cloudera®

Cloudera Reference Architecture for VMware vSphere with Locally Attached Storage

Version CDH 5.3

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: February 12, 2015

Table of Contents

Executive Summary

This document covers the architecture for running Cloudera Enterprise on VMware vSphere®-based infrastructure with locally attached Storage.

NOTE: This is a work in progress and details will change as software versions and capabilities change.

Audience and Scope

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

This document describes Cloudera recommendations on the following topics:

- 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 (this is the software that controls movement of VMs and storage on a vSphere 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 - this is what enables proper placement of data blocks and scheduling of YARN jobs in a Virtualized Environment wherein, multiple copies of any single block of data or YARN jobs (don't get placed/scheduled on VMs that reside on the same hypervisor host). The YARN component of HVE is still work in progress and won't be supported in CDH 5.3 (YARN-18). The HDFS component is supported in CDH 5.3.
JBOD	Just a Bunch of Disks (this is in contrast to Disks configured via software or hardware RAID with striping and redundancy mechanisms for data protection)

Job History Server	Process that archives job metrics and metadata. One per cluster.
LBT	Load-based teaming. - this is a teaming (LBT) policy that is traffic-load-aware and ensures physical NIC capacity in a NIC team is optimized.
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	Non-uniform memory access. Addresses memory access latency in multi-socket 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 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 vSphere.
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.

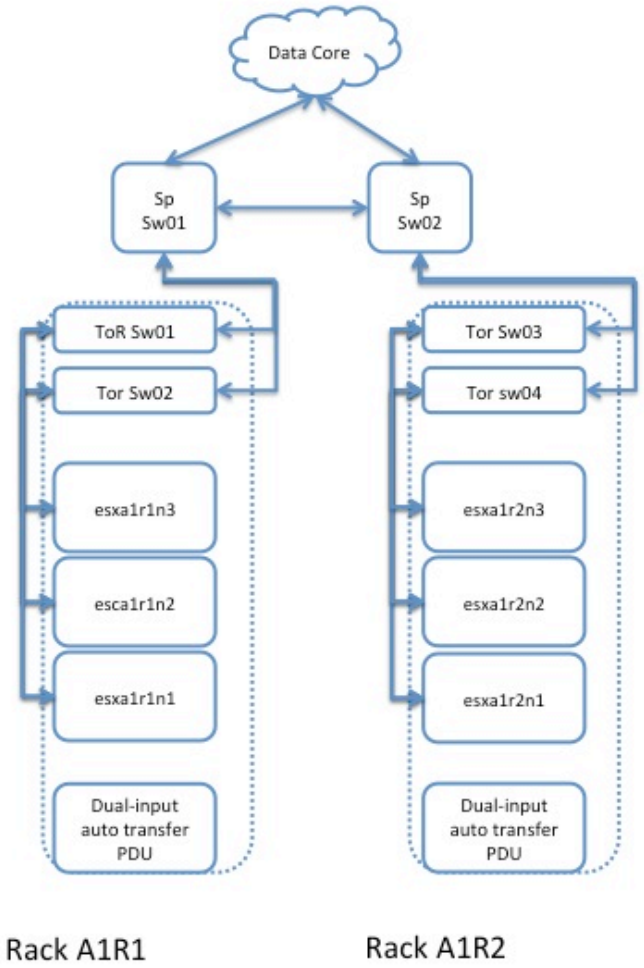
VM	Virtual machine.
vMotion	VMware term for live migration of virtual machines across physical hosts.
ZK	Zookeeper. A centralized service for maintaining configuration information, naming, and providing distributed synchronization and group services.

VMware-Based Infrastructure with Direct Attached Storage for HDFS

The storage subsystem described in this section is completely local (direct attached storage) on each vSphere host, and the VMs access the disks by one-to-one mapping of physical disks to vSphere VMFS virtual disks, or by raw device mappings (RDMs).

Physical Cluster Topology

Figure 1



Physical Cluster Component List

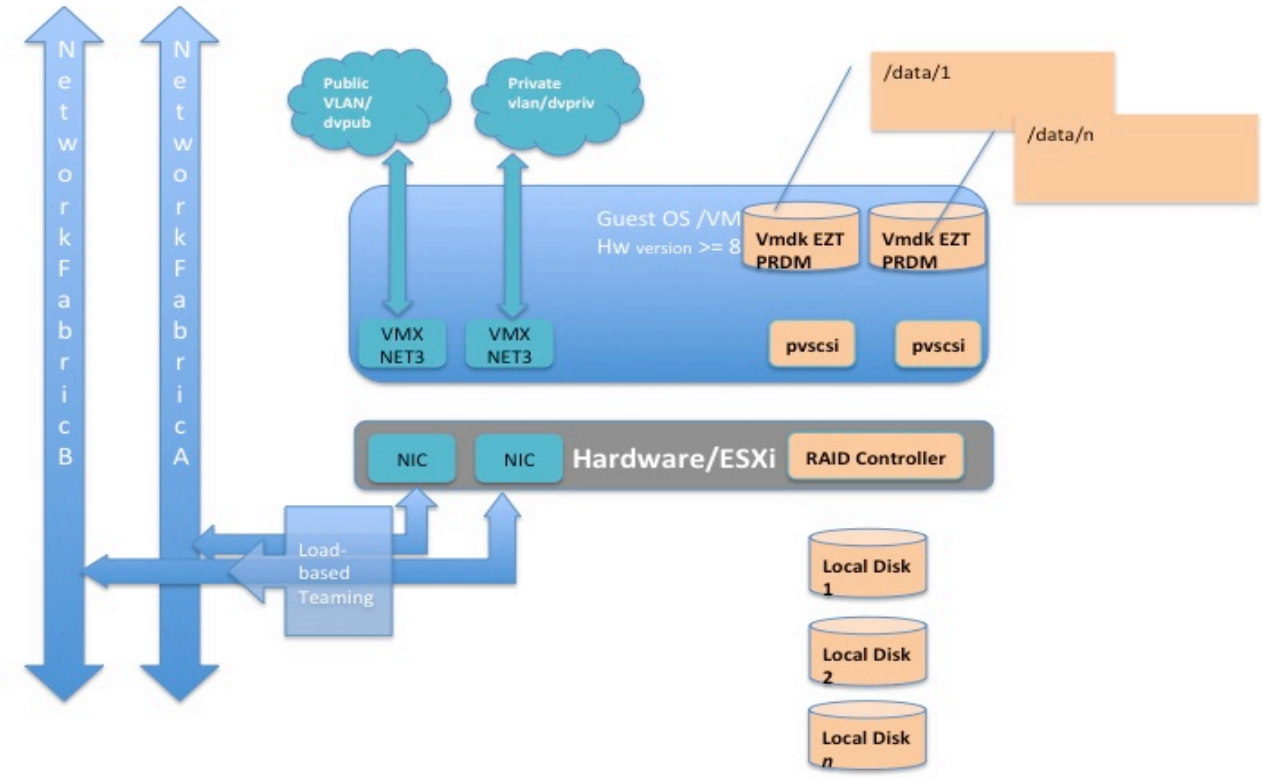
Table 1

Component	Configuration	Description	Quantity
Physical servers	Two-socket, 6-10 cores per socket > 2 GHz; minimally 256GB RAM.	ESX/VMware 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 cluster.	At least two per physical server.
Internal HDDs	Standard OS sizes.	The vSphere hypervisor requires little storage, so size is not important. These ensure continuity of service on server resets. The number of drives depends on the server form factor and internal HDD form factor.	12-24 per physical server (including OS disks).
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.

Logical Cluster Topology

Logical VM Diagram

Figure 2



Do not allow more than one replica of an HDFS block on any particular physical node. This is ensured by configuring the Hadoop Virtualization Extensions (HVE).

The minimum requirements to build out the cluster are:

- 3x Master Nodes (VMs)
- 5x DataNodes (VMs)

The DataNode count depends on the size of HDFS storage to deploy. For simplicity, ensure that DataNodes cohabitate with YARN NodeManager roles. The following table identifies service roles for different node types.

Care must be taken to ensure that CPU and Memory resources are not overcommitted while provisioning these node instances on the virtualized infrastructure.

Craft 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, this can also be achieved fairly easily 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 MasterNode VM (reference the BDE CLI guide for more details)

Care should also be taken to ensure automated movement of VMs is disabled. There should be no DRS or vMotion of VMs allowed in this deployment model. This is critical as VMs are tied to physical disks and movement of VMs within the cluster will result in data loss.

Table 2

	Master Node	Master Node	Master Node	YARN NodeManager Nodes /HDFS Data Nodes 1..n
ZooKeeper	ZooKeeper	ZooKeeper	ZooKeeper	
HDFS	NameNode, Quorum Journal Node	NameNode, Quorum JournalNode	Quorum JournalNode	DataNode
YARN	Resource Manager	Resource Manager	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	
SOLR	Out of scope	Out of scope	Out of scope	Out of scope
HBASE	HMaster	HMaster	HMaster	RegionServer
Impala			Impala Statestore	Impalad (Impala Daemon)

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.

Table 3

Component	Configuration	Description	Quantity
Master Nodes: **two-socket with 6-10 cores/socket > 2GHz; minimally 128 GB RAM; 4-6 disks**	Do not house the active NameNode/RM node and the standby in the same chassis (if using blades) or in the same rack (if using rackmount servers).	These nodes 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/HDFS Data nodes: **two-socket with 6-10 cores/socket > 2GHz; minimally 128 GB RAM;** **no. of disks = no. of cores/2**	These are VMs that can be deployed as needed on the vSphere cluster, without over-subscription of either CPU or Memory resources. Configure CPUs along physical socket boundaries. According to vmware, one VM per NUMA node is advisable but up to two VMs per NUMA node can be deployed.	These nodes house the HDFS DataNode roles and YARN node managers, as well as additional required services. Adjust memory sizes based on the number of services, or provision additional capacity to run additional services. Disks can be provisioned either as PRDM or with a 1:1:1 mapping of physical disk:VMFS datastore:VMDK virtual disk.	TBD (based on customer needs).

The following table provides recommendations for storage allocation.

Table 4

Node/Role	Disk Layout	Description
Management/Master	1 x 200 GB OS. 4 x 1 TB drives (2+2 RAID 10) for database and HDFS metadata. 2 x 100 GB drives as JBOD for ZK and QJN each.	Drive capacity depends on the size of the internal HDDs specified for the platform. Cloudera recommends not fracturing the filesystem layout into multiple smaller filesystems. Instead, keep a separate "/" and "/var".
HDFS DataNodes	1 x 200 GB OS. n x 2 TB HDDs.	The drive capacity depends on the size of the internal HDDs specified for the platform. Avoid fracturing the filesystem layout into multiple smaller filesystems. Instead, keep a separate "/" and "/var". We recommend more spindles for higher throughput and IOPs.

VMware vSphere Design Considerations

Virtual Network Switch Configuration (vswitch)

Standard vswitches may be employed, which need to be configured for each ESXi host in the cluster. Key configuration parameter to consider is the MTU size to ensure that the same MTU size being 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.

Storage Group Configuration

Each provisioned disk is either -

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

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. Evaluate your environment and choose accordingly. In some cases, performance might be more important than power optimization.

Avoid memory and CPU over-commitment, and use large pages for Hypervisor (which is the default).

For network tuning, enable advanced features such as TSO, LRO, scatter gather, and so on interrupt coalescing.

Guest OS Considerations

Assume that the guest OS is a flavor of Linux.

Special tuning parameters may 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.x and CentOS 6.x 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 (should be enabled by default). Only VMXNET3 drivers at the Guest layer can leverage this.
- Similarly, 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 the 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 -- Typically if VMFS is already aligned, this is not necessary (TBD for Linux).

I/O 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. This varies and must be tested. Any performance gains must be quantified appropriately; for example, 1-2% improvement vs. 10-20% improvement).

Memory Tuning

Disable or minimize anonymous paging by setting vm.swappiness=0 or 1.

VMs that fit on a single NUMA node will get 100% local memory accesses automatically.

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 Cloudera's website.

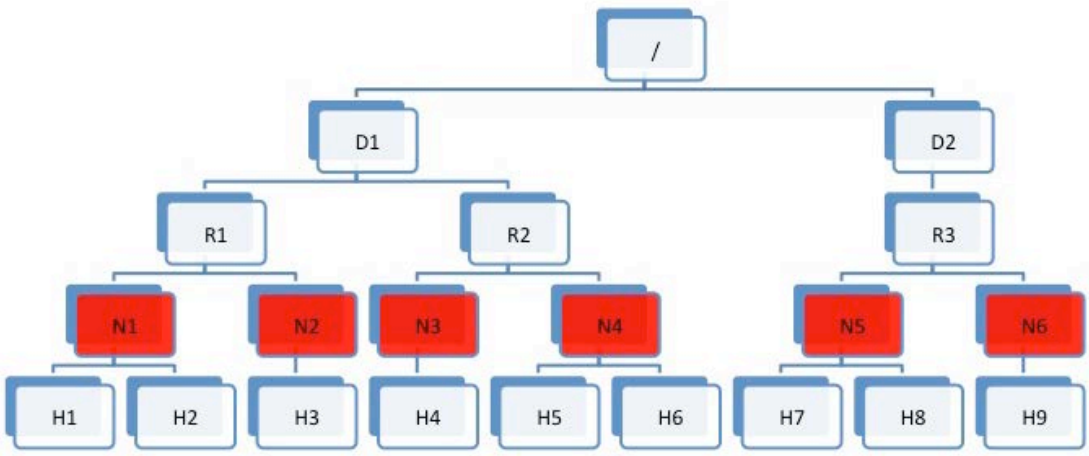
Enabling Hadoop Virtualization Extensions (HVE)

Referring to the HDFS-side HVE JIRA (https://issues.apache.org/jira/browse/HADOOP-8468), following are considerations for HVE:

1. 1. VMs on the same physical host are affected by the same hardware failure. In order to match the reliability of a physical deployment, replication of data across two virtual machines on the same host should be avoided.
2. 2. The network between VMs on the same physical host has higher throughput and lower latency and does not consume any physical switch bandwidth.
3. Thus, we propose to make hadoop network topology extendable and introduce a new level in the hierarchical topology, a node group level, which maps well onto an infrastructure that is based on a virtualized environment.

The following diagram illustrates the addition of a new layer of abstraction (in red) called NodeGroups. The NodeGroups represent the physical hypervisor on which the nodes (VMs) reside.

Figure 3



All VMs under the same node group run on the same physical host. With awareness of the node group layer, HVE refines the following policies for Hadoop on virtualization:

Replica Placement Policy

- No duplicated replicas are on the same node or nodes under the same node group.
- First replica is on the local node or local node group of the writer.
- Second replica is on a remote rack of the first replica.

- Third replica is on the same rack as the second replica.
- The remaining replicas are located randomly across rack and node group for minimum restriction.

Replica Choosing Policy

- The HDFS client obtains a list of replicas for a specific block sorted by distance, from nearest to farthest: local node, local node group, local rack, off rack.

Balancer Policy

- At the node level, the target and source for balancing follows this sequence: local node group, local rack, off rack.
- At the block level, a replica block is not a good candidate for balancing between source and target node if another replica is on the target node or on the same node group of the target node.

HVE typically supports failure and locality topologies defined from the perspective of virtualization. However, you can use the new extensions to support other failure and locality changes, such as those relating to power supplies, arbitrary sets of physical servers, or collections of servers from the same hardware purchase cycle.

Using Cloudera Manager, configure the following in safety valves:

- HDFS
  - hdfs core-site.xml:

```
<property>
    <name>net.topology.impl</name>
      <value>org.apache.hadoop.net.NetworkTopologyWithNodeGroup</
      value>
</property>
<property>
    <name>net.topology.nodegroup.aware</name>
    <value>true</value>
</property>
<property>
    <name>dfs.block.replicator.classname</name>
      <value>org.apache.hadoop.hdfs.server.blockmanagement.BlockP
      lacementPolicyWithNodeGroup</value>
</property>
```

- In mapred-site.xml, add the following properties and values:

```
<property>
      <name>mapred.jobtracker.nodegroup.aware</name>
      <value>true</value>
</property>
<property>
      <name>mapred.task.cache.levels </name>
<value>3</value>
</property>
```

Establish the Topology:

- Create a topology data file; for example:

```
192.168.x.1 /rack1/nodegroup1
192.168.x.2 /rack1/nodegroup1
192.168.x.3 /rack2/nodegroup1
192.168.x.4 /rack2/nodegroup2
192.168.x.5 /rack3/nodegroup1
192.168.x.6 /rack3/nodegroup2
```

- Create a topology script; the following is from http://ofirm.wordpress.com/2014/01/09/exploring-the-hadoop-network-topology/)

```
#!/bin/bash
HADOOP_CONF=/usr/local/hadoop-1.2.1/conf
echo `date` input: $@ >> $HADOOP_CONF/topology.log
while [ $# -gt 0 ] ; do
 nodeArg=$1
 exec< ${HADOOP_CONF}/topology.data
 result=""
 while read line ; do
   ar=( $line )
   if [ "${ar[0]}" = "$nodeArg" ] ; then
     result="${ar[1]}"
   fi
 done
 shift
 if [ -z "$result" ] ; then
#echo -n "/default/rack "
    echo -n "/rack01"
 else
   echo -n "$result "
 fi
done
```

- Place the topology data file and script in the same location on each node in your cluster. Specify the location using set topology.script.file.name (in conf/hadoop-site.xml).
  - The file must be an executable script or program.
  - You can also set topology.script.file.name using Cloudera Manager.

References

1. https://issues.apache.org/jira/secure/attachment/12551386/HVE%20User%20Guide%20on%20branch-1%28draft%20%29.pdf
2. http://www.vmware.com/pdf/Perf_Best_Practices_vSphere5.5.pdf
3. http://ofirm.wordpress.com/2014/01/09/exploring-the-hadoop-network-topology
4. http://www.vmware.com/files/pdf/products/vsphere/Hadoop-Deployment-Guide-USLET.pdf
5. https://issues.apache.org/jira/browse/YARN-18
6. https://issues.apache.org/jira/browse/HADOOP-8468
7. http://www.cloudera.com/content/cloudera/en/documentation.html
8. Virtualized Hadoop Performance with VMware vSphere 5.1- http://www.vmware.com/resources/techresources/10360
9. Benchmarking Case Study of Virtualized Hadoop Performance on vSphere 5 -- http://vmware.com/files/pdf/VMW-Hadoop-Performance-vSphere5.pdf
10. Hadoop Virtualization Extensions (HVE) -- http://www.vmware.com/files/pdf/Hadoop-Virtualization-Extensions-on-VMware-vSphere-5.pdf
11. https://www.vmware.com/support/pubs/vsphere-big-data-extensions-pubs.html