NEWSLETTERS

- 5 MINUTE BRIEFING: INFORMATION MANAGEMENT [LATEST ISSUE]
- 5 MINUTE BRIEFING: DATA CENTER [LATEST ISSUE]
- 5 MINUTE BRIEFING: MULTIVALUE [LATEST ISSUE]
- 5 MINUTE BRIEFING: ORACLE [LATEST ISSUE]
- 5 MINUTE BRIEFING: SAP [LATEST ISSUE]
- 5 MINUTE BRIEFING: BLOCKCHAIN [LATEST ISSUE]
- 5 MINUTE BRIEFING: CLOUD [LATEST ISSUE]
- IOUG E-BRIEF: ORACLE ENTERPRISE MANAGER [LATEST ISSUE]
- IOUG E-BRIEF: CLOUD STRATEGIES [LATEST ISSUE]
- DBTA E-EDITION

The New Data Analytics: Riding on Data Lakes, Data Warehouses, and Clouds

Oct 5, 2021

By Joe McKendrick

Page 1 of 5 next >>

What types of platforms are most viable for modern data analytics requirements? These days, there are a wide variety of choices available to enterprises, including data lakes, warehouses, lakehouses, and other options—resident within an on-site data center or accessed via the cloud. The options are boundless. It's a matter of finding the best fit for the business task at hand.

Choices will depend on the nature and purpose of the data. "There's no one-size-fits-all solution," said Teresa Wingfield, director of product marketing at Actian. A flexible, multi-platform approach is needed with a wide range of data capabilities —be it a data warehouse, data lake, transactional database, IoT database, or third-party data service, she noted.

The data platform best for analytics "depends on the type of information you want to analyze," said Matthew Adams, senior cloud architect at ZL Technologies. "If you hold large amounts of easily categorizable, structured data—sales metrics, warehouse items, or customer information—data warehouses are ideal. Structured reservoirs and warehouses cater to analytics by design as they are organically segmented and neatly packaged." In contrast, "if you want to analyze information that is less structured—emails, files, or messages—then a data lake is a better approach."

MORE CHOICES

There is a third way, observed Herb VanHook, vice president of enterprise CTO Services of BMC Software. Emerging data lakehouses address requirements for analyzing both structured and unstructured data, he said. "Lakehouses are a new construct that applies querying capabilities associated with structured data and data warehouses alongside the diverse workloads that data lakes use to process many formats. Many cloud providers offer a set of different platform capabilities, including a data warehouse with querying, a data lake with analytics processing, and managed infrastructure, to run these platforms."

The data lakehouse architecture offers a compelling option, "since it combines the best qualities of data warehouses and lakes to provide a single solution for all significant data workloads," including SQL analytics, business intelligence, data science, and AI, agreed Joel Minnick, vice president of marketing at Databricks.

Data warehouses "have been used in the past for lagging indicators, such as sales of the last quarter per product line," said Prashant Kelker, partner for digital strategy and solutions at ISG. "The focus is now shifting to leading indicators, for example, what could the forecasted sales be per product line? Leading indicators are making way for either judgment or prediction algorithms. Data lakes and data warehouses are merging into concepts like data lakehouses."

Both the data lake and data warehouse are not perfect solutions, Minnick pointed out. "They have contrasting benefits that often require organizations to deploy both architectures, which is costly to maintain, complicated, and causes information silos that slow down decision making. Data lakehouses provide the low-cost, open standards support of a data lake with data warehouse-like performance, reliability, quality, and scale. Lakehouses can support both structured and unstructured data, including video, audio, and text."

In addition, the range of choices extend well beyond the bounds of enterprises, often easily accessible via APIs, which support "a number of easily integrated data orchestration platforms," said Pragyansmita Nayak, chief data scientist at Hitachi Vantara Federal. "Open and easily extensible architectures enable this data exchange without significant development and learning."

NEWSLETTERS

- 5 MINUTE BRIEFING: INFORMATION MANAGEMENT [LATEST ISSUE]
- 5 MINUTE BRIEFING: DATA CENTER [LATEST ISSUE]
- 5 MINUTE BRIEFING: MULTIVALUE [LATEST ISSUE]
- 5 MINUTE BRIEFING: ORACLE [LATEST ISSUE]
- 5 MINUTE BRIEFING: SAP [LATEST ISSUE]
- 5 MINUTE BRIEFING: BLOCKCHAIN [LATEST ISSUE]
- 5 MINUTE BRIEFING: CLOUD [LATEST ISSUE]
- IOUG E-BRIEF: ORACLE ENTERPRISE MANAGER [LATEST ISSUE]
- IOUG E-BRIEF: CLOUD STRATEGIES [LATEST ISSUE]
- DBTA E-EDITION

The New Data Analytics: Riding on Data Lakes, Data Warehouses, and Clouds

Oct 5, 2021

By Joe McKendrick

<< back Page 2 of 5 next >>

In all scenarios, business requirements should be the ultimate determinant, and this is very much the case with data analytics platforms. "Ultimately, it is always about how well business challenges, such as customer churn, fraud detection, or security threat perception, are addressed," said Sri Raghavan, director of data science and advanced analytics, product marketing for Teradata. "Data platforms that are also accompanied by a palette of analytics capabilities—algorithms, visualizations, workflows—that can be used by a wide range of personas will always be dominant."

There are considerable choices above and beyond the data environment itself. "Today's analytics requirements range from real-time and time-series-based analysis right down to standard BI," said Kelker. Some analytics processes are too expensive to be done in the cloud due to ingestion costs, and, therefore, they require edge AI solutions with local algorithms and tinyML (tiny machine learning).

"The days of master data management are over," Kelker stated. The focus has moved away from data models toward algorithms, he noted. "The explosion of

external data fields intermeshed with internal data makes this increasingly difficult to design upfront. Modern concepts, such as streaming architectures and data meshes, are blending the worlds of data storage and analytics together."

The main issue is that "organizations need real-time, actionable insights to inform critical decision making," said Scott Gnau, VP of data platforms at InterSystems. "Seamless, cross-silo access to the right data at the right time is difficult due to increasing complexity and latency challenges. Scalable, high-performance data platforms that connect distributed data to the composable stack need to serve as the foundation of modern analytics strategies."

DATA WAREHOUSES, LAKES, AND EVERYTHING IN BETWEEN

What kind of role is emerging for today's data warehouses, and how have data lakes shaped this role? It's important that "data warehouses and data lakes operate in unison if businesses want to stay ahead of the game," said Adams. "Data warehouses typically ingested information from relational databases that was then extracted by business intelligence tools for further analysis," said Adams. Data lakes have created "an undercurrent for warehouses," whereby they are now able to store all business data—from contacts, user information, documents, pictures, logs—or any data the business and its users generate, he noted. "While having a breadth of diverse information in data lakes makes transforming data a more difficult task, it gives organizations a wealth of information that was previously inaccessible."

A converged data warehouse-lake architecture is the best path forward for supporting increasingly complex analytic data environments, Raghavan said. "Data lakes, or data swamps, require robust solutions to understand, search, and analyze the data in a context-sensitive manner, while not losing the associated lineage and provenance information," he pointed out. "Data warehouses have been rearchitected to meet the emerging need of analytics that is near real-time and can handle large volumes of data."

Raghavan also pointed out that "today's data warehouses have become high-efficiency, super-compute clusters where not only are ETL processes used to deliver clean data but also combined with state-of-the-art feature engineering and modeling capabilities to deliver high performance models and operationalizations at scale." Data lakes have contributed to these super data warehouses "by simply increasing the volume and the breadth of data that could be ingested into a data warehouse. The presence of a loosely coupled compute-storage architecture ensures that subsets of the data can be selected for ETL [processes] and more production-ready work within the warehouse."

The New Data Analytics: Riding on Data Lakes, Data Warehouses, and Clouds

Oct 5, 2021

By Joe McKendrick

<< back Page 3 of 5 next >>

Structured data repositories such as data warehouses may be essential within businesses turning to data-intensive approaches such as AI and machine learning. "Many analytics jobs running in an environment like Spark create structured features from unstructured data," said VanHook. "As features like this feed more machine learning models, as well as traditional reporting, the need for a structured data store to hold the composite features emerges. You can imagine traditional data warehouse capabilities evolving to hold feature tables and becoming a high-performance feature store that can drive both the training and inference activities of machine learning models. At the bottom line, warehouses become the querying point, while lakes are the analytics point."

The data within data warehouses is generally trusted as the central version of truth because it's highly curated and processed, said Anjan Kundavaram, chief product officer at Precisely. "For analytics, the structured format of data warehouses makes it easier for standardized access, queries, and reporting. The predetermined structure also offers ready-to-use, clean data that is ideal for organizations that need to conduct operational analysis or reporting."

The platform "should have a data fabric to drive data flow orchestration and automation to deliver information and intelligence to users," Winfield said. "The platform will also need shared management and security services and support for a range of clients to meet the application development requirements for different users—including data engineers, data scientists, business analysts, and business users."

Data fabrics also offer a way to bring these environments closer together to deliver analytics as needed. "Today's data warehouses are collecting immense amounts of data—more than may have been anticipated when these technologies were originally implemented," Gnau said. "While data lakes have helped organize this raw data into central repositories, they still are not typically

involved in operational and transactional data flows. This is where modern data architectures, such as data fabrics, come into play. Not only do data fabrics effectively organize the datasets into fields that help identify the most actionable and high-quality resources but each one [also] tends to meet a unique, IT-driven purpose. Without a well-orchestrated architecture, the data remains either inaccessible and wasted or not efficiently addressable, regardless of where it sits within the data lake or warehouse."

TO THE CLOUD—IN MOST CASES

Are analytical platforms such as data warehouses, lakes, or lakehouses going to the cloud? Are there scenarios where on-premise approaches are still preferable? A recent survey of IT leaders found that the majority, 53%, see hybrid or multi-cloud data warehousing as one of the most important data warehousing-related trends of this year—more than any other trend. The question isn't really about "why" to use cloud anymore, said Minnick. "Increasingly, we're seeing customers now ponder 'which' clouds." Minnick noted that the majority of Databricks' enterprise customers work with at least two cloud providers today. "As a result, it's become much more important that organizations adopt solutions that offer a consistent experience for their employees, regardless of where the data resides."

<< back Page 3 of 5 next >>

PUBLICATIONS & REPORTS

BUILDING A CULTURE OF TRUST IN A COMPETITIVE ECONOMY: 2021 SURVEY ON DATA QUALITY

EBOOK: TOP 10 PRINCIPLES OF A CLOUD BACKUP SERVICE

WHY YOUR POSTGRESQL DATABASES SHOULD LIVE ON AMAZON AURORA

MODERNIZING DATA MANAGEMENT FOR THE HYBRID, MULTI-CLOUD WORLD

SLASH CLOUD BACKUP COSTS WITH QUEST® QORESTOR®

The New Data Analytics: Riding on Data Lakes, Data Warehouses, and Clouds

Oct 5, 2021

By Joe McKendrick

<< back Page 4 of 5 next >>

Moving to the cloud may only be a matter of keeping up with the rest of the environment. "Data, regardless of the type, is already on the cloud," said Raghavan. "Every enterprise customer that I have worked with has petabytes of information stored in the cloud and has multiple clouds for disaster recovery, fault resistance, geographical duplication, as well as workload segmentation, among other things." However, there are still reasons why a part of an enterprise's workload will continue to stay on-premise, he continued. Security, for one, may require keeping data within the enterprise walls. This does not assume that there are no risks of security breaches in the on-premise world, Raghavan stressed. It is just that companies feel a bit more at ease with the perception of complete control of their on-premise installations.

There are other reasons to at least keep analytics data on-premise as well—such as "data transit costs, latency for real-time apps, and cost at scale," said VanHook. "This is true in cases where compliance requires complete control of the data infrastructure, as well as micro data centers on the edge that preprocess substantial amounts of data from remote locations." On-premise warehouses are also still desirable for large-scale, purpose-built, high-performance infrastructures for specialized applications, VanHook said. "Examples of this include a real-time data warehouse for a customer-facing service and the real-time routing app for a delivery service."

Still, the mass migration of data environments continues toward the cloud. While on-premise data management "offers the ultimate control over the location of data, hardware, software, and who has access, it can be expensive and difficult to scale," said Kundavaram. On-premise data management requires constant maintenance, time-consuming upgrades, and, potentially, outage management, Kundavaram continued. "On the other hand, companies that operate their data lakes and data warehouses in the cloud see a nimbler experience. Operating in the cloud allows businesses to accumulate massive amounts of data and scale

their operations accordingly. Data lakes and warehouses that operate in the cloud can manage multiple data streams simultaneously while backing up this data automatically."

WHAT'S AHEAD

Automation is making data analytics better. Looking into the immediate and longer-term future, convergence between analytical data platforms will continue, enhanced by automation capabilities, industry observers state. "Automation is drastically changing the ways that we consume and process data," said Adams. "Machine learning mitigates many of the difficulties associated with applying data science and review principles when processing and storing data. If machine learning gets to the point where information is streamed while processed, data lakes will transition into data warehouses where analytics can be conducted in real time. Effectively, machine learning holds the potential to structure the unstructured and open up a world of new analytic potential."

<< back Page 4 of 5 next >>

PUBLICATIONS & REPORTS

BUILDING A CULTURE OF TRUST IN A COMPETITIVE ECONOMY: 2021 SURVEY ON DATA QUALITY

EBOOK: TOP 10 PRINCIPLES OF A CLOUD BACKUP SERVICE

WHY YOUR POSTGRESQL DATABASES SHOULD LIVE ON AMAZON AURORA

MODERNIZING DATA MANAGEMENT FOR THE HYBRID, MULTI-CLOUD WORLD

SLASH CLOUD BACKUP COSTS WITH QUEST® QORESTOR®

The New Data Analytics: Riding on Data Lakes, Data Warehouses, and Clouds

Oct 5, 2021

By Joe McKendrick

<< back Page 5 of 5

While these forces converge, "we believe that the days of maintaining both data warehouses and data lakes are limited," said Minnick. "The lines between data engineering, data analytics, and data science are becoming increasingly blurry. The people in these fields are working more closely together than ever, driving the need to have everyone collaborating on the same data." As a result, he predicted, "a data lakehouse will be the norm since it builds on the open data lake, where most businesses already store the majority of their data. The data lakehouse also adds the transactional support and performance of a data warehouse needed for analytics that data lakes never delivered. The unified approach of the lakehouse will enable better collaboration with all data team members to work from the same data instead of their own siloed versions."

Kelker also sees movement toward data mesh and streaming architectures playing an increasing role in analytics environments. "A lot of this design will be impacted by forces that pull in both directions—5G and multi-access edge computing concepts can suddenly make connected use cases with large data possible. At the same time, the advancement of edge AI, tinyML, and smarter processing chips with low power will remove the need for all data to be transmitted to the cloud. Both forces are competing with how much data needs to be transferred to and from the cloud."

Overall, there will be capabilities around "data and analytics commoditizing in the cloud," said Stijn "Stan" Christiaens, co-founder and chief data citizen at Collibra. "While the drivers for that are strong and growth is fast, it will still take time for organizations to fully adopt and migrate to a modern infrastructure while at the same time 'keeping the lights on' [in] the old." He also foresees more focus on privacy-enhancing technologies such as encryption that keep patterns in data alive for analytics or synthetic data, as well as greater focus on real-time use cases, for example, with time-series approaches.

Wingfield sees a rise in containerized analytics architectures, which makes analytics capabilities "more composable so that they can be more flexibly combined into applications." In addition, data lake and data warehouse architectures "will also need to be containerized to meet the resource demands associated with big data, artificial intelligence, machine learning, streaming analytics, and other resource-intensive decision intelligence tools that are straining older data lake architectures."

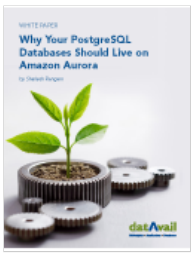
There will also be a push to democratize data analytics as these platforms open up and become more ubiquitous within enterprises. "The initial personas who were doing a lot of work on data lakes were architects and hardcore data scientists who were able to create complex analytic pipelines with multiple programming frameworks," said Raghavan. "This was a hard-to-replicate core competency that resulted in a hard-to-justify mystique." This will change, with more use cases, more types of users, and more applications derived from data lakes, said Raghavan.

<< back Page 5 of 5

PUBLICATIONS & REPORTS

BUILDING A CULTURE OF TRUST IN A COMPETITIVE ECONOMY: 2021 SURVEY ON DATA QUALITY

EBOOK: TOP 10 PRINCIPLES OF A CLOUD BACKUP SERVICE



WHY YOUR POSTGRESQL DATABASES SHOULD LIVE ON AMAZON AURORA

MODERNIZING DATA MANAGEMENT FOR THE HYBRID, MULTI-CLOUD WORLD

SLASH CLOUD BACKUP COSTS WITH QUEST® QORESTOR®