Skyscape Cloud Services Deploys Hadoop in the Cloud on VMware vSphere®

TECHNICAL CASE STUDY

V1.0/MAY 2015

Table of Contents

Introduction

Skyscape Cloud Services has been using tools such as VMware vSphere® and the VMware vSphere Big Data Extensions™ tool for deploying and managing virtualized Hadoop clusters since early 2014. This paper describes the deployment of a number of virtualized Hadoop clusters in production on vSphere as a cloud-based service at Skyscape's hosted sites in the United Kingdom. It gives the business background motivating this set of implementation technologies and a view of the technical features, as well as some specific customizations done.

Business Background

Skyscape (www.skyscapecloud.com) is a UK company that was founded in 2012 to provide cloud computing services through the UK Government's G-Cloud initiative. Delivering public services online is fundamental to shifting the government's approach to interacting with its citizens and businesses. Skyscape's offered services comprise infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS). Those services are provided through the Skyscape Cloud Alliance, which comprises VMware, Cisco, EMC, QinetiQ, and Ark Data Centres, along with Skyscape. The Ark Data Centres are based in the towns of Corsham and Farnborough in the South East of the United Kingdom and are 80 miles apart from each other. Skyscape delivers its services with no startup costs and no minimum contracts. The end user pays for only what they use. This produces an easy and predictable cost model for the end-user community.

Skyscape's Hadoop in the Cloud service is a PaaS implementation of the Hadoop data platform, built on the Skyscape UK Government–accredited IaaS environment. More information on the IaaS offering can be found at http://www.skyscapecloud.com/what-we-do/infrastructure-as-a-service/compute/.

As the volume, velocity, and variety of data generated by organizations increase, the cost of retaining this data over the longer term must be carefully balanced against the opportunity to exploit the potential value of big data analysis. Hadoop offers economic data warehouse capabilities through which all data is stored and accessible, while also offering a massively parallel processing (MPP) framework that supports various methods for analyzing and interrogating the data.

The Skyscape solution enables organizations to rapidly deploy, experiment with, and prove the value of Hadoop-based solutions without having to invest in the cost, time, and risk associated with purchasing and provisioning infrastructure, platforms, and licenses.

Hadoop solutions typically require an investment in dedicated hardware. Many organizations want to avoid this capital expenditure along with related operating expenses such as hardware maintenance and costs of power and cooling. Customers can use Skyscape Hadoop in the Cloud without any of this outlay and with no minimum commitment. The users pay for the amount they use, when they use it.

To achieve these business goals, Skyscape decided that virtualizing the Hadoop platform was the first key step in hosting that platform on the cloud.

Why Virtualize Hadoop on vSphere?

Skyscape chose to use vSphere as the strategic platform for hosting its Hadoop-based applications for several reasons:

- To provide Hadoop clusters as a service to the end users and the development community, reducing time to insight into the data
- To quickly deploy, scale up, and remove clusters through automation
- To capitalize on the most suitable hardware and storage to make use of the Hadoop design ethos, thereby lowering costs to the end user
- To provide a multitenant platform with support for heterogeneous Hadoop distributions

  NOTE: vSphere offered support for multitenancy through the security and performance isolation inherent in sets of virtual machines that exist in separate resource pools. vSphere also offered resource usage isolation by providing controls to mitigate contention for hardware resources.

- To optimize the use of existing VMware investments by extending Hadoop services to the current Skyscape cloud platform

The Project

This Hadoop virtualization project began in 2014. After some development and testing time, it went into production in February 2015 with multiple servers supporting the Hadoop workload and five different customer organizations as users or *tenant* customers.

The attraction of big data to the government's end users was viewed in different ways. Various public sector customer projects enabled by the Skyscape cloud have similar themes:

- Aggregation of disparate datasets – This breaks down into two categories:
  - The ingestion of common datasets from public and commercial sources
  - Aggregation of datasets from in-house applications, providing cross-departmental visibility
- Exposure of datasets – Customers have a common objective to expose their datasets through front-end applications to make the data visible to one of two communities:
  - The general public, so chargeable services can be created and value derived from those services
  - Other companies in the same field, so cross-pollination of data can then occur between internally developed applications

The Application Architecture

Skyscape provides turnkey Hadoop deployments that are architected to be hosted on a single multitenant platform while still providing end users with performance, security, and autonomy.

The core platform provides the shared hardware and services required to support multiple Hadoop clusters. Figure 1 provides a high-level representation of the architecture.

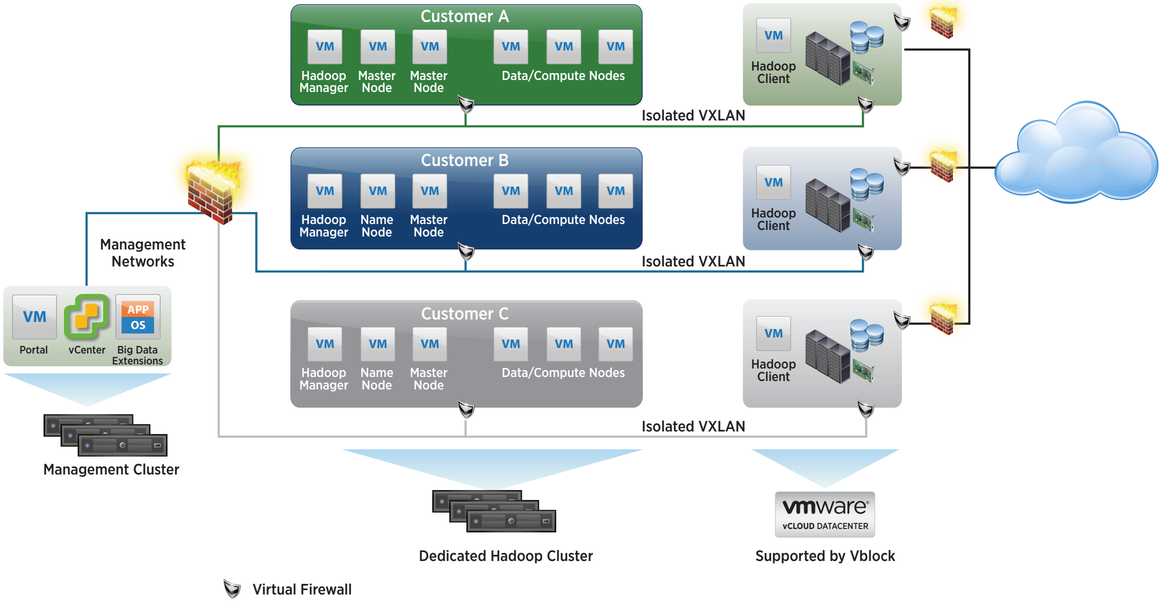

Figure 1. Architecture for the Skyscape Hadoop in the Cloud Service

The infrastructure is grouped into two logical areas:

1. A client area deployed on Skyscape's existing cloud platform, offering IaaS (as shown on the right side of Figure 1)
2. A data area where the Hadoop virtual machines are hosted on hardware specially chosen to support the "commodity" nature of the Hadoop design approach (as shown in the middle of Figure 1)

There is also a management cluster shown on the far left of Figure 1. This cluster contains the management tools that the Skyscape technical staff uses to administer the Hadoop in the Cloud infrastructure.

Client Services

To provide security isolation, each customer or tenant is provided with a *dedicated* virtual network on its own virtual extensible LAN (VXLAN) and firewall infrastructure. VXLAN is part of the product set comprising vSphere and VMware vShield Manager™. The VMware NSX™ networking portfolio of products is also being closely considered for this functionality in the future. After a VXLAN fabric has been created, users can consume isolated Layer 2 networks—also called "virtual wires"—on demand.

These components enable ingress and egress of datasets from the Internet or customer site via virtual private network (VPN) to provide users with a means of accessing their dedicated Hadoop infrastructure. The Hadoop client components are hosted on Skyscape's existing cloud platform, which is provisioned on VCE Vblock compute and storage. The end user is given full control over these client components. End users also have the ability to deploy additional virtual machines from the Skyscape catalog. The end-user client virtual machines have high-performance access to their associated Hadoop cluster.

Hadoop Services

A second virtual firewall is deployed and managed by Skyscape to control traffic in and out of the customer's Hadoop cluster from the client services previously described. A dedicated VXLAN wire with a scope that spans both the Hadoop and cloud clusters is used to provide transit of data between the customer's client area and the Hadoop cluster.

A third virtual wire is dedicated to Hadoop data network traffic. This VXLAN scope is limited to the Hadoop clusters, which are supported by a pair of 10GB Arista switches per rack. Each physical VMware ESXi™ host has two 10GB ports dedicated to support Hadoop data configured as active/standby to ensure that system performance is not impacted by oversubscribed network links during a failover situation.

The Hadoop clusters are architected to provide both shared and local storage. Shared storage is provided using VMware Virtual SAN™ to enable horizontal scaling. Local storage is provided by 2TB nearline SAS disks. These local (direct-attached) disks for Hadoop data are organized as JBOD and are presented to the vSphere and vSphere Big Data Extensions environment as one datastore per disk. VMware vSphere VMFS is used as the file system that is contained in those datastores. A single VMDK file is deployed to each datastore to enable consistent performance by providing dedicated access. The storage design here is geared to preserving the I/O pattern as sequential and not causing it to assume a random nature.

Shared storage, implemented using Virtual SAN, is used to host virtual machine operating system (OS) partitions, whereas local storage is provisioned to support the Hadoop data.

Figure 2 shows the logical layout of the various components of Hadoop on a typical customer's cluster. The Hadoop DataNode and NodeManager roles or daemons are held together on the guest OS of each "worker" virtual machine, and their storage needs are separated from those of the guest OS. The master roles—ResourceManager, NameNode, and ZooKeeper—are hosted on virtual machines of their own.

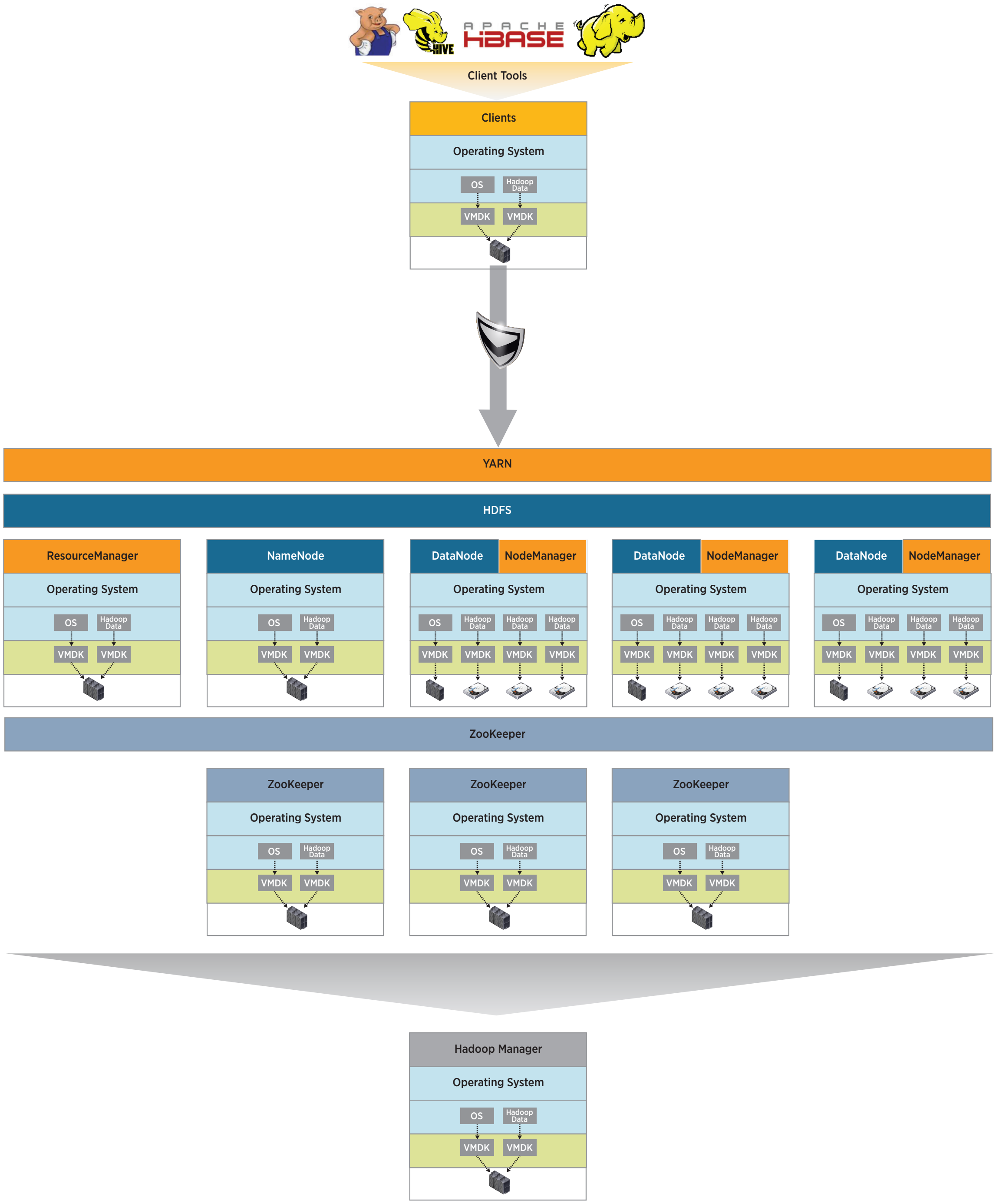

Figure 2. Software Layout: Each Multilayered Block Is A Virtual Machine Containing Hadoop Processes

Each customer's Hadoop cluster is provisioned with the following Hadoop roles, on their own virtual machines. The term "shared storage" as used here refers to a Virtual SAN implementation that supports the guest OS disks for all Hadoop roles other than those of the client virtual machines.

- Two master servers – These servers are deployed to shared storage with VMware vSphere High Availability (vSphere HA) enabled.
- Three ZooKeeper servers – These servers are deployed to shared storage with vSphere HA disabled.
- Three data/compute servers – The guest OS for these virtual machines is deployed to shared storage, and the Hadoop data is held on dedicated local disks (direct-attached) assigned to each virtual machine. The initial automated deployment provides three data/compute nodes. After the cluster is deployed, however, the customer can scale out that deployment to contain more nodes.
- One Hadoop manager – This is deployed to shared storage with vSphere HA enabled.

The Skyscape infrastructure provides shared DNS and NTP services to support the deployment of multiple clusters.

Configuration and Sizing of Virtual Machines

In the cloud provider environment that is Skyscape's main business, it is challenging to accurately model resource usage without having the actual customer use cases available ahead of time. Skyscape's key motivation was to make sure that the balance of resources provided in the cluster was appropriate for the workload while not leaving any wasted resources. The following were the key factors that were considered when estimating utilization:

1. Ensuring that the local storage would be fully utilized with a good balance of CPU, RAM, and network I/O for Hadoop processing
2. Covering the Hadoop cluster overheads—for example, the quorum servers, VMware vShield Edge™ requirements, and so on

Skyscape used two key factors to determine suitable virtual machine sizes:

1. The Hadoop cluster size – Every Hadoop cluster requires some overhead to run it. For example, the RAM, CPU, and storage and network bandwidth required varies depending on whether the infrastructure must support a 10TB cluster or a 200TB cluster. In the larger case, the NameNode process needs more RAM to reference the metadata for the larger number of blocks; the data/compute nodes need more compute capacity to process the larger dataset. As a cloud provider, however, Skyscape cannot create a perfectly balanced one-off cluster for each customer, so instead the company defines two cluster sizes that can be deployed with predetermined overheads.
2. The number of customer or tenant Hadoop environments to be deployed, the size of the clusters, and the amount of data stored in the clusters – This is important because Skyscape found that six medium-sized clusters with a minimal data/compute footprint of 3 nodes requires more overhead than a single medium-sized cluster with 18 data/compute nodes.

Estimating these values enables the implementer to understand how many resources will be required to cover the overheads and therefore how much spare capacity there is to be shared among the data/compute nodes.

Skyscape currently provides two clusters sizes: medium and large.

Tables 1 and 2 describe the minimal resources needed to run a medium and a large cluster **not** including data/compute—that is, the overheads required to support the storage and processing of Hadoop data.

MEDIUM				
Node Type	**RAM**	**vCPUs**	**Shared Disk (GB)**	**Local Disk (TB)**
Virtual Firewalls	16	4	9	0
NameNode	16	4	100	0
ResourceManager	16	4	100	0

Table 1. Cluster Size – Medium

LARGE				
Node Type	**RAM**	**vCPUs**	**Shared Disk (GB)**	**Local Disk (TB)**
Virtual Firewalls	16	4	9	0
NameNode	32	8	200	0
ResourceManager	32	8	200	0

Table 2. Cluster Size – Large

For the purposes of modeling, Skyscape estimated the amount of data to be stored in each cluster size and the number of customers—that is, tenants—anticipated to be supported for the various cluster sizes. A *customer* or *tenant* here is the organization that is purchasing Hadoop services from Skyscape.

The following are the initial sizes of Hadoop clusters Skyscape used for modeling and an initial estimation of the number of customers used for each cluster size:

- Medium cluster – 10TB usable – 10 customers used in the initial sizing model
- Large cluster, light use – 50TB usable – 5 customers used in the initial sizing model
- Large cluster, heavy use – 200TB usable – 2 customers used in the initial sizing model

There were two important factors in the process of estimating utilization for these different clusters:

1. The resources required to support the clusters
2. The amount of local disk space required to support the anticipated volume of usage

Understanding these factors enabled Skyscape to calculate the number of physical machines needed to support the Hadoop local storage capacity and from that determine how much CPU, RAM, and network and shared storage capacity would be available in aggregate across the vSphere clusters. Calculations were also made on how much of that capacity was to be set aside for overheads. The resources left over after factoring in the overheads were assigned to the data/compute nodes to enable them to process Hadoop data. The details of these calculations are beyond the scope of this paper and would be site specific in another implementation.

Hadoop Versions Supported

Currently the Skyscape Hadoop in the Cloud service offering supports Hortonworks HDP with the Apache Ambari tool for management of clusters. Each end-user or customer Hadoop cluster is provisioned with its own private Ambari Server in its own virtual machine, so that management of that new cluster at the Hadoop level is then completely independent of managing any other cluster. Work is ongoing to extend the service to also support Cloudera Distribution including Hadoop (CDH) in the same fashion.

Hortonworks HDP 2.1.3 software is currently deployed using the Ambari 1.6.1 tool.

The Skyscape Hadoop-as-a-Service Portal

Skyscape has its own homegrown Web portal that is used by the company's customers to manage their IaaS functionality, to log and manage support calls, and to get billing information. It is now also in use for provisioning clusters of suitable types for its customers. The work to expose the functionality of the Hadoop application managers, such as Ambari and Cloudera Manager, through the portal is still in the development stage at the time of writing. The underlying automation framework is constantly being worked on and developed.

Skyscape's technical staff customized the cloud environment to allow more than one tenant to have their own vendor-specific management console—the Ambari console, for example—provisioned with the target Hadoop cluster. This gives the end user control over their own Hadoop clusters in the style supported by the Hadoop vendor and enables them to see in a way that is familiar to them how various aspects of the cluster are performing through that console.

Customizations for Increased Flexibility

This section details some very useful technical customizations that Skyscape staff made to the base implementation to improve it for their cloud-service users. These are provided as nuggets of information for future deployers to help them learn new practices in virtualizing their own Hadoop clusters.

Using the Application Managers

The base virtual machines are deployed using the vSphere Big Data Extensions feature for VMware vCenter™. This is described in more detail in the next section. Skyscape decided to deploy the Hadoop software distribution directly into the virtual machines using the Hadoop application manager tools, initially Ambari, to enable greater flexibility in the following areas:

- Hadoop services – It was important to give customers control over deployment of additional Hadoop services they might want to use to support their application.
- User management – The end-user administrators needed control over creating additional users.
- Maintenance mode – Skyscape needed to be able to show its customers when maintenance operations were being carried out by its own staff. Some details of this type of notification will be given in a subsequent section.
- Gathering of cluster utilization statistics – Administrators, for example, needed to know the amount of consumed storage to understand when more physical hardware might be required.

Use of vSphere Big Data Extensions and APIs

The use of APIs is essential to Skyscape's provision of infrastructure to its customers. To be efficient and to reduce human error, the company must be able to automate tasks. This section provides a high-level description of some of the APIs that were used for creation of Hadoop clusters.

One of the key tools that is central to the automation of cluster creation for Skyscape is vSphere Big Data Extensions, which works with vCenter to provide the capability to create a set of virtual machines at certain sizes and to place software of one's choosing into those virtual machines. vSphere Big Data Extensions enables a vSphere administrator to provision Hadoop clusters onto virtual machines in a user-friendly way.

vSphere Big Data Extensions is available as a free download to VMware vSphere Enterprise Edition™ licensed users. It consists of a management server and a template virtual machine that are deployed as one unit, or virtual appliance, into the vCenter environment. After this vSphere Big Data Extensions VMware vSphere vApp™ is set up and configured with storage and network resources, it can then be used through a graphical user interface (GUI) as part of the vSphere Web Client, through its own command-line interface (CLI) or through its APIs. Figure 3 shows an outline of the vSphere Big Data Extensions architecture, with cloning of the preshipped template virtual machine, to form the basis for a new Hadoop cluster.

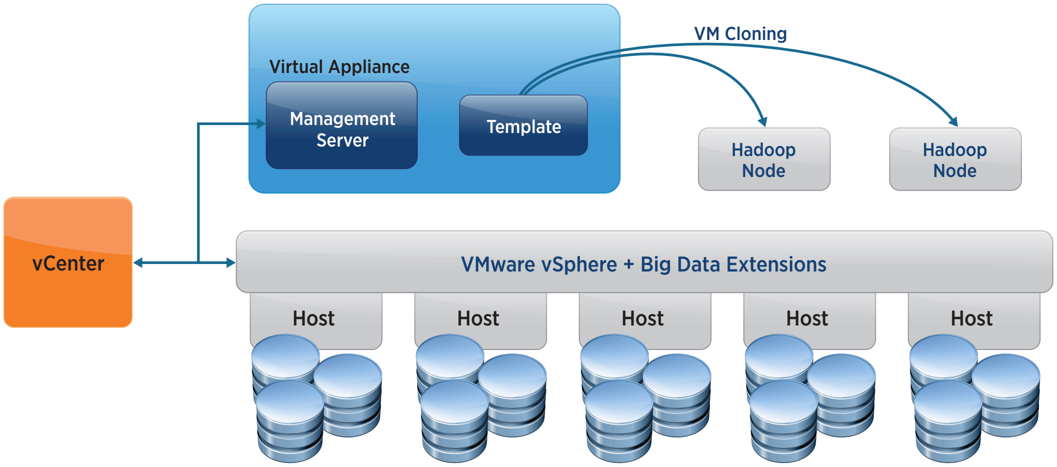

Figure 3. Architecture of vSphere Big Data Extensions

vSphere Big Data Extensions enables users or administrators to customize the various virtual machines that they create, either interactively through their own GUI or by means of using their CLI with a cluster configuration specification file. There are a number of samples of these JSON-based cluster specification files available in the "samples" directory on the vSphere Big Data Extensions management server seen in Figure 3.

The Skyscape architects took one sample cluster specification file from vSphere Big Data Extensions and customized it to create the two Skyscape Hadoop cluster definitions that are used to build predefined clusters. The predefined clusters then conform to the correct standards, enabling solid deployments while aiding the capacity management process. Skyscape has its own cloud automation framework that underpins its homegrown management portal. As part of the automation framework, each product or service that is supplied to users has modules that can perform the tasks needed by the Skyscape staff to administer the full system.

In the case of vSphere Big Data Extensions, the following are among the operations:

- Create a cluster
- Retrieve a cluster definition
- Delete a cluster
- Add a node to an existing cluster

These methods that are used within the Skyscape automation framework are all standalone and are targeted for a specific job. This enables other business units to consume these functions to build their own workflows to meet their needs. The goal here is to remain flexible while reducing risk to the core infrastructure by maintaining control of the interface to all services. Strict controls and testing processes are in place to make sure that the modules written to interface with infrastructure operate in an efficient and robust manner so they can be confidently utilized by the other business areas. The vSphere Big Data Extensions APIs are used internally in the Skyscape portal to provision the virtual machines and then make inquiries about them. This functionality is separated from that of the Hadoop manager tools utilized by the end users. In this way, administrators have more control over the Hadoop services they are delivering.

Host Server Maintenance

As a secure cloud provider, it is essential that Skyscape continue to perform maintenance tasks on the infrastructure over time. One issue the company has encountered is that moving away from shared storage for the Hadoop nodes has meant that administrators can no longer rely on VMware vSphere Distributed Resource Scheduler™ (vSphere DRS) to migrate work from one server to another. They cannot use VMware vSphere vMotion® for the data/compute nodes due to the utilization of local storage to balance the system or to evacuate a server when maintenance, such as patching or firmware upgrades, is required. Skyscape solved this issue by writing code that orchestrates the rebooting of its servers. This custom innovation code works by conducting the following operations:

1. Connects to the VMware vCenter Server™ and queries the ESXi server host that the administrator wants to perform maintenance on, to retrieve a list of all the current virtual machines on that host
2. Migrates all virtual machines that are not running the DataNode or ZooKeeper daemons to another host in the cluster
3. Determines which customer clusters are affected if the DataNode and ZooKeeper virtual machines are shut down
4. Retrieves connection details for the Ambari managers for the affected clusters from the configuration management database (CMDB)
5. Connects to each application manager and instructs Ambari to put the virtual machine into Ambari's maintenance mode so the customer can see that work is being performed on their system

   Entering a host into Ambari maintenance mode achieves two things:

   a. Suppresses any alerts from the Nagios management system (deployed in the provisioned virtual machines)

   b. Graphically represents the fact that an administrator is doing planned work on the server, so users do not need to worry
6. Powers off the virtual machine and enters the ESXi host into maintenance mode
7. After the maintenance work is completed, takes the ESXi host out of maintenance mode and powers on the virtual machines associated with the host
8. Reconnects to each Ambari server to start all the roles on those virtual machines
9. Finally, after the services have started, takes the Hadoop environment within the virtual machines out of maintenance mode in Ambari

SSL Replacement

As part of Skyscape's service delivery, the Hadoop in the Cloud product was subjected to an extensive set of tests to validate the security of the system. This involved design review, configuration review, and penetration testing. As part of this work, Skyscape replaced all the default self-signed SSL certificates with valid Skyscape SSL certificates while also removing all weak ciphers.

Template Updates

Skyscape staff also made a few changes to the virtual machine templates that vSphere Big Data Extensions uses, following guidance from various Hadoop distribution vendors:

- Installed and configured NTP
- Disabled Transparent Huge Pages (THP) compacting
- Added the following to the bottom of the file /etc/sysctl.conf
  - vm.swappiness = 1
  - vm.overcommit_ratio = 100
- Increased the network MTU to 8,950 (50 bytes is reserved for VXLAN headers)

Conclusion

Providing Hadoop cluster creation and management on VMware vSphere has greatly improved the way that big data applications are delivered in the Hadoop in the Cloud offering from Skyscape. The costs of dedicated hardware as well as those for management and administration for individual clusters have been greatly reduced by sharing pools of hardware among various user communities. Developers and other staff can now obtain a Hadoop cluster for testing or development without concerning themselves with the underlying hardware. Virtualization has also improved the production Hadoop cluster management environment, enabling more operational control over resource consumption and better management of the trade-offs in performance analysis. Virtualizing Hadoop workloads has become the standard for this important part of Skyscape's business. It has enabled IT to become a service provider to the business.

References

1. *Hadoop in the Cloud* – Services from Skyscape – Datasheet
http://www.skyscapecloud.com/wp-content/uploads/2015/01/NEW_Hadoop-in-the-cloud_datasheet_v2.pdf

2. VMware vSphere Big Data Extensions
http://www.vmware.com/bde

3. *Scaling the Deployment of Multiple Hadoop Workloads on a Virtualized Infrastructure* – White Paper by Intel, Dell, and VMware
http://www.intel.es/content/dam/www/public/us/en/documents/articles/intel-dell-vmware-scaling-the-deployment-of-multiple-hadoop-workloads-on-a-virtualized-infrastructure.pdf

4. *Virtualized Hadoop Performance with VMware vSphere 5.1*
http://www.vmware.com/resources/techresources/10360

5. *Virtualized Hadoop Performance with VMware vSphere 6 on High-Performance Servers*
http://www.vmware.com/resources/techresources/10452

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