Research University Adopts Innovative Solution for Hadoop Cluster Management

StackIQ

CASE STUDY

Challenge

- Accelerate Hadoop cluster deployment with the goal of building and managing reliable Hadoop clusters
- Enable automated, scalable, repeatable processes for Hadoop deployment, monitoring, and fixes

Solution and Benefits

StackIQ Enterprise Data Hadoop cluster deployment and management software.

- Automated, consistent, dependable deployment and management of Hadoop
- Simplified operation that can be quickly learned by those without extensive experience in Hadoop cluster administration
- Reduced downtime due to fewer configuration errors
- Lower total cost of ownership for Hadoop clusters
- Fast deployment

StackIQ Enterprise Data management software introduces reliable, automated solution for Hadoop clusters running research applications at Johns Hopkins University and reduces deployment time from days to half an hour.

Today, distributed Big Data computing architectures feature flexible, affordable commodity server clusters and many open source software components. This environment has been a boon to university researchers due to its flexibility, vendor independence, and lower cost. At Johns Hopkins University in Maryland, systems administrators have further optimized their environment with a simplified, automated solution from StackIQ for managing Apache Hadoop clusters—StackIQ Enterprise Data—which has made Hadoop deployments of all sizes more reliable, highly flexible, less costly, and much faster. The product provides a massively scalable, open source Hadoop platform for storing, processing, and analyzing large data volumes and also optimizes the management of underlying cluster infrastructures of any size.

Challenge: Avoid Complex and Time-consuming Hadoop Administration for Greater Efficiency, Workload Agility, and Cost Savings

The systems professionals within the Human Language Technology Center of Excellence at Johns Hopkins University support a diverse user base of researchers on the Maryland campus and at other institutions. The research at the Center encompasses a wide array of disciplines, including computer science, engineering, statistics, and linguistics. A 50-node Hadoop cluster provides high-performance computing for research projects related to speech recognition, social media, crowd sourcing, and machine translation. These include charting the spread of disease through text analysis of Twitter posts, and identification of voices through analysis of speaking patterns.

The servers in the 50-node Hadoop cluster are a mixture of IBM and Dell blades running CentOS, which were originally used for running compute cluster applications via Grid Engine, an open source batch queuing system used for accepting, scheduling, dispatching, and managing the remote and distributed execution of large numbers of standalone, parallel, or interactive user jobs. The Center recently decided to add Apache Hadoop for on-demand, distributed, Big Data applications.

"When we looked at adding Hadoop applications to the cluster, including Apache Accumulo (a sorted, distributed key/value store), alongside Grid Engine, the complexities mounted," recalls Scott Roberts, senior systems

administrator at the Human Language Technology Center of Excellence. "We didn't want to have dedicated machines for Grid Engine and Hadoop and we didn't want to spend weeks configuring individual servers with different applications manually."

Administrators at the Center looked at DevOps tools such as Chef and Puppet, which are billed as being capable of automating the configuration and management of infrastructure in cloud computing environments or across the servers in traditional data centers. These tools, which use programming or proprietary scripting languages, rely on high administrator skill levels, are time-consuming to use, and still entail a lot of manual work.

Hearing about the open source Linux cluster provisioning and management solution developed at the San Diego Supercomputer Center at the University of California, the Human Language Technology Center of Excellence's IT manager Benjamin Shayne and Scott Roberts were intrigued. They were drawn to the promise of building Hadoop clusters quickly and easily and then being able to tear them down, all with centralized command and control features. Research into a possible solution led to StackIQ Enterprise Data, which features a wide variety of enterprise-class features and the ability to support the Center's high speed interconnects.

If we had to build a new Hadoop cluster from scratch without StackIQ, it would take up to a week!

Solution: StackIQ Enterprise Data

The Center tested StackIQ Enterprise Data in November 2011 on 16 servers. The proof of concept was a success, with the product providing the tools, processes, and procedures for efficiently deploying and managing Hadoop infrastructure.

StackIQ Enterprise Data gives the Center administrators three key components.

- **The Hortonworks Data Platform,** which makes it easier than ever to integrate Hadoop with existing data architectures and includes Hadoop software including Hadoop Distributed File System (HDFS), MapReduce, Pig, Hive, HBase, and Zookeeper plus open source technologies that make the Hadoop platform more manageable, open, and extensible.
- **StackIQ Hadoop Manager,** which manages the day-to-day operation of the Hadoop software running in the clusters. It allows for Hadoop clusters of all shapes and sizes, including heterogeneous hardware support, parallel disk formatting, and interactive HDFS and MapReduce dashboards.
- **StackIQ Cluster Manager,** which manages all of the software that sits between bare metal and a cluster application such as Hadoop. A dynamic database contains all of the configuration parameters for an entire cluster. This database is used to drive machine configuration, software deployment using a unique Avalanche peer-to-peer installer, management, and monitoring.

"If we had to build a new Hadoop cluster from scratch without StackIQ, it would take up to a week," says Scott Roberts. "With StackIQ Enterprise Data, once the Hadoop configuration is deployed, I just have to network boot the machines to configure the entire cluster. It takes just half an hour to get the servers up and running from bare metal."

StackIQ Enterprise Data was deployed in the Center's 50-node cluster in January 2012. Administrators can deploy multiple applications on the cluster's heterogeneous devices with the software's fully-automated solution based on an easy-to-use Graphical User Interface (GUI) or powerful command line interface (CLI). StackIQ Enterprise Data enables customized configurations, fixes, patches, and changes to individual nodes through a flexible XML framework. A BitTorrent-like installer (Avalanche) simplifies and accelerates large-scale deployments. Configurations can be reused for fast deployment of additional Hadoop clusters.

Scaling and Modifying the Cluster to Meet User Needs On-demand

The Center's Hadoop cluster is growing as more researchers take advantage of its computing power and administrators provide those resources to users with true on-demand need. The cluster now contains 60 hosts, 1328 cores, 4.5 terabytes of RAM, and 26 TB of storage. Applications span analytics software from IBM, the MATLAB fourth-generation programming language from MathWorks, and in-house creations in Java, Python, and Perl.

In March 2012, the Center used StackIQ Enterprise Data's management interface to switch from the default first in/first out job scheduler feature to the Hadoop Fair Scheduler, which interweaves jobs across cluster resources. Users now enjoy fairer allocation of Hadoop cluster resources to their work with each job they submit.

Scott Roberts is convinced that the Center made the right decision in StackIQ Enterprise Data. "Configuring machines in the cluster is easy, straightforward, and takes only minutes," he says. "We deploy and re-deploy Hadoop alongside Grid Engine and shift resources on an as-needed basis, optimizing both the cluster resources and the use of our time."

More Information

To find out more about StackIQ Enterprise Data go to:
www.stackIQ.com/stackiq-enterprise-data

StackIQ
4225 Executive Square #1000
La Jolla, CA 92037
858.380.2020
888.400.3966 fax
info@stackIQ.com