NetApp Improves Customer Support by Deploying Cloudera Enterprise

Overview

NetApp creates storage systems and software that help customers around the world store, manage, protect, and retain one of their most precious assets: data. To better support its customers, NetApp offers AutoSupport, an integrated and efficient monitoring and reporting technology that constantly checks the health of NetApp systems. Customers leverage the My AutoSupport portal on the NetApp Support site for proactive systems management capabilities and insight into storage configuration, capacity, utilization, efficiency, and health-check information.

Business Challenge: Processing Massive Volumes of Data

AutoSupport collects over 600,000 data transactions weekly, consisting of unstructured logs and system diagnostic information. Approximately 40% of that data is transmitted during an 18-hour period each weekend, creating the potential for I/O bottlenecks that could affect service-level agreement (SLA) windows.

As the NetApp customer base is expanding, AutoSupport data is growing at approximately 7 TB per month. Related storage requirements are doubling every 16 months. NetApp's AutoSupport team proactively identified the need to upgrade its storage environment in order to accommodate continued growth.

NetApp's CIO Cynthia Stoddard explained, "I sit on thousands of customers' data and what I do with that data is essential to the company. I need to react and help customers do more with their systems."

AutoSupport needed a Big Data storage and analytic solution that would allow it to: store, manage, and analyze increasing volumes of unstructured data; gain insights from these large, complex datasets; and scale for continued growth.

"NetApp AutoSupport data is extremely valuable to us," said Marty Mayer, Director, NetApp AutoSupport. "Our customers depend on us to utilize the data to respond in a timely manner when potential problems and issues arise. Additionally, we actively analyze customer storage system data for fitness and health checks that help optimize the investment our customers have made in NetApp."

The Solution: NetApp Open Solution for Hadoop

The team focused on storage solutions that support the Apache Hadoop open-source software designed for data-intensive distributed applications. Other criteria included scalability, high performance, and rich analytics capabilities. The AutoSupport group conducted a proof of concept on numerous technologies, evaluating them for key functions including: parsing; extract, transform, and load (ETL) capabilities; and data warehousing. The team concluded that the NetApp Open Solution for Hadoop—comprised of NetApp storage coupled with Cloudera Enterprise—surpassed the other solutions in providing the ability to more deeply monitor customer solutions, and executing previously impossible data processing jobs and complex queries. The solution also provided an overall lower total cost of ownership (TCO) than other big data platforms.

NetApp™

Industry

Data Storage

Location

Global

Technologies in Use

- **Hadoop Platform:** NetApp Open Solution for Hadoop
- **Storage Systems:** NetApp E2600, FAS2040 HA
- **Operating System:** Data ONTAP 8.0
- **Protocols:** NFS, SAS

Business Applications Supported

- AutoSupport operations: monitoring, troubleshooting & health checks of customer storage systems

Big Data Scale

- 28-node Cloudera Enterprise cluster
- 600,000+ incoming data transactions processed weekly
- Data volumes growing 7 TB per month
- Storage requirements doubling every 16 months

Hadoop Impact

- Database query on 24 billion records reduced from 4 weeks to less than 10.5 hours
- Previously impossible database [KB1] query on 240 billion records runs in less than 18 hours

NetApp and Cloudera joined forces to offer organizations a solution that is highly scalable with enterprise storage features that improve reliability and performance and reduce costs. NetApp Open Solution for Hadoop customers can leverage the big data platform to accelerate adoption of analytic applications that deliver real-time results across intense data and computational workloads. Cloudera is a major enabler for the enterprise adoption and production use of Apache Hadoop. Cloudera Enterprise allows companies to manage the complete operational lifecycle of their Apache Hadoop systems with deep visibility into their CDH clusters. It also automates the ongoing system changes needed to maintain and improve the quality of operations.

System Architecture

The NetApp AutoSupport team deployed NetApp Open Solution for Hadoop, which includes a 28-node cluster of Cloudera Enterprise on four NetApp E2600 storage systems and a NetApp FAS2040 system.

"The NetApp Open Solution for Hadoop system offers us the scalability and flexibility we need to effectively support our growing client base and rapidly expanding data stores," says Mayer. "In addition, because the NetApp system addresses our parsing, ETL, and data warehousing needs in a single, comprehensive solution, it reduces our total cost of ownership, freeing up budget for other customer-focused projects."

The system offers high availability and high performance for even the most demanding AutoSupport workloads. Its balanced performance will sustain the high read-and-write throughput requirements of the system's data-intensive, high-bandwidth applications such as the weekend reporting that offers visibility into the health of hundreds of thousands of customer storage systems.

As a customer-facing organization, the NetApp AutoSupport team depends on the reliability of its storage environment 24x7x365. Data ONTAP® 8 on the FAS2040 storage system eliminates the single point of failure common in traditional Hadoop clustered deployments and instead offers full redundancy and automated path failover, along with online administration.

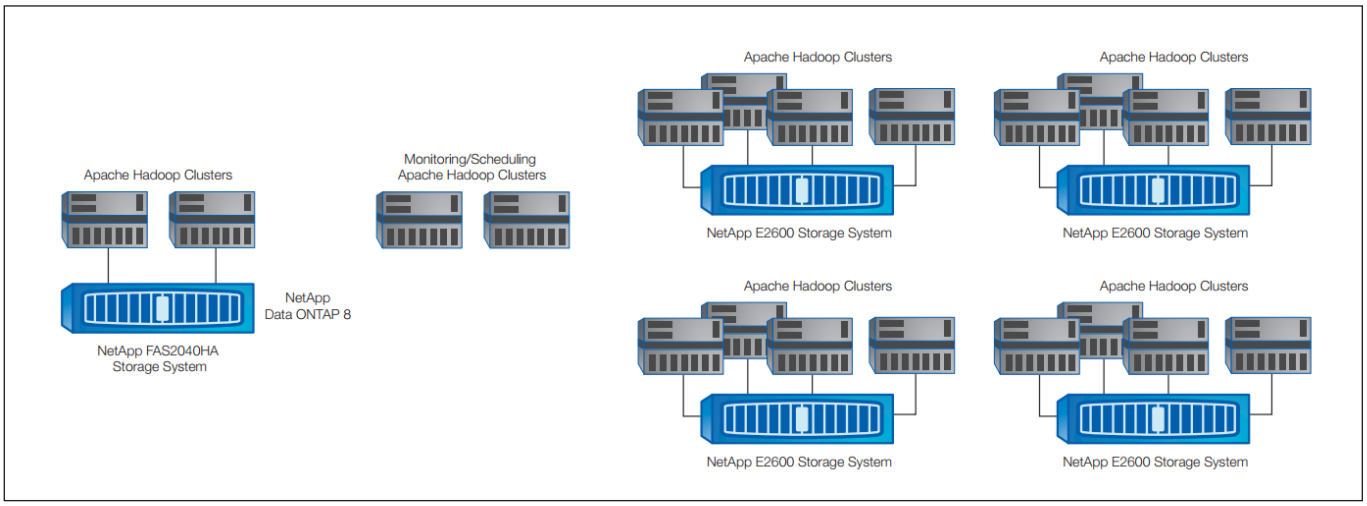

Figure 1) The NetApp AutoSupport team deployed a NetApp Open Solution for Hadoop, which includes Hadoop clusters on NetApp E2600 storage systems and a NetApp FAS2040 system running Data ONTAP 8.

Impact: High Performance for Big Bandwidth Applications

AutoSupport predicts significant performance gains running on the NetApp Open Solution for Hadoop. By supporting even the most bandwidth-intensive applications, the solution will enable the NetApp AutoSupport team to help meet stringent SLAs for parsing and loading data. In one case, AutoSupport wanted to correlate disk latency when a disk was hot with the type of manufacturer disk in order to identify whether there was a

relationship between the two. The report requires a query of 24 billion records, which took four weeks to run on the incumbent environment. On the massively parallel NetApp Open Solution for Hadoop, that query returns in 10.5 hours. **That's a 64x query performance improvement.**

"The productivity and customer service benefits enabled by the NetApp solution are significant," says Mayer. "Running the NetApp Open Solution for Hadoop gives us the ability to turn an unwieldy data explosion into a highly manageable environment. It also will allow us to perform deeper analytics than before, which will provide better monitoring and troubleshooting of NetApp customer storage systems."

In another case, AutoSupport was unable to run a pattern matching query that would help detect bugs. The report required a query of 240 billion records, which was simply too large to run with the existing infrastructure. Through performance testing on the NetApp Open System for Hadoop, the team was able to run the high-bandwidth query and achieve results in less than 18 hours, using just 10 data nodes.

Hadoop Impact: Rich Analytics for Large Datasets

AutoSupport provides skilled resources focused on supporting NetApp customers through analyzing log data for insight into system health. The NetApp Open Solution for Hadoop system enables parallel processing for structured and unstructured data and allows AutoSupport applications to work with thousands of nodes and petabytes of data. "This highly efficient CDH processing will enable our AutoSupport team to quickly mine hundreds of terabytes of data, for real-time analysis of customer system event and performance data and rapid resolution of any issues," says Mayer.

Weekly AutoSupport logs composed of large, complex datasets offer the team insight into information such as storage capacity trending by customer, country, and other criteria. The team also runs ad-hoc queries such as investigating an error alert across an entire system set to identify the source, allowing the team to take proactive measures to prevent impact to customer systems. In addition, through automatic AutoSupport system analysis of deep pools of data, the NetApp Technical Support team is immediately notified of critical alerts. The NetApp Open Solution for Hadoop is designed to accelerate these and other AutoSupport queries and processes.

NetApp customers can also benefit directly from the high performance of the NetApp Open Solution for Hadoop system, which offers deeper analytics capabilities than before. With more in-depth visibility into configuration and system health data through the My AutoSupport portal, customers can be more proactive and achieve greater efficiencies with their NetApp storage.

Hadoop Impact: Managing a Holistic Storage Environment

The AutoSupport team will efficiently manage its big data environment centrally on the NetApp Open Solution for Hadoop system. The solution will provide high levels of throughput and scalability to meet the growing data demands and intensive performance requirements of AutoSupport applications. Storage managers can efficiently manage huge stores of data in the integrated, unified AutoSupport storage environment, which helps NetApp customers optimize the performance and utilization of their storage solutions.

"The NetApp Open Solution for Hadoop is a holistic rather than point solution," says Kumar Palaniappan, enterprise architect, NetApp. "By design, the NetApp solution offers tight integration between our Hadoop applications and NetApp storage, thereby allowing us to maximize storage utilization and reduce our overall capex and opex. Hadoop analytics will empower us to transform NetApp storage data into business insight for our team and customers."

---

1 **Boulton, Clint. CIO Journal.** June 12, 2012. *NetApp CIO Uses Big Data to Assess Product Performance.* http://blogs.wsj.com/cio/2012/06/12/netapp-cio-uses-big-data-to-assess-product-performance/?mod=google_news_blog