book

WHY YOUR NEXT DATA WAREHOUSE SHOULD BE IN THE CLOUD

Sponsored by: talend

tdwi
Transforming Data With Intelligence™

EXPERT Q&A

CLOUD DATA WAREHOUSING: WHAT YOU NEED TO KNOW

Why are enterprises choosing data warehouses in the cloud? What are the advantages and what kinds of data warehousing and ETL projects should an enterprise choose first? Where does a cloud data warehouse fit within an enterprise's analytics? For answers, we turned to Ciarán Dynes, vice president of products at Talend.

TDWI: What are some of the drivers behind why companies are choosing a cloud data warehouse?

Ciarán Dynes: Many LOBs (lines of business) are forced to become more data-driven to justify their expenditures and doing so requires considerable analysis about how they're connected to top-line revenue. There are several complex variables and data points that go into determining the link between spending and revenue and these data points are scattered across multiple systems. These LOBs can't wait for central IT to provision a data warehouse for them so they can start analyzing their data. The only solution is to turn toward a cloud data warehouse that can be provisioned relatively quickly and provides the necessary horsepower to start crunching the numbers.

When should customers choose a cloud database versus a cloud data warehouse?

Cloud databases and cloud data warehouses are not necessarily mutually exclusive. Both enable you to run traditional relational databases in the cloud. However, a cloud database can be used for online transaction processing (OLTP) as well as for lightweight operational reporting. A cloud data warehouse is architected for analytics workloads on very large data sets. It

can distribute workloads across multiple nodes and leverage several optimizations to provide massive improvements over traditional databases. A dedicated cloud data warehouse is a great option if your data and query complexity grows or if you want to prevent your heavy-duty analytics workloads from interfering with the performance of your OLTP workloads.

What are some of the technical challenges in migrating data to the cloud?

In addition to terabytes worth of historical data accumulated over the years that must be moved to the cloud, many on-premises data warehouses contain stored procedures. Migrating these stored procedures to a cloud data warehouse can take a long time. Moreover, you may have to combine some of the more complex transformations contained in legacy data warehouses with newer data sources before pushing them into the cloud data warehouse.

Which initial use cases should companies adopt to get started with a cloud data warehousing project?

We are at a time when digital transformation initiatives are taking precedence as a way to stay competitive. Marketing departments are entrusted with leading these initiatives and need to make the right choices about their strategy. With visibility into multiple customer touchpoints across website visits, social media interactions, marketing automation systems, in-person events, and webinars, marketing analytics is one of the most critical use cases to adopt for an initial cloud data warehousing project. There are numerous systems in use in a marketing department, and each system generates a tremendous amount of data points that could be analyzed within a cloud data warehouse.

How should CIOs think about cloud data warehousing within the context of a broader analytics initiative?

First, even before beginning to think about cloud data warehouses, CIOs need to determine the top priorities for analytics for the year. These analytics priorities need to benefit as many departments as possible within the company, without encumbering IT too much, while at the same time allowing self-service to end users within each department.

Next, the CIO should meet with the line of business (LOB) IT personnel who are assigned to each department and find out how comfortable they are with basic database administration skills. The benefit of cloud data warehouses is that they don't require many cycles to be spent on performance tuning because these aspects are taken care of by the cloud provider. However, LOB IT personnel should be comfortable with basic SQL queries so they don't consume the time of central IT.

Based on a combination of self-service benefits to end users, data requirements, and LOB IT skillsets, the CIO should consider standardizing across all departments with a single cloud data warehousing provider that fulfills all three criteria.

How do cloud data warehousing and big data intersect?

Big data has enabled the world of unstructured data sources to be tapped for any sort of intelligence. This is why you see all the open source projects for streaming, batch, and machine-learning use cases. Cloud data warehousing can be a conduit for bringing the world of structured data from legacy on-premises data warehouses together with these newer big data sources. In this manner, companies can uncover insights that they previously would not have been aware of.

IN PRAISE OF ELASTICITY

Stephen Swoyer

Get ready for the elastic data warehouse—which is what, exactly?

There's a new EDW in town. No, it isn't the enterprise data warehouse. That's *so* last millennium. Nor is it the extended data warehouse, an attempt to bridge the SQL and NoSQL worlds.

It's the elastic data warehouse, which is—what exactly?

In a basic sense, the elastic data warehouse is a data warehouse in the cloud, or DW-as-a-service.

That isn't quite it, however, argues Kent Graziano, senior technical evangelist with cloud data warehousing specialist Snowflake Computing Inc.

According to Graziano, there's a world of difference between a conventional, on-premises database and a database that's been designed with the benefits of the cloud—especially elasticity—in mind.

"It requires a different approach to loading the data, querying the data, and moving the data around for data warehouse workloads. Because of the nature of it, you need to have the ability to do different kinds of things without physically touching the hardware, so there's a service layer in there that allows you to manage things remotely via the cloud," he explains.

"You think of Salesforce.com as a cloud-based CRM system, so something akin to that, but optimized purely for dealing with the

data. It allows you to access, load, and scale your data, scale your workloads, in the cloud, without a lot of messing around with the hardware."

Graziano says Snowflake was designed from scratch for the multi-tenant cloud. It's a massively parallel processing (MPP) database, which means it distributes data across multiple clustered nodes. This makes it a powerful query-processing platform, one that's notionally comparable to MPP DW-as-a-service offerings from Amazon Inc. (Redshift), Microsoft Corp. (Azure SQL Data Warehouse), and Teradata Corp. (Teradata Cloud). For Graziano, however, these and other MPP cloud data warehouse services are insufficiently elastic.

Elasticity Explained

"Elasticity" on Snowflake's terms means a system that's designed not just with the advantages but with the constraints of the cloud model in mind. The foremost of these—multi-tenancy—is the type of thing that cuts both ways. On the one hand, multi-tenancy makes it possible for multiple workloads to coexist on the same physical hardware, simultaneously sharing access to virtualized compute, storage, and network resources.

Multi-tenancy is necessary in order for elasticity to be possible. Think of elasticity as that property—unique to the cloud—that permits a subscriber to scale up or scale down compute or storage capacity as needed. Business conditions change? Scale up or scale down on demand. Need to improve query responsiveness for certain groups or users? Scale up by adding more nodes. The beauty of MPP is that it can be predictably scaled: add four nodes to a four-node MPP cluster and you'll roughly double its performance. The beauty of MPP in the multi-tenant cloud is that you can add extra compute capacity at negligible cost. It's also much cheaper to turn off capacity when you no longer need it.

That's the good. The bad is that classic multi-tenancy can be hostile to decision support workloads. The reasons for this are complicated, having largely to do with the characteristics of analytic workloads, which are often both computationally intensive and involve lots of disk writes. In a multi-tenant context in which resources are virtualized, and compute and storage resources aren't "local" in the sense of an on-premises MPP configuration, the performance and responsiveness—the availability—of a data warehouse system could be impacted, right?

Yes and no, Graziano parries. If you're talking about a database system that wasn't originally designed for multi-tenancy, yes, he says, performance is likely to be impacted. Snowflake, he argues, uses several techniques to mitigate potential issues. For example, for primary storage, Snowflake uses Amazon's Simple Storage Service (S3) as a persistence layer. However, it also uses an SSD layer for caching data, as well as for temp space. The faster SSD layer helps to offset S3's latency.

What the Cloud Data Warehouse Changes

Will the performance of a multi-tenant MPP cloud data warehouse—even an "elastic" data warehouse, such as Snowflake—be roughly comparable to that of an on-premises MPP data warehouse system? Graziano says it will, but he's hardly a disinterested observer. More likely, performance and other availability characteristics will be impacted by the vicissitudes of the cloud model. In moving data warehouse workloads to the cloud, you're going to sacrifice control over some of the features (such as performance and availability) that you were able to tweak in an on-premises environment. (It's telling that most cloud data warehouse providers do not offer granular, performance-based service level agreements. Teradata's cloud offering is an exception.)

On the other hand, the cloud offers a slew of advantages vis-à-vis physical, on-premises implementations. There's elasticity, for starters, which radically changes how you plan for, budget for, procure, and maintain a data warehouse system. "In the traditional data warehousing world, whether it's your traditional on-premises databases or even your pre-packaged data warehouse appliances, there's actual hardware constraints on how many nodes you can buy, or how much disk you can buy, and you have to do all of that up front. With elastic data warehousing, you don't need to do that. You don't need to preallocate or prepurchase a certain amount of disk or a certain amount of compute power. It makes things a lot easier," Graziano argues.

In addition, the DW-as-a-service model eliminates data warehouse maintenance. DW-as-a-service a la Snowflake, Amazon, Microsoft, and Teradata aims to eliminate most of the

tedious upkeep associated with the conventional, on-premises data warehouse model. In a sense, DW-as-a-service eliminates the problem of data warehouse obsolescence, too. System hardware doesn't have to be upgraded or replaced. That is done in the background by the service provider.

"You don't have to be an expert administrator, a systems administrator, a database administrator, to deal with this and make it work," Graziano concludes. "What are we going to do when we need to go from 100 to 1,000 users? We're going to spin up more compute clusters, that's what we're going to do. On the infrastructure side, you don't have those planning and budgeting problems anymore."

Stephen Swoyer is a contributing editor to *Upside*.

CLOUD DATA WAREHOUSES: MORE THAN JUST COST SAVINGS

There's much more to cloud data warehouses than saving capital and operating expenses.

Cloud is the most important thing to happen to data warehousing (DW) since its inception.

The cloud model lowers the barriers to entry—especially cost, complexity, and lengthy time-to-value—that have traditionally limited the adoption and successful use of data warehousing technology. It permits an organization to scale up or scale down—to turn on or turn off—DW capacity as needed. It's fast and easy to get started with a cloud data warehouse. Doing so requires neither a huge up-front investment nor a time-consuming (and no less costly) deployment process. Even if the cloud data warehouse isn't quite risk-free, it's close.

The cloud data warehouse is a new and fundamentally different technology offering and to get the most out of it will require a new and fundamentally different kind of thinking. Unfortunately, the best practices we use to design and build on-premises DW systems will not translate en masse to the cloud. To copy an existing data warehouse over to the cloud—a "lift and shift" operation—is to fail to optimize for the advantages (and constraints) of the cloud model.

"The reason the cloud data warehouse is becoming so popular is because it takes a lot less time to provision. It's also much cheaper than the traditional on-premises data warehouse, and it takes a lot less time to administer and monitor. In fact, cloud data warehouse services automate as much provisioning, administration, and monitoring as possible for subscribers,"

says Ashwin Viswanath, director of cloud product marketing with data integration vendor Talend. "This is a radical advantage, as is the new world the cloud data warehouse opens up for early adopters. I'm talking about new workloads, new use cases, and the unprecedented ability to analyze data at scale."

For these and other reasons, experts presume that traditional, on-premises data warehouses will gradually cede ground to cloud data warehouse services. "This is not going to happen overnight. It's going to take some time," Viswanath argues, noting that companies can do several things to prepare for (and get the most out of) this transition. The most important thing is to identify one or more candidate use cases for the cloud data warehouse, he says.

Viswanath cites customer success analytics and sales operations analysis as two common concrete use cases. Think of the former as predictive analytics projecting customer churn for software-as-a-service (SaaS) vendors. (If it's trivially easy for subscribers to spin up SaaS services, it's only slightly harder for them to *switch* SaaS providers.) The latter lends itself to sales pipeline analysis and other predictive endeavors. Because the data used in these and other kinds of analytics is primarily sourced from other cloud services, it makes sense to shift these workloads into the cloud—especially to the degree that the historical transaction data that feeds on-premises DW systems is slowly and steadily shifting to the cloud, too.

"Over time, a larger portion of the historical data that's traditionally been stored in the on-premises data warehouse will migrate to the new cloud data warehouse," he says. "The cloud data warehouse is also going to be used to populate some of the newer types of data that would not typically have gone into a traditional data warehouse."

The cloud data warehouse is uniquely positioned for the emerging class of advanced analytics workloads that combine traditional, transactional data with data of different types and provenance. Advanced analytics uses data from cloud and social media services; sensors, embedded devices, and other machine signalers; subscription and open data sets, and so on.

The Cloud Data Warehouse Is a Platform for Business Transformation

To recap: the cloud data warehouse is an elastic resource. You can scale it up and down (or turn it on and off) as needed. The cloud data warehouse also delivers rapid time to value: subscribers can be up and running in weeks instead of months or years as with on-premises DW projects. Finally, the cloud data warehouse largely eliminates the risks endemic to the costly and time-consuming on-premises data warehouse paradigm. You don't have to budget for and procure hardware and software. You don't have to set aside a budget line item for annual maintenance and support. In the cloud, the cost considerations that have traditionally preoccupied data warehouse teams—budgeting for planned and unplanned system upgrades—go away. For these reasons, on-premises data warehouse workloads will continue to shift to the cloud.

In the same way, new data warehouse workloads will *originate* in the cloud. Thanks to its combination of low-cost storage and processing capacity, the cloud data warehouse is well suited for the new and different types of data used in advanced analytics workloads. It's an ideal platform in which to combine and analyze transactional data from core operational systems with data from RESTful cloud APIs, the social web, and the streaming data generated by sensors.

There's another advantage—the cloud data warehouse can process massive amounts of data at relatively low cost. In the on-premises world, the most powerful and scalable data warehouse systems are also prohibitively costly. These systems are able to perform massive parallel processing (MPP), which makes it possible to scale them by adding extra compute and storage nodes. If you add two extra nodes to an existing two-node MPP system, you basically double its performance. In practice, it is extremely costly to research, procure, tune, manage, and upgrade an on-premises MPP data warehouse system. In the cloud, you don't have to: Amazon, Microsoft, Snowflake Computing, Google, and other vendors take care of that for you. These providers make MPP performance a realistic option for a much larger pool of potential users.

Even if you don't opt for MPP performance in the cloud, you can still spin up *multiple data warehouse clusters at relatively low cost*. One DW cluster can support your business intelligence reporting and analytics use cases. Other clusters can support field operations analysis, customer churn analysis, and other kinds of predictive and advanced analytics.

This just isn't possible in the on-premises data warehouse model.

"Many cloud data warehouse vendors only really talk about one benefit of the equation: it's cheaper, cheaper, cheaper. Beyond just cost, however, there is a lot of value in looking at the cloud data warehouse as a platform on which to really build data-fueled apps and use cases," says Viswanath. "In the cloud data warehouse, just by its nature, you can deploy many separate clusters dynamically. This enables you to crunch massive amounts of data at scale. What that means is that whole new use cases—and whole new worlds—are opened to you. One example is machine learning use cases that can predict the next set of actions your customers might take."

The economics of data warehousing in the cloud make it possible to store, manage, and analyze data at a scale never before imagined. As the complexity of your analytics workloads increases, you can dynamically spin up new compute clusters—or add new nodes to existing compute clusters. As new nodes are added, the cloud service itself takes care of rebalancing the data warehouse. The time-consuming process of provisioning, testing, and tuning data warehouse performance goes away.

Data Warehousing in the Cloud Isn't Completely Turnkey

Cloud data warehouse services automate or accelerate many aspects of designing, deploying, managing, and tuning a data warehouse system. They can't and won't replace the need for human skill and ingenuity, however. The knowledge and expertise of data architects and other skilled professionals will be as important as ever, Viswanath argues.

"You need to understand the entire organization's requirements with respect to different lines of business. For example, a marketing department might run Marketo, Google Analytics, and other services. It could be beneficial for the marketing department to combine this data with product data or transactional data that's stored in an on-premises data warehouse. This is why you need to do some knowledge transfer around the apps the different lines of business are using. You need to figure out the kinds of data they want to measure within these apps—along with how this data might be combined with data from other sources," he points out.

"This means joining data that originates from a service within their own domain—e.g., Marketo or Google Analytics—with data from a traditional data warehouse, such as a certain product SKU, or a certain price-point from that SKU. That way, they can do better analytics."

This is first and foremost a data integration (DI) problem. One thing the cloud data warehouse model doesn't substantially accelerate is DI—not on its own, at least. The good news is that third-party cloud DI services make it much easier (and cost-effective) to integrate data, at least in comparison with the on-premises DI model. Cloud DI does away with upfront investments in hardware and software, as well as ongoing maintenance and support. It can likewise accelerate the pace of DI development, which is by far the biggest drag on data warehouse and analytics projects. For example, Talend's Integration Cloud features a wizard-driven, self-service user experience, with built-in support for popular SaaS services such as Marketo, NetSuite, and Salesforce.com, along with connectors for cloud platforms such as Amazon Web Services (AWS), Google Cloud, and Microsoft Azure. It also provides connectors for integrating data from on-premises applications and services and it has the capacity to enforce data quality and governance rules for traditional, strictly structured relational data and new polystructured data types.

Enforcing governance and security with respect to strictly structured data is a well-understood problem. For example, most cloud DW services are able to enforce user- and role-based access control restrictions and auditing for relational data. Some support data lineage tracking, too.

Enforcing governance and security standards with respect to polystructured data is a much harder problem, however. "The first thing to remember is that unstructured data doesn't have any schema, at least not in the way that conventional

structured data does. To really get some sort of intelligence out of it, you need to first give it some kind of schema," Viswanath says.

"This is a very special kind of data integration challenge. That's when you need quality integration tools. A cloud data warehouse is not going to give you a schema all by itself," he continues. "You need a data integration tool that can parse the unstructured data, give it some sort of schema, and then feed it into the cloud data warehouse for further analysis."

He uses the example of streaming sensor data. "First you have to ingest the data. There are a lot of open source alternatives out there, such as Kafka, or cloud services such as Amazon Kinesis. Using either of these, you then have to process the data—to give it schema—using Spark, for instance," Viswanath explains. "Once you give it some kind of schema, you can analyze it in the cloud data warehouse to gain further insight. One way the cloud data warehouse plays a critical role here is it enables you to take that unstructured data and merge it with other forms of structured data to get value out of it. When you do so, you get more insights."

Cloud providers must address another important consideration for certain sectors. Governance, data quality, and especially *security* are particularly salient issues for certain highly regulated verticals, as well as for all government organizations. This is why AWS and other cloud services maintain dedicated zones—the equivalent of a virtual private cloud—in which these customers can spin up isolated instances of cloud services. The public cloud is a multi-tenant environment. This means subscribers share access to the same virtualized resources. For example, if two companies subscribe to the same cloud service, they could be running in the same hardware context, albeit in separate virtual contexts. In the virtual private cloud, the applications, workloads, and services of each subscriber are restricted to a particular hardware context. This context isn't shared with any other "tenants." This scheme addresses a common regulatory requirement that would otherwise prevent an organization from taking advantage of the cloud for DW, DI, and other workloads.

"We see a lot of the more highly regulated industries, such as pharmaceutical companies and financial services companies, adopting this virtual private cloud. It's in the interests of providers to make it as easy as possible for organizations in these highly regulated industries to create their own virtual private clouds using a public cloud service," Viswanath notes.

Conclusion

On-premises workloads are shifting and will continue to shift to the cloud. Over time, the cloud data warehouse will effectively supplant the on-premises warehouse as the focal point of decision support and analytics. In point of fact, *most* on-premises IT workloads will shift to the cloud, which makes the DW shift inevitable. Consider the "data lake," which describes a vast repository—a reservoir, so to speak—for enterprise data of all types. The original data lake was conceived for the on-premises enterprise. Increasingly, however, organizations are spinning up data lakes in the cloud, attracted by the low cost (Amazon's Simple Storage Service, or S3, is a hugely popular option in this regard) and the suitability of the cloud for workloads and use cases of all kinds.

"Basically, what you're seeing is an emerging cloud-first strategy. Let's look at what the adoption curve looks like. Initially, you'll have a line of business that will deploy a cloud analytics tool, Salesforce Wave Analytics, for example. That's really only the first step," Viswanath argues. "Then they discover that their data needs are really intensive and they can't just use a cloud analytics tool by itself. They also need a cloud data warehouse, because the volume of data is just tremendous."

In most cases, they'll also need a robust cloud DI service. After all, it's important to have a streamlined process that connects the traditional on-premises data warehouse—which will persist for the foreseeable future—the cloud data warehouse, and the SaaS and PaaS applications used by the line of business. "You need that streamlined process. The traditional data warehouse and centralized IT aren't going away overnight. For some time to come, organizations will need to enrich cloud data and data from other non-traditional sources with transactional data and product data," he continues. "Cloud integration technology is the key towards streamlining this data integration process."

www.talend.com

Talend's integration solutions allow data-driven organizations to gain instant value from all their data. Through native support of modern big data platforms, Talend takes the complexity out of integration efforts and equips IT departments to be more responsive to the demands of the business, at a predictable cost. Based on open source technologies, Talend's scalable, future-proof solutions address all existing and emerging integration requirements. Talend is privately-held and headquartered in Redwood City, CA. For more information, please visit www.talend.com and follow us on Twitter: @Talend.

IAIT Lab Test of Talend Integration Cloud

The Data Warehouse in the Cloud – What You Need to Know

tdwi.org

TDWI is your source for in-depth education and research on all things data. For 20 years, TDWI has been helping data professionals get smarter so the companies they work for can innovate and grow faster.

TDWI provides individuals and teams with a comprehensive portfolio of business and technical education and research to acquire the knowledge and skills they need, when and where they need them. The in-depth, best-practices-based information TDWI offers can be quickly applied to develop world-class talent across your organization's business and IT functions to enhance analytical, data-driven decision making and performance.

TDWI advances the art and science of realizing business value from data by providing an objective forum where industry experts, solution providers, and practitioners can explore and enhance data competencies, practices, and technologies.

TDWI offers five major conferences, topical seminars, onsite education, a worldwide membership program, business intelligence certification, live webinars, resourceful publications, industry news, an in-depth research program, and a comprehensive website: tdwi.org.