The Future of Analytics:

Leveraging Data Lakes and Data Warehouses

By Joe McKendrick

Both data warehouses and data lakes offer robust options for ensuring that data is well-managed and prepped for today's analytics requirements. However, the two environments have distinctly different roles, and data managers need to understand how to leverage the strengths of each to make the most of the data feeding into analytics systems.

Data warehouses are repositories of structured, transformed data configured for specific applications. They serve as central locations for integrated data from one or more disparate sources, said Ryan Wisnesky, co-founder of Conexus. "They store current and historical data and are used for creating trending reports such as annual and quarterly comparisons. A data warehouse is highly transformed and data is not loaded to the data warehouse until the use for it has been defined."

Typically, data warehouses "support functions that are used to create reports, understand trends, and make more tactical decisions that address day-to-day and short- to medium-term business activity," said Sri Raghavan, director of data science and advanced analytics product marketing at Teradata. Data lakes on the other hand, which typically see a lot of analytics activity, are used to investigate, discover insights, and address a more holistic set of business challenges. They usually require data and analytics functions that are not a part of the data warehouse environment, Raghavan noted.

Data lakes draw in data from all sources, whether for defined or unspecified

Both data lakes and data warehouses can be supported at the same time since it is not so much a question of which product you should use for your data but instead a matter of having purpose and intent regarding how you're going to use it.

purposes. They serve as repositories for raw, unprocessed data straight from data sources, and this data may reside in the lakes until needed at a future time. While a data warehouse may be more akin to a city water supply, a data lake "is more like a body of water in its natural state," said Wisnesky. "Data flows from the streams—the source systems—to the lake. Users have access to the lake to examine, take samples, or dive in. Data lakes retain all data. All data is loaded from source systems. No data is turned away."

The good news is that both environments can be supported at the same time. "In some cases, enterprises are operating an open data lake right alongside of the data warehouse," said Dave Mariani, co-founder and chief strategy officer at AtScale. The choice, he noted, often depends on the business case at the end of the data funnel. "It's not so much a question of which product you should use for your data; rather, it's a matter of having purpose and intent around how you're going to use your data and being able to do something with it that is the gold standard," agreed Nima Negahban, CTO and co-founder at Kinetica.

DATA-DRIVEN FUNCTIONS

It's important that enterprises understand which functions are applicable to either type of environment. Data warehouses, for one, are traditionally seen as systems of records—implying the data within these environments is well-organized, mapped, supported, and has some level of quality, said Kim Kaluba, senior manager of data management solutions at SAS. Data warehouses best support CRM-, ERP-, EDW-, and MDM-type initiatives which require stable and trusted data for decision-making functions, she said.

At the same time, Kaluba continued, data lakes "offer inexpensive options to traditional database systems." They expedite processing and function as more of a sandbox or investigational environment for data. Since data lakes are rarely managed and supported to the degree of the data warehouse, Kaluba added, "the data functions or business needs they best support include exploratory analytical functions where raw, unrefined, and large data is used to test new algorithms, identify insights, and answer questions."

The bottom line is that data warehouses "are best suited to providing high-performance, ad hoc analytics whereas data lakes are more suitable for use cases where raw data access is required," said Mariani. "Data warehouses are ideal for analytics because the data is usually cleaned and normalized. In addition, data warehouse architectures are optimized for analytics. In contrast, a data lake is best utilized as a landing zone for raw data for use in downstream applications and data warehouses. Data lakes are optimal for data science workloads, where access to granular data is needed."

A new vision of a hybrid environment, called the 'lakehouse,' provides a structured transactional layer to a data lake—allowing many of the use cases that would traditionally have required legacy data warehouses to be accomplished with a data lake alone.

The low-cost availability of storage enables enterprises to increasingly use data lakes, agreed Chris Bergh, CEO of DataKitchen. "A data lake utilizes simple storage to retain the organization's critical data. Data analysts commonly understand data lakes as a repository for raw data. Processed data can also be deposited into a data lake, allowing it to be more easily combined with other data." That's because "in its native and isolated form, accessing data is difficult." Imagine a new analytics project that needs to work with data from a series of databases containing CRM, ERP, syndicated, and sales-channel data, he noted. "Accessing data in each of these repositories is time-consuming and requires authorization and specific skills. Collocating data all in one place makes it much easier to work with. The data lake serves as the common repository for the various data sources, greatly simplifying the job of transformation."

THE FUTURE OF DATA WAREHOUSES

What is the future of the data warehouse in the emerging real-time, data-driven enterprise? How is its role changing, and how does it fit into the picture? "The need for warehouses hasn't changed much; however, now they are being accessed through the cloud in many instances," said Wisnesky. The problem, he said, is that cloud platforms can create interoperability problems by becoming a new type of silo, "especially given that ELT technologies encourage deferring schema construction."

Data warehouses, which once focused on historical data, are also taking on real-time duty. Machine learning and AI modeling allow data warehouses to operationalize those models so that the gap between an activity, such as a customer purchase, and the response, in the form of product recommendation, is a matter of seconds as opposed to days or weeks, Wisnesky said.

Data warehouses can also handle much larger datasets as they speed through rapid analysis. "Computing power and memory have advanced to the point that data warehouses can process much larger and more complicated datasets,” said Mariani.

REAL-TIME AND UNIFIED ANALYSIS

The data warehouse has adapted by moving from on-premise to the cloud, and it will continue to adapt, he noted. However, when it comes to real-time processing, data lakes present a better choice. “For real-time workloads, data warehouses are not ideal because even this new generation of data warehouse requires that data be loaded, thereby introducing latency,” Mariani said.

While the traditional data warehouse “focused on the first mile of ingesting and storing data for analysis, a modern data warehouse both ingests and stores data, and analyzes that data in real time as it is received,” said Negahban. “Modern data warehouses will deliver real-time analysis on incoming data streams, while incorporating all of an organization’s data and applying cutting-edge location intelligence and machine learning-powered predictive analytics.” Data warehouses of the future that process data in real time and unify analysis of the data in different formats—such as relational, geospatial, graph, and time series—at scale will benefit from increased accuracy and detail for customers across industries, Negahban noted.

Ultimately, the success of data warehouses going forward comes down to “semantics, semantics, semantics,” said Wisnesky. “In 2020 and beyond, the new challenge for data warehouses is how to best internalize the domain semantics in a way that provides the most value to users. For example, a data warehouse that automatically knows that two entities—say Pete and Peter—are actually the same can internalize that fact so that anyone who queries the warehouse will be made aware of the fact that there are two references to the same real-world entity. Similarly, a warehouse that automatically knows what risk is because it has internalized an ontology such as the Financial Industry Business Ontology can provide semantic query capability to users. We see lightweight knowledge graphs—as opposed to decades-old semantic web technology—as being the harbinger of semantics in 2020.”

An emerging generation of ETL visualization tools may also increase the value of data warehouses into the future. “The visualized ETL process that is essential to integrate data from multiple source systems, especially the legacy systems, is the technology having the most positive impact on enterprises’ ability to compete on data,” according to Alex Ough, senior CTO architect at Sungard Availability Services. “Machine learning models, along with the frameworks used to train the models, have improved significantly. These technologies have made it easier for less skilled individuals to train models with high accuracy. However, data engineering is still very complex and time-consuming, as many of the processes need to be done manually, especially when there are multiple sources of truth with duplicated data in legacy systems. In most cases, pre-processing data requires a deep knowledge of SQL or other programming languages to define relationships among the source data, remove duplicates, and clean mistyped data to improve data quality. Having top-notch ML algorithms and frameworks is useless if you cannot prepare quality data.”

‘Data lakes can be just as suitable as traditional data warehouse systems for analytical processes and data-driven initiatives if they are grounded in a comprehensive data strategy supported by data governance and data management processes.’

TIME FOR A DATA ‘LAKEHOUSE’?

While data lakes may be more ideally suited for fast-paced, real-time requirements, they can be more cumbersome to manage than data warehouses. For example, it’s difficult to automate the way they are used. “If an enterprise relies on the cloud, data ingestion into a cloud data lake is usually a laborious process, given the immutable nature of such systems,” said Raghavan. Data workflows need to be built and managed with a view toward smooth orchestration across multiple environments, including multi-cloud and hybrid cloud scenarios, while dealing with some environment-specific differences in governance, metadata management, and user experiences across different environments, Raghavan added.

In addition, data governance and quality is another challenge with data lakes. “Appending and modifying data is hard, jobs fail without notification, and keeping historical versions is costly,” pointed out Joel Minnick, vice president of product marketing for Databricks. In addition, it’s difficult to handle large metadata catalogs, and consistent performance “is elusive due to small file problems and complicated partitioning.” Finally, it’s a constant headache to maintain data quality, he added.

Mark Fernandes, managing partner at Sierra Ventures, said he has often seen companies build a data lake and quickly start ingesting large amounts of data. “Soon after, the lake turns into a swamp with a lack of visibility and compromised data quality,” he said. “End users don’t feel confident in the data, and analytics projects come to a halt. Data lake technology stacks based upon Hadoop can be complicated and challenging to manage, especially when you start migrating to the cloud and integrating the various tools needed for data ingestion, quality, data management, governance, preparation, etc. Lastly, many data lake projects fail because they weren’t built with a business use case in mind.”

Minnick said a new vision of a hybrid environment, called the “lakehouse,” is emerging, which “provides a structured transactional layer to a data lake to add data warehouse—like performance, reliability, quality, and scale. It allows many of the use cases that would traditionally have required legacy data warehouses to be accomplished with a data lake alone.” A lakehouse architecture also can support “unstructured data like video, audio, and text, as well as structured data that has traditionally been the domain of legacy systems," Minnick explained.

SUPPORTING DATA-DRIVEN INITIATIVES

Establishing the best environments for supporting data-driven initiatives using AI, machine learning, and IoT is a learning process, industry observers noted. "Data lakes can be just as suitable as traditional data warehouse systems for analytical processes and data-driven initiatives if they are grounded in a comprehensive data strategy supported by data governance and data management processes," said Kaluba. "This ensures that the data inside of the data lakes is reliable for organizational decisioning processes."

Enterprises are constantly looking for ways to use data across the business to build smart analytical applications that drive competitive advantage.

It may be more efficient to keep storage and compute separate as well. "Decoupling of storage and compute reduces costs and improves scalability," advised Fernandes. "Data can be stored in a cloud environment like AWS S3, and compute clusters can be spun up as needed to run workloads or queries. This type of elastic provisioning and pay-per-use are key requirements for modern data warehouses. Enterprises are also looking for integrated data governance and self-service data access to support various downstream applications, including artificial intelligence and machine learning use cases."

In addition, it's important not to rush into anything. Companies that move into AI before mastering the fundamentals—whether their data is in a data lake or a data warehouse—will end up paralyzed, Mariani cautioned. "Not only do you need to be good at data engineering and business analytics, you also need to embrace advanced automation. Lack of automation means people manually use the keyboard to process pipelines, do ETL, move data, and create downstream assets, which does not scale. All of these activities can and should be automated."

Enterprises are constantly looking for ways to use data across the business to build smart analytical applications that drive competitive advantage, said Negahban. "Traditional data warehouses do not address the need to integrate data across all aspects of the business—custom applications, IoT, or analytics dashboards. A modern data warehouse should provide a full set of APIs to embed analytics in applications. By taking an API-first approach to building data-driven applications, a modern data warehouse is able to present data at any point of user interaction, giving the business the flexibility to use the tools, apps, and platforms it prefers across departments."

EMERGING APPROACHES

An emerging discipline, DataOps, may also help bring greater order to data lake or data warehouse management. "Imagine a 50-person team managing numerous large integrated databases for a big insurance or financial services company," said Bergh. "Their customers—colleagues in a business unit—have lots of questions that drive new and updated analytics, but the data team can't keep up due to heavyweight processes, serialization of tasks, overhead, difficulty in coordination, and lack of automation. They need a way to increase collaboration and streamline the many inefficiencies of their current process without having to abandon their existing tools. DataOps automates the orchestration of data to production and the deployment of new features, both while maintaining impeccable quality. DataOps can be incredibly beneficial to both data lake and data warehouse agility in large data teams."

An emerging discipline, DataOps, may help bring greater order to data lake or data warehouse management.

When it comes to the best ways to manage data lakes to support data-driven initiatives, "first identify your key use cases, your key business sponsors, and organize your data initiative to achieve use-case success," said Fernandes. "Then create a solid foundation for your data lake by leveraging an agile and flexible DataOps approach to automate processes, standardize governance, and provide self-service access to the data. DataOps optimizes the full data cycle by controlling data sprawl and managing the entire supply chain of data from ingestion to consumption," he noted. "Finally, use augmented data management approaches and a unified platform to enable and govern data lake/store functions, such as cleansing, deduplication, data classification, and gain visibility and insights about the data lake's health and usage."

Organizations should also consider looking into the hybrid lakehouse approach, Minnick advised. "It builds on the best qualities of data warehouses and data lakes to provide a single solution for all major data workloads and supports use cases from streaming analytics to BI, data science, and AI. Historically, companies have been forced to create data silos with legacy data warehouses and data lakes, and use them separately for BI and AI use cases. This results in information inequality, high costs, and slower operations. By combining all the data onto the same open, high-performance, low-cost platform, the entire organization is able to move faster and make better decisions."

For those relying more on data lakes as core enterprise repositories, Wisnesky suggested that enterprise data managers build data models. "A data lake is a data storage device. The data stored in it still has underlying meaning, even if that meaning isn't formalized as a data warehouse schema. Automation is driven by formalization. The best-managed data lakes are actually data warehouses of data warehouses." ■