Enterprise Strategy Group | Getting to the bigger truth.™

ESG SHOWCASE

The Next Frontier in Analytics: Welcome to the Data Lakehouse

Date: October 2022 **Author:** Mike Leone, Senior Analyst

ABSTRACT: The increasingly significant, strategic, and essential role of data in driving organizational excellence has put intense pressure on traditional data architectures. To unlock the true, long-term value of data, organizations need to go far beyond the designs and capabilities of data warehouses and modern data lakes. Enter the data lakehouse, an architecture that delivers the best of data warehouses and data lakes.

Introduction: Unlocking and Unleashing the Full Value of Data

No one needs to be told about the value of all kinds of data—structured, unstructured, and semi-structured—to organizations across all sectors, markets, industries, and geographies. The development of more data sources, devices, applications, and services has led to an explosion in data volume, variety, and velocity, giving organizations more weapons to use in making smarter, faster decisions. Having more data at their fingertips means business stakeholders can identify opportunities, evaluate solution options, and monitor results—over and over, if necessary. Making better decisions for today and for the future is every organization's goal, with data the linchpin to organizational excellence and success.

There's just one problem: It is harder than ever to get through all that data to find the true nuggets of insight. Unless organizations can locate, analyze, and extract the right data, they won't be able to drive results, from determining which new products to introduce to selecting the right price points to attracting buyers without sacrificing profit margins. This requires a more modernized, forward-looking data architecture that can promote the performance, scale, simplicity, and quality requirements of the business. While data warehouses and data lakes certainly remain as valuable tools in helping organizations mine data for better analytics, the dramatic changes in data volume, variety, and velocity—not to mention all the "dark data" that many organizations don't even know exists—have put pressure on organizations to use new data architectures designed specifically for improved analytics of all that data.

Challenges of Traditional Data Architectures

Let's start with data warehouses, which have been around for decades and still are considered essential elements in a data-driven business strategy. Organizations have long relied on them as a staple of analytics and business intelligence, but several challenges have presented themselves, including:

- Inefficiency and high costs of traditional data warehouses in terms of continuously growing data volumes. ESG research shows the top challenge organizations face relating to their data management and analytics initiatives is the cost of data services, data tools, and/or supporting infrastructure, including support costs.[1]

[1] Source: ESG Complete Survey Results, *The State of DataOps*, August 2022.

- Inability to handle unstructured data such as audio, video, text documents, and social media posts.
- Insufficient data processing in support of emerging technologies such as machine learning.
- Inability to support all data types (i.e., unstructured data) with high volume, variety, and velocity.

As a result, data lakes were introduced as an updated and more purpose-built alternative to data warehouses. Still, data lakes have their own shortcomings. Designed as a repository to store huge amounts of raw data in its native formats, data lakes don't require data transformation prior to populating data in the lake, meaning there isn't any schema for data to fit into. Instead, data lakes are often used as a landing zone, making them accessible for data science and exploration purposes. But while suitable for storing massive, non-curated data, data lakes lack some critical features, such as:

- Lack of standard organization leads to data stagnation problems (i.e., data swamps).
- Lack of native support for transactions.
- Additional tools and techniques are required to leverage business intelligence and reporting.
- Issues related to poor data quality, reliability, and integrity are common, as are issues with data security and governance. In fact, ESG research notes that data quality and data security/governance are considered among the greatest data lifecycle challenges faced by organizations, each cited as problems by 46% of organizations.[2]

The net result is that data lakes, while a promising technology compared to legacy data warehouses, have not borne full fruit, in many cases resulting in unacceptable return on investment. Because both data lakes and data warehouses have shortcomings, many organizations have decided to use both together (e.g., one big data lake and multiple, purpose-built data warehouses). This often results in increased complexity and costs, as data should be kept consistent between the two systems.

Things to Consider in a Modern Data Architecture

Organizations want a solution that enables them to curate all data based on specific use cases and ensures high quality and trust in the data. This solution must allow data teams to structure the data according to quality levels and define roles and responsibilities. According to ESG research, organizations most commonly cited reliability, performance, and security/governance as the most important capabilities or attributes of technologies used to support their data initiatives.

Another key consideration for modern data architectures is openness in the data ecosystem. Specifically, a modern data architecture should encourage the use of open interfaces and formats, which are critical to enabling interoperability and preventing dependency on any single vendor. ESG research notes that 95% of organizations leverage open source technology for data analytics purposes. The need to support open source technology directly ties to the need to support diverse data ecosystems, too. ESG research says that 79% of organizations work with at least five technology vendors to support their data initiatives. This puts pressure on organizations to ensure that disparate vendor tools can properly access data in a trusted, reliable, and consistent manner.

In today's IT environment, which is built around agility, flexibility, and scalability, organizations need to have a modern data architecture that can be deployed either in the cloud or on-premises, perhaps in a co-location environment. Many

[2] Source: ESG Research Report, *Cloud Analytics Trends*, March 2022. All research references in this white paper are taken from this research report unless otherwise noted.

organizations don't want to move all their workloads to the cloud for a variety of reasons, such as cost, perceived security risks, and more. Also, some organizations might run into data sovereignty issues that limit or even prevent moving some data exclusively to a cloud platform.

Finally—and perhaps most importantly—modern data architectures must promote simplicity. Keep in mind that there is a substantial and growing shortage of data professionals, such as data scientists, data architects, database administrators, and more. With these skills shortages, organizations need help building these systems to ensure high levels of availability, reliability, performance, and cost-effectiveness, while also empowering stakeholders to utilize the data that matters to them via self-service.

Why a Data Lakehouse Makes Sense

In order to address and overcome the limitations of data warehouses and data lakes, organizations are increasingly turning to a new technology: data lakehouses. This emerging form of data structure is an open architecture, combining the best of both data warehouses and data lakes. Notably, data lakehouses implement similar data structures and data management features normally seen in a data warehouse but do so directly on top of low-cost storage systems in an open format. Key features organizations should look for in a data lakehouse include:

- Transaction support: In an enterprise lakehouse, many data pipelines will often be reading and writing data concurrently. Support for ACID transactions ensures consistency, as multiple parties concurrently read or write data, typically using SQL.

- Schema enforcement and governance: The lakehouse should have a way to support schema enforcement and evolution, supporting data warehouse schema architectures such as star/snowflake-schemas. The system should be able to reason about data integrity, and it should have robust governance and auditing mechanisms.

- Business intelligence support: Lakehouses enable the use of BI tools directly on the source data. This reduces staleness, improves recency, reduces latency, and lowers the cost of having to operationalize two copies of the data in both a data lake and a warehouse.

- Storage decoupled from compute: This means storage and compute use separate clusters; thus, these systems are able to scale to many more concurrent users and larger data sizes. Some modern data warehouses also have this property.

- Openness: The storage formats they use are open and standardized, such as Parquet, and they provide an API so a variety of tools and engines, including machine learning and Python/R libraries, can efficiently access the data directly.

- Support for diverse data types ranging from unstructured to structured data: The lakehouse can be used to store, refine, analyze, and access data types needed for many new data applications, including images, video, audio, semi-structured data, and text.

- Support for diverse workloads: These include data science, machine learning, SQL, and analytics. Multiple tools might be needed to support all these workloads, but they all rely on the same data repository.

- End-to-end streaming: Real-time reports are the norm in many enterprises. Support for streaming eliminates the need for separate systems dedicated to serving real-time data applications.

- Security and governance: Security and access control; data governance capabilities such as auditing, retention, and lineage; and tools that support data discovery such as data catalogs and data usage metrics should be prioritized.

Dell’s Approach to Data Lakehouses

As a major IT infrastructure and architecture partner for enterprises across a wide range of industries, geographies, IT frameworks, and organizational sizes, Dell Technologies has put together a data lakehouse solution to help those organizations get more out of their data. Dell Validated Designs for Analytics is a data lakehouse that enables all data formats to be used in an open environment for higher data quality, faster performance, tighter security, and improved governance.

Dell Validated Designs for Analytics forms the basis of an architecture, supporting the cohesive integration analytics, BI, real-time data applications, data science, and machine learning. It is designed using unified engineering and development techniques to support data-driven use cases and built upon the concept of openness to deliver high data quality, while obviating the need to continually copy and move data between environments and data sources. This improves cost efficiency and data quality, while simplifying data management. The validated design facilitates self-service that is essential for on-demand data workloads for analytics and business intelligence. It includes support for a wide range of hardware and software platforms, including Spark, Kafka Delta Lake, Robin Cloud Native Platform, and Dell’s own PowerEdge servers, PowerSwitch networking, and PowerScale scale-out network attached storage.

The Bigger Truth

Although both data warehouses and data lakes have played—and, in some cases, continue to play—an important role in helping organizations pull together data for key decisions, the relentless pace of business and the astounding growth of data has made it clear that organizations need a new option for faster and more efficient data analytics. By moving to a data lakehouse structure, organizations can achieve greater data quality, support for the full range of data types, better security, stronger governance, and improved cost efficiency compared to legacy data architectures.

Dell Technologies Validated Designs for Analytics is a powerful, efficient data lakehouse option that organizations looking to take their data analytics to a new level should consider. The Dell solution is particularly viable for organizations that want to experience the many benefits of running essential analytics workloads in the cloud, but also want to retain key workloads in an on-premises environment as business conditions warrant.

All product names, logos, brands, and trademarks are the property of their respective owners. Information contained in this publication has been obtained by sources TechTarget, Inc. considers to be reliable but is not warranted by TechTarget, Inc. This publication may contain opinions of TechTarget, Inc., which are subject to change. This publication may include forecasts, projections, and other predictive statements that represent TechTarget, Inc.’s assumptions and expectations in light of currently available information. These forecasts are based on industry trends and involve variables and uncertainties. Consequently, TechTarget, Inc. makes no warranty as to the accuracy of specific forecasts, projections or predictive statements contained herein.

a division of TechTarget

Enterprise Strategy Group is an integrated technology analysis, research, and strategy firm that provides market intelligence, actionable insight, and go-to-market content services to the global IT community.

www.esg-global.com

contact@esg-global.com

508.482.0188