BUSINESS GROUP

VMware business units across the enterprise.

KEY CHALLENGES

Deliver a scalable, cost-effective way for VMware business units to create their own data analyses while eliminating multiple, separate database instances for greater efficiency in production.

SOLUTION

Consolidate 7 Oracle data marts into one, federated data mart in the cloud using Greenplum from Pivotal.

Business Transformation through IT Transformation: VMware IT Tackles Big Data in the Cloud

Big data is here to stay. Gaining actionable business insight out of all that data presents huge opportunities for companies and poses unique challenges for IT. VMware IT was no exception as it sought to quickly deliver business information to users, at scale, with a minimum of IT overhead. VMware's business units had multiple independent data marts delivering custom reports in a "hub and spoke" fashion that siloed information and consumed resources. This model would not be sustainable as VMware continued to rapidly grow.

The Challenge

"VMware's growth demanded that we shift from a 'hub and spoke' approach to a shared enterprise data model," said David McMath, IT director and executive program sponsor for the Magellan initiative. "This was necessary to ensure users had timely access to business information at scale, but also because there was no way IT would be able to keep up as we grew."

VMware IT evaluated options for a federated data mart that would scale with user needs and meet big data requirements. They needed to be able to serve both operational and analytical needs of multiple departments worldwide. For economy, agility and because VMware is committed to the cloud, IT considered migrating its existing Oracle instances but found it would require the deployment of Oracle RAC, which meant dramatically higher cost and complexity. Furthermore, the Oracle database is an SMTP database designed fundamentally for OLTP queries; increasingly the user demand was for OLAP queries, so IT examined newer technologies purpose-built for analytics and data warehousing.

"Users had expressed dissatisfaction with the 'organically grown' Oracle data marts as it was," said McMath. "We knew that as we get more into the need for big data analysis, reporting delays would only grow if not addressed at a fundamental level."

The Solution

The decision was made to evolve to a centralized electronic data warehouse and to move to a database that used a massively parallel processing (MPP) database to better handle unstructured data queries.

Enter Pivotal Greenplum

Pivotal's Greenplum rose to the top of the list for several reasons. It is a purpose-built, dedicated analytic data warehouse designed to extract value from business data.

vmware®

"Now our business users can add their own data and create their own analysis without IT assistance."

David McMath
IT Director,
VMware

KEY TAKEAWAY

First cloud instance of Greenplum delivers capability for big data analysis and a "single source of truth."

Pivotal Greenplum Database manages, stores and analyzes up to petabytes of data in large-scale analytic data warehouses. VMware owns 31 percent of Pivotal, resulting from the March 2014 acquisition of Pivotal by VMware parent EMC. Greenplum uses an MPP architecture and flexible column-and row-oriented storage. It also leverages fully parallel communication with external databases and Hadoop to continually harness data (Apache™ Hadoop® is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers).

The problem: Greenplum hadn't been proven in the cloud. Yet.

In the old model, the extract, transfer and load (ETL) process was taking at least 24 hours and sometimes as many as 28. To improve efficiency, IT would load two batches in staggered intervals, but it was cumbersome, time-consuming and not a great use of smart people. Today, Greenplum manages that process for 6,000 users, up to 200 concurrently, in about 7 hours.

The old model required "two hops," as McMath puts it: first each group's requested data had to be pumped out to its data mart, then the reporting process took place in that "silo." Overall, IT has consolidated 7 data marts into one.

"Now you don't have to reinvent the wheel, with finance doing one report, marketing doing another," he explained. "No more 'shadow data marts."

This led to the standardization of data and the elimination of more than 600 duplicate data elements, McMath added. Moreover, the burden on IT has been greatly reduced and won't increase as the load rises: today they are realizing 100 percent faster load times, and the job of copying data for specific business units has been eliminated.

Benefits

- Use case for Greenplum in a private cloud proves to be less costly to purchase and operate.
- Slash ETL times by more than two-thirds, from 28 hours to 7.
- Provide more up-to-date data to analysts.
- Eliminate "shadow data marts" and support better business insight from big data through a federated data mart and a "single source of truth."
- Eliminate costly IT overhead involved in managing custom reports and loads.
- Delight users with a self-service "sandbox" where they can create their own data sets and perform analyses with no impact on production.
- Achieve 100 percent improvement in data delivery to the desktop.
- Improve scalability by 500 percent by leveraging the linear scalability of VMware's platform.

Self-Serve a Boost for Users and IT

The real breakthrough in data delivery, however, comes in the form of self-service. This means data on-demand for users with no IT overhead. IT has enabled users to work with data in a "self-service sandbox," which means they can "play around" without impacting production.

"That may sound like we did it for fun, but what self-service sandboxing really means is that business professionals can easily import data and ask different questions to gain the insight needed for today's problems," said McMath. "Consider the impact of social data, for example. It's one thing to work with Census data and inquire how many people over a certain income live in a certain geography; it's quite another to cross-reference Twitter feeds with Facebook comments to gain insight into how customers feel about your brand by analyzing syntax, semantics or other variables."

Simply put, big data, social information and other trends are making it possible — and necessary — to ask different questions of business data. Cloud-based data warehousing delivers that functionality at scale while minimizing cost and IT overhead. Greenplum in the cloud is 500 percent more scalable than the traditional hardware-based digital computing appliance (DCA), which would have required a significant hardware upgrade of an additional DCA, McMath noted. By acting independently from production workloads, business users can self-serve to custom mixes of data that allow them creative new ways to test out hunches and mine for "business gold."

"Now our business users can add their own data and create their own analysis without IT assistance," said McMath.

IT has delivered Greenplum to 6,370 users with 50 – 200 of them concurrent worldwide at any one time. Cloud deployment supports cost-effective daily data refreshes and 100 percent faster delivery to the desktop. The scope of the project included 1,750 tables, more than 43,000 columns of data and 108 total terabytes.

"The benefits of this architecture to our business are transformational," McMath said. "We now have a single source of truth, with consolidated DSS data marts into a single enterprise data warehouse. We've cut load times in half and have improved report rendering by over 15 percent. That means faster answers to business questions."

"The so-called 'Internet of Everything' is going to depend on the ability to handle unstructured data," McMath continued, citing that Hadoop support was an essential criterion in selecting Greenplum.

Deployment and proof of concept presented some challenges, of course. Greenplum's load times (ETL) were 3 times faster right out of the box, but report times doubled from the previous solution. IT decided to go with two clusters to eliminate overlap and found that indexing the data gave a big boost in performance, according to McMath.

Internally dubbed the "Magellan" project, this big data initiative was a voyage into unchartered waters and is now the world's largest Greenplum data warehouse in the cloud, noted McMath.

"It took Ferdinand Magellan three years to circumnavigate the world," said McMath. "Our Magellan team took just 9 months to circumnavigate and consolidate VMware's enterprise data, and I'm really proud of them."

VMWARE ON VMWARE

As the leading proponent of our own products, VMware is committed to passing on the lessons learned by our internal IT group in applying virtualization and cloud management technology to solve business challenges.

vmware®

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com