DELL Technologies intel

Analytics and AI in a massive data lake

The Dell Technologies IT organization delivers data-driven business insights to employees and customers around the world.

DELL Technologies Data Analytics | United States

Business needs

The Dell Technologies IT organization needs a fast, highly scalable data lake and analytics environment to deliver insights to users around the world.

Solutions at a glance

- VMware® Tanzu™ Greenplum® database
- Cloudera® distribution of Apache™ Hadoop®
- Dell EMC PowerEdge servers with Intel® Xeon® processors and NVMe memory
- Dell PowerSwitch networking
- Dell EMC PowerScale and PowerFlex data storage

Business results

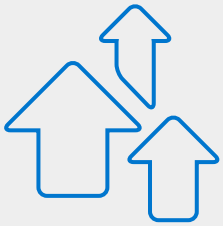
- Driving operational efficiencies and cost savings
- Improving the customer experience
- Increasing revenue though data-driven sales and marketing
- Showcasing the impact of digital transformation

In the Dell IT data lake, an analytics query on a 7-billion-row table can come back in

7 seconds
or less

With a planned upgrade, the analytics environment will move data at a rate of

2.5 TB
per second



An enterprise data lake

As a global technology company with approximately 165,000 employees, Dell Technologies generates an enormous amount of data on a daily basis. To help the company and its customers gain value from this deluge of bits and bytes, the Dell IT organization manages a massive data lake and a world-class set of tools for data analytics, machine learning, deep learning and artificial intelligence.

At the heart of this data environment is a Greenplum database, an open-source massively parallel data platform for structured data analytics, machine learning and AI. VMware Tanzu Greenplum is based on the PostgreSQL relational database management system and the open source Greenplum Database project. This leading-edge platform allows the Dell IT organization and users from around the world to rapidly create and deploy models for complex applications, from sales and marketing to predictive maintenance, cybersecurity and more.

This same data environment also includes an Apache Hadoop platform from Cloudera. The Cloudera distribution of Hadoop enables the Dell IT organization to store and process enormous amounts of unstructured and semi-structured data from a wide variety of sources. In a typical use case, this raw data gets parsed in Hadoop into a structured format, and that structured data then gets pumped into the Greenplum database, so business and IT users can consume it in analytics applications.

As for consumption of data, the Dell IT team uses several best-of-breed technologies in the execution tier of the data lake. These technologies enable more efficient use of subsets of the larger pool of data for specific types of analysis, along with faster response times compared to running queries on larger sets of data. These technologies include:

- **PostgreSQL** — the number one open source database that is extremely capable of handling many tasks very efficiently (or managing concurrency)
- **MongoDB** — an in-memory document database
- **Cassandra** — a database that is particularly good for querying large amounts of structured or semi-structured data, such as sensor logs
- **SingleStore** — an in-memory database for both row and column data that provides very low latency responses to structured data
- **Neo4j** — an open-source, NoSQL, native graph database

Enabling digital transformation use cases

In the mid-2010s, the Dell IT organization chose Greenplum as the basis for the company's data lake. In an indicator of the platform's success, this database quickly grew to hundreds of terabytes of data accessed by users from around the world.

The use cases for the analytics database are all over the map — literally. The data is used by Dell Technologies employees and customers in the Americas, Europe, the Middle East, Asia and other geographic regions, according to Darryl Smith, chief data platform architect and distinguished engineer at Dell Technologies.

"Marketing is one of our biggest customers," Smith says. "They have about 460 terabytes of their own data within the database. And in this case, IT is working in conjunction with marketing — we do some of the work and they do some of the work. But most of the analytics itself is developed from the marketing business unit."

In one such initiative, the marketing team worked with IT to develop an analytics solution that uses predictive modeling to help the sales team leverage customer data to offer the right products and the best experiences to individual customers — all at the right time. This solution, known as the Customer Engagement Platform (CEP), supports direct mail and email campaigns, as well as customer calls from sales reps.

Another big user of the system is the Dell Technologies Digital Supply Chain organization. It runs data analytics focused on making the company's supply chain as efficient as possible by tracking quality defects, monitoring vendor parts availability and more.

"We've also got customer-facing apps that have analytics running in our Greenplum database," Smith says. These include a customer support application called MyService360 and the Dell EMC CloudIQ monitoring solution.

A look at the Dell IT analytics environment

For an inside look at the Dell IT organization's data analytics journey and its experiences with the VMware Tanzu Greenplum database, watch the video "Greenplum Your Way: Why Dell IT Continues to Choose Greenplum." This video includes details on the most recent upgrade to the analytics environment, which leverages the latest version of Greenplum combined with the fastest NVMe Dell EMC hardware available. This upgrade led to a 15X performance boost from the already-fast analytics system.

Cutting costs with predictive maintenance

Dell realized significant cost savings from the use of a predictive maintenance application run in the Dell IT data lake. This application everages data from sensors on connected devices, lab test results, and vendor technical specs to determine the likelihood of a system drive failure based on factors ranging from usage to data center environmental conditions. By proactively replacing drives, Dell avoided more than $100 million in costs associated with repeatedly bringing technicians into its data centers for one-off fixes on failed drives.

MyService360 serves as the customer's one-stop-shop for services information, including personalized data and actionable insights forDell Technologies products. CloudIQ, in turn, provides cloud-based monitoring and advanced analytics for data storage resources. It combines machine intelligence and human intelligence to give storage administrators the insights they need to take quick actions and more efficiently manage their data environments.

"We've got probably 100 or more different analytical apps that are running currently in the data lake and on our Greenplum platform," Smith says.

Amazing results

The Dell IT data lake is continually evolving, in terms of technology, performance and capacity. With recent upgrades, the environment is now fully running Dell EMC PowerEdge servers with NVMe drives with local disk. Each of the NVMe drives is capable of 2 gigabytes-per-second throughput — which is more than twice as fast as SSD drives and roughly 10 times faster than spinning disk.

"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."

With its rich mix of technologies, the data lake environment delivers some amazing results. An analytics query on a 7-billion-row table might come back in 7 seconds or less, while 98 percent of the queries on the Greenplum database come back in less than a second.

"And we typically run at least 6 million queries per day, so it is a pretty busy database," Smith says.

While the performance numbers are stunning, they don't begin to tell the full story of the business impact of the VMware Tanzu Greenplum analytics environment. The use of powerful analytics is helping Dell Technologies achieve better customer service, improve customer satisfaction, reduce costs, increase operational efficiency and enhance security, Smith says.

"It doesn't matter what you want to get out of it — you can achieve any one of those business goals with the right data analytics and artificial intelligence tools," Smith says.

One of Smith's favorite examples of the data-driven results from the analytic environment stems from the Dell Customer Account Lifecycle Management application, which streamlines the management of service-contract renewals.

"Since we implemented that application, we've been pulling in an additional $200 million a year," Smith says. "We're not wasting our time calling people who don't want to renew their service contracts, and we're calling people we never would have thought of who actually buy it. And so the improved efficiencies led us to bring in more dollars."

Another example is better customer satisfaction numbers stemming from the analytics-driven Dell EMC CloudIQ Platform. This application gives users of Dell EMC Unity storage arrays the ability to see performance metrics and get health checks on their systems, so they can find and fix any performance bottlenecks.

"And so as a customer, I'm buying a better product," Smith says. "I've got better management of my storage arrays. And that leads to better customer satisfaction numbers for Dell Technologies."

Learn more about Dell EMC advanced computing

Unlock the value of data with artificial intelligence

Share this story