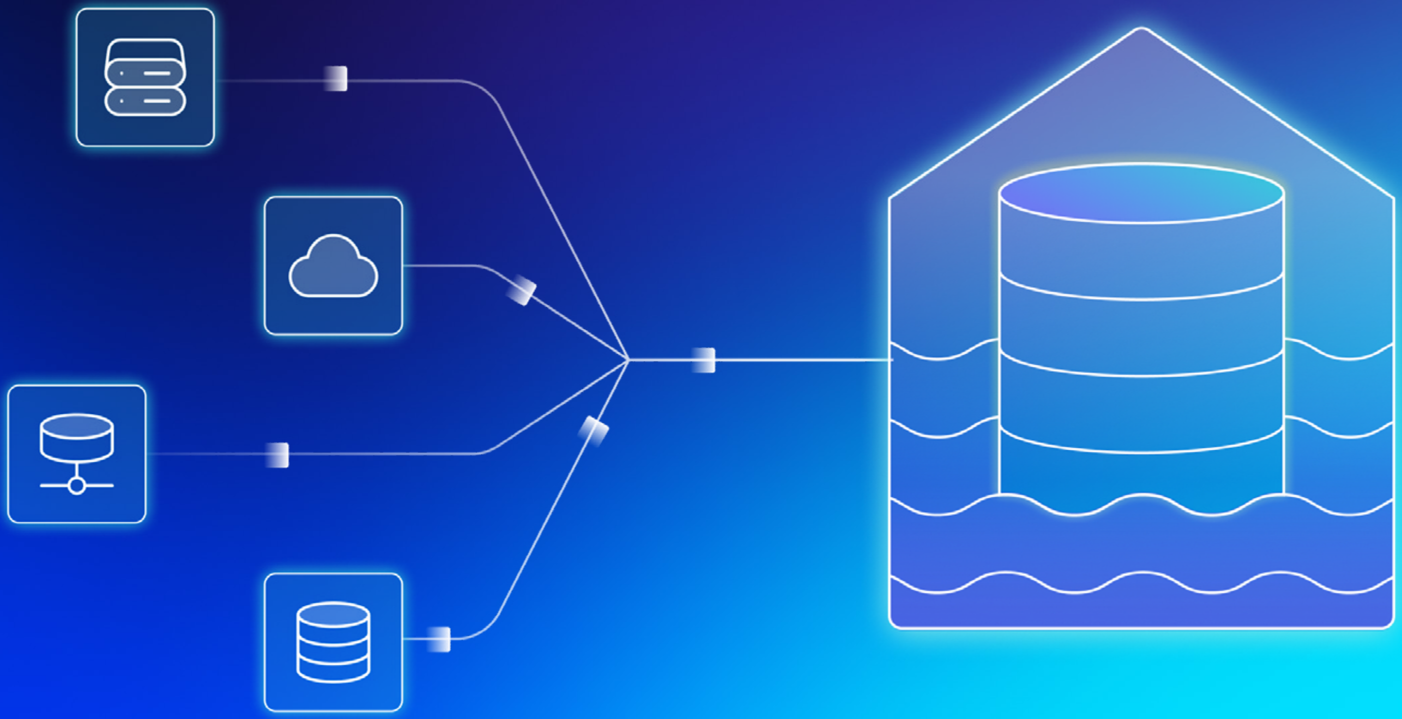
It’s Time to Consider a Hybrid Lakehouse Strategy

Author: **David Loshin,** President of Knowledge Integrity, Inc.



accelerated by intel.

dremio

Introduction

For many organizations, cloud computing has revolutionized analytics by lowering the cost barrier to entry for big data computing, both in terms of massive data volumes and high-performance. The benefits of the cloud's flexibility and scalability has motivated organizational stakeholders to adopt cloud computing for their enterprise application and data management strategies. Now that organizations have gained experience both migrating existing applications and developing new ones in the cloud, many are recognizing that cloud adoption doesn't always deliver the anticipated benefits, and that a cloud-only strategy isn't always the optimal choice.

Many organizations find that actual costs for cloud storage, transfer, and egress quickly exceed original estimates, catching stakeholders by surprise. Additionally, as data becomes distributed across multiple cloud providers and various SaaS and PaaS environments, accessing this data grows increasingly complex. In some cases, data latency across these distributed, multicloud landscapes can significantly impact application performance. To address these performance challenges and meet business data needs, some organizations are considering leaving some data on-premises and implementing a hybrid strategy.

The bottom line? Even if a company is not constrained by the limitations of the cloud, migration projects require planning and may take time. Since it is likely that organizations will need to operate in a hybrid scenario for 5-10 years as they consider their long-term cloud adoption strategy and tactics, companies should become more familiar with alternative platform strategies to meet their current and their future needs.

This paper reviews the evolution of the enterprise analytical data landscape, motivations for migrating to the cloud, the cloud's impact on the evolution of data analytics architecture paradigms, and considers how limitations of cloud computing affect meeting organizational expectations for performance and cost management. The paper examines factors that may inform data platform architecture decisions and suggests that a hybrid data lakehouse strategy composing both on-premises and cloud data management may help to inform the design and implementation of an enterprise data lakehouse to optimally modernize Hadoop data lakes, empower the creation of data products, prepare data for Artificial Intelligence- and Machine Learning-based applications, and address overall business requirements.

Dremio's Hybrid Lakehouse for the Business supports 3rd Gen Intel Xeon Scalable processors and are at the core of strong, capable lakehouse platforms—on-premises and in the cloud— for the data-fueled enterprise.

Key features and capabilities include the following:

- Infused with Intel® Crypto Acceleration, enhancing data protection and privacy by increasing the performance of encryption-intensive workloads, while reducing the performance impact of pervasive encryption.
- Built-in AI acceleration, end-to-end data science tools, and an ecosystem of smart solutions.
- Engineered for the demands of cloud workloads and to fuel a wide range of XaaS environments.
- Fueled by Intel® Software Guard Extensions (Intel® SGX), which protects data and application code while in use from the edge to the data center and multi-tenant public cloud.
- Built-in workload acceleration features include Intel® Deep Learning Boost (Intel® DL Boost), Intel® Advanced Vector Extensions 512 (Intel® AVX-512), and Intel® Speed Select technology (Intel® SST).

Platform Evolution: Data Warehousing, Cloud Migration, and the Concept of the Data Lake

Analytical environments have evolved from the on-premises data warehouse, which has been the predominant platform for reporting and analytics over the past few decades. Enhancing analytical application performance was the primary motivation behind the data warehouse paradigm. By copying data from source systems and moving it to a database platform optimized for analytics, the data warehouse aimed to separate operational and analytical workloads, preventing them from competing for resources.

But there are disadvantages to the data warehouse approach. First, data warehouse models are defined early in their development, imposing limits on the data sources that are loaded into the warehouse, and complicating the ability to add new sources that may benefit the downstream users. Second, downstream, data users are constrained to only being able to use the data sets and corresponding data elements that the modelers chose to make available for analysis. Third, business data users requiring the support of data and technology teams to facilitate the development of applications are often faced with a development bottleneck due to limited IT resources.

The development of the open-source Hadoop environment, featuring its distributed file system (HDFS), significantly lowered the barrier to entry for organizations adopting high-performance computing. This innovation enabled the storage of large data volumes on a distributed computing and storage platform, paving the way for the concept of the data lake—a repository for storing vast data sets in their raw form. New data sets can be easily added to the data lake and accessed through defined structured interfaces, such as Hive external tables, making them readily available for downstream use.

Some self-service techniques enhance productivity by reducing some dependence on IT support, empowering users to manage data more independently. The data lake has the ability to grant users access to raw data before it is processed and filtered for inclusion in a data warehouse, offering greater flexibility for developing reports and analytics. It also promotes data reuse by enabling multiple users to access and work with the same data sets simultaneously.

Cloud computing was the first platform to have amplified the advantages of data lakes. With virtually unlimited storage and the ability to more easily decouple storage from computing resources, cloud-based data lakes provide organizations the first step in being able to have greater flexibility and efficiency in their data lake environments. This ability to decouple compute and storage was soon adopted by on-premises storage vendors, further extending the benefits of the data lake model.

Data Lake to Data Lakehouse

Data lakes provide flexibility but also have some notable drawbacks. To remain useful, data lakes require constant oversight and governance, such as accurate documentation and cataloging of the data resources and monitoring their quality and consistency. Data Lakes are often challenging to work with for business users. Accessing data as individual files breaks the more friendly “table” model that databases and data warehouses provide. And because the data sets are stored in their original form, data users are required to apply their own data preparation techniques to make those resources usable. Data analysts must provide their own end-user tools to access, query, and analyze the data. Importantly, data lakes are not designed to provide consistent data structures and the ACID (atomicity, consistency, isolation, and durability) support that can lead to inconsistent and poorly performing analyses.

A data lakehouse, however, is an architectural paradigm intended to provide the scalability and flexibility of a data lake while empowering data users with the capabilities, ease-of-use and governance of a data warehouse. In a data lakehouse, data sets are not accessed as individual files. Instead, they are presented as tables with the same guarantees and transactional consistency features (such as ACID compliance) that one would find in a data warehouse but stored in a data lake. The lakehouse paradigm integrates a catalog of its data resources and provides an access layer simplifying the ability to find, understand, and use the data in the lakehouse.

Facing facts: Considering a Hybrid Strategy, Limitations of the Cloud

There are sound arguments for migrating corporate data and applications to a cloud-based data lakehouse. Some of the motivating factors are financial, such as the desire to reduce capital acquisition budgets and maintenance costs to lower platform total cost of operations. Other factors are focused on aspects of system performance, such as improved scalability and improved operational performance. Still other factors are associated with aspects of enterprise platform management such as consolidating data operations to a single platform and improving data security and protection.

The cloud's virtually unlimited storage has made the cloud an attractive choice for the data lakehouse. In turn, many organizations are assiduously migrating their data environments and applications to the cloud, with the hope that the cloud's elastic computing, effectively unlimited storage, flexibility enabled through the data lakehouse paradigm, and breadth of new technology utilities will lower barriers to digital transformation while embracing more sophisticated types of reporting and analytics.

In some cases, however, those charged with executing a cloud modernization effort might not fully understand some inherent challenges and drawbacks of adopting a cloud-only strategy. Organizations that have opted to migrate their entire data footprint to the cloud may find reasons to regret that decision for a number of reasons including, but not limited to:

- **Spiraling costs.** In some scenarios, system architects and managers may not completely understand the total cost of operations for storing, managing, accessing, and using data in the cloud. Although cloud service providers charge little (or sometimes no) fees for moving data into cloud storage, there are different fee structures for using that data, including varying fees for different classes of storage, charges per access via API, and bulk data egress charges. The complexity of cloud services pricing models coupled with ungoverned cloud service management hides these types of expenses, in some cases allowing cloud storage costs to spiral out of control.
- **The risk of vendor lock-in.** There is a perception of vendor neutrality associated with decisions about cloud data platforms. Yet the choices made associated with adopting cloud data storage, management, and component architectures still pose the risk of creating a dependency on one or more proprietary vendor products of services. Vendor lock-in can impact system and application scalability, constrain innovation, lead to increased costs, and ultimately can affect the organization's ability to meet business needs.
- **Interoperability difficulties in multicloud data landscapes.** In environments that have allowed the organic transition to a variety of instances

of cloud computing and storage services hosted by multiple cloud service providers, Software as a Service (SaaS), and Platform as a Service (PaaS) capabilities, the organization may face management, maintenance, and interoperability challenges related to blending the use of these different multicloud environments.

- **Gaps in data protection.** Although the cloud providers have been diligent in firming up their data security and protection capabilities, improper classification and neglectful governance opens the door for data loss or security breaches. Despite the efforts of the cloud service providers, there are still risks associated with data security and protection of sensitive personal and corporate information.

Lack of governance. Organizations may be subject to a variety of both internally-imposed data constraints (such as data quality, data consistency, and data currency expectations) and externally-imposed data constraints (such as national records management laws, data sovereignty limitations, mandated data privacy laws, and data quality laws).

- **Access Latency impacting performance.** Accessing data from across a multicloud distributed data landscape is subject to network bandwidth constraints, and as the data volumes grow, delays due to the data latency becomes unacceptable for meeting performance service levels.

Framing Organizational Performance Requirements

These challenges and limitations imply that despite the enthusiasm for migrating data to a cloud-based data lakehouse, there are many valid reasons for keeping corporate data on-premises. To ensure that your corporate lakehouse implementation supports the organization's ability to meet data user needs, consider establishing data user performance and usability criteria that should guide the data architecture and the data landscape, such as:

- **Cost management.** Organizations need effective overall visibility into their enterprise data landscape infrastructure to set criteria for, oversee, and manage the costs associated with data storage and computing needs of growing communities of data analysts and data scientists. This implies that there is value in developing a data storage and management cost model that takes into account the varying costs of types of cloud storage, data transfer, and data egress that can be configured based on the volumes of data stored and accessed by enterprise applications. Using this cost model can alleviate issues caused by being overwhelmed by opaque pricing structures, misunderstanding how the separation of data from compute allows for data storage volumes to remain high even through computing instances have been deleted, or the need to control the costs of the high computational demands implied by the rush to deploy AI capabilities.[1]

[1] Grant Gross, "Rising Cloud Costs Leave CIOs Seeking Ways to Cope," CIO, 08/27/2024, accessed 09/29/2024 via https://www.cio.com/article/3496509/rising-cloud-costs-leave-cios-seeking-ways-to-cope.html

- **Openness.** When possible, use open standards, formats, and technologies, since this will reduce the dependence on proprietary technologies that are tightly coupled with vendor products and services.
- **Data access flexibility.** Data users should not be forced to become experts in the syntax and processes for accessing data across a wide variety of file and storage formats. They should also not be confined to a particular proprietary format based on the selection of a vendor product or service. Establish criteria for data lakehouse flexibility to hide the particulars of accessing data sources, promote simplicity for self-service access, and avoid the risk of vendor lock-in to proprietary file and storage formats.
- **Sensitive data protection.** Protecting against the unauthorized release of sensitive information is not limited to observing data privacy laws. The scope of data sensitivity extends to ensuring against leakage of valuable information such as intellectual property, trade secrets, and financial intelligence. Therefore, ensure that there are proper controls in place for authenticating data users, authorizing data access, and specify criteria and methods for monitoring against any type of unauthorized data access.
- **Data governance.** As more organizations jump on the AI bandwagon, they forget that the quality of data used to train and fine-tune AI models can have disastrous effects on the trustworthiness of the AI application’s outputs. It may be difficult to ensure the consistency of data resources that are distributed across a multicloud landscape. Specifying data governance criteria allows one to monitor the quality and consistency of enterprise data.
- **Data access speed.** There are two facets of data latency that impact operation and analytical system performance. One is the delay or time lag between the time that a data set is created or updated and when that data is available to be used by the downstream applications. The other facet is the time it takes for available data to be accessed and delivered to the requesting application. Specify a minimum level of service for data availability and data access speed.

Considerations: Benefits of a Hybrid Lakehouse Approach

Organizations need to determine the extent to which their enterprise data architectures balance the commitment to cloud data modernization with the realities of the cloud’s limitations. Some applications will perform better when the data resources are situated close to where they are being used, suggesting that an on-premises deployment might be a better decision. Even if an organization is not subject to the cost or performance constraints associated with cloud storage services, migration projects require planning and time to execute. And if it is likely that organizations will operate in a hybrid on-premises/multicloud scenario for 5-10 years as they consider their long-term cloud adoption, an optimal approach today is to develop a strategy for a **hybrid data lakehouse.**

A hybrid data lakehouse is an architecture that encompasses data seated in both multi cloud and on-premises environments that addresses the drawbacks of a cloud-only approach while satisfying defined performance requirements, such as:

- **Cost-consciousness.** A hybrid data lakehouse allows you to choose the most appropriate environment in which to store data. Infrequently accessed data might be better suited to cold storage in the cloud, but it might be more cost-effective to store frequently accessed data on-premises to reduce repeated charges for data access/egress.
- **Implementation flexibility.** Employing open standards and formats for storing data not only frees your architecture from vendor lock-in and reliance on proprietary formats, it also simplifies data and application portability.
- **Access flexibility.** Providing a semantic layer on top of a variety of underlying data resources for simplifying data access obviates the need for a data user to know the details of any of the physical data layouts, data formats and data structures, or the details of the data transformations and pipelines for producing the data in the data lakehouse.
- **Data protection.** Choosing to store data on-premises helps to comply with legal and regulatory directives such as data sovereignty or data privacy laws.
- **Governance and oversight.** The hybrid lakehouse approach allows for centralization and management of data policies for data access control, controls for monitoring for unauthorized data access, and controls for monitoring the quality, consistency, and currency of data.
- **Reduced data latency and improved performance.** Look for hybrid lake architectures that encompass techniques to reduce or hide data latency, such as materializing and caching frequently accessed data using a columnar memory layout in alignment with a query engine designed to optimize query execution.

Apache Iceberg: An Open Standard Table Format for the Hybrid Lakehouse

It is important to develop a strategy for a hybrid lakehouse when there are valid reasons for keeping some corporate data on-premises. There are two key components to building a hybrid data lakehouse: an open table format that allows defining of tables with ACID guarantees, and an enterprise data catalog that tracks these tables and the location of their metadata so different engines and tools can operate on those tables. Open table formats provide a metadata layer around files in the data lakehouse that allows those tables to be used in the same way (with the same ACID guarantees) as tables in a data warehouse are used.

Apache Iceberg is an open table format for high-performance computing that can be implemented across different on-premises storage environments (such as NetApp StorageGRID, Pure Storage FlashBlade, VAST Data Platform or MinIO) or cloud object storage systems (such as AWS S3, Google Cloud Storage, Azure Blob Storage, or object storage provided by other cloud service providers). That means that Apache Iceberg ensures transactional consistency between applications, allowing for read isolation, concurrent write transactions, and atomicity when adding or removing data from the lakehouse environment. Iceberg’s schema evolution allows changes to a table to be tracked over time. Iceberg maintains full history with clear history lineage, which allows for rollback to prior versions of tables as well as enabling users to query the data at prior states of its evolution. Apache Iceberg’s metadata enables a unique composition of features like hidden partitioning and partition evolution, which makes partitioning tables for enhanced performance flexible and easier to use for data engineers and data analysts. Since open-table formats store their metadata as files on your data lake, large-volume data sets can be maintained using the Apache Iceberg open table format in a way that is decoupled from computing resources, which means that you can use a variety of query engines to access the same data resource.

These features that give the data lake the capabilities of a data warehouse (ACID compliance, improved governance and management as data sets change over time, efficiency in managing large-scale tables, and the ability to employ different query engines) make Apache Iceberg a natural format for the hybrid data lakehouse architecture. Apache Iceberg is quickly becoming the dominant open table format for the data lakehouse and is supported by other key open-source technologies such as Project Nessie and Apache Polaris. The enterprise data catalog allows you to manage multiple domains containing hierarchies of data resources. This enables centralized access control management at the domain level and federated data sharing.

There is no need to abandon the data lakehouse paradigm just because corporate data remains on-premises. A hybrid data lakehouse architecture built using open standards like Apache Iceberg, Nessie, and Polaris provides all the benefits of a cloud lakehouse while providing the flexibility to manage data where it makes the most sense, either in the cloud or on-premises. The hybrid lakehouse supports accessing data in different cloud environments as well as data remaining on-premises while making data access transparent to the data users in ways that do not impact operational performance. Optimized storage layouts facilitate rapid access but also employ data virtualization techniques allowing for caching of frequently accessed data. And integrated governance supports measuring compliance with defined data policies.

Summary and Next Steps

There are benefits and drawbacks to having data in the cloud and on-prem, so you should carefully weigh these considerations before any migration begins. There is a definite value proposition for retaining data on-premises when circumstances indicate that a cloud strategy purely for the sake of "moving to the cloud" will not meet the organization's needs. Therefore, take these steps to review your cloud modernization strategy and determine whether there are any requirements to maintain a hybrid data landscape:

- **Evaluate and balance costs.** Cloud costs for data storage, retrieval, and egress accumulate as the data volumes increase. The total cost of data operations will not only be a function of the data volume itself, but rather a function of the frequency and volume of data accessed. Assess the current cloud infrastructures' total costs of operation and project those costs according to the anticipated data volume growth. Set a limit on the budget available for all of the aspects of cloud storage costs. When the cost exceeds the limit, consider which data resources would be better suited to manage on-premises.
- **Solidify operational performance requirements.** Response times for critical applications should drive the specification of service level performance criteria. No matter how fast the cloud service provider's network is, it will never be wide enough to accommodate massive data transfers. If those critical applications need to run on-premises and need to access a lot of data, consider whether that data should be repatriated to your in-house data center.
- **Identify contextual business constraints.** Determine whether there are any business constraints that would prevent managing data in the cloud. Examples include laws about data sovereignty, heightened levels of data sensitivity requiring governed data protection, or data traceability requirements.
- **Simplify, simplify, simplify.** Evaluate the complexity of accessing data distributed across a multicloud landscape. If the data users need to jump through multiple hoops to access the data from across the cloud landscapes, consider whether data access can be simplified by adopting a semantic layer on top of a hybrid data lakehouse.
- **Vendor and Solution Evaluation:** Conduct a thorough assessment of vendors in the market, focusing on solutions that align with your specific requirements. Evaluate each vendor's ability to deliver a robust hybrid lakehouse solution that meets critical needs, including query performance on the data lake, openness, flexibility, self-service capabilities, and effective data management through a hybrid catalog. Prioritize vendors that demonstrate strong alignment with these key capabilities and offer solutions that can support your long-term data strategy.

Processes and practices for data management and application performance should not be penalized in those cases where data needs to remain on-premises. Evaluating the outcomes of these steps could help indicate whether they should rethink their cloud data deployment strategy and instead consider a hybrid on-premises and cloud data lakehouse. Feel confident that a hybrid data lakehouse built using an open standard like Apache Iceberg can enable your organization to support the development of a transparent semantic layer that allows on-premises data to seamlessly interoperate with data in the cloud.

About the Author

David Loshin, president of Knowledge Integrity, Inc., (**www.knowledge-integrity.com**), is a recognized thought leader and expert consultant in the area of data strategy, information risk, and information innovation. David is a prolific author regarding information management best practices as the author of numerous books and papers, 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. David is a frequent invited speaker at conferences, web seminars, and sponsored websites and channels. David is also an Associate Research Engineer at University of Maryland's Applied Research Laboratory for Intelligence and Security.
David can be reached through LinkedIn, or email via loshin@knowledge-integrity.com.

About Dremio