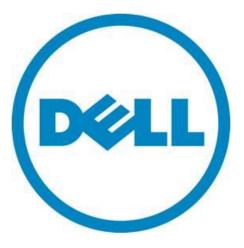
Dell Reference Configuration for Hortonworks Data Platform 2.4

A Quick Reference Configuration Guide

Kris Applegate Solution Architect Dell Solution Centers



Executive Summary

This document details the configuration set-up for Hortonworks Data Platform (HDP) software on the PowerEdge R730XD platform. The intended audiences for this document are customers and system architects looking for information on configuring Hortonworks Data Platform Hadoop clusters within their information technology environment for big data analytics.

The reference configuration introduces the server set-up that can run the Hortonworks stack. The document will only focus on hardware configuration; it will not go into detail about Hadoop solution components or resiliency, performance, or software considerations. This document does not focus on best practices or complete architecture for a Hortonworks Data Platform Solution

Dell developed this document to help streamline configuration for the Hortonworks Data Platform software and demonstrate functional compatibility.

THIS WHITE PAPER IS FOR INFORMATIONAL PURPOSES ONLY, AND MAY CONTAIN TYPOGRAPHICAL ERRORS AND TECHNICAL INACCURACIES. THE CONTENT IS PROVIDED AS IS, WITHOUT EXPRESS OR IMPLIED WARRANTIES OF ANY KIND.

© 2016 Dell Inc. All rights reserved. Reproduction of this material in any manner whatsoever without the express written permission of Dell Inc. is strictly forbidden. For more information, contact Dell. Dell, the DELL logo, and the DELL badge are trademarks of Dell Inc. Intel and Xeon are registered trademarks of Intel Corp. Red Hat is a registered trademark of Red Hat Inc. Linux is a registered trademark of Linus Torvalds. Other trademarks and trade names may be used in this document to refer to either the entities claiming the marks and names or their products. Dell Inc. disclaims any proprietary interest in trademarks and trade names other than its own.



Reference Configuration

Hortonworks Data Platform is available on both for Linux and, in partnership with Microsoft, on Windows. This initial configuration will target deployment on bare-metal servers running Redhat and Redhat-Compatible releases 7.2+.

Server Roles

Master Roles(s) – As the name implies, these servers are responsible for the majority of the state of the cluster.

- HDFS NameNodes own the block tables for the shared, distributed Filesystem,
 There is a primary node, a standby node, and a witness server that enables high availability.
- YARN Resource Manager Nodes decide what compute task to divide up amongst the workers.
- HBase Master This node is responsible for all the metadata changes that inside the NoSQL (Not-only-SQL) database that resides alongside Hadoop.
- Spark History Server This node tracks all previously run Spark-on-YARN jobs run on the server.

Worker Role(s) – This is where the heavy lifting gets done. Whether disk, compute, or memory, these are your units of scale. The great part of about these is that they scale linearly and when you add disk capacity you can also add memory and CPU cores

- HDFS DataNodes These own the distributed Hadoop Filesystem that is made up
 of the many local disks.
- YARN NodeManagers These own the CPU Processing shares that are allocated by the YARN ResourceManager.
- HBase RegionServers These own the replicated chunk of the NoSQL databases that reside on any given node.
- Spark Workers/Clients These nodes execute all Spark (in-memory processing) jobs on each data node.

It should be noted that often times many of these services are collocated on the same machine. In many of the reference configurations, this is absolutely the intention. This reduces complexity and allows you to do more with less nodes. However, as you move towards production, you'll want to start isolating some of these functions onto discreet nodes in order to reduce failure domains.

Edge Role(s) – Edge nodes lie on the perimeter of the dedicated Hadoop network and bridge the Hadoop environment with the production IT environment. Edge nodes enable external users and business processes to interact with the cluster. Additional edge nodes may be added to the Hadoop cluster as external access requirements increase.



Ambari Manager Role(s) – The Ambari management node is where the Ambari server resides. The Ambari management node runs the configuration management processes, web server software, monitoring software (open-source project Nagios) and performance monitoring (open-source project Ganglia) software. In a production environment, the Ambari server should run on a dedicated node; however, for the purposes of this document, Ambari server was installed on the edge node.

Node Type Considerations

- Infrastructure Nodes
 - o Name Node and HA Witness Node
 - o Edge Node(s)
 - Staging Edge Node
 - Streaming Edge Node
- Data Node(s)
 - o General Purpose Data Node
 - o High Performance Data Node
 - o High Capacity Data Node

Infrastructure Nodes

We separate the types of nodes into two major categories, Infrastructure nodes (Master, Edge, and Ambari roles) and Data nodes (Worker roles). Let's start with detailing the hardware profiles of the Infrastructure nodes:

Name Node(s) and HA Witness Node

These servers should be hardened to reduce downtime and provide enough storage performance to facilitate fast metadata lookups and rapid job assignments. The HA witness node is more or less an independent file share server that can act to break a tie during a failover scenario. The reason that we recommend using the same hardware config as the Name Nodes is so that if needed, he has the right hardware to assume the role of a Name Node should the original Name Nodes be unrecoverable. It should be noted that while an HA failover is automated, the act of moving services to the HA witness permanently would be a manual task. The recommended disk layout should be as follows:

Table 1. Name Node and HA Witness Node Disk Layout

Function	Disks	Type
Operating System	2	RAID 1 (Mirror)
Zookeeper Journal	1 (Optionally SSD)	Non-RAID or RAID 0
NameNode Journal	1 (Optionally SSD)	Non-RAID or RAID 0
HDFS Metadata	2	RAID 1
Database Storage	4	RAID 10

The recommended hardware configuration for these nodes is detailed below. Though not required, it's been our experience that many customers prefer to keep the platform type



the same across their Infrastructure nodes and Data Nodes. These nodes share all the master roles outlined earlier.

Table 2. Name Node and HA Witness Node Hardware

Component	Hardware Option
Platform	Dell PowerEdge R730XD (12-Drive Option with Flexbay)
Processor	2x Intel Xeon E5-2650 v4 2.2 GHz (12-Core)
RAM (min)	128 GB
Network Daughter	Intel X520 Dual-port 10GbE + I350 Dual-port 1GbE (LACP
Card	Bonded)
Add-in PCI-E	None
Network Card	
Disk (Hot-Plug)	8x 1TB 7.2K RPM SAS 12Gbps (Data) – See above table
Disk (Flexbay)	2x 600GB 10K RPM SAS 12Gbps (OS) – RAID 1 (Mirror)
Storage Controller	Dell PowerEdge RAID Controller (PERC) H730

Optional Upgrades:

CPUs – The 12-core processor here should provide plenty of power for most workloads. Only the extreme fringe use-cases may need to look at faster processors.

Disks – You could look to 10K SAS for your data drives as well as look to write-intensive SSDs (like the Intel S3710) for your journal drives. As of the publishing of this paper, SSD drives for the Flexbay are getting cost competitive to the 10K SAS drives, and thus you may want to leverage those for speed/power consumption.

Network – No real need to go faster that bonded 10GbE, but if you upgrade to 40 or 100GbE on the data nodes, you may want to be able to use the same cards / switches on these nodes. You may also want to multi-home these nodes to the data and corporate networks by adding a 2nd Intel X520 Dual-port 10GbE Server Adapter to a PCI-E slot. **Memory** – 128 GB of RAM should be plenty for most average clusters. As you scale to the 100+ node mark or you use a small block size for HDFS, you may want to look at upgrading this.

Edge Node(s)

Edge nodes are the primary interface through which data traverses in and out of the cluster. As such, they can vary wildly from customer to customer and use-case to use-case. Generally they fit two profiles:

Staging – Data is moved to this node and then will be moved into the cluster resulting in a two-stage process. In this case, there would need to be a relatively large amount of storage local to this node to hold the staged data.

Streaming – In this case, the data will simply be passing through this node, not residing on local storage, but solely passing through the processor/memory.

With so much variation, this is an excellent place to engage you Dell Customer Solution Center subject-matter experts to help determine how many Edge nodes of which type you should go with. The recommended hardware configuration for these both staging and streaming Edge nodes is detailed below. Though not required, it's been our experience



that many customers prefer to keep the platform type the same across their Infrastructure nodes and Data Nodes.

Table 3. Staging Edge Node Hardware

Component	Hardware Option
Platform	Dell PowerEdge R730XD (12-Drive Option with Flexbay)
Processor	2x Intel Xeon E5-2650 v4 2.2 GHz (12-Core)
RAM (min)	128 GB
Network Daughter	Intel X520 Dual-port 10GbE + I350 Dual-port 1GbE (LACP
Card	Bonded 1 Data 1 Corporate)
Add-in PCI-E	Intel X520 Dual-port 10GbE Server Adapter (LACP Bonded 1
Network Card	Data 1 Corporate)
Disk (Hot-Plug)	12x 4TB 7.2K RPM SAS 12Gbps (Data) – RAID 5 or RAID 10
Disk (Flexbay)	2x 600GB 10K RPM SAS 12Gbps (OS) – RAID 1 (Mirror)
Storage Controller	Dell PowerEdge RAID Controller (PERC) H730

Table 4. Streaming Edge Node Hardware

Component	Hardware Option
Platform	Dell PowerEdge R730XD (12-Drive Option with Flexbay)
Processor	2x Intel Xeon E5-2650 v4 2.2 GHz (12-Core)
RAM (min)	128 GB
Network Daughter	Intel X520 Dual-port 10GbE + I350 Dual-port 1GbE (LACP
Card	Bonded 1 Data 1 Corporate)
Add-in PCI-E	Intel X520 Dual-port 10GbE Server Adapter (LACP Bonded 1
Network Card	Data 1 Corporate)
Disk (Hot-Plug)	2x 800GB Intel S3710 Write-intensive SATA (Data) — RAID 1
	(Mirror)
Disk (Flexbay)	2x 600GB 10K RPM SAS 12Gbps (OS) – RAID 1 (Mirror)
Storage Controller	Dell PowerEdge RAID Controller (PERC) H730

Data Node(s)

Data nodes are the workhorses of any Hadoop cluster. They are responsible for running the worker roles outlined in the section above. The fact that they couple compute and storage means that depending on the use-case they could be storage-heavy, compute-heavy, or mixed ratio. This is another excellent place to engage you Dell Customer Solution Center subject-matter experts to help determine what the mix of storage and processing needs of your data nodes will be. Below we will detail 3 node types:

General Purpose Data Node – This is the average configuration that include large form-factor (LFF) 3.5" drives for the data and two drives for the OS in the rear Flexbay. **High Performance Data Node** – This node has small form-factor (SFF) 2.5" disks and includes SSDs for leveraging Hadoop's heterogeneous storage tiering. **High Capacity Data Node** – This node is on our special R730XD chassis that includes a massive 16x 3.5" + 2 2.5" disks in 2U. It is the same 12-drive chassis as the General Purpose configuration, but it also includes a 4-drive mid-bay for more 3.5" disks (still hot-swappable).



Table 5. General Purpose Data Node Hardware

Component	Hardware Option
Platform	Dell PowerEdge R730XD (12-Drive Option with Flexbay)
Processor	2x Intel Xeon E5-2650 v4 2.2 GHz (12-Core)
RAM (min)	128 GB
Network Daughter	Intel X520 Dual-port 10GbE + I350 Dual-port 1GbE (LACP
Card	Bonded)
Add-in PCI-E	None
Network Card	
Disk (Hot-Plug)	12x 4TB 7.2K RPM SAS 12Gbps (HDFS) – Non-RAID or RAID 0
Disk (Flexbay)	2x 600GB 10K RPM SAS 12Gbps (OS) – RAID 1 (Mirror)
Storage Controller	Dell PowerEdge RAID Controller (PERC) H730

Table 6. High Performance Data Node Hardware

Component	Hardware Option	
Platform	Dell PowerEdge R730XD (24-Drive Option with Flexbay)	
Processor	2x Intel Xeon E5-2690 v4 2.6 GHz (14-Core)	
RAM (min)	512 GB	
Network Daughter	Intel X520 Dual-port 10GbE + I350 Dual-port 1GbE (LACP	
Card	Bonded)	
Add-in PCI-E	Intel X520 Dual-port 10GbE Server Adapter (LACP Bonded)	
Network Card		
Disk (Hot-Plug)	 20x 1.2 TB 10K RPM SAS 12Gbps (HDFS Disk Tier) – 	
	Non-RAID or RAID 0	
	 1x 800GB Intel S3710 Write-intensive SATA (Spark 	
	Scratch) – Non-RAID / RAID 1	
	3x 800GB Intel S3710 Write-intensive SATA (HDFS SSD	
	Tier) – Non-RAID / RAID 1	
Disk (Flexbay)	2x 600GB 10K RPM SAS 12Gbps (OS) – RAID 1 (Mirror)	
Storage Controller	Dell PowerEdge RAID Controller (PERC) H730	

Table 7. High Capacity Data Node Hardware

Component	Hardware Option
Platform	Dell PowerEdge R730XD (12+2+4 Drive Option with Flexbay and
	Mid-bay)
Processor	2x Intel Xeon E5-2650 v4 2.2 GHz (12-Core)
RAM (min)	128 GB
Network Daughter	Intel X520 Dual-port 10GbE + I350 Dual-port 1GbE (LACP
Card	Bonded)
Add-in PCI-E	None
Network Card	
Disk (Hot-Plug)	12x 4TB 7.2K RPM SAS 12Gbps (HDFS) – Non-RAID or RAID 0
Disk (Flexbay)	2x 600GB 10K RPM SAS 12Gbps (OS) – RAID 1 (Mirror)
Disk (Mid-bay)	4x 4TB 7.2K RPM SAS 12Gbps (HDFS) – Non-RAID or RAID 0
Storage Controller	Dell PowerEdge RAID Controller (PERC) H730

Optional Upgrades:



CPUs – With the advent of Spark and other in-memory processing technologies, the need for more CPU horsepower is increasingly likely. Make sure to work with your Dell Technical Sales teams to determine the appropriate size processors given thermal and power restrictions.

Disks – More and more customers are expecting higher I/O performance out of their data nodes. Some customers have started to move to all 10K RPM disks and even into SSDs. The other option, detailed in a latter section is using HDFS tiering. From a capacity standpoint, you'll want to be observant of creating too large of failure domains. The reason you don't see the 8-10TB drives above is that would be a really large chunk of data to lose at once. You should balance your need for capacity with your willingness to assume risk. **Network** – As 10GbE per-port costs drop and 40-100GbE switching becomes more mainstream, many customers are opting to increase the bandwidth to the Datanodes. This can also help in reducing rebuild time in the case of whole-node failures. **Memory** – 128 GB of RAM is probably the starting point for conversations around Datanode memory amounts. With the rise in Spark (and other in-memory storage and processing engines) we are seeing customer push this to 256-512GB routinely.

Node Count Recommendations

Dell recognizes that use-cases for Hadoop range from small development clusters all the way through large multi petabyte production installations. It's a good idea to leverage the Dell Customer Solution Centers subject matter experts to help determine you exact needs. As a starting point, three cluster configurations can be defined for typical use:

Minimum Development Cluster – The minimum development cluster is targeted at functional testing and may even be built from existing equipment; however, the performance of these types of clusters can be significantly less as development clusters typically do not benefit from the highly distributed nature of HDFS.

Recommended Small Cluster – The recommended small cluster is a good starting point for customers taking the initial steps for running HDP in production. A small cluster provides some layers of basic resiliency that is expected in today's production IT world.

Recommended Production Cluster – The recommended production cluster configuration provides dense storage and compute capacity, coupled with high degree of resiliency. The production cluster allows for an adequate number of data nodes to demonstrate the performance benefits of distributed storage and parallel computing.



Table 8. Recommended Cluster Sizes

Node Type	Minimum Development Cluster	Recommended Small Cluster	Recommended Production Cluster
NameNode	1¹	2 ²	2
HA Node	0	0	1
Edge Node(s)	0	1	1
Data Nodes	3	5	10
1 GbE Switches (Dell Networking S3048-ON) ³	1	1	1
10 GbE Switches (Dell Networking S4048-ON) ⁴	1	2	2
Rack Units	8U	19U	31U

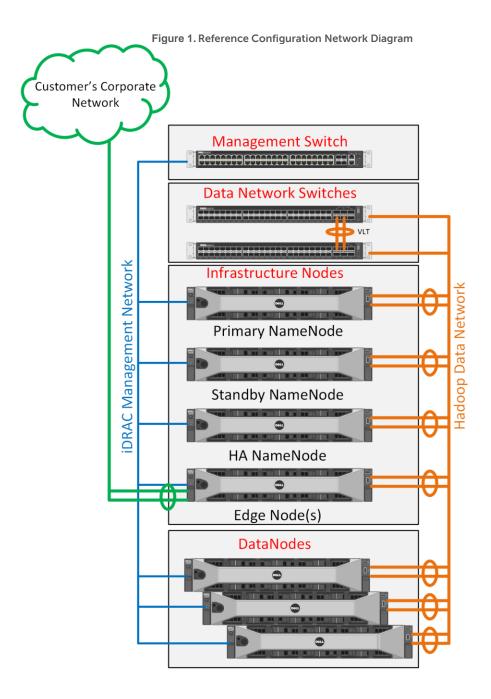
¹ In this case a single node serves as the NameNode Edge Node and runs all the master roles. There is no failover or resiliency here. It would also require a complete rebuild of the cluster to scale this into production in any way.



² The roles would be evenly distributed amongst the two NameNodes. However, the failover would not be automatic. It would require user intervention to failover and to failback.

³The 1GbE switches here are used only for access to the Dell Remote Access Controllers (iDRAC) for outof-band management.

⁴The 10GbE switches we recommend have some important attributes such as non-blocking backplanes and dedicated per-port packet buffers. Work with you Dell Networking Specialists to find the right switch for your environment.





Tested Configuration

For the purposes of this document, a production-level Hadoop cluster was deployed as recommended in **Table 8**. The specific software revisions used in the test are shown in **Table 9**. The PowerEdge R630 and R730XD hardware configurations we tested are shown in **Table 10** and **Table 11**. The hardware listed was used as it was immediately available. This reference configuration used Terasort / Teragen as the testing method. This is a great tool for stressing IO (network, CPU, and memory) but each customer's workload will exhibit different I/O and service usage profiles. Since the scope of this document is only to validate Dell PowerEdge Hardware with Hortonworks 2.4, that basic functional I/O testing would suffice. Other benchmarks such as HiBench, BigBench, and TPCx-BB may be used to perform further validations as customer demand dictates.

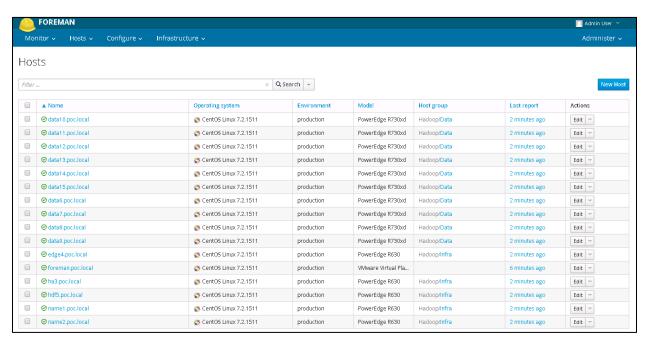


Figure 2. Foreman Hosts List

Important OS Provisioning Considerations

A couple of extremely (and well documented) exceptions should be made for the cluster.

- While firewalld serves a valuable purpose, due to the close network for the Hadoop backend, it can be bothersome. For my test cases I disabled it all together.
- In turn I disabled SELinux on all hosts per the install guide.
- A good provisioning subsystem is very handy. I find Foreman
 (http://theforeman.org) to meet my needs, Tools such as Bright Computing or even homegrown kickstart / pxelinux scripts should do the work as well.
- Format the data drives as dictated in the Hortonworks Installation guide (I used ext4). I created mount-points under /data/n (where n is the drive number).



Figure 3. Hortonworks Master Role Assignment

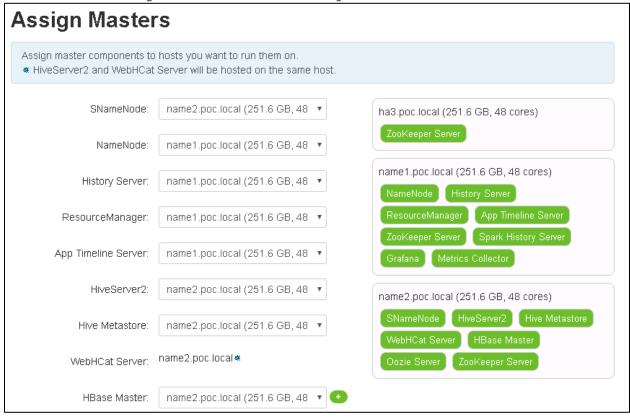
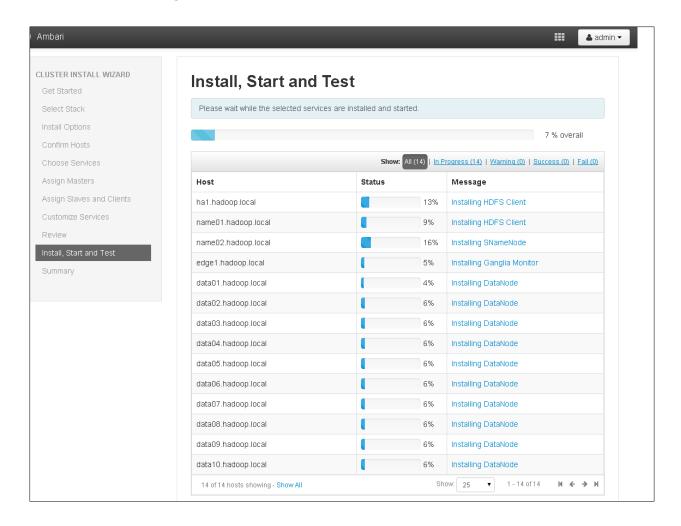




Figure 4. Ambari Node Installation





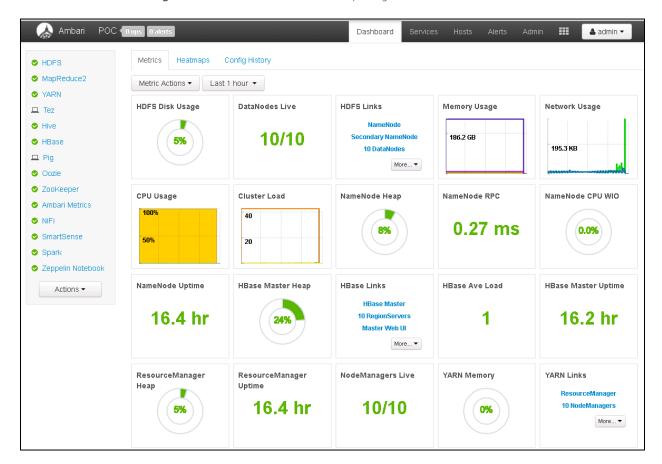


Figure 5. Ambari Cluster Performance Reporting

Table 9. Software Revisions (As Tested)

Component	Revision
CentOS Linux	7.2.1511
Hortonworks Data	2.4
Apache Hadoop	2.7.1.2.4.2.0-258
Foreman	1.12
Oracle JDK	1.7.0_67-b01



Table 10. As-tested Infrastructure Node Hardware

Component	Hardware Option
Platform	Dell PowerEdge R630 (10 2.5" Drive Option)
Processor	2x Intel Xeon E5-2680 v3 2.5 GHz (12-Core)
RAM (min)	128 GB
Network Daughter	Intel X520 Dual-port 10GbE + I350 Dual-port 1GbE (LACP
Card	Bonded)
Add-in PCI-E	None
Network Card	
Disk (Hot-Plug)	10x 600GB 10K RPM SAS 12Gbps (Data) – As laid out in Table 1
Storage Controller	Dell PowerEdge RAID Controller (PERC) H730P

Table 11. As-tested Data Node Hardware

Component	Hardware Option	
Platform	Dell PowerEdge R730XD (12+2+4 Drive Option with Flexbay and	
	Mid-bay)	
Processor	2x Intel Xeon E5-2650 v3 2.3 GHz (10-Core)	
RAM (min)	256 GB	
Network Daughter	Intel X520 Dual-port 10GbE + I350 Dual-port 1GbE (LACP	
Card	Bonded)	
Add-in PCI-E	None	
Network Card		
Disk (Hot-Plug)	 10x 4TB 7.2K RPM SAS 12Gbps (HDFS) – Non-RAID or RAID 0 	
	 1x 800GB Intel S3700 Write-intensive SATA (Spark Scratch) – Non-RAID / RAID 1 	
	1x 800GB Intel S3700 Write-intensive SATA (HDFS SSD)	
	Tier) – Non-RAID / RAID 1	
Disk (Flexbay)	2x 600GB 10K RPM SAS 12Gbps (OS) – RAID 1 (Mirror)	
Disk (Mid-bay)	4x 4TB 7.2K RPM SAS 12Gbps (HDFS) – Non-RAID or RAID 0	
Storage Controller	Dell PowerEdge RAID Controller (PERC) H730P Mini	

Dell Customer Solution Centers

The Dell Customer Solution Centers are a global network of connected labs that allow Dell to help customers architect, validate and build solutions. With multiple footprints in every region, they can help you understand anything from simple hardware platforms, to more complex solutions. These engagements range from an informal 30-60 minute briefings, through a longer half-day workshop, and on to a proof-of-concept that allow customers to kick the tires of their solution prior to signing on the dotted line. Customers may engage with their account team and have them submit a request to take advantage of these services for no charge.



Links

Dell Customer Solution Centers – http://hortonworks – http://hortonworks.com/products/hdp/
Hortonworks Data Platform - http://hortonworks.com/products/hortonworks-sandbox/
Foreman - https://theforeman.org/
Installation Video - https://youtu.be/zesOkO5000U

