Big Data Without Big Headaches: Managing Your Big Data Infrastructure for Optimal Efficiency
The Growing Importance, and Growing Challenges, of Big Data

Big Data is hot. Highly visible early adopters such as Yahoo, eBay and Facebook have demonstrated the value of mining complex information sets, and now many companies are eager to unlock the value in their own data. Many choose open source platform Hadoop to create infrastructures of cost-effective, scalable and resilient commodity hardware and software. But, few fully anticipate the complexities of managing a large, multi-vendor Big Data infrastructure.

What are Hadoop (Big Data) Administrators up against?

New nodes and clusters are being added to the network at a breakneck pace. And, multi-vendor environments can make system and application problems difficult to isolate and resolve, often leading to costly downtime. Accurately determining the capacity needed to support rapidly changing business requirements can also be problematic. And, finally, today’s leaner IT organizations frequently lack the time, skills and resources necessary to manually monitor the infrastructure and effectively manage processes.

Where does Big Data infrastructure management go from here?

To reduce complexity and allow centralized management of multi-vendor environments, the next generation of Big Data infrastructure solutions need to be vendor-agnostic. In addition, processes, monitoring and incident response must be increasingly automated to boost speed and eliminate human error, thereby increasing system uptime.

In the pages that follow, we’ll examine several specific use cases common to Big Data environments. We’ll look at the challenges they present, as well as the best way to address them.
Managing Mixed Hadoop Distributions

Today’s Hadoop Administrators are primarily focused on running siloed, single-vendor Big Data platform stacks. However, as Big Data environments grow, hardware and software are frequently added based on cost and capabilities, with vendor preference playing a subordinate role. The result can be a large, sprawling, heterogeneous environment.

The challenges:

Systems can quickly grow to encompass too many clusters or nodes to effectively manage with existing resources. This situation is exacerbated by a plethora of incompatible hardware and software. Hadoop Administrators are forced to separately monitor individual, proprietary management tools, and manage dozens of individual clusters simultaneously. This lack of integration can result in performance degradation and prolonged system and application outages that can inhibit productivity, as well as generate financial repercussions.

The need:

What administrators want is an agnostic, heterogeneous management solution that helps manage a complex mix of Hadoop distributions. The ideal solution would have the ability to display an aggregated, real-time view of all infrastructure elements—including clusters, nodes and jobs—on a single dashboard. It would also be highly scalable to effectively manage the hundreds—or even thousands—of nodes required to meet evolving business needs.

“Our observation is that companies typically experiment with cluster sizes of under 100 nodes and expand to 200 or more nodes in the production stages. Some of the advanced adopters cluster sizes are over 1,000 nodes.”

Susheel Kaushik, Senior Director of Technical Product Marketing at Pivotal.

“The largest Hadoop networks increased their number of clusters 42% from 2012-2013”

The solution:

CA Big Data Control Center (BDCC) is being specifically designed to help Hadoop Administrators more efficiently and effectively manage large, complex, multi-vendor environments.

According to current plans, CA BDCC will offer you a host of real-world benefits, including:

- **The ability to manage Big Data infrastructures** from virtually any vendor
- **A unified view into any configuration** so that Hadoop Administrators experience less overall resource demand associated with ongoing system health monitoring and problem management
- **The capability to manage complex Hadoop infrastructures** comprised of hundreds—or even thousands—of nodes
- **An aggregated view** across the entire range of functions, including cluster, job, alert, system, process and resource management, as well as reporting and automation

**See how CA BDCC will help to solve real world challenges**

Manage all Big Data environments from one place

Secure access to Big Data (Hadoop)

Operationalize, manage & secure Hadoop environments

- Single, consistent management UI experience
- Simplified heterogeneous environment management
- Multi-vendor Hadoop management domains
Hadoop Process Management/Automation

One of the principle day-to-day responsibilities of a Hadoop Administrator is performing frequently used Hadoop operations. But, this obligation can be considerably more complex than it sounds when administrators are overseeing complex, multi-vendor environments comprised of hundreds of nodes or clusters.

The challenges:
The larger and more complex the infrastructure, the greater the risk that manual administration of common Hadoop operations will introduce costly human errors that degrade job and system performance. Today’s IT departments often lack sufficient time and resources to manually execute operations throughout sprawling Hadoop infrastructures. Furthermore, many IT organizations lack the in-depth skills and experience to support increasingly complex Big Data infrastructures.

The need:
The strain on IT resources could be greatly relieved by the introduction of reliable automation of common Hadoop operations built on best practices. This would include automated actions for handling monitored system activities requiring attention, as well as a flexible approach for configuring and activating pre-determined, automated responses when alerts are triggered.

One potential customer explained the lack of reliability of existing restart mechanisms and the need for reliable automation:

“We get alerts from Zabbix saying workflow is slow or the job failed. Some workflows automatically try and restart after the failure but, if it doesn’t start, we [have to] manually start it.”
The solution:

CA Big Data Control Center (CA BDCC) is being designed to support unparalleled levels of automation in even the largest and most complex Big Data environments.

According to current plans, you will be able to:

- **Create automated, one-click Hadoop operations** built on proven best practices
- **Define a flexible alert** threshold configuration that allows you to set desired response levels to informational- or problem-related alerts
- **Detect and remediate negative trending activities** “on-the-fly” with alert-triggered automation
- **Invoke automated problem resolution** through custom scripts
Hadoop Resource Management/Reporting

Management relies upon Hadoop Administrators to assess various Hadoop system resources and ensure the capacity to meet current and future requirements for rapidly changing business needs. Lacking the necessary bandwidth to meet demand can carry a substantial opportunity cost. On the other hand, scaling up too soon can be costly and inefficient. It falls on the Hadoop Administrator to accurately determine when to expand and by how much in order to accommodate projects in the pipeline.

The challenges:
A big part of the challenge facing Hadoop Administrators is getting maximum productivity out of the existing infrastructure. The larger and more complex the system, the more difficult it can be to get an accurate read on true capacity utilization. The same is true of poorly integrated system-wide resource utilization monitoring. Misconfigured and poorly-tuned system environments can lead to late recognition of the need to increase capacity, or just as easily, unnecessarily premature and expensive system expansion.

The need:
Hadoop Administrators are under significant pressure to get the most out of their existing systems, while at the same time, accurately predict the need for additional capacity in real time. They are looking for a tool capable of monitoring and managing large, complex infrastructures and providing comprehensive, integrated reports on capacity and system resource utilization.
The solution:

CA Big Data Control Center (CA BDCC) is being designed to simplify monitoring of a large, complex Big Data infrastructure to extract accurate and meaningful information about the utilization of system resources.

According to current plans, with CA BDCC, you’ll be able to:

- **Report on different system resources** and their utilization by jobs and applications
- **Track resource utilization** at a job definition/application level
- **Generate detailed reports** with time-series data to aid Hadoop capacity planning
- **Carry out metrics monitoring**, resource management, alert management and automation capabilities
- **Answer key questions**, such as “How much disk space/CPU/network IO is being consumed?”, “Which jobs are consuming these system resources?” or “What is the trend looking like?”
- **Identify and mitigate** negative Big Data infrastructure trends
Hadoop Job Management & Monitoring

As Hadoop infrastructure grows, so does the number of related jobs. The role of a Hadoop Administrator is to effectively manage and monitor all of the active jobs, ensuring that they conclude successfully and that the system continues to run efficiently.

The challenges:
The sheer size of today's Big Data infrastructures, combined with the complexity of managing jobs running on different vendors’ hardware and software from cluster to cluster, makes it extremely difficult to effectively track job details. The failure to keep tabs on every aspect of every job introduces the risk that job, application and system performance can degrade, resulting in costly inefficiency and downtime.

The need:
Hadoop Administrators want to be able to efficiently monitor, and simultaneously manage, a wide range of Hadoop jobs. They want a tool that generates reliable data for optimizing job execution environments.
The solution:
CA Big Data Control Center (CA BDCC) is being designed to make the simultaneous management of multiple jobs easier and more governable.

According to current plans, with CA BDCC, you will be able to:

- Manage different Hadoop job definitions and monitor their individual health
- Manage and monitor Hadoop jobs much more easily
- Assess and optimize your job execution environments
Hadoop System Health Monitoring & Alert Management

Just because a system hasn’t crashed doesn’t mean it’s operating at peak performance. Hadoop Administrators are responsible for monitoring and maintaining overall system health in a way that detects and corrects issues that can lead to inefficiencies, and larger issues down the road.

The challenges:

Due to the increasing complexity of many Hadoop-based Big Data infrastructures, it can be very difficult to efficiently and effectively monitor overall system health. Lack of integration between heterogeneous components requires multiple monitoring solutions to keep track of the entire system. The end result is increased difficulty detecting system issues before they compromise the entire infrastructure.

The need:

Administrators want to track key system metrics in real time, ideally from a single dashboard. In addition, they demand monitoring solutions that provide early warning of negative trends that can lead to system slowdowns, errors and outages.
According to current plans, CA BDCC will allow you to:

- **Closely monitor** the health of Hadoop clusters, nodes and jobs
- **Create custom thresholds** for triggering alerts
- **Aggregate alerts** generated by other manager instances within the Big Data environment
- **Generate unique alerts** associated to an expanded set of monitored system metrics
- **Send an appropriate alert** (with details) prior to system degradation when an issue is detected
- **Increase system uptime** for greater operational efficiency
- **Reduce manual periodic system monitoring** to improve administrator productivity
- **Analyze the detailed metrics** provided within alerts to improve overall performance

**The solution:**

CA Big Data Control Center (CA BDCC) is being designed to provide easy access to key metrics related to system health, as well as an efficient means to quickly address problems as soon as they are detected.
What if you could improve enterprise operational efficiency significantly?

CA Technologies is designing a unique Big Data infrastructure management solution for complex, heterogeneous environments. According to current plans, CA BDCC will aggregate and consolidate management of the entire Hadoop Big Data environment, helping you to:

- Manage your entire Big Data environment from a single, consistent management user interface
- Monitor, detect and resolve issues quickly, with visibility across all Hadoop clusters
- Obtain key insights into your Hadoop resources to make intelligent decisions regarding capacity
- Automate actions based on best practices to reduce downtime and resource demands
Why rely on CA for your Big Data infrastructure management solution?

CA BDCC is being specifically designed to help answer the challenges Hadoop Administrators are facing today—the challenges that come with managing large, complex, multi-vendor Big Data environments. CA BDCC will provide vital tools to help keep your system up and running at peak levels of performance, offering you a wide range of unique benefits, including:

- **Heterogeneous System Management**, which will allow Hadoop Administrators to manage all of your Hadoop clusters from a single UI, to provide consistent management regardless of vendor
- **Resource Reporting**, which will allow Hadoop Administrators to view and assess overall system health and receive recommended changes based on proven best practices
- **Process Management** with built-in automated operations, which will help to increase overall Big Data infrastructure efficiency health
- **Job Monitoring**, which will enable Hadoop Administrators to view jobs in progress, to make changes in resources in real time and to identify and resolve issues as they arise, for increased efficiency of processes and reduced downtime
- **Intelligent Alert Management**, which will notify the Hadoop Administrator when negative trending issues are detected across the Big Data environment—down to the cluster and node—before the trend adversely affects system performance, thereby streamlining complex infrastructure management and helping to eliminate the need for redundant resources

Learn more about the value of CA Big Data Control Center
CA Technologies (NASDAQ: CA) creates software that fuels transformation for companies and enables them to seize the opportunities of the application economy. Software is at the heart of every business, in every industry. From planning to development to management and security, CA is working with companies worldwide to change the way we live, transact and communicate – across mobile, private and public cloud, distributed and mainframe environments.

Learn more at ca.com.