable advanced analytics while acknowledging that issues with the data lake need to be overcome. We then explain the differences between a data lake and a data lakehouse and how the flexibility of the lakehouse architecture allows you to enable optimized and governed data access to a wide variety of data assets managed across a multicloud data environment.

1 Recognize the limitations of the traditional data warehouse model

For the past three decades, the data warehouse has been the primary data architecture supporting enterprise reporting and analytics using structured data. This paradigm may be sufficient for organizations limited to descriptive and diagnostic analytics. However, as organizations augment their analytics strategies to include more advanced analytics that leverage semistructured and unstructured data assets, it is abundantly clear that the data warehouse model is insufficient to meet these organizations' future analytics needs.

The inadequacy of the data warehouse approach is magnified as organizations migrate their data and workloads to the cloud. Cloud platforms provide streamlined services for blending different varieties of data and offer greater algorithmic depth for advanced machine learning and artificial intelligence (AI) services. Increasingly savvy data analysts and data scientists wishing to adapt more sophisticated analytics are constrained by the hurdles put up by the data warehouse's rigid framework, such as:

- **No support for semi/unstructured data.** A data warehouse supports SQL queries on structured data but has limited support (if any) for analyzing semistructured or unstructured data.

- **Functional limitations for advanced analytics.** The traditional data warehouse has limited or no support for advanced analytics such as machine learning without extracting the data.

- **Data staleness.** The typical "lock-step" batched population of a data warehouse means that its data is often not the most current.

- **Modeler bias.** Information technologists design the warehouse table structure following techniques developed to meet query performance demands. This leads to a bias associated with the IT team's decisions about which source data elements are (or are not) included.

- **Data integration effort.** The need to configure data access streams or extract data from existing sources and organize data pipelines for preparation and integration into the warehouse creates an artificial dependency for complicated ETL, ELT, or other data preparation and integration efforts.

- **Lack of flexibility.** For more sophisticated data scientists, the data warehouse limits the ability to interoperate with emerging popular open source tools or cloud-native services for machine learning and AI.

- **Vendor lock-in.** Data warehouses are often tightly coupled with proprietary database management platforms, and the organization's warehouse capabilities are limited by what the vendor provides.

The data warehouse architecture has served organizations well. However, organizations that cannot augment their data analytics architecture to address its limitations will rapidly find themselves unable to remain competitive.

2 Acknowledge the value of integration with semistructured and unstructured data assets

Organizations are collecting more diverse data assets for analytics, specifically semistructured and unstructured data (from a variety of sources including Internet of Things (IoT) devices and social media), and "a majority of data (80% to 90%, according to multiple analyst estimates) is unstructured information."[1] TDWI research indicates that the "primary use cases for data lakes include supporting source data staging, advanced analytics, and extending the data warehouse to process and store newer data types (such as unstructured data)."[2]

Although "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,"[3] only 18 percent of companies responding to a Deloitte survey have "taken advantage of unstructured data (such as product images or customer audio files) or comments from social media."[4]

Semistructured and unstructured data power machine learning and deep learning use cases. Whether the task is transforming unstructured data into searchable text or more sophisticated scenarios (such as tagging and classifying data assets, natural language processing, entity identification and extraction, identifying and analyzing social networks associated with text streams, or sentiment analysis), machine learning and artificial intelligence techniques are how many organizations are going to derive significant value from data going forward.

Although the data warehouse remains a solid platform for reporting and analyses using structured data, it cannot adequately support integration and processing of unstructured data. Organizations wanting to analyze all types of data must consider data architectures that, through governance and standards, blend the data warehouse's capabilities for conventional analytics, services for advanced analytics workloads, and the flexibility and low-cost storage of data lakes.

[1] Tam Harbert, "Tapping the power of unstructured data." Feb 1, 2021, MIT Sloan School of Management, https://mitsloan.mit.edu/ideas-made-to-matter/tapping-power-unstructured-data.
[2] *TDWI Best Practices Report: Unified Platforms for Modern Analytics*, 2021, https://tdwi.org/bpreports.
[3] *TDWI Best Practices Report: Building the Unified Data Warehouse and Data Lake*, 2021, https://tdwi.org/bpreports.
[4] Tom Davenport, Jim Guszcza, Tim Smith, Ben Stiller, "Analytics and AI-driven enterprises thrive in the Age of With," July 25, 2019, https://www2.deloitte.com/us/en/insights/topics/analytics/insight-driven-organization.html.

3 Understand the differences between a data lake and a data lakehouse

Cloud-based data lakes avoid the data warehouse's deficiencies in managing a variety of data types by providing a storage platform to accommodate semistructured and unstructured data assets. Because data lakes typically don't have the proper levels of governance and oversight to ensure that shared data assets are suitable for use, this leaves the hard work of analysis to the data consumer, who is often not familiar enough with the tools and techniques for utilizing those semistructured and unstructured resources.

Complications emerging from an ungoverned data lake include the following:

- **Lack of data awareness.** As an ad hoc dumping ground for data, the data lake's lack of a catalog describing each data object's metadata, contents, organization, and location complicates the ability to find and use data.
- **Unacceptable data quality.** The absence of data validation and other data controls can severely impact the perception of data trustworthiness and usability of data assets in the data lake.
- **Data access issues.** Data consumers must have simplified methods for accessing data lake data assets, especially when an access method may not be clear, without forcing every consumer to know the details of each data set's level of structure and organization.
- **Data usability.** The lack of data lake organization limits utility when data consumers don't know the optimal methods for querying a combination of structured, semistructured, and unstructured data.
- **Consistency.** As data streams flow into the data lake, there is the risk that there will not be a coherent consistent view among the sets of applications accessing the data.
- **Data protection.** Access controls protecting against unauthorized access must be put into place before allowing general access to data sets shared via a data lake.
- **Query performance.** Aside from the general lack of SQL support, query performance can be seriously impacted by the large number of small files, the need for repeated accesses to the same data, and minimal indexing and partitioning.

The data lakehouse approach is a framework adopting standardized system designs (relying on open standards such as Delta Lake, Hudi, and Iceberg) that intends to mitigate these issues by blending the benefits of a data lake with the structures and data management features of the data warehouse.

The capture and publication of data asset metadata in a data catalog improves data awareness, while increased governance, transaction consistency, and auditing supports improved quality and consistency. Schema enforcement coupled with query and BI tool support layered on top of data sets in a data lake simplify data accessibility while enabling query optimization.

4 Leverage ACID transaction semantics to establish consistency and trust

A commonly used computational paradigm for big data employs the lambda or kappa architecture, which is designed to combine batch processing and stream processing to provide greater throughput and reduced latency in delivering results when processing massive data volumes. As developers and programmers have transitioned into big data management and analytics, there are some who are unfamiliar with the drawbacks of the lambda architecture, specifically in its complexity (managing multiple code bases) and expectations of eventual consistency (when computed results are not always consistent with incoming streamed data).

Many developers are unaware of the scope of the consistency issue. Although learning the basics of transaction semantics is prescribed in any college computer science databases class, many data professionals and analytics consumers have transitioned into the discipline with neither a formal computer science nor a database background. The outcome is that analysts that have solely used batch-loaded data warehouses are unfamiliar with the value of ACID transactions when transitioning to analyzing data on top of a lambda architecture.

ACID-compliant transactions comply with four properties:

- **Atomicity,** in that collections of tasks grouped together as a single transaction either succeed or fail completely, leaving no doubt that some of the tasks completed while others did not.
- **Consistency,** asserting that the database remains valid across each executed transaction.
- **Isolation,** in that concurrent execution of more than one transaction leaves the system in a state as if the transactions were executed sequentially.
- **Durability,** in that once a transaction has been committed it remains committed in the presence of any issues or system failures.

Populating and accessing a data lakehouse engineered with a development framework and optimized structured transactional layer supporting ACID transactions addresses some key drawbacks of the data lake. First, it reduces the complexity of the lambda architecture by allowing developers to focus on a single code base for their applications. Furthermore, enforcing the ACID properties helps to establish data consistency, thereby increasing the level of trust in the data accessed from the data lakehouse.

5 Consider a decentralized architecture

One critical benefit of the data lakehouse strategy is that it overcomes what we referred to earlier as “modeler bias” associated with the traditional data warehouse model. Historically, data engineers designed the data warehouse with schemas using predefined data models intended to alleviate the performance challenges of querying data organized to support transaction processing. Although the data warehouse schema was arranged to speed aggregate queries (such as sums and counts), the data warehouse developers often made decisions about what data to include (or exclude) in the warehouse model.

Between these decisions by fiat and the processing dependencies emerging from the need for complex data pipelines to flow data via sequences of transformations into a cloud-based data warehouse (such as increased storage, increased costs for data movement, diminished ability to govern replicated data, and inconsistency of data), the data consumers are limited to the data elements preselected for population.

The concept of the data lake somewhat loosens those constraints by enabling access to the source data sets in their original formats. However, extracting data and moving that copy of the source data into a data lake does not eliminate the aforementioned issues. Each time data sets are extracted from a source and the extract is moved to another location, it creates opportunities for inconsistencies to creep into the data.

One alternative is a decentralized, federated data lakehouse architecture. In this approach, each functional area can build and oversee its own set of information domains and maintain control over its quality and usability. A logical lakehouse configuration can be created and superimposed over the repositories to connect federated information domains. The transparency of the lakehouse approach simplifies data access without imposing technical constraints on each information domain. Access to these domains facilitates cross-functional reporting and analytics driving corporate strategic decisions. At the same time, accessing and using data in place is less costly than copying it, moving it, and storing it multiple times.

6 Institute governance to centralize data awareness and protection

If anything, the main objective of a data lake (in general) and a data lakehouse in particular is data democratization—establishing a unified platform environment allowing streamlined data access for a variety of downstream data consumers and empowering more sophisticated analytics using machine learning and AI services and data science tools. Expanding the pool of data consumers poses two critical challenges.

First, ungoverned organization of data assets in the data lake prevents the developer teams from rapidly assembling applications and APIs enabling access, suggesting a need to raise awareness of the data assets that are accessible. Second, even if there are well-defined access methods and APIs, there is still a need to enforce data policies associated with access control, data protection, and compliance with imposed business rules. For example, rules such as the HIPAA Privacy Rule restrict what data can be viewed, by whom, and under what circumstances that information may be shared.

These issues are really two sides of the same democratization coin: enabling access to those privileged to see the data while simultaneously preventing unauthorized data exposure to those who have not been granted those rights. These issues become muddied when subjecting the data to analytics. What data assets do data scientists need to produce analytical models, and what are the best ways to ensure the appropriate degree of protection for sensitive data?

Institute a comprehensive governance framework that organizes and catalogs data while classifying those data assets according to a taxonomy specifying levels of sensitivity. Define data protection policies with fine-grained controls that can be imposed on data assets managed within the data lakehouse. At the same time, devise a taxonomy for data consumers aligned with limitations on use. Integrate directives for data encryption and masking where necessary.

These governance tactics will simplify data protection by focusing on the mappings between these taxonomies and seeking out tools that can transform those high-level data protection policies into implementation directives. Doing so will simultaneously facilitate compliant data sharing and enable governed development of advanced analytics models.

7 Seek opportunities for optimization

Organizations migrate their data and computing to the cloud to take advantage of scalability while leveraging the anticipated savings of cloud economics. However, a naïve approach to migration runs the risk of having the opposite effect: decreased performance and increased costs when the implementers are not savvy about how cloud computing really works. A data lakehouse architecture can help finesse both of these potential issues.

Use a data lakehouse strategy to optimize data utility across a number of dimensions:

- **Faster development time.** By providing standardized methods and APIs for data access, the data lakehouse architecture streamlines development of advanced analytics applications, speeding time to value.

- **Storage optimization.** A lakehouse strategy that federates data from multiple functional units allows data to be used in place without extracting data to a separate data lake or loading it into a data warehouse. This reduces the amount of replicated data copies, thereby reducing the storage demands.

- **Performance optimization.** A properly configured data lakehouse can leverage the same algorithms for query compilation and optimization for structured data. A data lakehouse encourages moving computation to the data instead of the conventional approach of moving data to the computing resources, which also reduces data latency and speeds execution time.

- **Cost optimization.** Cloud service providers charge clients for storage and data movement based on data volume. Data lakehouses that use storage formats that take advantage of compression minimize data storage volume. Reduced storage volume significantly lowers the volume of data stored and moved, resulting in lower overall cloud services costs.

- **Time optimization.** Using a data lakehouse architecture simplifies the way data sets are managed within the data lake, and the semantic layer on top of data in a federated data lakehouse allows data analysts to quickly come up to speed in accessing and using all data that has been effectively published into the lakehouse.

Afterword

Cloud-based data strategies continue to evolve, with each generation adopting properties that address the flaws and drawbacks of previous attempts. Digital transformation cannot succeed when attempting to use legacy architectures.

When transitioning to the cloud, look for technology partners that can provide platforms that support:

- Reduced burden for management and administration through the use of a data catalog
- Simplified development by reducing the dependency on the lambda architecture
- Optimized data access with reduced data latency through a managed access layer
- Reduced costs for data movement and data storage
- Standardized access for structured and unstructured data

The data lakehouse supports these objectives, and transitioning to a data lakehouse strategy enables data democratization, empowers data analysts and data scientists, and avoids the pitfalls and limitations of legacy data warehouses and data lakes.

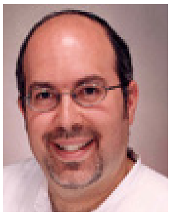
About our sponsor

databricks

Databricks.com

Databricks is “the data and AI company.” More than 5,000 organizations worldwide—including Comcast, Condé Nast, H&M, and over 40 percent of the *Fortune* 500—rely on the Databricks Lakehouse Platform to unify their data, analytics, and AI. Databricks is headquartered in San Francisco, with offices around the globe. Founded by the original creators of Apache Spark, Delta Lake, and MLFlow, Databricks is on a mission to help data teams solve the world’s toughest problems.

About the author



David Loshin, president of Knowledge Integrity, Inc., (www.knowledge-integrity.com), 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 frequently 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 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 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.

tdwi
Transforming Data
With Intelligence™

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

E info@tdwi.org

tdwi.org

© 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.