JUNE 2021

4 Practical Tips for Building a Data Lake With VMware Tanzu Greenplum

These best practices can simplify the process of setting up and scaling a data lake.

NEXT ▶▶

Most large enterprises have already built data lakes, and they are seeing a number of benefits as a result of becoming more data driven. Many have reduced costs, improved customer service, increased sales and margins, and identified new business opportunities that might have otherwise gone unnoticed. As a direct result of integrating more data and analytics into their decision-making processes, organizations are becoming more optimized, more profitable, and more competitive within their markets.

By contrast, some small and midsize enterprises haven't yet fully adopted data lakes. They might have run early experiments that were unsuccessful, or they may have believed that creating a data lake was simply impractical for an organization their size. And some may have thought their current data warehouses and business intelligence applications were sufficient. As a result, they aren't realizing the data lake advantages that their larger counterparts are seeing, and some are even falling behind competitively.

Despite the obvious disadvantages, being slower to deploy a data lake does have one big advantage: These smaller organizations can learn from the experiences of those that have gone before. While the larger early adopters had to pioneer new processes and technologies, often stumbling along the way, these smaller companies have the opportunity to set up data lakes more quickly and smoothly by leveraging the lessons learned by those larger enterprises.

Despite the obvious disadvantages, being slower to deploy a data lake does have one big advantage: These smaller organizations can learn from the experiences of those that have gone before.

Dell Technologies was one of those early data-lake adopters. It built an incredibly fast data lake based on Dell EMC infrastructure, VMware Tanzu Greenplum, Hadoop, and several other technologies. In an effort to help other organizations speed their own data-lake deployment, it has released many of the details related to that project. The case study provides valuable tips that other organizations can use as they build and scale their own data lakes.

Background: The Dell Technologies Story

You may have seen the news that VMware, which is majority owned by Dell Technologies, purchased Pivotal Software and its Greenplum database in 2019. But Dell's involvement with Greenplum preceded that purchase by many years. When Dell was looking for a database for its internal data lake project, it selected Greenplum ahead of other options.

Darryl Smith, chief data platform architect and distinguished engineer at Dell Technologies, explains that Greenplum gives the organization the best performance for the cost. In fact, the Greenplum database is three-times faster than a competing option for one-tenth of the cost.

And speed is really important to Dell. "Our database is currently performing at a rate of a terabyte per second of I/O," Smith says. "This year, we will add another 72 PowerEdge servers, and then we will be running at roughly 2.4 terabytes per second of I/O with a 2.5-petabyte database."

To put that in context, employees accessing the Greenplum database receive the results of their queries in less than a second 98% of the time. Even extremely large queries are incredibly fast. For example, a query on a 7-billion row table takes about seven seconds to process. Impressively, the database is providing that level of service while serving approximately 6 million queries per day.

Your organization may never need to scale to quite that level. But through the process of creating this very large, very fast database, Dell learned four lessons that apply to all organizations building data lakes, no matter their size.

Tip 1. Enabling Self-Service Can Revolutionize Your Business.

For Dell Technologies, one of the most transformative aspects of the data lake was that it empowered staff in the different business departments to meet their own needs. "Originally the problem we wanted to solve was that IT does not have the capacity to keep up with the analytical requirements of the business," says Smith. "It was really all about data democratization and getting the data into the hands of the business to be able to experiment with it on their own without having to wait for IT to build them a report."

Adopting technologies that made it possible for business users to do analytics on their own lifted a burden off the IT team while delivering valuable, actionable insights to the business team. As their needs changed, the business users could quickly create new reports that answered new questions and helped them find new opportunities.

These early efforts quickly grew as other departments and other regions of the world began to see what was possible with the data lake. Today, Dell employees from the Americas, Europe, the Middle East, Asia, and other areas all rely on the Greenplum database. "We've got probably 100 or more different analytical apps that are running currently in the data lake and on our Greenplum platform," notes Smith.

And those applications are generating real savings and revenue for the company. For example, the Dell Customer Account Lifecycle Management application, which streamlines service-contract renewals, is generating an additional $200 million per year thanks to the data lake.

In addition, a predictive maintenance application uses data from the Greenplum database to determine the likelihood that hardware in Dell's data centers will fail. By using that data to help them proactively replace drives before they fail, the company saved more than $100 million.

The data lake also directly benefits Dell customers. The company's MyService360 customer service application and its Dell EMC CloudIQ monitoring solution both rely on the data lake as well. The data lake allows Dell to provide valuable services to its customers very quickly, which in turn, improves customer satisfaction and leads to more repeat business.

And those are just a few of the hundreds of different applications all being fed by the data lake.

Tip 2. You Need More Than One Technology.

Hadoop has become so closely associated with data lakes that some organizations wrongly assume that Hadoop is all they need. The truth is that while Hadoop is very good at some things, it is not great at everything that you want a data lake to do.

Smith says he has seen some organizations try to do all their analytics on Hadoop, but analytics processing can take up to 24 hours because Hadoop doesn't deal well with structured data. "You really should pick the technology that is going to satisfy your use case."

He adds, "You can't build a data lake, you can't do data analytics, with one technology. So, use the right tools for the job."

In Dell's case, the right tools included the Cloudera implementation of Apache Hadoop for unstructured data and the VMware Tanzu Greenplum database for structured data. It also uses Apache Spark and Apache Kafka for data services. And in the execution tier, it relies on a host of different databases, including PostgreSQL, MongoDB, Cassandra, SingleStore, and Neo4j.

Your data lake might not need all these different solutions, but you should not expect that a single database will be able to meet your needs adequately. You don't want to lock yourself into one technology that can't meet all your needs.

Tip 3. You Need the Right Hardware.

Of course, the right software is only part of the solution. You also need the right hardware to create and scale your data lake.

The Greenplum architecture incudes Dell EMC PowerEdge R740xd vSAN Ready Nodes featuring Intel Xeon scalable processors for compute and Dell EMC PowerScale family Isilon H500 storage.

During its 10-year journey of building a data lake, Dell Technologies has added many more nodes and upgraded them as faster technology became available. "As we grew and added in more use cases, not only did we add more storage, which was necessary, and more memory and CPU, we focused on how we could lower the latency of the I/O so that our queries could run a lot faster." During this part of its journey, Dell added in NVMe, which improved performance about 15x. It was "crazy fast," says Smith.

Of course, Dell also needed a fast network to support queries coming in from all over the world. "High-speed networking is very important," Smith notes. "Basically, we have 128 ports all capable of 100 Gbit."

With this infrastructure in place, Dell Technologies had the right combination of technologies to meet its needs then and to evolve into the future. The company's data lake continues to grow as business units find new use cases for it. To keep up, the IT team continues to add nodes and make improvements, and that investment has yielded big financial results for the company.

Tip 4. Start Small and Grow.

The size and performance of Dell Technologies' data lake may seem overwhelming if your organization doesn't yet have a fully functional data lake. But it's important to remember that Dell started very small.

"The project started back in 2010, and the goal was to solve a few use cases," Smith explains. "This allowed our business units to do things they could never do before."

As word spread about what the data lake could do, more business units expressed interest in trying out the data lake for themselves. Smith says, "We set up self-service workspaces to allow people to really do proof of concepts in terms of: 'What can this data do for me? What kind of insights could I get?'"

When those initial experiments yielded good results, the data lake began to grow and expand. The team added more hardware and data solutions, always matching the technologies they chose to the needs of each particular use case.

Based on Dell's experience, Smith recommends that other organizations follow a similar path. The risks and effort required for a small project are relatively low. By picking one good use case and proving it out, your team can quickly gain experience, much like Dell did. With scale-out hardware and the right software in place, you can easily expand your data lake over time.

It's Easier Than You Think

Data lakes can be very complex, but fortunately, you don't have to build yours from scratch.

Dell Technologies and Greenplum have published a reference architecture so that you don't have to figure out which pieces and parts you will need to make the data lake functional. Dell Technologies has also made public many of the best practices it discovered with its own data lake deployment. Following those best practices can ease the process of building and scaling your own data lake.

The sooner you get started, the sooner you can start enjoying the benefits of self-service analytics fueled by a very fast data lake.

Product Spotlight: Dell Technologies VMware Tanzu Greenplum Reference Architecture

Deploying the Greenplum reference architecture offers many benefits for small and midsize enterprises. It can help reduce data silos, integrating information across your organization so that everyone is working from the same trusted dataset. That can speed data-science initiatives, providing the data you need to fuel advanced analytics and machine-learning applications.

On the technology side, the architecture offers outstanding total cost of ownership. It leverages open-source software and industry-standard hardware to keep costs low.

And at the same time, the solution offers very, very fast performance. It relies on high-performance Intel® Xeon® Scalable processors and Dell EMC PowerEdge servers to meet the highest levels of demand. It's also flexible and scalable. The solution features simple building blocks in several different configurations to meet various use cases. Organizations can choose the blocks that meet their unique needs and quickly add nodes when they need to scale.

Dell Technologies Greenplum Architecture Blocks

	Fast Block	Balanced Block	Dense Block	Super-Dense Block
Best for:	Fast performance for concurrent workloads	Ideal ratio of price to performance	Data lakes with large storage needs	Data lakes with extremely large storage needs
CPU	96 cores	96 cores	96 cores	96 cores
RAM	1.5 TB	1.5 TB	1.5 TB	1.5 TB
Primary	19.2 TB	38.4 TB	76.8 TB	153.6 TB
Mirror	19.2 TB	38.4 TB	76.8 TB	153.6 TB
Temp	12.8 TB	12.8 TB	25.6 TB	51.2 TB
Full Rack Raw Primary	134.4 TB	268.8 TB	537.6 TB	1.072 PB

About Dell Technologies

No matter how basic or advanced your current data-management practice, we're here to help you leverage the full potential of your organization's data. With Dell Technologies Consulting Services, Professional Services, and our extensive partner network, we work with companies at all stages of data maturity to plan, implement, and optimize the people, process, and technology changes needed to unlock your data capital and support advanced technologies, like artificial intelligence and machine learning.

See ways you can innovate with data.

Accelerate Data Science & AI Pipelines

The Intel® oneAPI AI Analytics Toolkit gives data scientists, artificial intelligence (AI) developers, and researchers familiar Python tools and frameworks to accelerate end-to-end data science and analytics pipelines on Intel® architectures. The components are built using oneAPI libraries for low-level compute optimizations. This toolkit maximizes performance, from preprocessing through machine learning, and provides interoperability for efficient model development. Learn more and download.