GREENPLUM DCA DELIVERS PETASCALE PERFORMANCE TO MEDIAMIND

> "Thanks to our data warehousing project with Greenplum, we can now offer new products to our customers. We can offer them better value based on the larger volumes of data that we are able to analyze."
>
> SAYS EFI COHEN, VICE-PRESIDENT TECHNOLOGY AT MEDIAMIND.

ESSENTIALS

Industry
Media

Business Challenge

- Find a scaleable architecture
- Simplify the data process flow
- Obtain greater flexibility in the data analysis

Solutions

- Greenplum Data Computing Appliance
- Data Domain Back-up

Results

- Architecture scales linearly
- MediaMind can offer new products to its customers

MediaMind recently replaced its existing data warehousing solution with Greenplum's Data Computing Appliance, to ensure scalability for the big data volumes to come.

MediaMind (a division of DG) is a global provider of digital advertising solutions that optimize the use of media, creative content and data for enhanced campaign performance. MediaMind captures and analyzes the entire flow of what happens with online advertising: from the number of page views and click-throughs to the conversion rates. In 2010, MediaMind realized that the volume of data collected was growing year over year, outstretching the capacity and performance of its existing platforms.

"The amount of data we analyzed was doubling every year," Cohen explained. "We realized that our in-house solutions were no longer able to scale up and would not be able to process the volume of data predicted for 2012. So we had to double the number of servers, the number of cores, the volume of storage,.... By then, the cost associated with this was really getting out of hand." Besides, MediaMind's existing solution was only able to process several terabytes of data and could not scale into the petabyte range. To give an example of the size of the data: the largest table holds 90 days of ad impressions, and almost 350 billion rows. Compressed the database is 24 TB, the equivalent of over 100 TB of raw data.

While defining the new project, MediaMind also wanted to simplify the data process flow and obtain greater flexibility in the data analysis. "The old system only stored aggregated data, so we only stored what we knew we were going to need for the analysis. It was impossible to cater to the needs of customers who asked for customized measurements," said Cohen. "So, in the new system, we wanted to be able to store all the raw data instead of the aggregated data, thus enhancing the granularity of our analyses." Cohen calls the data MediaMind captures 'semi-structured': there's structured data like the number of ad impressions, conversion rates... but also unstructured data like URLs.

Wanted: MPP

From the onset of the project, MediaMind knew it needed a massively parallel processing system and the company made a long list of potential providers, eventually narrowing it down to just three vendors who were asked to perform a Proof of Concept. To establish the shortlist, MediaMind looked at crucial aspects, such as the track record of the companies, the maturity of their solution, their pricing and service level agreements. MediaMind also performed a large number of reference calls to get feedback from existing clients.

To drive down the operational cost of the new solution, MediaMind only invited companies to the Proof of Concept whose technology and methodology fitted with the experience of its staff. "We had always been working with relational databases and SQL, and having to retrain everyone completely was something we wanted to avoid," argued Cohen.

Real-life data for Proof of Concept

For the Proof of Concept, the three contestants were given a large set of raw data – 50 TB, worth 90 days of activity - on which they had to test a number of real-life use cases so that they could simulate the actual queries MediaMind usually runs. “Essentially, we wanted to see resource utilization and response times on different sets of data. We wanted to test the linear scalability.” The participants in the PoC first needed to test on a week’s worth of data, then on two weeks, then on four.… To level the playing field, all three participants had to use equivalently priced hardware platforms.

The Proof of Concept came up with a clear winner: Greenplum’s Data Computing Appliance. “Their performance was extremely good,” testified Cohen. “We were also extremely pleased with the team that worked on the project. We are a long-time customer of EMC, so we know what to expect.” Another thing that tipped the balance in favor of Greenplum was the conceptual difference with one of the other contestants: the other solution was more of a black box that had everything pre-tuned. “A good idea in itself, but we wanted to perform further fine-tuning and parameterization. Greenplum allows that in the way that we were used to working with standard databases. The system is very flexible. I liked that approach.”

Core of data infrastructure

Once implementation of the Greenplum system had started, a lot of attention was given to training MediaMind-personnel on the correct use of Massively Parallel Processing. “MPP demands a completely different architecture, and Greenplum provided the necessary resources to train our database administrators. It took us a while to realize that we had to change our way of working and thinking about such projects because MPP is so radically different. The main challenge was really shifting our mindset.”

Over the course of the project a number of minor technical, hardware-related issues had to be resolved, but Cohen is very upbeat about the experience of partnering with Greenplum and EMC, who brought in several specialists from different countries to assist MediaMind.

Since April 2012, the new solution has been in full production. “We have shut down the old system. We are quite pleased with the data availability, the uptime, the SLAs… the Greenplum DCA really meet our needs.”

Greenplum is at the core of MediaMind’s data infrastructure. The solution consists of one and a half racks of DCA for the production system (with 6 standard DCA modules, plus another rack for disaster recovery installed at a different location (with 4 standard DCA modules). The production cluster is backed up using EMC Data Domain which integrates nicely with GP leveraging advanced de-dup algorithms.

Because of its mission-critical nature, ETL (Extract, Load and Transform) processes are running 24/7, with zero downtime, in parallel to routine maintenance operations (such as backup, DB maintenance operations which are performed online).

Business value

When it comes to business value, Cohen points out that data is now available quicker, while it is also possible to perform more detailed queries. “This is a very important project for us. It allows us to build new product offerings based around the extra analyses we can perform on our data. We can now offer products to customers that really differentiate us from our competitors. In a second phase, we are going to look at adding advanced analytical capabilities.”

CONTACT US

To learn more about how EMC Greenplum products, services and solutions can help you realize big data opportunities visit us at www.greenplum.com.

EMC Greenplum
Division Headquarters
1900 South Norfolk Street San Mateo, CA 94403
Tel: 650-286-8012
www.greenplum.com