cloudera®

Cloudera Enterprise Reference Architecture for VMware Deployments with Isilon-based Storage

Important Notice

Cloudera, Inc.
1001 Page Mill Road, Building 2
Palo Alto, CA 94304-1008
info@cloudera.com
US: 1-888-789-1488
Intl: 1-650-843-0595
www.cloudera.com

Release Information

Date: November 19, 2015

Table of Contents

Executive Summary

This document is a high-level design and best-practices guide for deploying Cloudera Enterprise on a VMware vSphere®-based infrastructure with a shared storage back end.

This document describes the architecture for running Cloudera Enterprise on VMware vSphere-based infrastructure with shared Isilon-based storage.

Audience and Scope

This guide is for IT architects who are responsible for the design and deployment of virtualized infrastructure and a shared storage platform in the data center, as well as for Hadoop administrators and architects who will be data center architects or engineers or who collaborate with specialists in that space.

This document describes Cloudera recommendations on the following topics:

- Storage Area Network considerations
- Storage array considerations
- Data network considerations
- Virtualization hardware/platform considerations
- Virtualization strategy for the Cloudera software stack

Glossary of Terms

Term	Description
DataNode	Worker nodes of the cluster to which the HDFS data is written.
DRS	Distributed Resource Scheduler. The software that controls movement of VMs and storage on a VMware cluster.
HBA	Host bus adapter. An I/O controller that is used to interface a host with storage devices.
HDD	Hard disk drive.
HDFS	Hadoop Distributed File System.
High Availability	Configuration that addresses availability issues in a cluster. In a standard configuration, the NameNode is a single point of failure (SPOF). Each cluster has a single NameNode, and if that machine or process became unavailable, the cluster as a whole is unavailable until the NameNode is either restarted or brought up on a new host. The secondary NameNode does not provide failover capability. High availability enables running two NameNodes in the same cluster: the active NameNode and the standby NameNode. The standby NameNode allows a fast failover to a new NameNode in case of machine crash or planned maintenance.
HVE	Hadoop Virtualization Extensions. Enables proper placement of data blocks and scheduling of YARN jobs in a virtualized environment in which multiple copies of any single block of data or YARN jobs are not placed/scheduled on VMs that reside on the same hypervisor host. The YARN component of HVE is work in progress and won't be supported in CDH 5.4 (YARN-18).

JBOD	Just a bunch of disks. In contrast to disks configured through software or hardware with redundancy mechanisms for data protection.
Job History Server	Process that archives job metrics and metadata. One per cluster.
LBT	Load-based teaming. A teaming policy that is traffic-load aware and ensures physical NIC capacity of a NIC team is optimized.
LRO	Large receive offload. A technique used to improve throughput of network connections by coalescing multiple incoming packets from a single stream into a large receive buffer before passing them up the networking stack.
LUN	Logical unit number. Logical units allocated from a storage array to a host. This looks like a SCSI disk to the host, but it is only a logical volume on the storage array side.
NameNode	The metadata master of HDFS essential for the integrity and proper functioning of the distributed filesystem.
NIC	Network interface card.
NIOC	Network I/O Control.
NodeManager	The process that starts application processes and manages resources on the DataNodes.
NUMA	Nonuniform memory access. Addresses memory access latency in multisocket servers, where memory that is remote to a core (that is, local to another socket) needs to be accessed. This is typical of SMP (symmetric multiprocessing) systems, and there are several strategies to optimize applications and operating systems. vSphere ESXi can be optimized for NUMA. It can also present the NUMA architecture to the virtualized guest OS, which can then leverage it to optimize memory access. This is called vNUMA.
PDU	Power distribution unit.
QJM **QJN**	Quorum Journal Manager. Provides a fencing mechanism for high availability in a Hadoop cluster. This service is used to distribute HDFS edit logs to multiple hosts (at least three are required) from the active NameNode. The standby NameNode reads the edits from the JournalNodes and constantly applies them to its own namespace. In case of a failover, the standby NameNode applies all of the edits from the JournalNodes before promoting itself to the active state. Quorum JournalNodes. Nodes on which the journal services are installed.
RDM	Raw device mappings. Used to configure storage devices (usually logical unit numbers (LUNs)) directly to virtual machines running on VMware.

RM	ResourceManager. The resource management component of YARN. This initiates application startup and controls scheduling on the DataNodes of the cluster (one instance per cluster).
SAN	Storage area network.
SIOC	Storage I/O Control.
ToR	Top of rack.
TSO	TCP segmentation offload.
VM	Virtual machine.
vMotion	VMware term for live migration of virtual machines across physical hosts.
ZK	ZooKeeper. A centralized service for maintaining configuration information, naming, and providing distributed synchronization and group services.

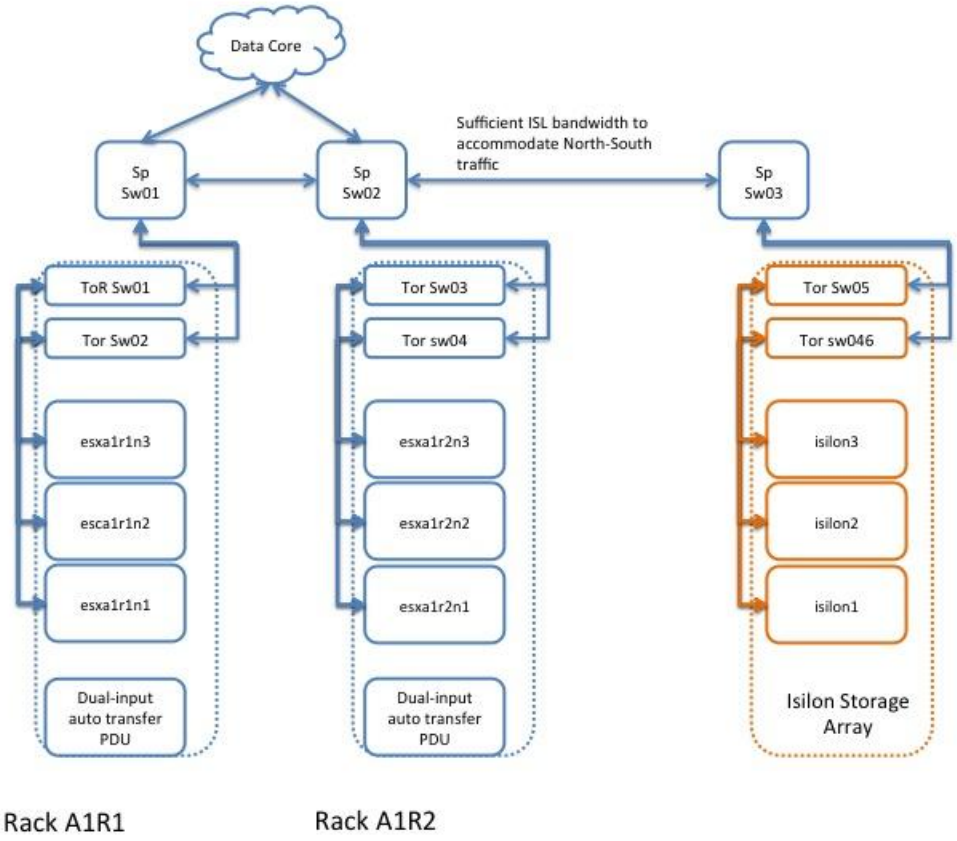
Isilon Distributed Storage Array for HDFS and VMware-based VMs as Compute Nodes

This model decouples the HDFS DataNode functionality from the YARN NodeManager and other components of Cloudera Enterprise.

In this architecture, Isilon acts as the HDFS/storage layer, and the VMs only provide the compute resources needed.

Considerations for a storage component are not required; however, from a vSphere design perspective, the storage component must be factored into the distributed vSwitch design. This is noted in VMware vSphere Design Considerations.

Physical Cluster Topology



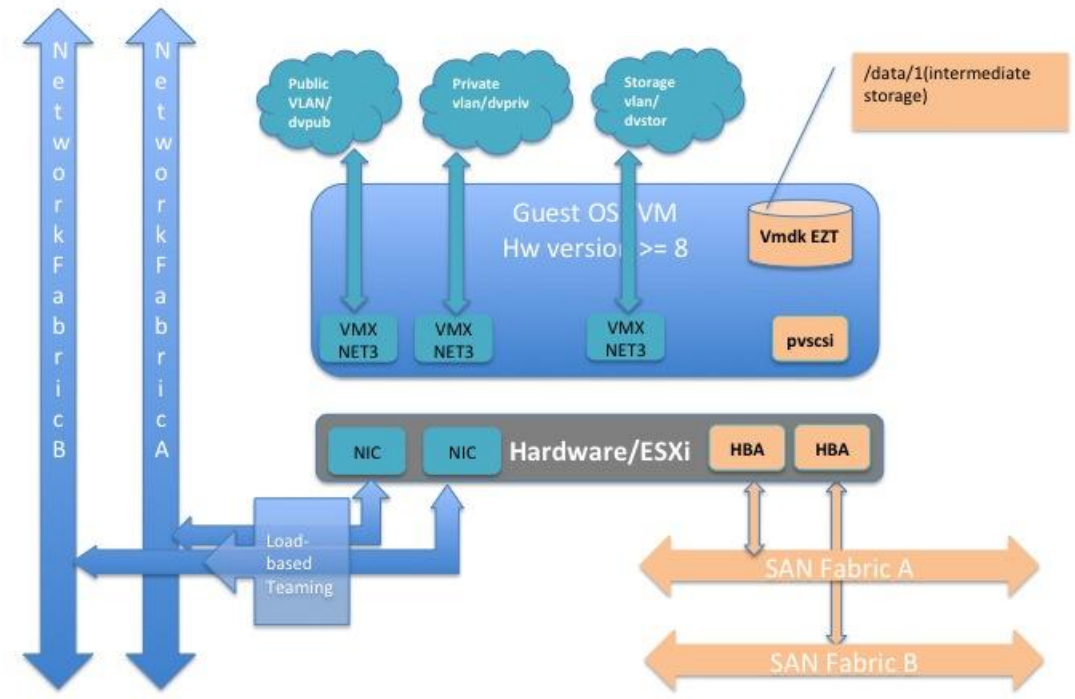
Note: In each rack of compute nodes, EMC recommends interspersing Isilon storage nodes connected to the respective ToR switches (if possible). For example, for two racks of compute nodes, distribute the Isilon storage nodes between the first and second rack, sharing the ToR switches. The Isilon storage nodes share an InfiniBand back end to provide better front-end performance (10 GB Ethernet).

Physical Cluster Component List

Component	Configuration	Description	Quantity
Physical servers	Two-socket, 6-10 cores per socket > 2 GHz; minimally 256 GB RAM.	vSphere hosts that house the various VMs/guests.	TBD (based on cluster design).
NICs	Dual-port 10 Gbps Ethernet NICs. The connector type depends on the network design; could be SFP+ or Twinax.	Provide the data network services for the VMware vSphere cluster.	At least two per physical server.
Internal HDDs	Standard OS sizes.	The ESXi hypervisor requires little storage, so size is not important. These ensure continuity of service on server resets.	Two per physical server.
Ethernet ToR/leaf switches	Minimally 10 Gbps switches with sufficient port density to accommodate the VMware cluster. These require enough ports to create a realistic spine-leaf topology providing ISL bandwidth above a 1:4 oversubscription ratio (preferably 1:1).	Although most enterprises have mature data network practices, consider building a dedicated data network for the Hadoop cluster.	At least two per rack.
Ethernet spine switches	Minimally 10 Gbps switches with sufficient port density to accommodate incoming ISL links and ensure required throughput over the spine (for inter-rack traffic).	Same considerations as for ToR switches.	Depends on the number of racks.

Note: Low-latency workloads are subject to network latency, because all data traffic between compute nodes and HDFS (Isilon-based) is north-south.

Logical Cluster Topology



For the YARN NodeManagers, data protection at the HDFS level is not required, because the VMs are running only the compute part of the cluster.

The minimum requirements to build out the cluster are:

- Three master nodes (VMs)
- The number of compute nodes/worker nodes (VMs) depends on cluster size

The following table identifies service roles for different node types.

Create Distributed Resource Scheduler (DRS) rules so that there is strong negative affinity between the master node VMs. This ensures that no two master nodes are provisioned or migrated to the same physical vSphere host. Alternately, you can do this when provisioning through vSphere Big Data Extensions by specifying "instancePerHost=1", which asserts that any host server should have at most one instance of a master node VM. (See the BDE CLI guide (PDF) for more details.)

	Master Node	Master Node	Master Node	YARN NodeManager nodes 1..n
ZooKeeper	ZooKeeper	ZooKeeper	ZooKeeper	
YARN	ResourceManager	ResourceManager	History Server	NodeManager
Hive			MetaStore, WebHCat, HiveServer2	
Management (misc)	Cloudera Agent	Cloudera Agent	Cloudera Agent, Oozie, Cloudera Manager, Management Services	Cloudera Agent
Navigator			Navigator, Key Management Services	
Hue			Hue	
HBASE	HMaster	HMaster	HMaster	RegionServer
Impala			StateStore, Catalog	Impala Daemon

NOTE: Low-latency workloads are subject to network latency, because all data traffic between compute nodes and HDFS (Isilon-based) is north-south traffic.

The following table provides size recommendations for the VMs. This depends on the size of the physical hardware provisioned, as well as the amount of HDFS storage and the services running on the cluster.

Component	Configuration	Description	Quantity
Master nodes: **two-socket with 6-10 cores/socket > 2 GHz; minimally 128 GB RAM; 4-6 disks**	VMs or bare metal. If VMs, do not house the ResourceManager node and the standby in the same chassis (or blade chassis if using blades).	Nodes that house the Cloudera master services and serve as the gateway/edge device that connects the rest of the customer's network to the Cloudera cluster.	Three (for scaling up to 100 cluster nodes).
YARN NodeManagers: **two-socket with 6-10 cores/socket > 2 GHz; minimally 128 GB RAM**	VMs that can be deployed as needed on the vSphere cluster, without oversubscription of either CPU or memory resources.	Nodes that house the YARN node managers and additional required services. Adjust memory sizes based on the number of services, or provision additional capacity to run additional services.	TDB (based on customer needs).

The following table provides recommendations for storage allocation.

Node/Role	Disk Layout	Description
Management/Master	• 1 x 200 GB OS • Swap partition <= 2 GB • 4 x 500 GB data	Avoid fracturing the filesystem layout into multiple smaller filesystems. Instead, keep a separate "/" and "/var".
YARN NodeManager nodes	• 1 x 200 GB OS • Approximately 20% of total DFS storage (in this case Isilon storage) needs to be provisioned as intermediate storage on NodeManagers. The storage can either be NFS mounts from the Isilon storage array or SAN-based VMFS/VMDK storage. Distribute the 20% of capacity evenly across all the NodeManager nodes, each with its own mount point and filesystem.	Avoid fracturing the filesystem layout into multiple smaller filesystems. Instead, keep a separate "/" and "/var". For example, for 10 TB of total storage in Isilon, 2 TB is needed for intermediate storage. If storage is SAN-based, for 20 nodes, reserve 100 GB LUNs/datastores/VMDKs to each node.

Note: The OS and intermediate storage referenced in the table above can reside on SAN-based platforms such as the EMC vBlock. There are no restrictions in terms of storage back end or deployment model for either the OS bits or intermediate storage.

Supportability/Compatibility Matrix

CDH	Cloudera Manager	OneFS	Supported
5.4.4 and higher (HDFS 2.6)	5.4	7.2.0.3	All services except Navigator

*Navigator support is contingent on iNotify and fsmanage functionality being added into OneFS.

Environment Sizing and Platform Tuning Considerations

Start with the following guidelines for compute node sizing and selection. The number of Isilon nodes depends on required storage capacity and back-end performance considerations. Work with the Cloudera and EMC sales teams to determine back-end requirements.

- Default option—Cloudera and EMC recommend a starting configuration with a ratio of 2:1 for compute nodes to EMC Isilon storage nodes. So, if the Isilon backend has four storage nodes, use eight compute nodes.
- Heavy IO option—When higher IO performance is required, Cloudera and EMC recommend a 1.5:1 ratio for compute nodes to Isilon storage nodes. So, for four storage nodes in the backend, use six compute nodes.

Note: These estimates are provided as a guideline. Cloudera recommends running a pilot with a preliminarily sized cluster, and then fine-tuning the requirements based on empirical data (corresponding to specific workloads).

VMware vSphere Design Considerations

Network Switch Configuration

Employ standard vswitches and configure them for each ESXi host in the cluster. The key configuration parameter to consider is the MTU size, ensuring that the same MTU size is set at the physical switches, guest OS, ESXi VMNIC, and the vswitch layers. This is relevant when enabling jumbo frames, which is recommended for Hadoop environments.

Disk Multipathing Configuration

Disk multipathing (DMP) policy uses round robin (RR). The storage array vendor might have specific recommendations. Not every vendor supports RR, so use an appropriate DMP algorithm.

This has little impact on this deployment model. Use best practices for provisioning storage for the OS drives and intermediate drives.

Storage Group Configuration

Each provisioned disk is mapped to either:

- One vSphere datastore (which in turn contains one VMDK or virtual disk), or
- One raw device mapping (RDM)

NOTE: In this case, this is only relevant for the OS disks and any intermediate storage disks that might be provisioned in the cluster.

Storage Configuration

Set up virtual disks in “independent persistent” mode for optimal performance. Eager Zeroed Thick virtual disks provide the best performance.

Partition alignment at the VMFS layer depends on the storage vendor. Misaligned storage impacts performance.

Disable SIOC, and disable storage DRS.

vSphere Tuning Best Practices

Power Policy is an ESXi parameter. The balanced mode may be the best option. In some cases, performance might be more important than power optimization. Evaluate your environment and choose accordingly.

Avoid memory and CPU overcommitment. Use large pages for Hypervisor. For network tuning, enable advanced features such as TSO, LRO, scatter gather, interrupt coalescing, and so on.

Guest OS Considerations

Special tuning parameters might be needed to optimize performance of the guest OS in a virtualized environment. In general, normal tuning guidelines apply, but specific tuning might be needed depending on the virtualization driver used.

Generic Best Practices

Minimize unnecessary virtual hardware devices. Choose the appropriate virtual hardware version; check the latest version and understand its capabilities.

NIC Driver Type

VMXNET3 is supported in RHEL 6 and CentOS 6 with the installation of VMware tools.

- Tune the MTU size for jumbo frames at the guest level as well as ESXi and switch level.
- Enable TCP segmentation offload (TSO) at the ESXi level. (It should be enabled by default). This can be leveraged only by VMXNET3 drivers at the Guest layer.

- Other offload features can be leveraged only when using the VMXNET3 driver.
- Use regular platform tuning parameters, such as ring buffer size. However, RSS and RPS tuning must be specific to the VMXNET3 driver.

HBA Driver Type

Use a PVSCSI storage adapter. This provides the best performance characteristics (reduced CPU utilization and increased throughput), and is optimal for I/O-intensive guests (as with Hadoop).

- Tune queue depth in the guest OS SCSI driver.
- Disk partition alignment—If VMFS is already aligned, this is typically not necessary.

IO Scheduler

The I/O scheduler used for the OS disks might need to be different if using VMDKS. Instead of using CFQ, use deadline or noop elevators. Performance varies and must be tested. Any performance gains must be quantified appropriately (for example, 1-2% improvement vs. 10-20% improvement).

Memory Tuning

Minimize anonymous paging by setting vm.swappiness=1.

Consider using virtual NUMA (vNUMA). This exposes the NUMA architecture to the guest OS so that the guest OS can be tuned to leverage NUMA during scheduling. Virtual Hardware version 8 or later is required to leverage vNUMA.

Cloudera Software Stack

Guidelines for installing the Cloudera stack on this platform are nearly identical to those for bare metal. This is addressed in various documents on the Cloudera website.

To configure the Isilon service (instead of HDFS), follow the instructions at Managing Isilon.

References

1. Performance Best Practices for VMware vSphere® 5.5 (PDF)
2. Exploring the hadoop network topology (blog)http://ofirm.wordpress.com/2014/01/09/exploring-the-hadoop-network-topology/
3. Virtualized Hadoop Performance with vSphere 5.1http://www.vmware.com/resources/techresources/10360
4. Hadoop Deployment Guide for vSpherehttp://www.vmware.com/files/pdf/products/vsphere/Hadoop-Deployment-Guide-USLET.pdf
5. vmware BDE Command line Guidehttp://pubs.vmware.com/bde-2/topic/com.vmware.ICbase/PDF/vsphere-big-data-extensions-22-command-line-guide.pdf
6. Cloudera Documentationhttp://www.cloudera.com/content/cloudera/en/documentation.html
7. EMC Hadoop Starter Kit -- Step By Step Guide To Quickly And Easily Deploy Hadoophttps://community.emc.com/docs/DOC-26892
8. EMC HSK 3.0 For Cloudera Enterprise