Market Pulse

WHITE PAPER

Optimizing Business Analytics by Transforming Data in the Cloud

Organizations need to adopt an extract, load, transform sequence to take full advantage of cloud data warehouses

Company executives understand they will gain valuable business insights and many other competitive advantages by analyzing the oceans of data now available to their organizations. However, this potential boon comes with a significant catch.

Before relevant data can be analyzed, it must first be collected from multiple sources, and then properly prepared, before big data analytics systems can process it. Given the escalating data volumes involved, the well-known data preparation stages—extract, transform, and load, or ETL—can be labor-intensive and costly, and may create major process bottlenecks.

Data volumes themselves, of course, are being driven up by billions of computers, smartphones, IoT devices, and other data generators across both business and consumer landscapes. Adding to the mix are millions of websites, ecommerce operations, global social media networks, and other promising sources of meaningful data.

SPONSORED CONTENT

CIO FROM IDG

MATILLION

The collective result of the world's digital data explosion is mindboggling. During 2018 alone, storage suppliers added more than 700 exabytes of storage capacity to the worldwide installed base, according to IDC's Global StorageSphere report.

These staggering data volumes promise to be rich sources of valuable business data. But the volumes can bring traditional ETL tools and processes to their knees.

Many organizations see cloud computing as a solution to their data collection and preparation problems, and with good reason. For starters, moving massive amounts of data into cloud data warehouses can provide open-ended scalability and lower costs compared with purchasing and deploying on-premises data warehouse servers.

Simply leveraging data warehouses in the cloud doesn't provide a full solution to the core ETL challenge, however. Organizations also need to take advantage of the cloud's scalability and processing power to orchestrate and automate data collection as well as data transformation—the most demanding element of the pre-analytics process.

To gain visibility into the current data collection, storage, and transformation scene, IDG recently surveyed more than 200 IT, data science, and data engineering professionals at North American organizations with at least 1,000 employees. The results from this survey provide a picture of organizations' growing reliance on placing data in the cloud, on the startling rise in both data volumes and data sources, and on the challenges and benefits associated with cloud, automated data transformation.

Tapping the Cloud for Data Warehousing

While we hear a lot about exabytes and global data growth, it's sometimes difficult to grasp how that translates to individual companies. Zooming in on company data growth, the numbers are still daunting. The IDG survey found that, on average, organizations' data volumes are growing at 63% per month. For 12% of the respondents, volumes are growing at 100% or more per month.

Faced with this growth, most organizations are relying on cloud storage to one degree or another. IDG found that 9 out of 10 organizations surveyed have already placed at least some data in cloud data warehouses.

Three cloud data warehouses led the pack in usage among the surveyed organizations. Amazon Redshift was being used by 54% of the respondents, Google BigQuery by 50%, and Snowflake by 26%.

All told among the respondents, 37% of organizational data was currently in cloud data warehouses, 35% was in on-premises data warehouses, and 25% was in offsite, non-cloud data warehouses. Those percentages will shift significantly going forward: virtually all of the respondents said they will migrate or continue to migrate data to the cloud over the next two years.

Survey respondents cited a dozen reasons for migrating their data to cloud platforms, with the top reason being a faster time to value for implementing analytics projects. That objective is critical because the respondents thought that cloud business intelligence (BI) and analytics could help them address more than a dozen business problems (see Figure 1).

FIGURE 1. **BUSINESS PROBLEMS ADDRESSED BY CLOUD BI AND ANALYTICS**

SOURCE: IDG

As noted earlier, however, simply placing data in cloud warehouses won't solve all of the business problems identified. For starters, just collecting data from large numbers of sources can be an operational and technical challenge. Once data has been collected and loaded into the cloud, the real work begins. The diverse data must be combined, normalized, structured, and transformed in multiple ways before it can be analyzed.

Moving from ETL to ELT to Maximize the Cloud's Value

It's fair to assume that many IT and business executives, surveying today's data landscape, may sometimes think, "Too much of a good thing!" Still, most companies are making a concerted effort to tap the innumerable sources of data now available to them.

Among the companies surveyed, more than 20% were drawing from 1,000 or more data sources to feed their BI and analytics systems. The mean number of data sources per organization was 400.

While potentially of great value, the number and diversity of data sources was a factor underlying several of the obstacles respondents cited as slowing or stalling their data analytics projects. For example:

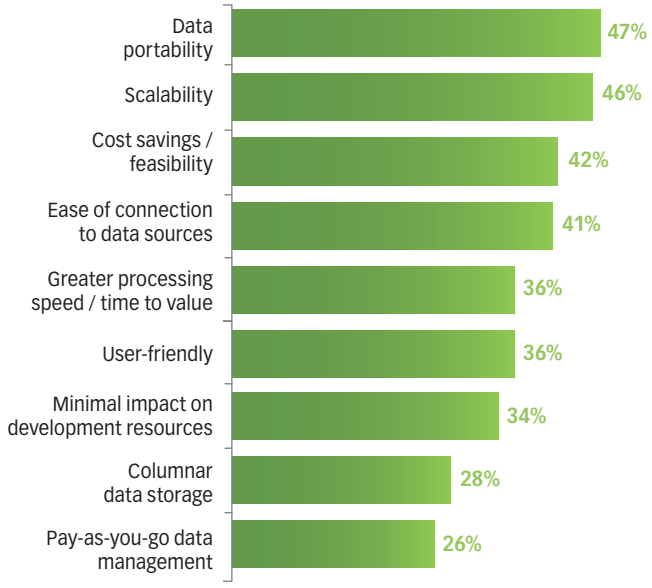
- 45% of those surveyed said navigating the complexities of the data environment was an obstacle
- 38% said manual coding of data pipelines was an obstacle
- 32% said connecting to multiple data sources was an obstacle

Clearly, organizations need connectors and other tools to ease the extraction of data from a wide variety of sources and expedite the loading of that data into cloud data warehouses.

More than 20% of companies surveyed were drawing from **1,000 or more** data sources to feed their BI and analytics systems. The mean number of data sources per organization was **400**.

Beyond the collection of raw data lies a more challenging need: preparing the amassed data for eventual analysis. More than 90% of those surveyed said it was challenging to some degree to make data available in a format usable for analytics. Of those, 57% said this data transformation task was either highly challenging or extremely challenging.

FIGURE 2. **POTENTIAL BENEFITS OF AUTOMATING DATA TRANSFORMATION IN THE CLOUD**



SOURCE: IDG

The best way to clear this hurdle is to load the data first and then transform it, rejiggering the traditional ETL (extract, transform, load) process into an ELT (extract, load, transform) process. Companies can then use cloud tools to automate processes and speed up the transformation of data for analytics using the full power, speed, and scalability of the cloud.

Most companies have yet to make the leap to ELT. Just 28% of those surveyed by IDG were loading data into the cloud and then using cloud-optimized tools to transform it. The remainder were either transforming data with non-cloud ETL tools (35%), or manually coding data into the needed format before loading it into BI/analytics systems (37%).

Again, these numbers are almost certain to change given that the survey respondents broadly recognized there are many benefits to be gained with cloud, automated data transformation. Data portability, scalability, and cost advantages lead the list of anticipated benefits of cloud ELT, as shown in Figure 2.

Matillion Enables ELT Data Integration and Transformation

Since 2015, when it launched its extract, load, and transform (ELT) solutions, purpose-built for cloud data warehouses, Matillion has grown its customer base to nearly 700

organizations operating in all industry sectors. Having previously been in the business of building and managing cloud data warehouses for its customers, the company launched its own ELT solutions after finding legacy tools wanting.

While the legacy ETL solutions can address and automate many required tasks, they often struggle to perform well when paired with cloud data warehouses. Even when they can work with data in the cloud, the existing tools typically aren't optimized for the particular characteristics and requirements of each data warehouse.

As a result, Matillion decided to create extraction, loading, and transformation solutions designed specifically for Amazon Redshift, Snowflake, and Google BigQuery. In addition to optimizing operations for each of these popular cloud data warehouses, Matillion's support for all three makes it easier for customers to shift data among them if desired.

To simplify and expedite data collection, Matillion offers a wide range of connectors built specifically for many popular data sources. These include connectors for Amazon Web Services and Microsoft Azure public clouds, for major databases, for social networks, and for a wide variety of additional sources.

Once data is loaded into a cloud data warehouse, data processes can be orchestrated and data can be automatically cleansed, standardized, and structured into formats required for subsequent data analytics. Additional features, from collaboration to security, are available in different editions designed to meet the needs of customers ranging from mid-sized businesses to global enterprises.