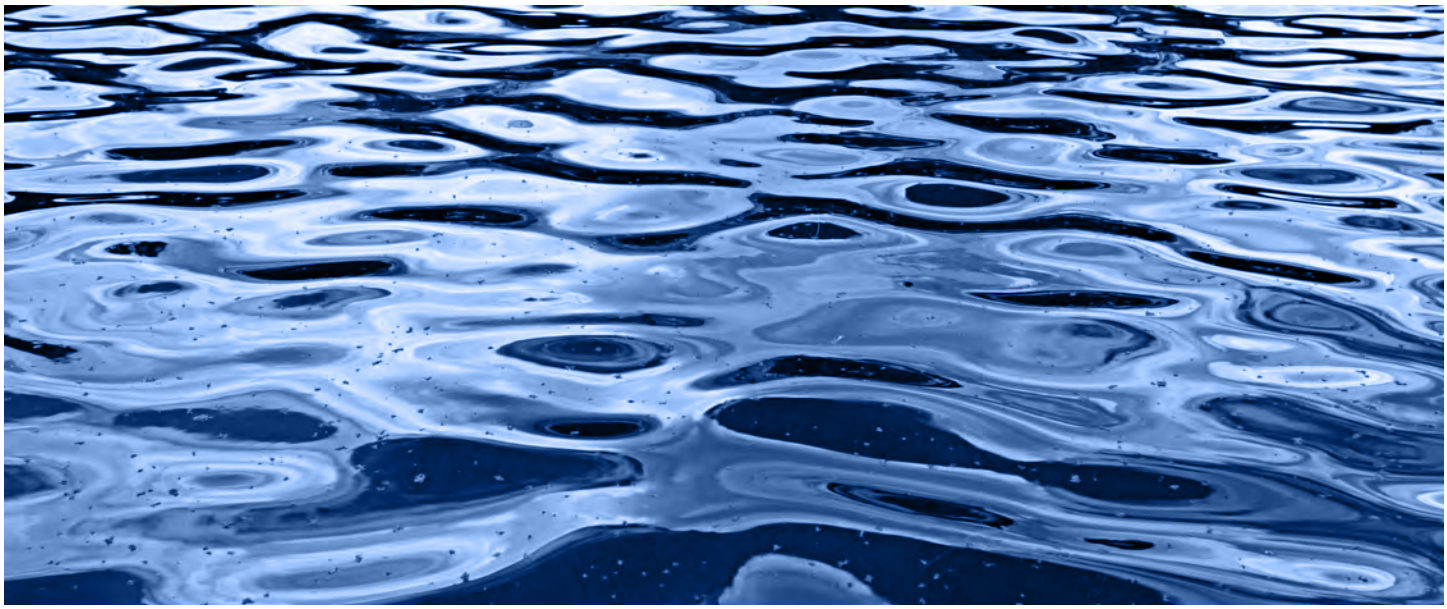
CHECKLIST 2023

Modernizing Your Data Warehouse and Analytics: Key Pillars of a Data Lakehouse

By Fern Halper, Ph.D.

databricks
+ableau from Salesforce
tdwi TRANSFORMING DATA WITH INTELLIGENCE™



Modernizing Your Data Warehouse and Analytics: Key Pillars of a Data Lakehouse

By Fern Halper, Ph.D.

To stay competitive and thrive in a constantly changing environment, organizations are collecting and analyzing larger amounts of diverse data. As part of this process, they often realize that their current data management environment is not sufficient for their needs. In a recent TDWI survey, for instance, the top technical challenge cited by respondents preventing them from moving forward with analytics was their data infrastructure.[1] Many survey respondents realize that their data warehouse will not support the move to more advanced analytics such as machine learning that may utilize high volumes of unstructured or semistructured data.

The five pillars of a data lakehouse:

1. Provides a unified architecture for diverse data and analytics
2. Supports a unified governance layer
3. Supports all analytics
4. Supports open source standards
5. Optimized for modern data implementations and use cases

This has, of course, prompted organizations to move to the data lake. For years, TDWI research has tracked the modernization and evolution of data warehouse architectures, as well as the emergence of the data lake design pattern for organizing massive volumes of analytics data. The core function of the data warehouse is to drive queries, reporting, dashboarding, and other decision-support analytics

[1] Unpublished 2023 TDWI survey.

on data that is structured, cleansed, and curated. The data lake, on the other hand, is meant to ingest raw data that could be used by data scientists and others who want to support and develop more advanced analytics. Initially built on the Apache Hadoop open source platform, many early data lake projects failed due to poor data management capabilities, poor performance, and lack of governance. Many data lakes evolved into large data storage platforms that weren't nearly as performant as an organized, micro-partitioned, higher-quality data platform.

The dual data warehouse/data lake architecture is also a siloed architecture, which means that data is often copied multiple times for the data warehouse and the data lake. Aside from the errors that this could cause, the copying process involves duplicate development and maintenance efforts.

In recent years, a new paradigm has emerged to address the deficiencies of both the data warehouse and the data lake—*the data lakehouse*—a combination of a data lake and a data warehouse that provides warehouse data structures and data management functions on low-cost platforms, such as cloud object stores. The lakehouse grew out of incremental technical advancements of columnar storage types, data access patterns, cloud adoption, highly parallel computing orchestration, and increased indexing capabilities. This has led to a set of previously unavailable platform capabilities. These new platforms have blurred the distinction between the traditional data warehouse and data lake. They support and manage large volumes of diverse data along with SQL, BI, AI, machine learning, and other advanced analytics on one common platform—typically in the cloud.

This TDWI Checklist Report examines what sets the data lakehouse apart from the data warehouse and the data lake and the five key pillars of the modern cloud data lakehouse. These pillars can serve as the requirements for evaluating lakehouse platforms.

1 Provides a unified architecture for diverse data and analytics

If your organization is considering a lakehouse, it is important to understand what it is and how it differs from data warehouses and data lakes to determine if it is the right fit for your organization.

- Unlike data warehouses that deal primarily with structured data in a predefined schema, the data lakehouse can store massive amounts of both structured and unstructured data such as streams, text, video, image, or audio in its original format.

- A traditional data warehouse supports primarily basic analytics such as reports and dashboards. A data lakehouse can support more advanced, compute-intensive, and iterative analytics such as machine learning and other forms of AI that operate on diverse data types.

- A data lake can store large amounts of raw semistructured and unstructured data but has no real structure and does not support querying. Data lakehouses support efficient querying and analysis. Additionally, whereas data lakes are not ACID compliant (atomicity, consistency, isolation, durability—a set of properties to ensure database transactions are processed reliably), most data lakehouses are. ACID compliance ensures that every transaction will either be

fully completed or fully rolled back for future attempts, without requiring the creation of new pipeline processes.

- Traditional data lake platforms were not designed to separate computing resources from storage. When these aren't decoupled, when one grows, the other must also grow. This can be costly, especially for analysis. Data lakehouses separate the two, allowing each to scale independently.

Data lakehouses typically operate in the cloud, which provides scalability and low-cost storage. Additionally, modern data management tools and services are often part of the data lakehouse to perform tasks such as inferring database schemas, automating pipelines, and finding data quality problems.

In most cases, the architecture implements data warehouse functionality over an open data lake file format.[2] The architecture may consist of layers. Different data types (structured, semistructured, or unstructured) are ingested into a storage layer that may be an open file format such as Parquet. Above this layer is a metadata layer that includes data management features such as file format descriptions. There may be a processing layer and a semantic layer that includes a data catalog.

Above this is the API layer that can support high-performance SQL or data frames (a two-dimensional data structure where data is stored in a tabular format, similar to a spreadsheet or even a database table) used in many data science tools. This feeds into the consumption layer where the analytics tools are located. Lakehouse architectures may vary, but the point is that the modern lakehouse provides a structured format, supports updates, inserts, and merges, and can be queried and used for analytics.

[2] See Michael Armbrust, Ali Ghodsi, Reynold Xin, and Matei Zaharia, "Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics," Conference of Innovative Data Systems Research (CIDR) 2021.

2 Supports a unified governance layer

A key pillar of the data lakehouse must be a unified governance layer. Data governance includes the policies and procedures implemented by organizations to ensure adherence to rules related to data and to instill confidence in the quality of their data. In TDWI research, data governance is a top priority and a top challenge in cloud environments.

Over the past few years, vendors have been ramping up solutions in this space. Data lakehouse solutions should support the following kinds of tools and services as part of a unified governance layer.

- **Data quality tools** ensure data's quality and trustworthiness for use in reports, queries, analytics, and other applications. Modern data quality management tools automate the profiling, parsing, standardization, matching, merging, correction, cleansing, and enhancement of data for delivery into enterprise data warehouses and other downstream repositories. They support data quality rule creation and revisions.

- **Data catalogs** play a key role in data management by making it easier for users to search for and find data from diverse sources. They supply semantic layers with descriptions about data. Some modern data catalogs include features for automated data cleansing, classifying sensitive data, and certification of data sets by their owners. Other solutions keep

track of changes to data schema or structure. Some catalogs capture an audit log of actions performed against the data which can help meet audit requirements.

- **Data lineage** solutions help determine how specific data items originated, how extensively they have been transformed and cleansed, and how widely they have been distributed. Some data catalogs provide data lineage.

- **Compliance** is a key part of data governance to meet organizational audit and compliance guidelines. This includes controls for data reuse, cross-jurisdictional data transfer, and data deletion in the cloud environment.[3] Data protection laws, including the General Data Protection Regulation (GDPR) in the European Union and the California Consumer Protection Act (CCPA), have introduced requirements for collecting and safeguarding personally identifiable information (PII). These laws have also made managing data use and data sharing more challenging, and they raise sovereignty concerns regarding the location of customer data because consumer rights vary by region.

- **Data security and protection** also need to be part of the data lakehouse platform. This includes fine-grained access controls as well as schema enforcement and evolution. It also includes the ability to protect sensitive data at rest, in motion, and in use. Data security and protection may involve using encryption or masking techniques.

3 Supports all analytics

Another key pillar of the data lakehouse is its ability to support *all* analytics including query, visual data exploration, and reporting as well as advanced analytics such as AI or machine learning.

TDWI research indicates that demand for modern analytics is strong. For example, in a recent TDWI survey, modern self-service BI, which might include features such as automatically surfacing insights, ranked as the top priority for analytics.[4] Machine learning wasn't far behind in terms of priority—over half of the respondents to that same survey stated that demand for ML had increased in the past year.

To facilitate a broad range of analytics, your data lakehouse should support:

- **Unified user access to all data.** The data lakehouse eliminates the silos between the data warehouse and the data lake. This means users don't need to jump from the data warehouse to the data lake for analytics needs that involve large amounts of data or different data types (see next point).

- **Diverse data types.** Machine learning and other advanced analytics often require high volumes of diverse data. For instance, an organization may want to marry structured data such as billing data with sentiment data available as unstructured text. That can be an issue for traditional platforms, but should be simple on a data lakehouse.

[3] For more about data management and compliance risks, see the Deloitte Cloud Computing Risk Intelligence Map.

[4] Unpublished 2023 TDWI survey.

- **Third-party data.** Modern analytics often makes use of enriched data sets that might include demographic/firmographic data, weather data, or industry-specific data. This data can help improve model accuracy. Some data lakehouse providers offer data marketplaces, where buyers and sellers can exchange data. Look for this feature.

- **Tools for multiple types of users.** Modern analytics often involves open source as well as commercial products, so the lakehouse must support both. If the data lakehouse is a unified platform to service a range of personas, those personas must be able to utilize the tools that they need to get their jobs done. That means that, for some organizations, it might be enough to support the data frame paradigm (defined earlier) that might be used in R or Python. For others, that might mean supporting APIs that work with other open source tools (such as Tensor Flow) or commercial tools.

- **Models in production.** To support advanced analytics such as machine learning, the data lakehouse must be able to process the data to score models once they are in production.

4 Supports open source standards

Vendor lock-in occurs when an organization becomes dependent on a particular vendor's products or services, making it difficult or expensive to switch to another vendor's products. Proprietary software can be a source of vendor lock-in because the software is often tightly controlled by the vendor. Open source standards are developed and maintained by a community of contributors. This promotes innovation and supports interoperability, meaning that organizations can move between implementations. Organizations adopt open source solutions because they can be free to use. They also like that it is in some ways future-proof as long as the community continues to be engaged.

Many organizations are particularly concerned about vendor lock-in on cloud data platforms—they do not want their data in a proprietary format that can't be owned or used by another processing engine. In addition to the data, many also want to potentially be able to take their code somewhere else.

Although open standards are not a strict requirement for a data lakehouse, a key pillar of data lakehouses on the market today is the use of open standards. Depending on your organization's values, you may wish to consider whether your platform should support open source standards, where data is in an open table format and you can bring the engine of your choice for different use cases.

Delta Lake, Apache Hudi, and Apache Iceberg are three popular open source data table formats designed to make it easier to manage large amounts of data in the data lakehouse. The data itself may be stored in a columnar database such as Parquet. These three projects provide various features, such as ACID transactions, data versioning, data quality, data governance, and time-travel queries (a type of database query that provides access to data as it existed at a previous point in time), to make building data lakehouses easier. These open source technologies can work with various big data frameworks and integrate into existing data pipelines.

5 Optimized for modern data implementations and use cases

The data lakehouse emerged from the need for a unified set of data to support all kinds of data and analytics. It is a modern platform that must support modern implementations and use cases. This is a pillar of the data lakehouse and includes support for

- **Multicloud environments.** In TDWI research, we see organizations utilizing more than one cloud platform (the average is two). That means that if an organization is storing its data on multiple clouds, then the cloud data lakehouse needs to unify the data from these environments into a single lakehouse platform that is abstracted from the end user. It also means that any governance solution must support a multicloud environment and provide visibility across multiple clouds, including unified metadata.

- **Batch *and* streaming data.** At TDWI, we're seeing more organizations use streaming and real-time data for many use cases. The data lakehouse must support the ingestion and processing for this real-time data.

- **Scalable SQL.** The modern lakehouse supports schema enforcement and enables SQL to run on open table formats. That SQL must also be sufficiently scalable to perform tasks such as handling different data types, supporting complex joins, optimizing distributed joins, and supporting many concurrent users.

- **Data sharing and collaboration.** Enterprises are evolving their data strategies to support reuse, sharing, and collaboration practices to derive more value from data. Data sharing has gained particular significance for organizations as they engage with customers, suppliers, and partners. Data sharing promotes transparency, efficiency, and innovation. Data lakehouses should enable data sharing and collaboration for use cases such as supply chains where organizations may want to share insights with their customers and partners (e.g., inventory level forecasts) to drive efficiencies.

The data lakehouse should also support use cases for data products that might be used to monetize data (e.g., via customer dashboards or applications). It should enable data engineers, data scientists, and ML engineers to share data (including streams), which is often a challenge, especially across disparate tools, security models, and data copies.

The data catalog was described earlier in this report terms of data governance. The catalog is also important for data understanding and sharing to centrally manage shared assets within and across organizations. The lakehouse must enable users to search for data in the catalog being shared.

Concluding thoughts

When rushing to migrate to cloud platforms (including cloud data warehouses and data lakes) to support more diverse data and more advanced analytics, an enterprise often creates its own set of problems with disjointed and duplicated data silos. A new paradigm—the data lakehouse—combines the capabilities of the data warehouse and the data lake. The data lakehouse offers unified management, schema enforcement, and optimization in a comprehensive metadata-driven platform.

Look for data lakehouse platforms that support the key pillars of this kind of architecture: unified governance, support for all data types and analytics as well as open standards, and support for modern data implementations including multicloud.

About our sponsors

databricks

databricks.com

Databricks is the data and AI company. More than 9,000 organizations worldwide—including Comcast, Condé Nast, and over 50% 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. To learn more, follow Databricks on Twitter, LinkedIn and Facebook.

tableau from Salesforce

tableau.com

Tableau helps people transform data into actionable insights. Explore with limitless visual analytics. Build dashboards and perform ad hoc analyses in just a few clicks. Share your work with anyone and make an impact on your business. From global enterprises to early-stage start-ups and small businesses, people everywhere use Tableau to see and understand their data.

About the author

Fern Halper, Ph.D., is vice president and senior director of TDWI Research for advanced analytics. She is well known in the analytics community, having been published hundreds of times on data mining and information technology over the past 20 years. Halper is also coauthor of several Dummies books on cloud computing and big data. She focuses on advanced analytics, including predictive analytics, machine learning, AI, cognitive computing, and big data analytics approaches. She has been a partner at industry analyst firm Hurwitz & Associates and a lead data analyst for Bell Labs. She has taught at both Colgate University and Bentley University. Her Ph.D. is from Texas A&M University.

You can reach her by email (fhalper@tdwi.org), on Twitter (twitter.com/fhalper), and on LinkedIn (linkedin.com/in/fbhalper).

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