Cloudera Data Platform on VMware Cloud Foundation Powered by VMware vSAN

Reference Architecture

vmware

CLOUDERA

Table of contents

Introduction

VMware Cloud Foundation™ is built on VMware's leading hyperconverged architecture, VMware vSAN™, with all-flash performance and enterprise-class storage services including deduplication, compression and erasure coding. vSAN implements hyperconverged storage architecture, by delivering an elastic storage and simplifying the storage management. VMware Cloud Foundation also delivers end-to-end security for all applications by implementing micro-segmentation, VPN (VMware NSX®), VM hypervisor, VMware vSphere® vMotion® encryption, AI-powered workload security and visibility (vSphere), and data-at-rest storage encryption (vSAN).

Technology Overview

Solution technology components are listed below:

- VMware Cloud Foundation
    - VMware vSphere
    - VMware vSAN
    - VMware NSX Data Center
- Cloudera Data Platform (CDP)

VMware Cloud Foundation

VMware Cloud Foundation is the ubiquitous hybrid cloud platform built on full stack hyperconverged infrastructure. Cloud Foundation provides a complete set of secure software-defined services for compute, storage, network, security, Kubernetes management, and cloud management. The result is agile, reliable, and efficient cloud infrastructure that offers consistent infrastructure and operations across private and public clouds.

VMware vSphere

VMware vSphere is VMware's virtualization platform, which transforms data centers into aggregated computing infrastructures that include CPU, storage, and networking resources. vSphere manages these infrastructures as a unified operating environment and provides operators with the tools to administer the data centers that participate in that environment. The two core components of vSphere are ESXi™ and vCenter Server®. ESXi is the hypervisor platform used to create and run virtualized workloads. vCenter Server is the management plane for the hosts and workloads running on the ESXi hosts.

VMware vSAN

VMware vSAN is the market leader in hyperconverged Infrastructure (HCI), enables low cost and high-performance next-generation HCI solutions, converges traditional IT infrastructure silos onto industry-standard servers and virtualizes physical infrastructure to help customers easily evolve their infrastructure without risk, improve TCO over traditional resource silos, and scale to tomorrow with support for new hardware, applications, and cloud strategies.

VMware NSX Data Center

VMware NSX Data Center is the network virtualization and security platform that enables the virtual cloud network, a software-defined approach to networking that extends across data centers, clouds, and application frameworks. With NSX Data Center, networking and security are brought closer to the application wherever it's running, from virtual machines to containers to bare metal. Like the operational model of VMs, networks can be provisioned and managed independently of the underlying hardware. NSX Data Center reproduces the entire network model in software, enabling any network topology—from simple to complex multitier networks—to be created and provisioned in seconds. Users can create multiple virtual networks with diverse requirements, leveraging a combination of the services offered via NSX or from a broad ecosystem of third-party integrations ranging from next-generation firewalls to performance management solutions to build inherently more agile and secure environments. These services can then be extended to a variety of endpoints within and across clouds.

Cloudera Data Platform (CDP) Private Cloud Base

CDP Private Cloud Base is the on-premises version of Cloudera Data Platform. This new product combines the best of Cloudera Enterprise Data Hub and Hortonworks Data Platform Enterprise along with new features and enhancements across the stack. This unified distribution is a scalable and customizable platform where you can securely run many types of workloads.

CDP Private Cloud Base supports a variety of hybrid solutions where compute tasks are separated from data storage and where data can be accessed from remote clusters, including workloads created using CDP Private Cloud Experiences. This hybrid approach provides a foundation for containerized applications by managing storage, table schema, authentication, authorization, and governance.

For details, see https://docs.cloudera.com/cdp-private-cloud-base/latest/index.html.

Validation Strategy

We validate VMware Cloud Foundation with vSAN can support Cloudera by deploying a Cloudera Hadoop cluster in a VMware Cloud Foundation workload domain, running representative workloads against the Hadoop cluster. This solution validation uses Dell vSAN ReadyNode™; however, this applies to other vSAN Readynode partners and Dell EMC VxRail. The test will ensure that VMware Cloud Foundation is able to meet Hadoop infrastructure requirements and validate design assumptions about the infrastructure.

Validation Environment Configuration

This section introduces the resources and configurations:

- Architecture diagram
- Hardware resources
- Software resources
- Network configuration
- VM and storage configuration
- Cloudera Hadoop Cluster configuration
- Monitoring tools

Architecture Diagram

VMware Cloud Foundation test environment is composed of a management workload domain and a workload domain as shown in Figure 1. We deploy all the VMs required for the Cloudera test cluster in the VI Workload Domain and all the other infrastructure VMs in the separate management workload domain.

The Cloudera solution uses all-flash servers running on the VMware Cloud Foundation 4.0 software suite for application workloads as well as the management components such as the vCenter Appliance. This provides a platform for a Cloudera environment using vSAN as the storage platform for workloads. vSAN scales naturally in the same fashion as Hadoop nodes thus Hadoop cluster can scale out with vSAN scaling out. Besides, Cloudera can leverage the SPBM (Storage Policy Based Management) in vSAN to manage storage in a flexible way.

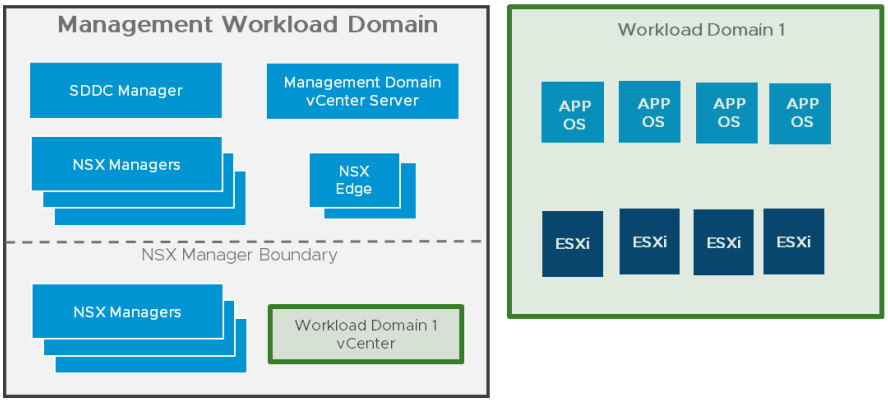

Figure 1. Hadoop on VMware Cloud Foundation Solution Architecture

Hardware Resources

In this solution, for the workload domain, we used a total of eight PowerEdge R640 each configured with two disk groups, and each disk group consists of one cache-tier NVMe and four capacity-tier SAS SSDs.

Each node in the cluster had the configuration per Table 1.

Table 1. Hardware Configuration

Number of Servers	8
Server	PowerEdge R640
CPU	2 x Intel(R) Xeon(R) Gold 6132 CPU @ 2.60GHz, 14 core each
Logical Processor (including hyperthreads)	56
Memory	512 GB
Cache device	2 x NVMe PM1725a 1.6TB SFF
Capacity device	8 x 1.75TB Samsung SSDs
Network	2 x 10GBps Intel(R) Ethernet Controller 10G X550

The VMware Cloud Foundation management workload domain hardware details are not provided since all Hadoop workload VMs ran on the workload domain and may not be relevant.

Software Resources

The software resources used in this solution are shown in Table 2.

Table 2. Software Resources

SOFTWARE	VERSION	PURPOSE
VMware Cloud Foundation	4.0	A unified SDDC platform that brings together VMware ESXi, vSAN, NSX and optionally, vRealize Suite components, into a natively integrated stack to deliver enterprise-ready cloud infrastructure for the private and public cloud. See *BOM of VMware Cloud Foundation 4.0* for details.
Guest Operating System	RHEL 7.7 x86_64	Operating system
Cloudera CDP Data Center	7.0.3	*Cloudera Runtime 7.0.3*

Network Configuration

Figure 2 shows the VMware vSphere distributed switches configuration for the workload domain. Two 10Gbps vmnics were used and configured with teaming policies. The NSX-T controllers resided in the management domain. The Hadoop virtual machines were configured with a VM network for management and Hadoop network for Hadoop traffic on the NSX-T Segments. Two different configurations were validated: traditional VLAN backed segment and overlay backed segments for VM, vSphere vMotion, and vSAN had a dedicated Portgroup configured as shown in Table 3.

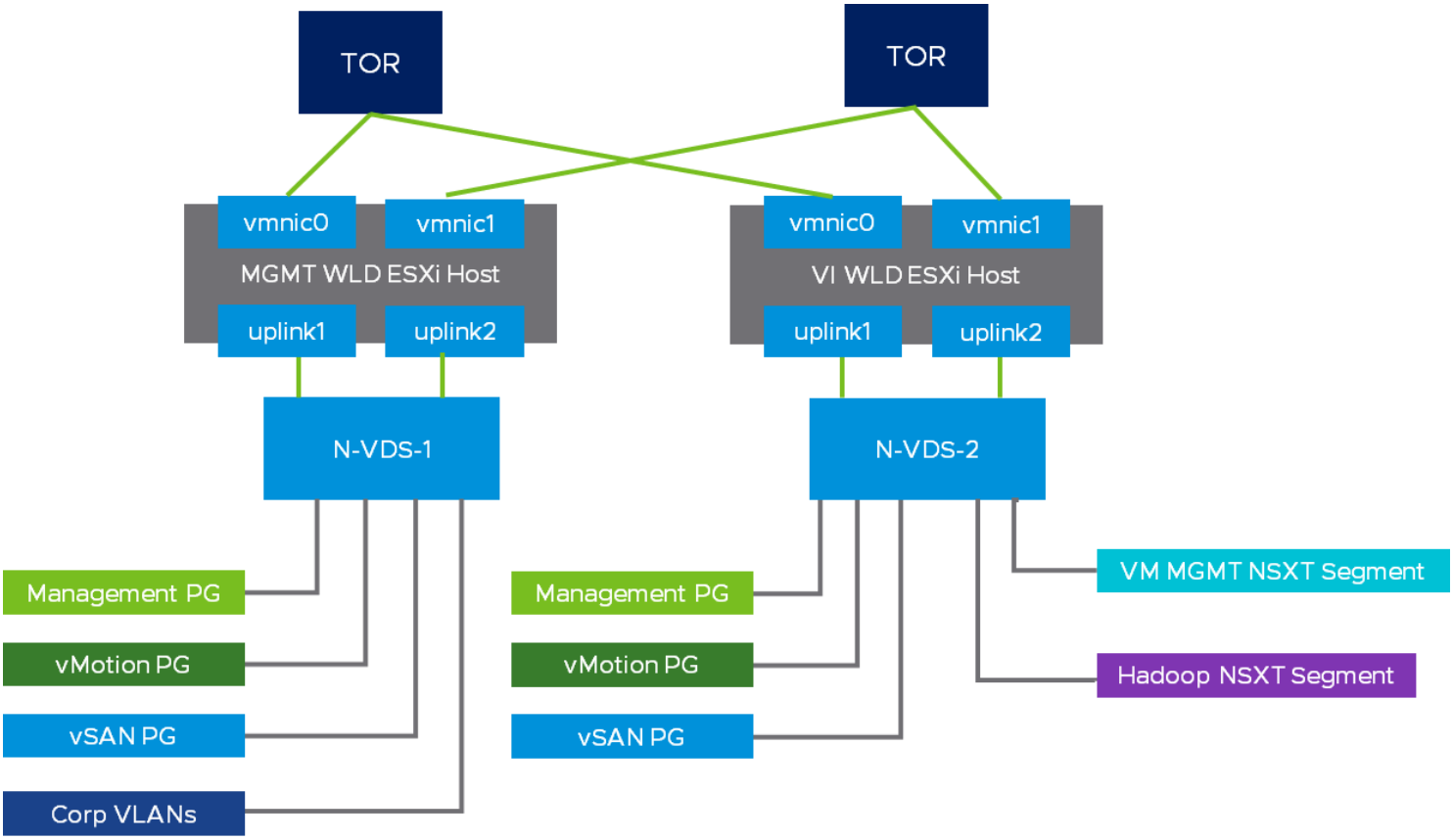

Figure 2. Network Configuration

Table 3. Virtual Distributed Switch Teaming Policy for 2x10 GbE Profile

Port Group	Teaming Policy	VMNIC0	VMNIC1
Management network	Route based on Physical NIC load	Active	Active
vSphere vMotion	Route based on Physical NIC load	Active	Active
vSAN	Route based on Physical NIC load	Active	Active
VM MGMT NSXT Segment	Load Balance Source	Active	Active
VM Hadoop NSXT Segment	Load Balance Source	Active	Active

Overlay Backed Segment Configuration: The test environment instantiated an NSX-T T0 and T1 router on an edge cluster comprised of two medium sized edges (4 vCPU and 8GB RAM). Overlay backed logical segments were created and attached to a T1 gateway (Figure 3). The segments were configured with NSX-T default parameters as shown in Figure 4 and Figure 5. Two overlay segments were created one for the management network for VM (10.159.228.1/26) and other for Hadoop traffic (192.168.20.1/24) both part of the vSphere distributed switch (Figure 6).

In this test environment, Hadoop VMs were located in a single vSphere cluster. It should be noted that during our validation, the Hadoop application traffic flew from east to west between cluster hosts with only minimal orchestration traffic transiting from north to south through the edges. If the Hadoop application traffic transits from north to south, ensure proper sizing of NSX Edge Nodes depending on the workload.

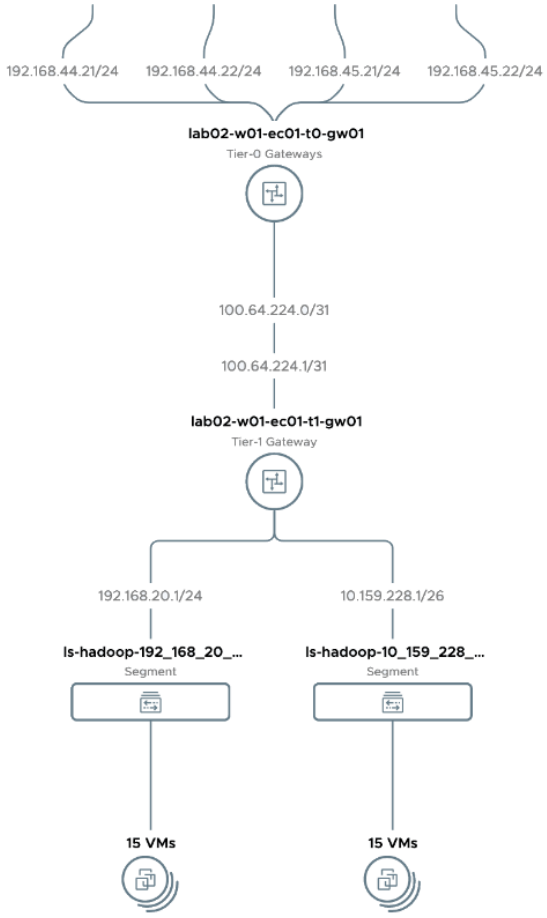

Figure 3. NSX-T Overlay Backed Segment Logical Topology

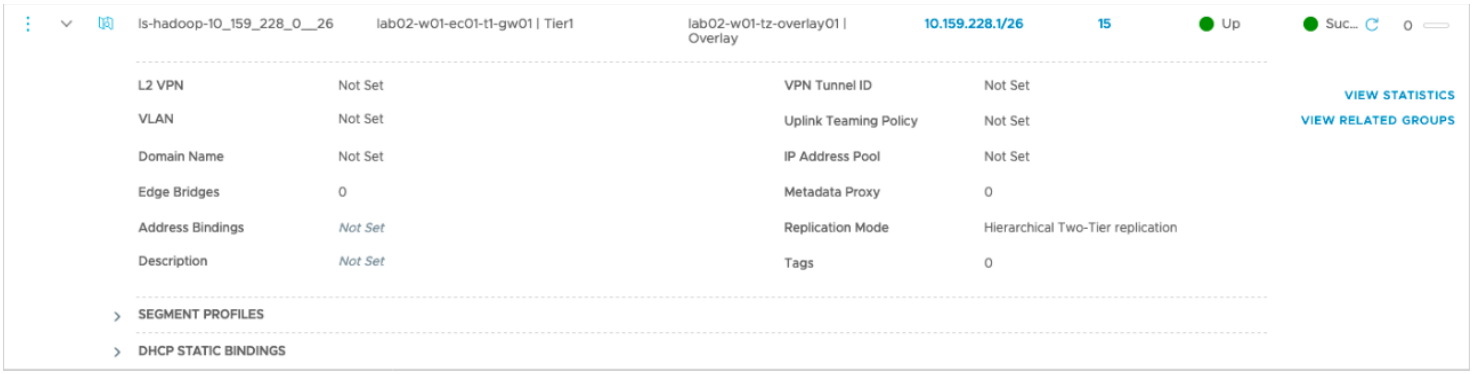

Figure 4. VM Management Network Segment Configuration

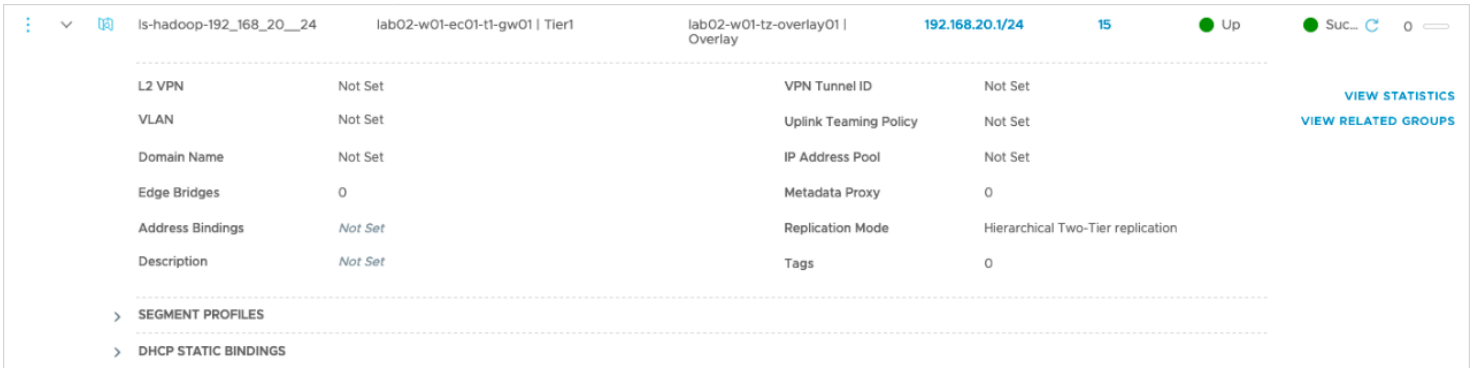

Figure 5. VM Hadoop Network NSX-T Segment Configuration

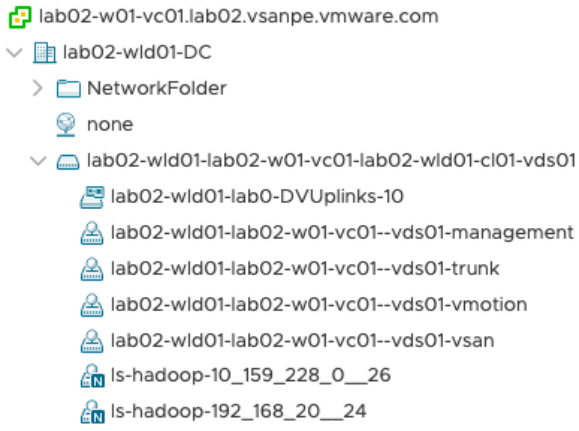

Figure 6. vSphere Distributed Switch with NSX-T Segment

VMs and Storage Configuration

vSAN Configuration

A workload domain with 8-node was deployed with VMware vSAN for storage. Each server was deployed with an identical configuration. Two disk groups were configured per host. Each disk group used one NVMe for the cache tier and four SSDs for the capacity tier, resulting in a datastore capacity of 111.78 TB. vSAN dedupe and compression was disabled.

Storage Policy Based Management (SPBM) allows you to align storage with application demands of the virtual machines. Below are some of the key SPBM parameters set for disks provisioned from vSAN datastore.

vSAN FTT (Failures to Tolerate): With vSAN FTT, availability is provided by maintaining replica copies of data, to mitigate the risk of a host failure resulting in lost connectivity to data or potential data loss. For instance, FTT=1 supports *n*+1 availability by providing a second copy of data on a separate host in the cluster. However, the resulting impact on capacity is that it is doubled.

Stripe width: Number of Disk Stripes Per Object, commonly referred to as stripe width, is the setting that defines the minimum number of capacity devices across which replica of a storage object is distributed.

All testing was carried with the default storage policy of Mirror (FTT=1), Stripe width =1 as shown in Figure 7 and Figure 8.

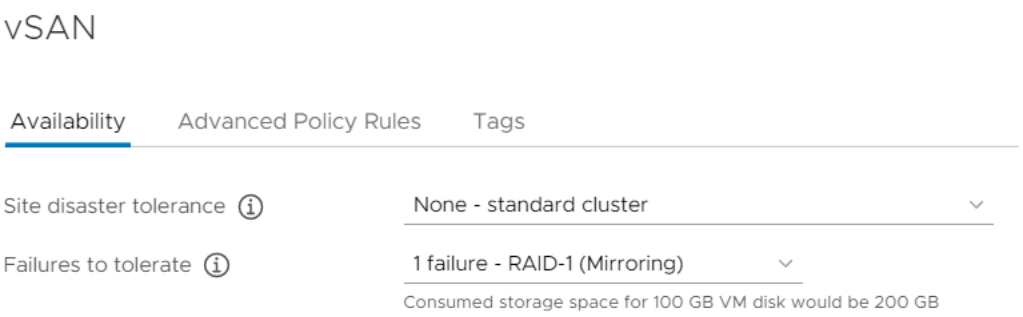

Figure 7. vSAN Storage Policy Availability Settings

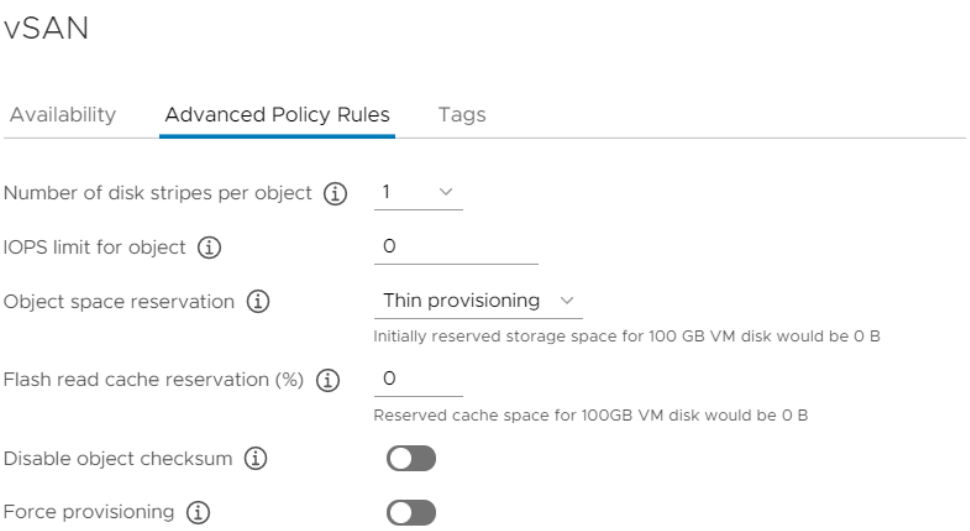

Figure 8. vSAN Storage Policy Advanced Settings

VM Configuration

Two VMs were installed on each server, each Hadoop VM with 14 vCPUs. Using this VMware *KB2113954*, the memory requirement of vSAN was calculated, for this configuration it is ~20% of the server memory, so the remaining 400 GB will be divided equally between the two VMs.

The OS disk will be placed on a dedicated PVSCSI controller and the data disks spread evenly over other three PVSCSI controllers. The six VMDKs on each worker VM formatted using the ext4 filesystem, and the resulting data disks will be used to create the Hadoop filesystem.

At vSAN level, FTT=1 provides data block redundancy; and at HDFS level, replication factor provides redundancy (dfs.replication) two different values: RF2 (Replication Factor =2) and RF3 (Replication Factor =3). Table 4 shows the usable capacity for both RF3 and RF2 configuration depending on the HDFS replication factor, the max HDFS file size is10.54 for RF3 and 15.82 for RF2.

Hadoop HDFS Configuration

The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. HDFS is designed to reliably store very large files across machines in a large cluster. The blocks of a file are replicated for fault tolerance. The block size and replication factor (RF) are configurable per file. Replication Factor is the number of copies a file is stored at Hadoop HDFS level. Two different Replication factors are tested RF3(Default) and RF2. In this testing, the HDFS block size was set to 256MB.

Note: In a single vSAN Cluster implementation, if the vSAN policy used is FTT=1, the guarantee is to tolerate only 1 host failure. Hence to improve availability when feasible for RF3, at least 3 vSAN clusters can be optimal so each HDFS copy is replicated across 3 different vSAN clusters.

Table 4. vSAN Storage Policy and Hadoop Replication Factor Configuration

Feature	vSAN FTT1 & Hadoop RF3	vSAN FTT1 & Hadoop RF2
vSAN replicas	2	2
HDFS replication factor	3	2
Data VMDKs per worker VM	6x 450 GB	6x 450 GB
HDFS configured capacity (12 Worker Nodes)	31.64 TB	31.64TB
HDFS Max file size at default replication factor	10.5 TB	15.82 TB
Total vSAN storage capacity	111 TB	111 TB
Used vSAN for HDFS	63.28	63.28

FTT=0 with vSAN host affinity (RPQ only) is not considered in this testing as it is not supported on VMware Cloud Foundation based deployment.

Cloudera Hadoop Cluster Configuration

As shown in Table 5, there are three types of servers or nodes in a Hadoop cluster based on the key services running in them.

- Gateway/Edge server: One or more gateway servers act as client systems for Hadoop applications and provide a remote access point for users of cluster applications. This also runs the Cloudera Manager components.
- Master server: Run the Hadoop master services such as the HDFS NameNode.
- Worker server (Data Node): Primarily run the resource intensive HDFS DataNode and other distributed processes such as Impala.

In this validation, two ESXi hosts ran infrastructure VMs to manage the Hadoop Cluster. On the first ESXi server, a VM hosted the gateway node running the Cloudera Manager and several other Hadoop functions. These two ESXi hosted the master VMs on which the active and passive Namenode and ResourceManager components and associated services ran. The active name node and Resource Manager ran on different servers for best distribution of CPU load, with the standby on each of the opposite master nodes. This also guarantees the highest cluster availability. For NameNode and ResourceManager high availability, at least three ZooKeeper

services and three HDFS JournalNodes are required. Two of the ZooKeeper and JournalNode services ran on the two ESXi hosts; the third set ran on the first worker node.

The remaining 6 ESXi hosted two Worker nodes per host running the HDFS DataNode and YARN NodeManager and other Hadoop services. As noted above, one of the Data Node VM also ran the Zookeeper and Journal service. With the very small CPU/memory overhead, these processes do not measurably impact the Datanode. However, for larger deployments with other roles running on the infrastructure VMs, it might be necessary to run three ESXi hosts for infrastructure servers, in which case the third ZooKeeper and JournalNodes may be run on one of the infrastructure servers. This placement is shown in Figure 9.

See *Cloudera Runtime Cluster Hosts and Role Assignments* for the recommended role allocations for different cluster sizes.

Table 5. Hadoop VM Configuration

	Gateway/Edge VM (CDP01)	Master VM (CDP02 and CDP03)	Datanode VM CDP04 to CDP15
Quantity	1	2	12
vCPU	14		
Memory	200 GB		
OS VMDK Size	250 GB	100 GB	100 GB
Data Disks	1x 100 GB	1x 100 GB	6x 450GB

The nodes were configured per *CDP Private Cloud Base Hardware Recommendations*.

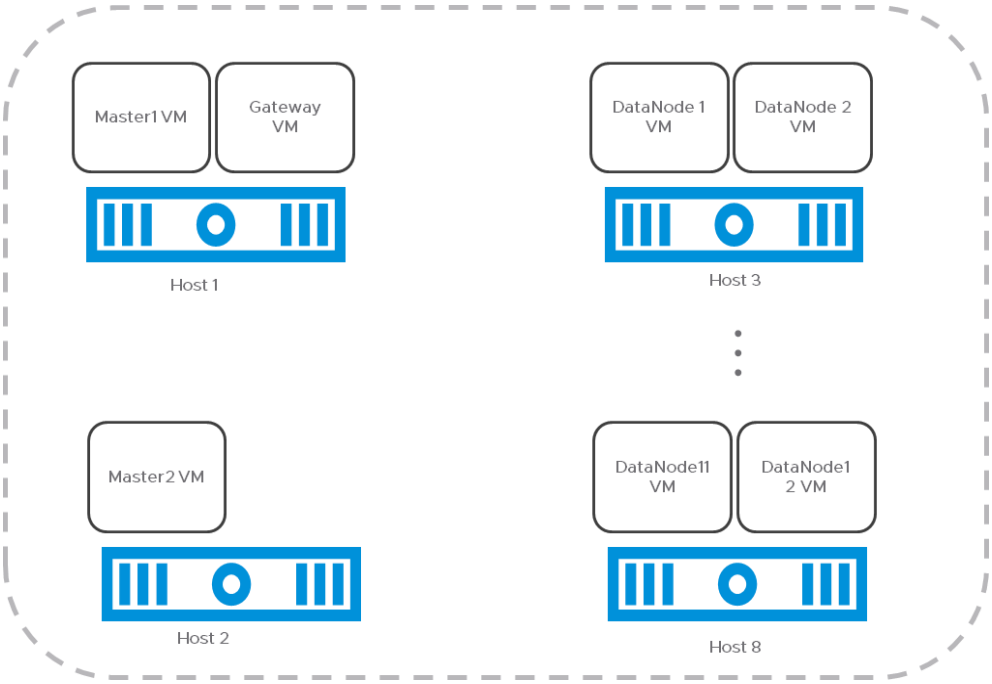

Figure 9. Hadoop VM Placement on ESXi Host

Table 6. Physical host, VM and HVE Node Group Mapping

ESXi Host (Physical Host)	RackID/HVENodegroup-ID	Hadoop VMs role
Host1	/rack1/physicalhost1	Master1 VM,Gateway VM
Host2	/rack1/physicalhost2	Master2 VM
Host3	/rack1/physicalhost3	DataNode 1 VM, DataNode 2 VM
Host4	/rack1/physicalhost4	DataNode 3 VM, DataNode 4 VM
Host5	/rack1/physicalhost5	DataNode 5 VM, DataNode 6 VM
Host6	/rack1/physicalhost6	DataNode 7 VM, DataNode 8 VM
Host7	/rack1/physicalhost7	DataNode 9 VM, DataNode 10 VM
Host8	/rack1/physicalhost8	DataNode 11 VM, DataNode 12 VM

Hadoop Virtualization Extensions (HVE), an open-source Hadoop add-on (*https://issues.apache.org/jira/browse/HADOOP-8468*) is used to prevent multiple copies of a given HDFS block from being placed on the same physical server for availability reasons. HVE adds an additional layer to the HDFS rack awareness, node group, to enable the user to identify which VMs reside on the same physical server. HDFS uses that information in its block placement strategy.

HVE on vSphere has been traditional used to group VMs on the same physical host into an HVE nodegroup. HVE configuration and mapping is done per Table 6. The procedure is provided in Appendix section "Hadoop Virtualization Extension Configuration".

VMware vSphere Distributed Resource Scheduler™ was placed in partially automated mode and it was made sure VMs are always hosted on the respective physical host. For details, see section Deployment Options for Hadoop on VMware vSAN.

Monitoring Tools

We used the following monitoring tools and benchmark tools in the solution testing:

vSAN Performance Service

vSAN Performance Service is used to monitor the performance of the vSAN environment, using the vSphere web client. The performance service collects and analyzes performance statistics and displays the data in a graphical format. You can use the performance charts to manage your workload and determine the root cause of problems.

vSAN Health Check

vSAN Health Check delivers a simplified troubleshooting and monitoring experience of all things related to vSAN. Through the vSphere web client, it offers multiple health checks specifically for vSAN including cluster, hardware compatibility, data, limits, physical disks. It is used to check the vSAN health before the mixed-workload environment deployment.

Cloudera manager

Cloudera Manager provides many features for monitoring the health and performance of the components of your clusters (hosts, service daemons) as well as the performance and resource demands of the jobs running on your clusters.

Platform Validation

Overview

Before the deployment, it is highly recommended to validate the performance capabilities of the intended platform. *HCIBench* is the preferred tool to validate both overall and I/O specific profile performance using synthetic I/O. HCIBench provides the ability to run user-defined workloads as well as a series of pre-defined tests, known as the EasyRun suite. When leveraging EasyRun, HCIBench executes four different standard test profiles that sample system performance and report key metrics.

Cloudera provides storage performance KPIs as the prerequisite of running Cloudera Hadoop on a given system. Also, Cloudera provides a tool-kit to conduct a series of performance tests including a microbenchmark and HBase. See *Cloudera Enterprise Storage Device Acceptance Criteria Guide* for detailed information.

Beyond synthetic IO testing, there are standard Hadoop benchmarks that simulate Hadoop workloads. These benchmarks may be used by customers as a starting point for characterizing their Hadoop clusters, but their own applications will provide the best guidance for choosing the correct architecture.

Workload Performance Comparison between RF2 and RF3

We validated the platform using some of the industry standard MapReduce benchmarks like TeraSort suite and popular Cloudera services. The benchmark was performed to compare Hadoop RF2 and RF3 configuration. The results for one of the MapReduce benchmark (Terasort Suite) and Cloudera service (Impala) were illustrated below.

Terrasort Suite Benchmark

The TeraSort suite (TeraGen/TeraSort/TeraValidate) is the most used Hadoop benchmark and ships with all Hadoop distributions.

TeraGen is a large block and sequential write-heavy workload. TeraSort starts with reads as the dataset is read from HDFS, then moves to a read/write mix as data is shuffled between task during the sort, and then concludes with a short write-dominated phase as the sorted data is written to HDFS. TeraValidate is a brief read-only phase.

The workload was run with 1TB and 3TB data size for both RF2 and RF3, and the elapsed time for each of the TeraSort suite phases was measured. The test methodology was provided in the Appendix section Test Methodology – Terrasort Suite. Table 7 shows the elapsed time difference (%) between the RF2 and RF3 configuration. Positive (+) % indicates RF2 is better than RF3 and Negative (-) % means RF3 is better than RF2.

TeraGen being write intensive to disk clearly has a performance advantage of over 50% when number of HDFS replicas is reduced. During the TeraSort and TeraValidate phases, the elapsed time is similar between RF3 and RF2 since there is no significant write to disk.

Table 7. TeraSort Suite Benchmark RF2 Versus RF3

Data size	Performance Advantage RF2 over RF3 (Elapsed Time % difference)		
	TeraGen	TeraSort	TeraValidate
1 TB	54%	4%	2%
3 TB	59%	- 0.5%	3.3%

Cloudera Impala with TPC-DS queries

The Apache Impala provides high-performance and low-latency SQL queries on data stored in popular Apache Hadoop file formats. It is composed of the Impala, Hive Metastore and Clients. For details, see *https://docs.cloudera.com/runtime/7.0.2/impala-overview/topics/impala-overview.html.*

TPC-DS derived queries were used to run on Impala for this validation. Test methodology is provided in the Appendix section Test Methodology - Impala based on TPC-DS queries. The SQL queries were generated for 1TB and 3TB datasets. All 90 queries were run, and the elapsed time was measured. The test was repeated for three times and the average of the total query time was calculated. Total query time is the time to take to complete all the 90 queries.

Table 8 shows % difference between the total query time for RF2 and RF3 configuration. Positive (+) % indicates RF2 is better than RF3 and Negative (-) % means RF3 is better than RF2. With this workload, the query performance difference between RF2 and RF3 is less than 5%.

Table 8. Impala TPC-DS Query Time RF2 Versus RF3

Data Size (Scale)	Performance Advantage RF2 over RF3 (Total query time % difference)
1 TB	-2.3%
3 TB	4.2%

Deployment Options for Hadoop on VMware vSAN

Single vSAN Cluster Hadoop Solution

In a single vSAN cluster solution, all the Hadoop nodes are deployed on a single vSAN cluster. If FTT=1 is used at vSAN layer, and RF3 is used at HDFS layer to protect against failure, in total there are six copies for each block (three copies on the HDFS layer and each copy of this is mirrored in vSAN layer). In this case, even though there are 3 copies at HDFS level, due to vSAN FTT=1, the Hadoop cluster can tolerate only one host failure because all three HDFS replicas might be placed on disks in the same physical host on the vSAN layer. Hence single vSAN cluster with FTT=1 and RF=3 does not guarantee availability for more than 1 host failure.

If there are multiple Hadoop VMs per ESXi host, HVE Hadoop Virtualization Extension is recommended.

Table 6 shows the simplest form of HVE grouping followed in traditional Hadoop on vSphere. One physical host is mapped to one node group. If vSphere HA is enabled, the Hadoop VMs may move to other available ESXi hosts in the vSphere cluster and potentially place the data nodes storing the same replica copies on the same physical host violating the rule. To avoid this, use vSphere DRS (Distributed Resource Scheduler) Host to VM affinity rules.

If vSphere HA is enabled instead of using one ESXi host to HVE nodegroup mapping, add VMs from multiple ESXi hosts in one HVE nodegroup configuration so there are more than one ESXi hosts in a HVE nodegroup as shown in Table 9. This will help virtual machines to restart in other available ESXi hosts within the same HVE nodegroup.

DRS affinity rule will help to specify how vSphere HA applies the rule during the virtual machine failover. VM-Host affinity rules are placed so a set of virtual machines under specific nodegroup stay on the same physical hosts avoiding the operational overhead of maintaining this separately by administrators.

Table 9. HVE Node Group Combining Multiple Physical Host

ESXi Host (Physical Host)	RackID/HVENodegroup-ID	Hadoop VM Placement (Virtual Machine)	vSphere VM Group	vSphere Host group
Host1	/rack1/group1	Master1 VM, Gateway VM	vmgroup1	Hostgroup1
Host2	/rack1/group2	Master2 VM	vmgroup2	Hostgroup2
Host3	/rack1/group1	DataNode 1 VM, DataNode 2 VM	vmgroup1	Hostgroup1
Host4	/rack1/group1	DataNode 3 VM, DataNode 4 VM	vmgroup1	Hostgroup1
Host5	/rack1/group2	DataNode 5 VM, DataNode 6 VM	vmgroup2	Hostgroup2
Host6	/rack1/group2	DataNode 7 VM, DataNode 8 VM	vmgroup2	Hostgroup2
Host7	/rack1/group3	DataNode 9 VM, DataNode 10 VM	vmgroup3	Hostgroup3
Host8	/rack1/group3	DataNode 11 VM, DataNode 12 VM	vmgroup3	Hostgroup3

Sample steps to create VM and host groups for the HVE node group:

As shown in Figure 10, create virtual machine to host rules so the VM is failed over to specific hosts in that group only, there are different choices while setting up DRS rules notice the use of "Must run on hosts in group" rule this will make sure VM is restarted on the available hosts in that group only. For more information on HA and DRS affinity rules, see *https://docs.vmware.com/en/VMware-vSphere/7.0/com.vmware.vsphere.avail.doc/GUID-E137A9F8-17E4-4DE7-B986-94A0999CF327.html*.

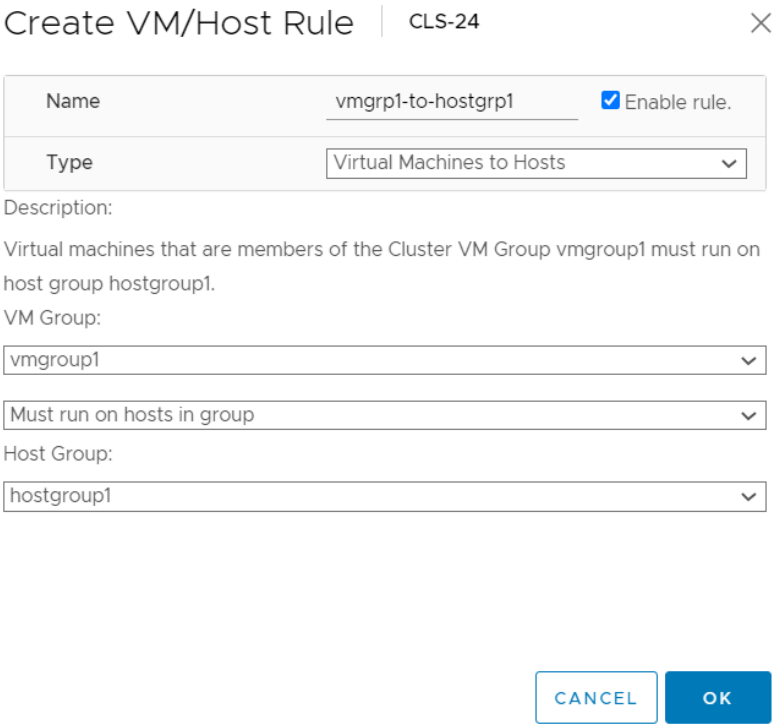

Figure 10. Create VM and Host Affinity Rules Using vCenter

Multiple vSAN Cluster Hadoop Solution

Like the single cluster solution, Hadoop solution can be deployed on multiple vSAN clusters and use HVE to make sure each copy of the HDFS replicas is stored on a different vSAN cluster. The number of vSAN clusters would depend on the Hadoop Replication factor. A pictorial representation is shown in Figure 11. Table 10 shows the summary of the various deployment options for Hadoop on vSAN storage.

Figure 11. Hadoop on Multiple vSAN Clusters

Table 10. Summary of Deployment Options, Benefits, and Tradeoffs

Single or Multiple vSAN Clusters	vSAN Policy	Hadoop Config Replication Factor (RF)	Benefits	Tradeoffs
Single vSAN Cluster	vSAN mirror to protect 1 failure (FTT=1)	RF = 3	• Simplified Day2 operations due to vSphere features (HA, DRS, vMotion) • vSAN SPBM advantage	• Some tradeoff in performance • Requires 1.5x storage compared to RF = 2 • Even though there are three HDFS replicas, it can protect from only single component (host, disk) failure, governed by vSAN FTT=1.
		RF = 2	• Capacity savings by reducing copies in HDFS • Improved write performance • Simplified Day 2 operations due to vSphere features (HA, DRS, vMotion) • vSAN SPBM advantage	• Minimal tradeoff in performance • Potential HDFS read optimization benefit lost due to reduction in HDFS copies
	vSAN mirror to protect 1 failure (FTT=1)	RF = 3 • Place HDFS Copies in different vSAN Cluster (at least 3 vSAN clusters)	• *HDFS copies in separate vSAN cluster supports to achieve independent Failure domain* • Simplified Day2 operations due to vSphere features (HA, DRS, vMotion) • vSAN SPBM advantage	• Some tradeoff in performance • Requires 1.5x storage compared to RF = 2

Multiple vSAN Clusters		RF = 2 • Place HDFS copies in a different vSAN Clusters (at least 2 vSAN clusters)	• HDFS Copies in separate vSAN cluster supports to achieve independent Failure domain • Capacity savings by reducing copies in HDFS • Improved write performance • Simplified Day 2 operations due to vSphere features (HA, DRS, vMotion) • vSAN SPBM advantage	• Minimal tradeoff in performance • Potential HDFS read optimization benefit lost due to reduction in HDFS copies

Production Criteria Recommendations

vCPU and Memory

Hadoop workloads are CPU and memory intensive, and the workloads require proper sizing of the VM vCPU and memory to achieve optimal performance. In vSphere environments running mixed workloads, the use of vCPU and memory reservations should be considered to ensure adequate compute resources.

Recommendation: Avoid CPU and memory overcommitment

Network

Hadoop workloads are network intensive and network port bandwidth is consumed by Hadoop VM traffic (application, HDFS traffic) and vSAN distributed storage traffic. Considering this size enough port with enough network bandwidth and choose network switches with non-blocking architecture with high buffers.

Recommendation: Use minimum 4 x 10Gbps port, preferably use larger bandwidth port like 25Gbps or higher.

vSAN FTT

The Number of Failures to Tolerate capability addresses the key customer and design requirement of availability. With FTT, availability is provided by maintaining replica copies of data, to mitigate the risk of a host failure resulting in lost connectivity to data or potential data loss.

Recommendation: FTT=1

vSAN RAID

vSAN has the ability to use RAID 1 for mirroring or RAID 5/6 for Erasure Coding. Erasure coding can provide the same level of data protection as mirroring (RAID 1), while using less storage capacity.

Recommendation: RAID 1

vSAN Dedupe and Compression

Deduplication and compression can enhance space savings capabilities; however, for optimal performance, we do not recommend enabling deduplication and compression.

Recommendation: Disable deduplication and compression

vSAN Encryption

vSAN can perform data at rest encryption. Data is encrypted after all other processing, such as deduplication, is performed. Data at rest encryption protects data on storage devices, in case a device is removed from the cluster. Use encryption per your company's Information Security requirements.

Recommendation: Enable encryption required by your company Information Security Policy.

vSphere DRS

DRS works on a cluster of ESXi hosts and provides resource management capabilities like load balancing and VM placement. DRS also enforces user-defined resource allocation policies at the cluster level, while working with system-level constraints.

Recommendation: DRS–partially automated

vSphere High Availability

vSphere HA provides high availability for virtual machines by pooling the virtual machines and the hosts they reside on into a cluster. Hosts in the cluster are monitored and in the event of a failure, the virtual machines on a failed host are restarted on alternate hosts.

Recommendation: HA Enabled, must use VM-Host DRS affinity rules to place VMs as per HVE configuration.

Conclusion

VMware Cloud Foundation delivers flexible, consistent, secure infrastructure, and operations across private and public clouds and is ideally suited to meet the demands of Hadoop. Using micro-segmentation, administrators can isolate traffic to a given set of consumers for workload and regulatory purposes. With SPBM, VMware Cloud Foundation can scale performance for both department and enterprise level clouds. Data-at-rest encryption meets both operational and regulatory compliance. CTO's and CFO's budget objectives can be achieved with dynamic provisioning, allowing enterprises to scale-up and scale-down as needed. Further VMware Cloud Foundation with VMware vSAN provides simplicity in management and Day 2 operations for Hadoop workloads.

Reference

- *CDP Private Cloud*
- *VMware vSphere*
- *VMware vSAN*
- *VMware NSX Data Center*

Appendix

Hadoop Virtualization Extension Configuration Procedure

In this example, HVE is configured for the requirement per Table 6.

Using Cloudera Manager Go to HDFS -> Advance configuration and update the following two configuration files: hdfs-site.xml and core-site.xml as shown in Figure 12 and Figure 13.

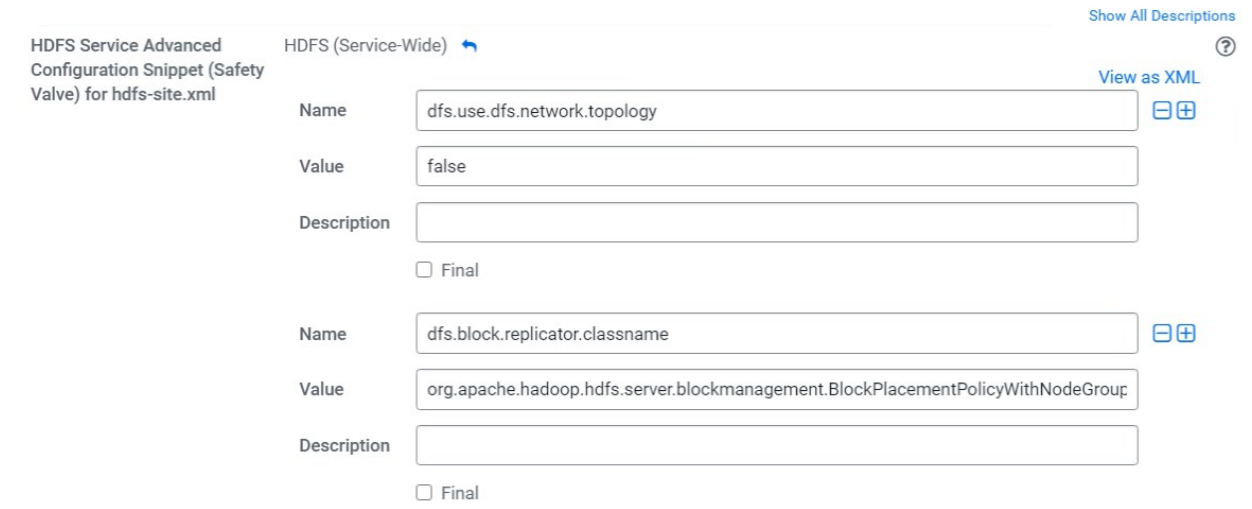

Figure 12. HDFS Service Advanced Configuration Snippet (Safety Valve) for hdfs-site.xml

XML View of hdfs-site.xml

```
<property><name>dfs.use.dfs.network.topology</name><value>false</value></property><property><name>dfs.block.replicator.classname</name><value>org.apache.hadoop.hdfs.server.blockmanagement.BlockPlacementPolicyWithNodeGroup</value></property>
```

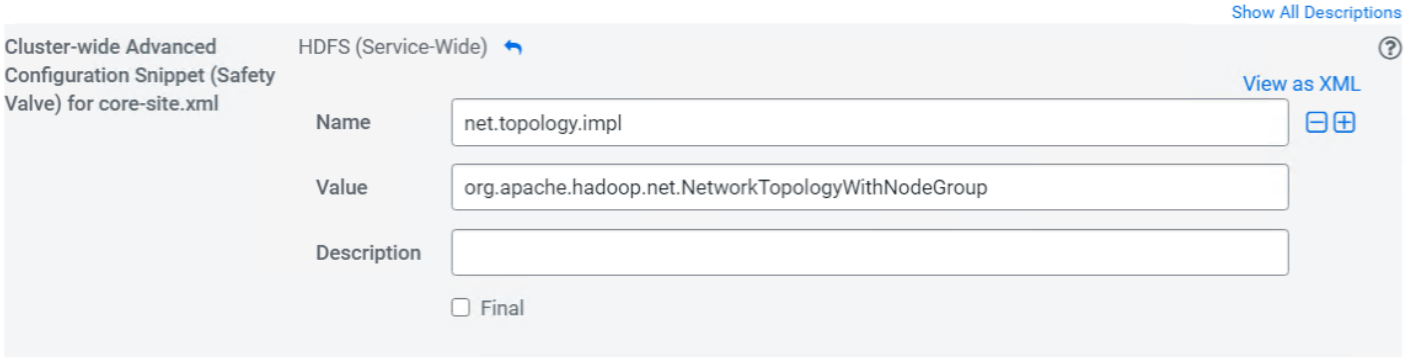

Figure 13. Cluster-wide Advanced Configuration Snippet (Safety Valve) for core-site.xml

XML View of core-site.xml

```
<property><name>net.topology.impl</name><value>org.apache.hadoop.net.NetworkTopologyWithNodeGroup</value></property>
```

Finally, using Cloudera manager map the Hadoop VMs to respective Node group,.Figure 14 from Cloudera Manager shows that VMs from each physical host (phy-hostxx) are assigned to a separate node group.

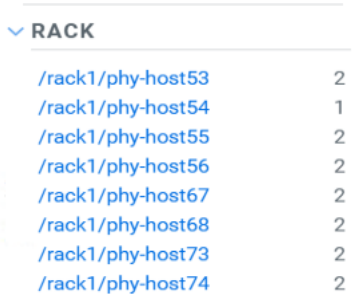

Figure 14. Hadoop Rack Configuration Map (Cloudera Manager Screenshot)

Test Methodology – Terrasort Suite

The TeraSort suite (TeraGen/TeraSort/TeraValidate) is the mostly used Hadoop benchmark and ships with all Hadoop distributions. By first creating a large dataset, then sorting it, and finally validating that the sort was correct, the suite exercises many of Hadoop's functions and stresses CPU, memory, disk, and network.

TeraGen generates a specified number of 100 byte records, each with a randomized key occupying the first 10 bytes, creating the default number of replicas as set by dfs.replication. In these tests, 10 and 30 billion records were specified resulting in datasets of 1 and 3 TB. TeraSort sorts the TeraGen output, creating one replica of the sorted output. In the first phase of TeraSort, the map tasks read the dataset from HDFS. Following that is a CPU-intensive phase where map tasks partition the records, they have processed by a computed key range, sort them by key, and spill them to disk. At this point, the reduce tasks take over, fetch the files from each mapper corresponding to the keys associated with that reducer, and then merge the files for each key (sorting them in the process) with several passes, and finally write to disk. TeraValidate, which validates that the TeraSort output is indeed in sorted order, is mainly a read operation with a single reduce task at the end.

TeraGen is a large block, sequential write-heavy workload. TeraSort starts with reads as the dataset is read from HDFS, then moves to a read/write mix as data is shuffled between task during the sort, and then concludes with a short write-dominated phase as the sorted data is written to HDFS. TeraValidate is a brief read-only phase.

1 vcore was assigned to each map and the reduce tasks. With 168 total cores available on the cluster, 168 1-vcore tasks could run simultaneously. However, a vcore must be set aside to run the ApplicationMaster, leaving 167 tasks. With this number of tasks, each task container was assigned 11 GB of memory to consume a total 1,848 GB in the cluster.

The test was run for 1 TB and 3 TB dataset. The commands to run the three components of the TeraSort suite (TeraGen, TeraSort, and TeraValidate) are shown below for 3 TB dataset.

TeraGen command

RF=3

time hadoop jar ~/hadoop-mapreduce-examples.jar teragen -Ddfs.replication=3 -Dmapreduce.job.maps=167 -Dmapreduce.map.memory.mb=11264 -Dmapreduce.map.cpu.vcores=1 30000000000 teragen3TB_input

RF=2

time hadoop jar ~/hadoop-mapreduce-examples.jar teragen -Ddfs.replication=2 -Dmapreduce.job.maps=167 -Dmapreduce.map.memory.mb=11264 -Dmapreduce.map.cpu.vcores=1 30000000000 teragen3TB_input

TeraSort and TeraValidate commands (For RF=3 and RF=2)

time sudo -u hdfs hadoop jar ~/hadoop-mapreduce-examples.jar terasort -Dmapreduce.job.reduces=167 -Dmapreduce.map.memory.mb=11264 -Dmapreduce.reduce.memory.mb=11264 -Dmapreduce.map.cpu.vcores=1 teragen3TB_input terasort3TB_output

time sudo -u hdfs hadoop jar ~/hadoop-mapreduce-examples.jar teravalidate -Dmapreduce.map.memory.mb=11264 terasort3TB_output terasort3TB_validate

Test Methodology - Impala based on TPC-DS queries

The Apache Impala provides high-performance and low-latency SQL queries on data stored in popular Apache Hadoop file formats. It is composed of Impala, Hive Metastore, and Clients. Impala service coordinates and executes queries received from clients. Queries are distributed among Impala nodes and these nodes act as workers executing parallel query fragments. Hive Metastore stores information about the data available to Impala. Clients are the interfaces which are typically used to issue queries. For details, see *https://docs.cloudera.com/runtime/7.0.2/impala-overview/topics/impala-overview.html.*

TPC-DS derived queries were used to run on Impala for this validation. The TPCDS KIT in this github repo (*https://github.com/cloudera/impala-tpcds-kit)* was used. The query templates and sample queries provided in this repo are compliant with the standards set out by the TPC-DS benchmark specification and include only minor query modifications (MQMs) as set out by section 4.2.3 of the specification.

At the time of running these tests, the repo supported 90 queries out of the 99 available in TPC DS KIT. The queries that were not supported are 8,9,14,23,24,38,44,45 and 87 (recently the repo has added support for these queries).

The SQL queries were generated for 1 TB and 3 TB datasets. All 90 queries were run and the elapsed time was measured. The test was repeated for three times and the average total query time was measured.

About the Author

Palani Murugan, Senior Solutions Architect in VMware Cloud Platform Business Unit wrote the draft with contributions from following members

- Chen Wei, Staff Solutions Architect in VMware Cloud Platform Business Unit
- Ali Bajwa, Director in Cloudera Partner Engineering
- Nijjwol Lamsal, Partner Solutions Engineer in Cloudera
- Harsh Shah, Partner Solutions Engineer in Cloudera

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