www.pipelinepub.com
Volume 17, Issue 11

Reshaping the Data Lake

By: Ori Reshef

The term 'big data' has been around since the 1990s and companies have certainly been prioritizing big data investments for almost as long. Still, according to a recent 2021 NewVantage Partners survey of Fortune 1000 executives, enterprises are continuing to struggle to derive value from their big data investments. Only 48.5 percent are driving innovation with data. Just 41.2 percent are competing on analytics. And only 24 percent have created a data-driven organization.

Over the past decade, enterprise data analytics attention has shifted away from the data warehouse architecture to the data lake architecture. There are various schools of thought on what the modern data lake stack and data lake architecture ought to look like. Many organizations have failed to realize ROI from their data initiatives, largely due to unplanned and unsustainable costs of modeling data and DataOps.

Despite the challenges, there are powerful reasons for organizations to care about achieving big gains in ROI from their data lake investments. Think about pharmaceutical companies looking for the next vaccine candidate. Or consider financial services firms striving to stay ahead of market fluctuations. Media firms want to discover which pieces of content each user is likely to binge next, and security teams want to conduct analytics on the security data lake with greater speed and precision. The common thread between all these strategic initiatives is the ability to analyze as much data as possible, with optimal flexibility and agility. Data users demand to run any query, whenever they need it. In this case, using data warehouse solutions will not deliver the needed agility and flexibility. The ability to quickly transition to a data lake architecture will deliver these benefits as well as a strategic competitive advantage.

Unlocking more powerful insights from data analytics is at the center of the data lake architecture paradigm shift. The ongoing demand for agile, more flexible data analytics to leverage big data investments has fueled the rise of data lakes and distributed SQL query engines like Presto and Trino. The power of data lakes to hold vast amounts of raw data in native formats until needed by the business, combined with the agility and flexibility of distributed engines in querying that data, promises organizations the ability to maximize data-driven growth.

Although they have bought into the analytics promise of data lake architecture along with its ability to provide cost effectiveness and efficiency, many organizations have yet to unlock the power of data lake architecture. Instead, they are utilizing it as little more than an aggregative storage layer.

The main value organizations derive from the data lake stack has three aspects:

1. It enables instant ease of access to their wealth of data, regardless of where it resides, with near zero time-to-market (no need for IT or data teams to prepare or move data).

2. It creates a pervasive, data-driven culture.

3. It transforms data into the digital intelligence that is a prerequisite for achieving a competitive advantage in today's data-driven ecosystem.

To create a modern data lake architecture that maximizes ROI, forward-looking data organizations are leveraging new data virtualization, automation and acceleration strategies and reaping the benefits. How can data organizations ensure that their modern data lake stack is analytics-ready? The following are key questions that data organizations need to ask themselves to ensure they are getting the most out of their big data investments.

How explorable is your data lake?

The biggest advantage of data lakes is flexibility. Allowing the data to remain in its native, raw and granular format means that data is not modeled in advance, transformed in flight, or at target storage. This is an up-to-date stream of data that is available for analysis at any time, for any business purpose. But data lakes only have meaning to an organization's vision when they help solve business problems through data democratization, reuse, and exploration by agile and flexible analytics. The access to the data lake provides a real force multiplier when it is used by companies thoroughly, across business units.

Is your data lake strategy living up to its potential?

Most organizations have the best of intentions to fully leverage the power of their data lake architecture. However, even after a successful implementation, many enterprises use the data lake on the fringes, running queries on a limited basis for ad hoc, high-value queries. Thus, they dramatically fail to use their data lake to its potential—and experience poor ROI as a result. There are several obstacles that prevent organizations from utilizing the power of their data lake stack,

all of which require organizations to rethink their data lake architecture to capitalize on their investment in big data and analytics.

Are you using compute resources effectively?

Research shows that 90 percent of compute resources are "wasted" on full scans. Traditional data lake query engines are based on brute force query processing, culling through all the data to return the result sets needed for application responses or analytics. The result is that SLAs are not sufficient to support interactive use cases andrealistically support only ad hoc analytics or experimental queries. To effectively support a wide range of analytics use cases, data teams have no choice but to revert back to optimized data silos and querying traditional data warehouses. This unnecessary leverage of widely excessive resources runs up significant costs.

Are you minimizing your DataOps and achieving observability?

Today's enterprises need deep and actionable workload-level observability to gain a comprehensive understanding of how resources are allocated among different workloads and users, how and why bottlenecks occur and how to allocate budgets accordingly. The workload perspective enables data teams to uniquely focus engineering efforts on meeting business requirements. To manage data analytics cost and performance efficiently, data teams should look to solutions that autonomously and continuously learn and adapt to users, the queries they're running, and the data being used.

Does your data lake stack have what it takes to be analytics-ready?

Manual query optimization is time-consuming, and backlog optimizations grow every day, creating a vicious cycle that diminishes the agility promise of data lakes. A lack of workload-level observability prevents data teams from identifying which workloads need priority based on business needs rather than on the needs of an individual user or query. Overcoming these obstacles to leveraging the power of the data lake demands a transition to an analytics-ready data lake stack, which is composed of:

- Scalable and massive storage (petabyte to exabyte scale) such as AWS S3
- Data virtualization layer that provides access to many data sources and formats
- Distributed SQL query engine such as Trino (PrestoSQL) or PrestoDB
- Query acceleration and workload optimization engine for performance and cost balance, to eliminate the disadvantages of brute force approach and their implications

Leveraging these tools, an organization can reap the benefits of agile data lake analytics that harness near-perfect data—with traditional data-warehouse-comparable performance and cost. As a result, a business no longer needs to adapt to existing data architecture, which limits which

queries can be run. Instead, the data architecture adapts itself to specific business needs, which are highly elastic and dynamic.

Are you leveraging indexing technology?

Indexing has been traditionally used in relatively small datasets. Recent innovation expands the usage of indexing to massive amounts of data. Indexing technology eliminates the need for full scans and can accelerate queries automatically without any overhead to query processing or any background data maintenance. This reduces the amount of data scanned by orders of magnitude. As an example, check out this benchmarking data:

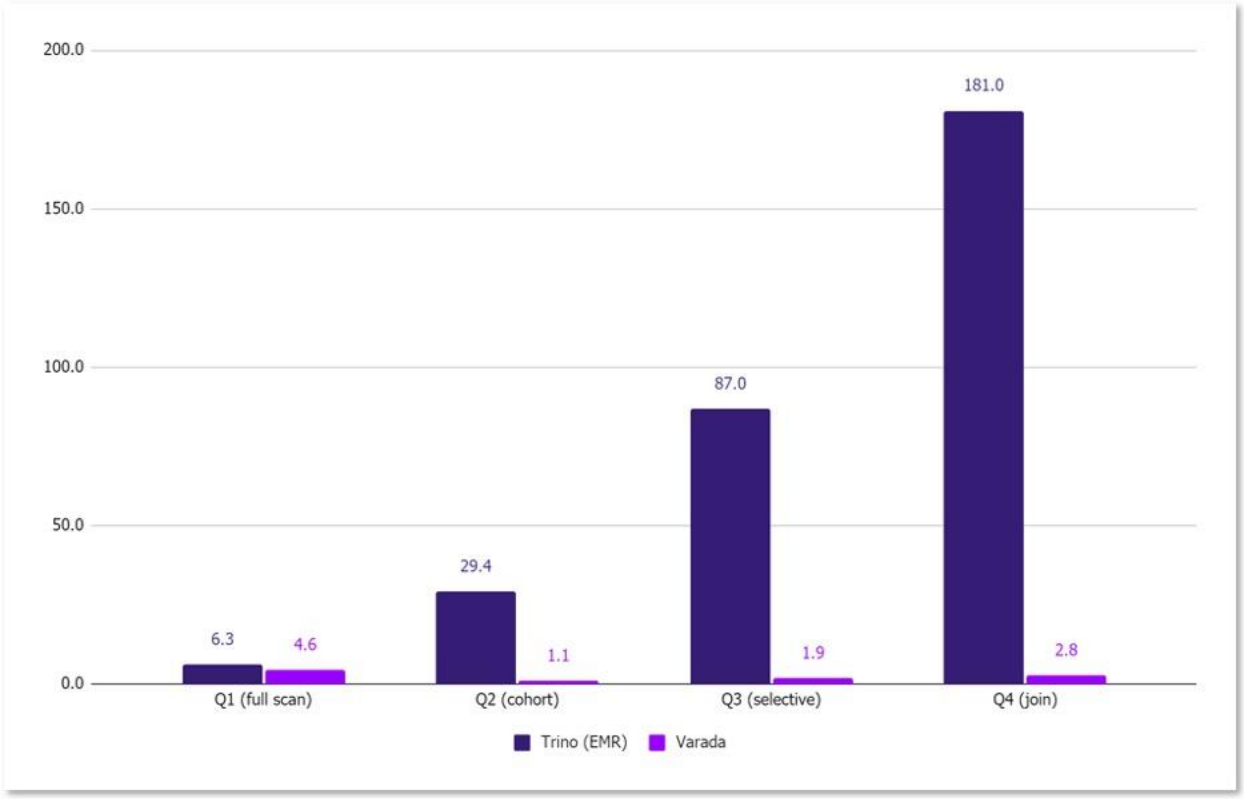

Figure 1: Trino vs. Varada Query Performance (seconds)

click to enlarge

The missing link in the data lake stack

Data lake query acceleration platforms are the missing link in your data lake stack. Sitting on top of your data lake and query engine, they serve as a smart acceleration layer on your data lake, which remains the single source of truth. The data lake becomes the business's mainstream data analytics platform, serving a very wide range of use cases and enabling enterprises to turn it into a strategic competitive advantage and achieve data lake ROI. Data also becomes a strategic asset, as businesses can use it to respond with agility to new opportunities through innovations that drive business growth and competitive advantage.