DIGITAL DIALOGUE

MODERNIZING ETL FOR FASTER CLOUD DATA MIGRATION

Cloud migration is a top priority for most organizations. However, the thousands of time-consuming and costly extract, transform, and load (ETL) jobs that organizations run using on-premises systems are a major barrier to rapid and transformative cloud migration. Organizations need to convert many of these to run in the cloud. Pressures on ETL are growing as organizations democratize data access and analysis and new users want to realize value by interacting with faster and more-voluminous data sources.

tdwi
Transforming Data
With Intelligence™

DIGITAL DIALOGUE

tdwi
Transforming Data
With Intelligence™

Introduction

ETL comprises processes for extracting, structuring, cleansing, validating, enriching, and loading data into a target system (such as a data warehouse) for users' reports, dashboards, and analytics. Traditionally, development and execution of ETL jobs is slow and heavily manual; the code is often hard to change and is poorly documented. ETL can require significant technical expertise as requirements become complicated. Inefficient ETL becomes expensive.

When ETL takes too long, the data loses value. It is therefore not surprising that 31 percent of organizations TDWI surveyed say that reducing ETL latency is one of their top modernization priorities.[1] Data-driven organizations want to empower users to do more of their own ETL development; 30 percent say it is a priority to improve self-service capabilities.

Fortunately, modern, cloud-native ETL services are available to help organizations leverage the power of the cloud's scalability for computation and capacity and increase self-service. If organizations choose serverless options, they leave machine resource allocation, provisioning, configuration, capacity planning, scaling, management, and maintenance to the cloud provider. Leading serverless ETL offerings can scale up and down automatically in response to demand.

However, organizations still face challenges when migrating to the cloud, primarily regarding complex ETL conversions. TDWI finds that organizations are also concerned about how they can control costs in the cloud. In a recent webinar, TDWI senior research director for business intelligence David Stodder spoke with Shiv Narayanan, product manager with Amazon Web Services (AWS), and Ranjith Ramachandran, big data lead with Wavicle Data Solutions, about strategies for modernizing ETL for faster cloud data migration. The roundtable discussion touched on a range of important topics, including how to achieve faster, more consistent and scalable ETL conversion and modernization for cloud migration. Here are highlights from that discussion.

Strategies for Achieving Benefits Sooner

The discussion began with a question many organizations are asking: what are the actual benefits customers are seeing by migrating to serverless ETL from traditional ETL tools that run on a dedicated server?

Shiv divided cloud options into three categories. "The first is self-managed. You install the software and you manage it, including activities such as making patches. The second is using a managed service. The biggest difference between the first and second options is that you are making the choices, but the cloud provider, such as AWS, runs and manages it on your behalf. You are still choosing the servers and making other choices such as how big a cluster you need and whether you want scaling features.

"Serverless computing is the third category. Here you don't concern

[1] See Figure 7 in 2021 *TDWI Best Practices Report: Modernizing Data and Information Integration for Business Innovation*, online at tdwi.org/bpreports.

yourself with infrastructure, management, or security decisions; you can focus on the business problems you are trying to solve."

Shiv then described five important benefits of a serverless ETL offering such as AWS Glue compared to traditional, on-premises ETL.

"The first is scalability. ETL jobs get their own dedicated clusters and can add speed by running in parallel. The second benefit is controlling costs, which TDWI noted is a major concern. For example, one of our customers, a rental car company, saw data volumes drop during the pandemic. The serverless option was cost-effective because billing scaled down as well.

"The third benefit is speed—that is, how fast you can deliver prepared data for users. With serverless ETL, you are not delayed by having to deal with infrastructure. The fourth is enabling innovation. Traditionally, companies have had to focus on solving problems such as data duplication as they write hundreds or thousands of lines of ETL code. Now, they are able to benefit from machine learning and automation in modern serverless ETL to do data deduplication. They can put their focus on business issues. Finally, the fifth benefit is having an open platform."

Webinar audience members asked the panelists how clients ensure that costs are controlled when there is so much going on. "The serverless option sounds like kind of a black box," one audience member said. "How do you forecast costs if it is consumption-based? How do you avoid runaway costs or unexpected increases?"

Shiv acknowledged the danger of cost overruns with a consumption-based model. However, he described a different experience with clients. Typically, he said, customers do not start on day one with a full production workload. They ease into it by going with smaller workloads in a more controlled way, where someone is monitoring costs and an enterprise can immediately understand fluctuations. There are also calculation tools available from companies such as Wavicle to perform a deeper analysis of potential cost drivers.

"Typically, clients onboard a few workloads in a development setting and turn on AWS Auto Scaling, which is a new feature available in AWS Glue Preview. Then you can avoid spending unnecessarily on things that are not working or nodes that you are not using.

"Once clients have a good estimate of costs based on the smaller-scale or limited workloads and use tools to put controls in place, they will onboard more workloads. They can fine-tune workloads using tools such as Wavicle's and AWS Cost Explorer, which lets you put checks in place, for example, to visually monitor whether a job is running too long and you might need to abort it. These methods and tools allow clients to optimize costs as they onboard more workloads."

Redesign or Migrate?

Panelists discussed whether data models and ETL require full redesign or whether they could just be migrated. "There are two options: 'lift and shift' or completely modernize," Shiv explained. "Lift and shift can be a good option, but with either choice, organizations have to get away from the notion that this whole thing is going to look exactly like what they are used to. The cloud is different; the techniques are different, especially when you have the ability to run jobs on several multi-clustered nodes. What we commonly see is that our customers go through a lot of analysis and find that they can keep their final data models the same. Users' dashboards can remain largely the same.

"What's different is you now have a modernized infrastructure. If before they had a staging layer for ETL, they can replace that with a cloud data lake. Customers can use serverless technologies to extract data; they may use change data capture (CDC) to improve data pipelines. For example, instead of using a batch workload, customers will use CDC to have a faster data ingestion pipeline and faster capturing of data updates.

"The common pattern we see most often is customers keeping their final data models intact and using the cloud to modernize the internals. They can reimagine the inner workings to drive higher business value."

Overcoming Latency and Complexity

The second part of the roundtable focused on how to reduce time-consuming and complex ETL development and how automation can make a difference.

"Cloud computing makes ETL pipelines completely different," said Ranjith. "The traditional way of manual coding and using tools that are not scalable from a licensing and manageability standpoint changes. There is also complexity and latency when companies create their own multiple ETL frameworks and require a range of skill sets to manage and modify jobs. With serverless tools such as AWS Glue, it is much easier to create ETL pipelines and automate code generation."

What do organizations need to understand the complexity of their ETL jobs and decide whether to migrate or redesign them? "It starts with understanding their current landscape," Ranjith explained. "At Wavicle, we assist with any priority areas and associated business pipelines and workloads. Based on that, we work with clients to evelop an overall road map from a business perspective. We leverage tools, including tools to assess current ETL environments. We perform a detailed inventory and analysis of all ETL jobs, who uses them, and other factors that cause complexity. Tools augmented with AI/ML can infer mappings, which speeds analysis. With the overall road map, we can get a holistic view of how to modernize and how long the project is going to take. This eases the migration journey."

An audience member asked whether Wavicle's project delivery approach was more agile or iterative. During the assessment, Ranjith said, Wavicle's methodology is to break ETL migration into subject areas. "We take a smaller piece and run it through multiple migrations. We can do an automated conversion and upload jobs for multiple subject areas and domains. This assessment and testing might take only a minimal amount of time because jobs for multiple areas and domains can run in parallel."

How can organizations streamline ETL code conversion and mapping? What about improving the testing and validating of new processes? According to Ranjith, "Migration starts with looking at the whole ETL landscape, understanding the complexities, and determining the most significant parts of the project. We define time frames for how soon we can achieve migration and assess the total cost of ownership (TCO) to understand the possible savings versus the cost of modernization.

"After the assessment, we begin the actual migration phase. This is where we take advantage of automation to convert legacy ETL to new ETL tools such as AWS Glue. We find that automation can reduce 70 to 90 percent of the migration time."

Ranjith noted the importance of data validation after migration. "We use automated tools to run data validation. We compare data generated by the old legacy ETL code with the new data generated by using modern ETL tools. We can set up a parallel environment to accelerate this step. By using automated testing and CDC tools, we can reduce up to 80 percent of the effort."

Road Map for Faster Cloud Migration

How can organizations make ETL conversion successful as part of their overall goal of achieving faster cloud data migration? Shiv underscored that the first step should be to understand your current ETL patterns fully.

"For example, you might have ETL jobs that are ingesting data from SAP applications into a staging area. You can use tools such as Wavicle's to analyze these jobs and divide them by their complexity, then you can choose the ones to start with and develop a proof of concept.

"AWS Glue supports a variety of approaches. If you want to continue to do batch-oriented data integration or want to shift toward more real-time data streaming using Apache Kafka or Amazon Kinesis, for example, Glue can support either strategy. Starting small with a pilot project that tests a few pipelines makes sense. You will have some lessons learned, which will be different for every organization. What's key is to leverage tool automation as you move into production and scale out. The important thing is to understand your current patterns, determine what modernization is the right fit, and then use automation to streamline conversion."

Beyond technology, organizations need to solve people and change management challenges. TDWI research indicates that for many organizations, one of the most important issues is how cloud conversion and migration affects the ability to hire and train personnel. Both Shiv and Ranjith noted that modernization actually makes an organization more attractive to current and potential employees.

"It improves overall morale because people are working with cutting-edge, cloud-native services," said Shiv. "Data engineers can take their skills and use them across platforms and languages. With Glue's user interfaces, you can expand from SQL or Python into machine learning use cases."

Shiv also noted that another big selling point is that many cloud-native services are built on open source technologies or the promise of an open platform that lets you build once and deploy anywhere to other platforms. "This improves hiring because the skill sets you need are more available than trying to find an expensive person who only understands proprietary ETL code."

A Final Word

Successfully modernizing ETL for faster cloud data migration requires analyzing the current state, choosing jobs that are appropriate for a small-scale proof of concept, and learning from the experience to guide expanded ETL conversion and growth in the number and complexity of jobs. In the roundtable discussion with Shiv and Ranjith, it was clear that following this approach will help organizations understand cost drivers, use tools effectively to monitor performance and ROI, and reduce errors and cost surprises.

Taking advantage of modern technology, including serverless options, is also key. Serverless ETL allows organizations to reduce delays, expenses, and the need to find skilled personnel who are capable of determining the appropriate cloud technology infrastructure. With serverless ETL, business users can set up transformations in a self-service fashion using graphical interfaces. At the other end of the spectrum, data engineers can use serverless ETL to launch standard and repeatable jobs for routine use cases quickly and easily. They can then devote their coding expertise to complex and challenging jobs that truly demand it.

Sponsored by:

wavicle
DATA SOLUTIONS

In collaboration with:

aws