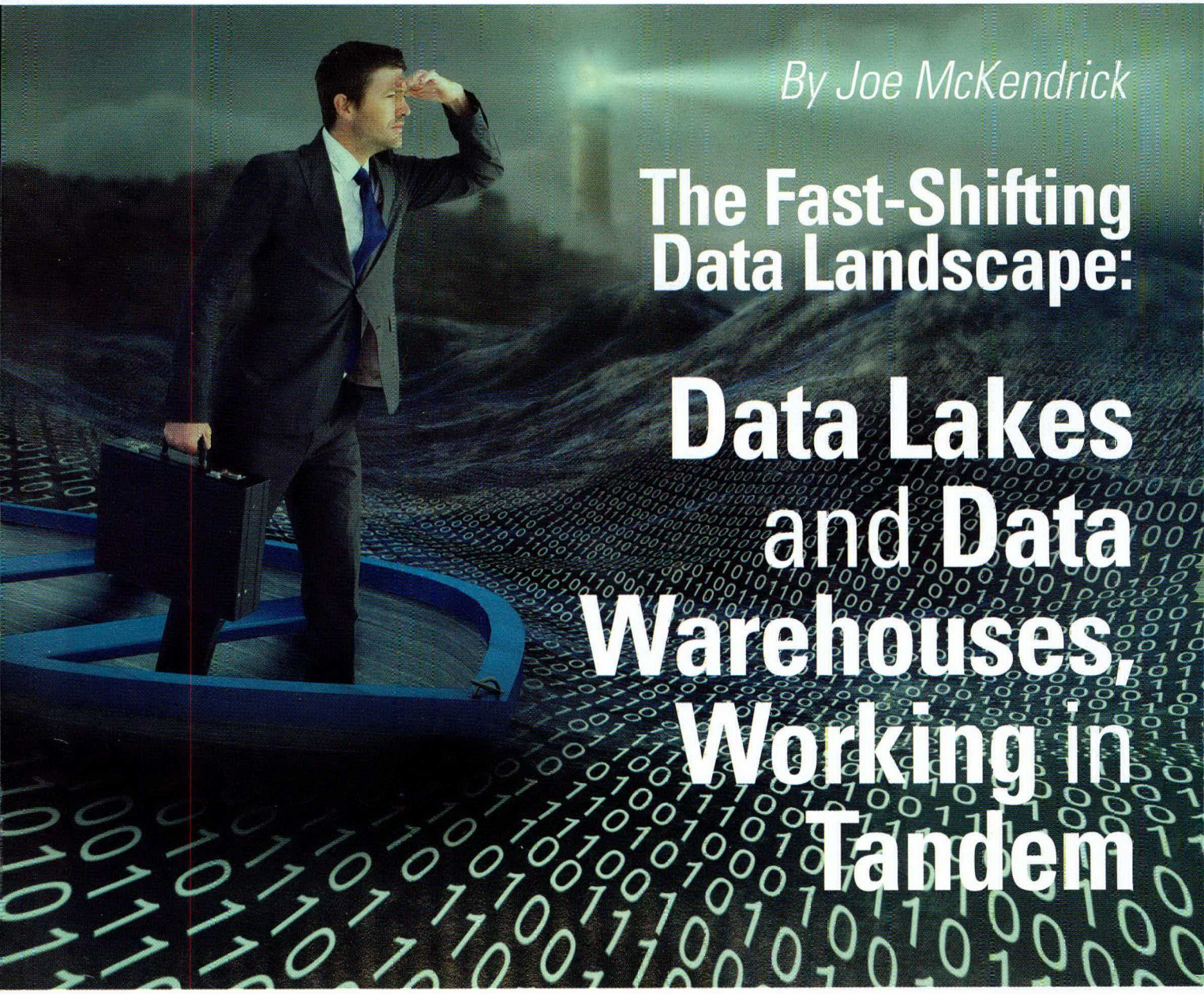

By Joe McKendrick

The Fast-Shifting Data Landscape: Data Lakes and Data Warehouses, Working in Tandem

While there have been many questions about the future of data warehouses, industry observers tend to agree that they are evolving and their role is just as crucial as ever.

At the same time, there is speculation about how data warehouses and data lakes will share the enterprise stage.

Data lakes—a place to store diverse datasets without having to build a model first—have seen adoption rise as data managers seek to develop ways to rapidly capture and store data from a multitude of sources in various formats. Overall, 38% of organizations in a recent Unisphere Research survey are employing data lakes as part of their data architecture, up from 20% in the 2016 survey. Another 15% are currently considering adoption.

At the same time, data warehouses aren't going away anytime soon because they "allow analysts to slice and dice the data to determine important insights derivable for a variety of purposes, such as driving better decisions, and laying out better organizational and pricing strategies," believes VoltDB CTO Dheeraj Remella. "Not all insights need to be utilized in a real-time manner."

Gerrit Kazmaier, EVP of HANA and analytics at SAP, agreed that data warehouses continue to play a vital role in enterprises. "Every enterprise needs a consistent and interconnected view of its data across its operations, customers, suppliers, and employees to understand its business. To make this possible, it is crucial to consider the key parts of the data value chain: the storage, quality and semantics, and usage of the data. Without a comprehensive, secure, consistent, and fast data management system, essential data loses its value and can even become a burden on a business."

Traditional data management systems are still deeply embedded within enterprises, added Kim Kaluba, senior manager of data management solutions at SAS. Such systems "will continue to be viable offerings because of their maturity and strong user community, as well as plentiful resources to support these environments."

Nonetheless, with emerging players and technologies, data warehousing solutions will have to get cheaper and more efficient, Kaluba observed.

For companies that don't have a data warehouse implementation, "the outlook appears quite different, with a 'data lake first' strategy generally taking precedence," said **Rob Small, principal consultant with Dell Technologies Consulting.** "For the time being, data warehouses will remain at the core of enterprises, serving as the source of truth for operational reporting and BI. What is changing drastically is the underlying technical and data management architecture." There is a broad, cross-industry trend to move away from the traditional vendor offerings and to migrate to different platforms and alternative data integration strategies to reduce cost, improve operational efficiency, and better support user needs around self-service analytics, Small said.

Data warehouses play a functional role in helping organizations to have data integrated in an understandable format that is efficient for queries and in capturing all business rules, added **Pedro Desouza, principal for solutions at Dell Technologies Consulting.** "The technology part is about how it is implemented. The functional part will remain, but the technology part will change dramatically with every new wave of technological innovations."

ENTER THE DATA LAKE

Are data lakes—enterprise repositories that capture and organize information before it is transformed for consumption—taking the place of data warehouses? Opinions vary across the industry as to how these two environments will mesh—or clash. According to some experts, there may be a great melding underway. "Data warehouses are in the process of becoming part of data lakes," said Desouza. "There will be no such thing as one or the other anymore."

"Data warehouses aren't necessarily being positioned against data lakes, they're working in tandem," Pete Brey, senior product marketing manager at Red Hat, concurred. However, he added, they organize data in different ways. Data lakes are an overarching technology that can accommodate all types of data—structured and unstructured. Data warehouses are specifically designed for structured data.

Not all industry leaders are convinced that data lakes are ready to storm the enterprise, however. "Data lakes are billed as the repositories for big data—data that is too large and diverse, semi-structured and unstructured, to be stored in relational databases," said Monte Zweben, CEO of Splice Machine. Proponents of data lakes positioned schema-on-read as an advantage so businesses no longer had to worry about the tedious process of defining which tables contained what data and how they are connected to each other—a process that took months and did not allow a single data warehouse query to be executed before it was complete, he said. "We believe the build-it-and-they-will-come philosophy of schema-on-read has failed."

SELECTING ONE OR THE OTHER

When is it preferable to rely on a data warehouse instead of a data lake for enterprise insights, and vice versa? It depends on the situation, as both bring advantages to the table. Data lakes, for instance, are a great source of primary data, said Jake Freivald, VP at Information Builders. "They're ideal for non-repeated processes where users already know the data's context, such as an AI model that looks at raw data to discover new patterns in it. A data warehouse has assumptions already built into its model, which would distort attempts to find new models."

Opinions** vary **across** the **industry** as to how **data lake** and **data warehouse** environments **will mesh—or clash.

Data warehouses have always been useful "to answer questions reliably, accurately, and in a predictable, timely manner," said Joe DosSantos, global head of data management strategy at Qlik. "They also provide an auditable transformation process with clear rules that allow users to get the key top-line information they need, ranging from revenue and KPIs to items related to regulatory reporting." However, he continued, many business questions are ad hoc or open to interpretation such as, "What do we think the impact of a storm will be?" and "What color will be the new black in fashion?" The ability to answer these types of questions requires experimentation with a volume and variety of data that are not often found in a data warehouse, as these systems are built with careful planning of a narrow set of business requirements, DosSantos explained. "Data warehouses are much better-suited to describe what has happened in the past than what's occurring now or in the future. Once enterprises uncover the initial information, then they can use real-time data to recognize and react to current conditions and potentially alter business decisions."

Look to data warehouses as "a source for consistent data and repeated processes, such as business reporting and dashboards," Freivald noted. "The consistency of data warehouses makes them perfect for trend analysis." In this way, with their different use cases, "data lakes and warehouses complement each other," he added. "New insights generated with raw data from a lake can inspire new dashboards and reports from the warehouse."

Data warehouses, conversely, are more suitable for storing structured, curated data in order to execute and operationalize dashboards and eventually perform predictive scoring, which requires a guarantee on the SLA and enables IT to operate the environment in a cost-effective manner, said Raghu Chakravarthi, SVP of R&D at Actian.

In addition, Chakravarthi sees data lakes as being well-suited for the fast ingestion and storage of vast amounts of uncurated data from various data sources. They can also handle a variety of data types very easily and provide connectivity while also supporting multiple language access and enabling data discovery, data preparation, model building, and model fitting, Chakravarthi added.

"The data contained in these data lakes can offer new insights and findings that are much different from the data usually collected in a data warehouse," according to Kazmaier. "When combining data from data lakes with more traditional data analytics, users can uncover business-critical insights that are not possible with traditional data warehouses or data lakes alone."

Data lakes also can act "as a staging area for data being prepared for a data warehouse," said Kazmaier. "The unique value that data lakes offer is to avoid upfront processing, categorizing, and filtering of data. Instead of having two separate tools, the ▶▶

combination of both is the ultimate solution: automatically feeding the data warehouse with data insights based on data out of the data lake and combining it with more structured data in the warehouse to get insights."

Data lakes "typically have much lower licensing costs per unit of capacity—typically either per-terabyte or per-node pricing," said Zweben. In addition, he noted, "Data lakes are typically easier to scale to very large sizes." In combination these characteristics mean that organizations need to be less selective in choosing what to retain. Data that might have been tossed out or been retained in cold storage can instead be stored in the data lake. At the same time, data warehouses will continue to be the primary area of data consolidation in organizations, he continued. "They provide more mature functionality—such as security, lineage, and transactional capabilities."

ENTER THE CLOUD

Of course, the rise of cloud computing is changing the equation for both data lakes and data warehouses—"drastically," said Small. "Public cloud providers continue to push the boundary in terms of richness of offerings and deployment flexibility. No modern data analytics strategy is complete without consideration of where public cloud plays a role."

The emergence of cloud and serverless computing "poses an interesting conundrum that will impact the way organizations store and access data," said Kaluba. It makes a data strategy even more important because it addresses "what data is stored in the cloud, the management and protection of data, how the information is going to be used, and by whom."

Increasingly, tying a data strategy to the cloud has become a necessity for many organizations. Scenarios that involve data exploration, data science, and prediction models benefit from elastic scaling and working with virtually unlimited hardware resources, said Kazmaier. Having a data lake and a data warehouse in the cloud with the ability to connect to other databases enables a simple gateway for all enterprise data. "This allows organizations to benefit from unlimited low-cost storage, highly flexible data lakes, and high-performance data warehouses for powerful, real-time business analytics."

But moving their data to the cloud is not an option for all companies, Kazmaier cautioned. As a result, a modern data warehouse solution must be able to work on top of on-premise systems while enhancing them with new data sources. "Three key areas need to be covered when looking at data governance in a cloud data warehouse. Data lineage ensures users know exactly where data is created, stored, maintained, and where it's going. The use of intelligent replication and virtualization ensures that data is refreshed in real time to create a single source of truth. Finally, roles and permissions ensure that each user has appropriate access levels."

There are complexities with data-driven cloud engagements as well, Zweben said. For example, he noted, if a company is interested in building a data infrastructure comprised of an OLTP database, an OLAP engine, and a machine learning workbench in the Amazon Web Services (AWS) cloud, this would require subscribing to the Amazon S3 storage layer, Redshift or Snowflake data warehouse solutions, Amazon RDS or Amazon DynamoDB, and one of nine machine learning engines such as Amazon SageMaker, depending on the use case, and then integrating all of them using AWS Glue, an ETL tool. This is a complex architecture that is expensive to build and operate and requires data movement across platforms that can result in poor business decisions if insights are drawn from stale data, Zweben said.

CHALLENGES

Data lakes may add a great deal of flexibility to an enterprise data strategy, but they are supported by fast-breaking technologies that require constant vigilance. "Designing data lakes that are future-proof is challenging, given the rate of new technologies coming to the market," said Desouza. "For example, GPU-accelerated databases are 10 times faster than in-memory databases that are 10 times faster than parallel columnar databases that are 10 times faster than traditional row-oriented databases. As prices drop, so does the balance among them in a data lake."

Another example of the changing technology foundation for data lakes is the evolution of massive parallel processing, exemplified by the rise of Spark versus MapReduce, Desouza pointed out. "Nowadays, it is difficult to justify the use of MapReduce; it used to dominate data processing only 5 years ago."

Governance is also a challenge, with the risk of data lakes becoming catch-all repositories for obsolete or irrelevant data. "Data lakes absorb data from many different places, and they're rarely cleaned up as people leave the company, reorganize, or abandon projects," said Freivald. "Someday, a data librarian or curator will be needed to help people find the relatively small amount of data they're looking for in the vast amount of data in the lake. The lack of governance limits the amount of trust that people can place in the information they get from data lakes."

However, said Kaluba, with a well-grounded data strategy, data lakes can serve many of the same purposes as data warehouses. "That data strategy should be supported by data governance and data management processes to ensure the data inside of the data lakes is reliable for decisioning processes needed by the organization. Otherwise, data lakes will continue to be the data dumping ground they are today."

SOLVING DIFFERENT PROBLEMS

The data warehouse and data lake each solve different business problems and impose their own unique challenges, said Anthony Roach, senior product manager of MarkLogic. "A data lake is a co-located storage solution, not a repository optimized to perform analysis and find business insights. The intention of the data lake is to store data from multiple sources in a single location in its raw form. The data warehouse is traditionally a relational system. By definition, that means the inbound data must be transformed into the warehouse model." The challenge of the data lake arises when you need to combine data from multiple sources. The data has been co-located, but in no way has it been harmonized or rationalized and curation is still required. "The data warehouse, on the other hand, offers a harmonized view of the data, but due to the transformation process, the original data has been altered or lost," Roach said.

The bottom line is that organizations shouldn't write off data warehouses—as they evolve, they are taking on new roles in digital enterprises. "The future of the data warehouse probably looks something like what Netflix has constructed," Brey pointed out. "Data is housed in a cloud object store; serialized in efficient binary formats like Parquet, ORC, or Avro; and schema and other metadata is stored in a surrogate system like the Hive metastore. This allows the use of a plurality of data processing and analytics engines."