Scalability and Flexibility Critical to the Modern Data Warehouse

Today's fast-paced world requires responsiveness in all types of interactions with customers, prospects, partners and employees. Organizations must be able to respond in the moment or risk missing the opportunity altogether. To do this, the systems and process an organization has in place must be agile enough to deal with the constant changes it faces. And, with the volumes of data that organizations process today, they must be able to access data in place to enable analysis and exploration without requiring the data be moved or copied.

Given this increasing reliance on data, I've identified five key requirements for modern data warehouses to assist organizations in the effective planning and selection of related technologies. The five key requirements are responsiveness, agility, access to data in place, cost-effective scale and flexible deployments. In this viewpoint, I will drill down on cost-effective scale and flexible deployments since these are often challenges for many organizations.

Let's start with the requirement of cost-effective scalability. Initially the cloud attracted attention because it allowed organizations to avoid capital outlays and because new systems could be provisioned quickly. These changes meant new workloads could be explored without large expenditures and then discarded if they didn't prove worthwhile. The cloud also made it easier to attain scale in computing infrastructures. Purchasing, installing and maintaining a physical computing cluster of 100 nodes is not only costly, but requires a significant amount of effort.

Instead of purchasing and operating in-house, those nodes could be rented from a cloud provider within a few minutes. Digitally native companies like Netflix used this to their advantage and accessed very large computing clusters only as they were needed. As big data became a mainstay of information architectures, organizations began to prefer using cloud-based systems for the computing clusters that manage this information. In fact, our research shows that one-third of organizations now use the cloud as their primary data lake platform. However, easily available computing resources led to an overreliance on brute-force scalability. In other words, organizations could just add another node to address performance or scalability problems.

For modest scalability requirements, brute force works, but as requirements increase, so do the costs and complexities. If these systems are designed with inefficiencies, they will not scale linearly. In other words, twice as many users or twice as much data will cost more than two times as much as it did previously. As a result, a project that started with a reasonable and modest budget might suddenly surprise its sponsors with significantly higher bills than expected, thus making it no longer economical

to continue the project using that platform. For cost-effective scalability, these systems must be designed from the ground up to scale linearly with increasing numbers of concurrent users and with increasing volumes of data. This makes it easier to predict costs and create economically sustainable projects. Organizations should evaluate performance and cost using numbers of concurrent users and data volumes that are based on projected real-world workloads.

We now turn to flexibility in deployment method. While the cloud is increasingly popular, most organizations still run a significant portion of their information architecture on-premises. Our research shows that in about two-thirds of organizations, data processing still occurs on-premises and more than 40% of organizations process data using a hybrid of cloud and on-premises solutions.

As a result, your solutions should support hybrid deployments spanning both on-premises and cloud infrastructure. Ironically, hybrid was once a defensive term, promulgated by on-premises vendors to make sure they appeared relevant, but now even the cloud stalwarts have announced plans to support hybrid deployments. It is the reality for many organizations today. Unfortunately, many hybrid deployments today offer little or no effective integration between cloud and on-premises instances.

Modern data warehouses also need to span multiple cloud providers and support standard hardware configurations. Vendor lock-in is a common concern when making technology purchases. Organizations want to know they can move from one vendor to another for a variety of reasons, but one of the major concerns is the potential pricing disadvantage that vendor lock-in creates. And organizations want to be able to rationalize systems—for instance, in the case of acquisitions—so that there are fewer different types of systems to manage. Or they may want access to new capabilities or complementary technologies that may be available on one platform but not another.

Organizations also require integration between different instances of the data warehouses. There should be a single unified management console across all instances, regardless of cloud provider, and regardless of whether those instances are on-premises or in the cloud. In this setup, the boundaries between instances would be nearly invisible, creating a single, common data fabric. You should be able to easily migrate data between instances and share data across those instances. And, ideally, when demand for resources increases (say for end-of-month reporting and analysis), the system should automatically burst from on-premises to the cloud or from one cloud provider to another for additional resources.

When evaluating modern data warehouse alternatives and the claims made by various vendors, organizations should consider the five key requirements I mentioned at the outset. Given the current state of the market, pay particular attention to cost-effective scale and flexible deployment options, as these two are critical when designing and deploying a modern data warehouse that meets your organization's current and future needs.

Dave Menninger – SVP and Research Director, Ventana Research

David Menninger is responsible for the overall direction of research on data and analytics technologies at Ventana Research. He covers major areas including artificial learning and machine learning, big data, business intelligence, collaboration, data science and information management along with the additional specific research categories including blockchain, data governance, data lakes, data preparation, embedded analytics, natural language processing (NLP) and IoT.