Dell EMC Hortonworks Hadoop Solution

Version 2.5

Contents

List of Figures

List of Tables

Trademarks

This document is for informational purposes only, and may contain typographical errors and technical inaccuracies. The content is provided as-is and without expressed or implied warranties of any kind.

Glossary

ASCII

American Standard Code for Information Interchange, a binary code for alphanumeric characters developed by ANSI®.

BMC

Baseboard Management Controller

BMP

Bare Metal Provisioning

Clos

A multi-stage, non-blocking network switch architecture. It reduces the number of required ports within a network switch fabric.

DBMS

Database Management System

DTK

Dell OpenManage Deployment Toolkit

EBCDIC

Extended Binary Coded Decimal Interchange Code, a binary code for alphanumeric characters developed by IBM®.

ECMP

Equal Cost Multi-Path

EDW

Enterprise Data Warehouse

EoR

End-of-Row Switch/Router

ETL

Extract, Transform, Load is a process for extracting data from various data sources; transforming the data into proper structure for storage; and then loading the data into a data store.

HBA

Host Bus Adapter

HDFS

Hadoop Distributed File System

HDP

Hortonworks Data Platform

HVE

Hadoop Virtualization Extensions

IPMI

Intelligent Platform Management Interface

JBOD

Just a Bunch of Disks

LACP

Link Aggregation Control Protocol

LAG

Link Aggregation Group

LOM

Local Area Network on Motherboard

NIC

Network Interface Card

NTP

Network Time Protocol

OS

Operating System

PAM

Pluggable Authentication Modules, a centralized authentication method for Linux systems.

RPM

Red Hat Package Manager

RSTP

Rapid Spanning Tree Protocol

RTO

Recovery Time Objectives

SIEM

Security Information and Event Management

SLA

Service Level Agreement

THP

Transparent Huge Pages

ToR

Top-of-Rack Switch/Router

VLT

Virtual Link Trunking

VRRP

Virtual Router Redundancy Protocol

YARN

Yet Another Resource Negotiator

Notes, Cautions, and Warnings

Note: A **Note** indicates important information that helps you make better use of your system.

Caution: A **Caution** indicates potential damage to hardware or loss of data if instructions are not followed.

Warning: A **Warning** indicates a potential for property damage, personal injury, or death.

This document 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.

Chapter 1

Dell EMC Hortonworks Hadoop Solution Overview

Topics:

This document details the architectural recommendations for Hortonworks Data Platform (HDP) software on the Dell EMC PowerEdge R730xd. 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.

Introduction

This reference architecture describes the Dell EMC server hardware and networking configuration recommended for running the Hortonworks Data Platform. This architecture is focused on hardware configurations, and does not go into details about the components in HDP or their applications.

Solution Use Case Summary

The Hortonworks Connected Data Platform helps customers create actionable intelligence to transform their businesses. Whether it's data-in-motion, data-at-rest, or modern data applications, HDP can power the future of data for any organization and any line of business with use cases including:

- Data Discovery
- Single View
- Predictive Analytics
- EDW Optimization

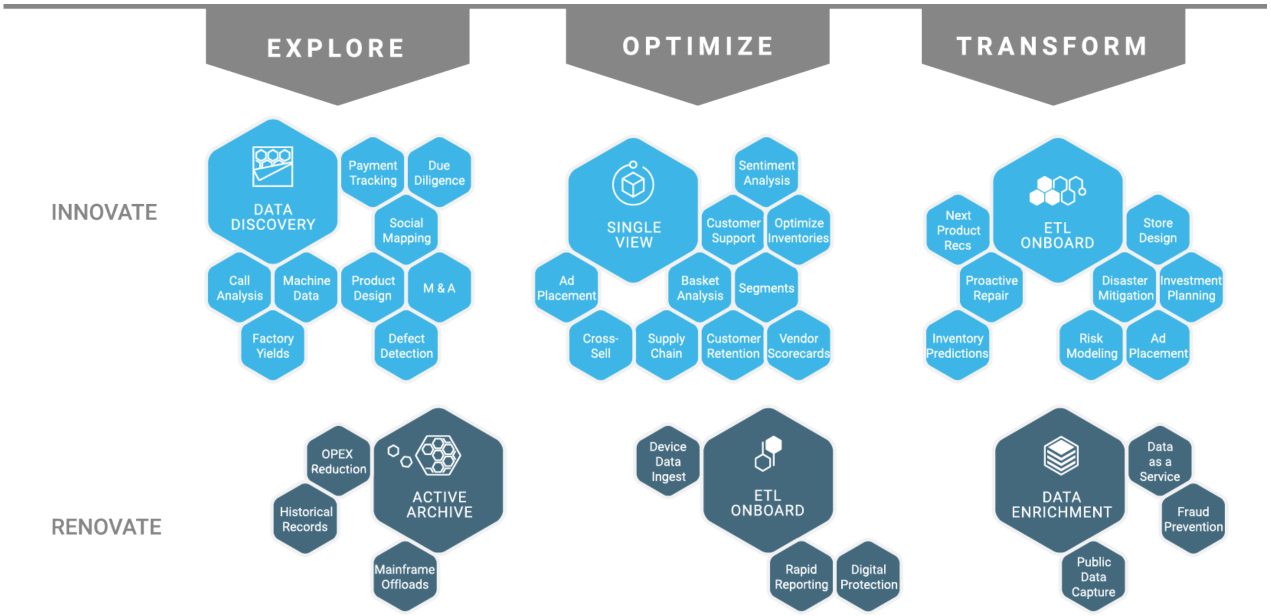

Figure 1: Hortonworks Data Platform Use Cases

Solution Components

This architecture combines the Hortonworks Data Platform with Dell EMC PowerEdge R730xd servers and Dell Networking switches to implement a complete Hadoop platform.

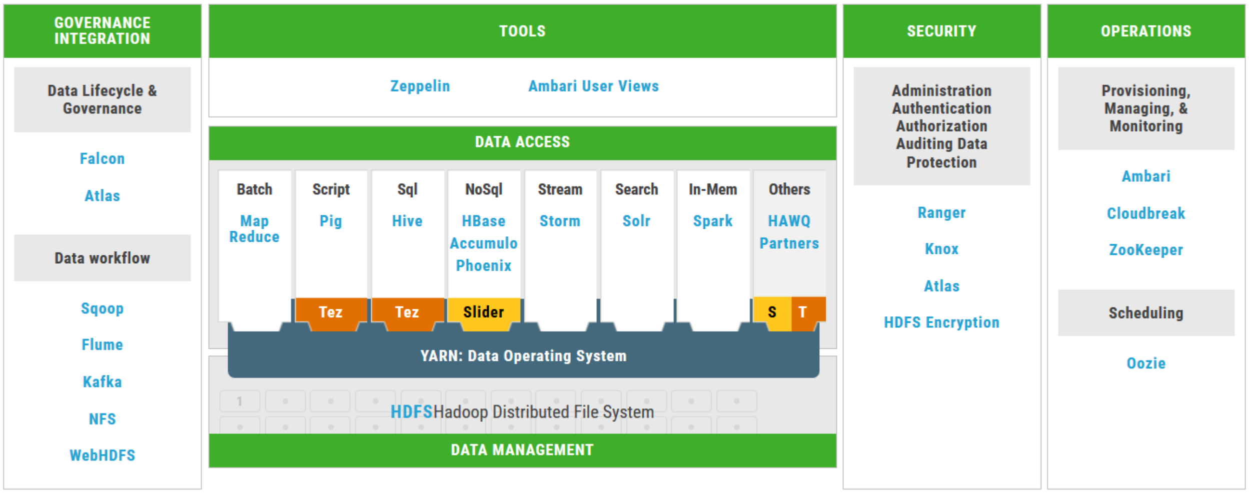

Figure 2: Hortonworks Data Platform Components

Chapter 2

Cluster Architecture

Topics:

- *Node Architecture*
- *Network Architecture*
- *Server Architecture*
- *Sizing Guidelines*

This chapter adresses the overall cluster architecture, including recommended server configurations, network fabric, and software role assignments.

Node Architecture

The Hortonworks Data Platform is composed of many Hadoop components covering a wide range of functionality. Most of these components are implemented as master and worker services running on the cluster in a distributed fashion.

In this architecture, we classify physical nodes into roles, and then map services to those roles. The assignment of services and roles to physical nodes is somewhat flexible, based on cluster workload. *Table 1: Cluster Physical Node Roles* on page 15 shows the physical node classification.

Table 1: Cluster Physical Node Roles

Physical Node Role	Required?	Server Hardware Configuration
Active NameNode	Required	Master
Standby NameNode	Required	Master
High Availability (HA) Node	Required	Master
Admin Node	Required	Master
Edge Node	Recommended	Master
Data Node 1 - N	Required	Data

The core HDP services are are listed in *Table 2: HDP Services* on page 15 .

Table 2: HDP Services

Service	Function	Master	Worker
HDFS	Hadoop distributed filesystem	Primary Namenode, Secondary Namenode	Data Node
YARN	Cluster resource management	YARN Resource Manager	YARN NodeManager
HBase	Column-oriented NoSQL Database	HBase Master	HBase Region Server
Spark	In-memory data processing engine	Spark Master, Spark History Server	Spark Worker
Ambari	Hadoop Cluster management	Ambari Server	Ambari Agent

Table 3: Service Locations on page 16 shows the recommended mapping of cluster services to physical nodes.

Table 3: Service Locations

Physical Node	Software Function
Active NameNode	NameNode Quorum Journal Node ZooKeeper HBase Master 2
Standby NameNode	Standby NameNode Resource Manager Quorum Journal Node ZooKeeper
HA Node	Standby Resource Manager Quorum Journal Node ZooKeeper HBase Master 1
Data Node(x)	Data Node NodeManager HBase RegionServer
Admin Node	Operational Databases (PostgreSQL) Ambari
Edge Nodes	Hadoop Client Applications

Network Architecture

The cluster network is architected to meet the needs of a high performance and scalable cluster, while providing redundancy and access to management capabilities.

The architecture is a leaf / spine model based on 10GbE network technology, and uses Dell Networking S4048-ON switches for the leaves, and Dell Networking S6000-ON switches for the spine. IPv4 is used for the network layer.

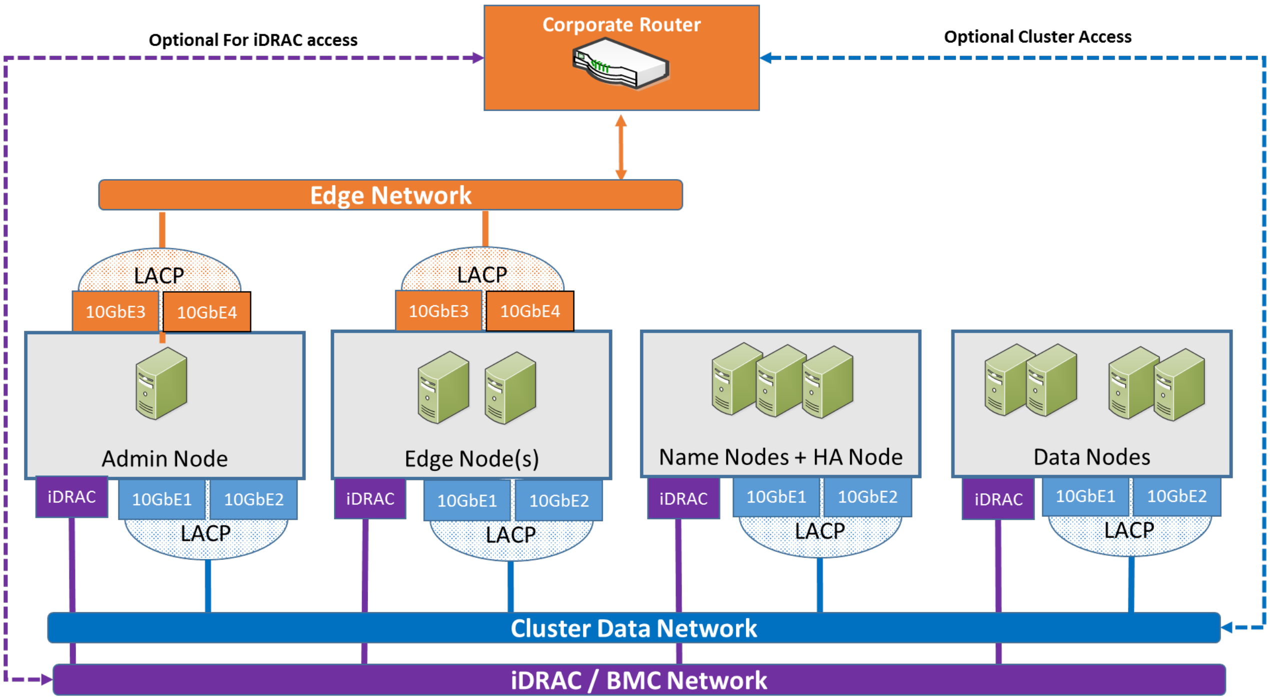

Figure 3: Network Connections

Cluster Networks

Three distinct networks are used in the cluster:

Table 4: Cluster Networks

Logical Network	Connection	Switch
Cluster Data Network	Bonded 10GbE	Dual top of rack (Pod) switches and aggregation switches
BMC Network	1GbE	Dedicated switch per rack
Edge Network	10GbE, optionally bonded	Direct to edge network, or via pod or aggregation switch

Each network uses a separate vLAN, and dedicated components when possible. *Figure 3: Network Connections* on page 17 shows the logical organization of the network.

Server Node Connections

Server connections to the network switches for the data network are bonded, and use an Active-Active LAN aggregation group (LAG) in a load-balance configuration using IEEE 802.3 Link Aggregation Control Protocol (LACP). (Under Linux®, this is referred to as 802.3ad or mode 4 bonding.)

The connections are made to a pair of Pod switches, to provide redundancy in the case of port, cable, or switch failure. The switch ports are configured as a LAG.

Connections to the BMC network use a single connection from the iDRAC port to a S3048-ON management switch in each rack.

Edge Nodes have an additional pair of 10GbE connections available. These connections facilitate high-performance cluster access between applications running on those nodes, and the optional edge network.

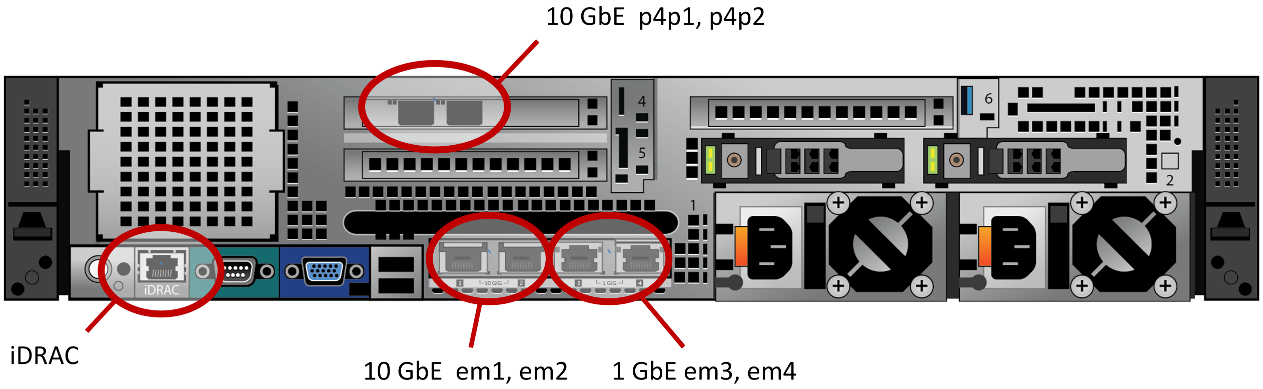

Figure 4: PowerEdge R730xd Node Network Ports

Server Architecture

We separate the server hardware configuration into two main categories:

- Master nodes
- Data Nodes

Edge nodes can use the Master node configuration, or a specialized configuration.

Master Nodes

Master nodes are used to host the critical cluster services, and the configuration is optimized to reduce downtime and provide high performance. The recommended configuration is listed in *Table 5: Server Hardware Configuration - Master Nodes* on page 18. The recommended disk layout is in *Table 6: Master Node Disk Layout* on page 18.

Table 5: Server Hardware Configuration - Master Nodes

Component	Hardware Option
Platform	Dell EMC PowerEdge R730xd (12-Drive Option with Flex Bay)
Processor	2x Intel Xeon E5-2650 v4 2.2 GHz (12-Core)
RAM (minimum)	256 GB
Network Daughter Card	Intel X520 Dual-port 10GbE + I350 Dual-port 1GbE
Add-in PCI-E Network Card	None
Disk (Hot-Plug)	8x 1TB 7.2K RPM SAS 12Gbps (Data)
Disk (Flex Bay)	2x 600GB 10K RPM SAS 12Gbps (OS)
Storage Controller	Dell EMC PowerEdge RAID Controller (PERC) H730

Table 6: Master Node Disk Layout

Function	Disks	Type
Operating System	2	RAID 1 (Mirror)

Function	Disks	Type
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

Note: 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.

Note: Disks – You can consider 10K SAS for your data drives as well as write-intensive SSDs (like the Intel S3710) for your journal drives. As of the publishing of this paper, SSD drives for the Flex Bay are becoming cost-competitive to the 10K SAS drives, so you may want to leverage those for speed/power consumption.

Note: Memory – 256 GB of RAM is adequate for clusters up to approximately 100 nodes. Larger clusters or additional services running on these nodes will require more memory.

Data Nodes

Data Nodes are the workhorses of the HDP cluster. Data Nodes combine compute and storage, so depending on the intended workload they can be optimized for storage-heavy, compute-heavy, or mixed loads

We provide three alternative data node configurations:

- *Table 7: General Purpose Data Node Hardware* on page 19– This is a mainstream configuration that includes large form-factor (LFF) 3.5" drives for the data and two drives for the OS in the rear Flex Bay.
- *Table 8: High Performance Data Node Hardware* on page 20– This node has small form-factor (SFF) 2.5" disks and includes SSDs for leveraging Hadoop's heterogeneous storage tiering.
- *Table 9: High Capacity Data Node Hardware* on page 20 – This node uses our Dell EMC PowerEdge 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 additional 3.5" disks (still hot-swappable).

A cluster using the *Table 7: General Purpose Data Node Hardware* on page 19 provides a good balance between processing and storage for a large variety of workloads, including large batch processing jobs and analytical workloads. This configuration will run HBase, Spark, and MapReduce workloads.

A cluster using the *Table 8: High Performance Data Node Hardware* on page 20 trades off total storage capacity for additional spindles. This configuration provides higher performance, especially for query oriented workloads using HBase, Accumulo, or HAWK that need interactive response.

Using the *Table 9: High Capacity Data Node Hardware* on page 20, a cluster can provide very deep storage of 64TB per node. This configuration is useful for situations requiring deep archival storage, and is only recommended for larger clusters where a single node failure will not impact total available storage capacity.

Table 6: Master Node Disk Layout on page 18 shows the recommended disk and filesystem usage.

Table 7: General Purpose Data Node Hardware

Component	Hardware Option
Platform	Dell EMC PowerEdge R730xd (12-Drive Option with Flex Bay)
Processor	2x Intel Xeon E5-2650 v4 2.2 GHz (12-Core)

Component	Hardware Option
RAM (min)	256 GB
Network Daughter Card	Intel X520 Dual-port 10GbE + I350 Dual-port 1GbE (LACP Bonded)
Add-in PCI-E Network Card	None
Disk (Hot-Plug)	12x 4TB 7.2K RPM SAS 12Gbps (HDFS) – Non-RAID or RAID 0
Disk (Flex Bay)	2x 600GB 10K RPM SAS 12Gbps (OS) – RAID 1 (Mirror)
Storage Controller	Dell PowerEdge RAID Controller (PERC) H730

Table 8: High Performance Data Node Hardware

Component	Hardware Option
Platform	Dell EMC PowerEdge R730xd (24-Drive Option with Flex Bay)
Processor	2x Intel Xeon E5-2690 v4 2.6 GHz (14-Core)
RAM (min)	256 GB
Network Daughter Card	Intel X520 Dual-port 10GbE + I350 Dual-port 1GbE (LACP Bonded)
Add-in PCI-E Network Card	Intel X520 Dual-port 10GbE Server Adapter (LACP Bonded)
Disk (Hot-Plug)	• 20x 1.2 TB 10K RPM SAS 12Gbps (HDFS Disk Tier) – Non-RAID or RAID 0 • 2x 800GB Intel S3710 Write-intensive SATA (Scratch) – Non-RAID / RAID 1 • 2x 800GB Intel S3710 Write-intensive SATA (HDFS SSD Tier) – Non-RAID / RAID 1
Disk (Flex Bay)	2x 600GB 10K RPM SAS 12Gbps (OS) – RAID 1 (Mirror)
Storage Controller	Dell PowerEdge RAID Controller (PERC) H730

Table 9: High Capacity Data Node Hardware

Component	Hardware Option
Platform	Dell EMC PowerEdge R730xd (12+2+4 Drive Option with Flex Bay and Mid-bay)
Processor	2x Intel Xeon E5-2650 v4 2.2 GHz (12-Core)
RAM (min)	256 GB
Network Daughter Card	Intel X520 Dual-port 10GbE + I350 Dual-port 1GbE (LACP Bonded)
Add-in PCI-E Network Card	None
Disk (Hot-Plug)	12x 4TB 7.2K RPM SAS 12Gbps (HDFS) – Non-RAID or RAID 0
Disk (Flex Bay)	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

Table 10: Data Node Disk Layout

Function	Disks	Type
Operating System	2 (flex bay)	RAID 1 (Mirror)
HDFS Data	12 (or 16)	Non-RAID or RAID 0

Note: 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 EMC Technical Sales teams to determine the appropriate size processors given thermal and power restrictions.

Note: 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.

Note: Network – As 10GbE per-port costs drop and 40-100GbE switching becomes more mainstream, many customers are opting to increase the bandwidth to the Data Nodes. This can also help in reducing rebuild time in the case of whole-node failures.

Note: Memory – 256 GB of RAM is probably the starting point for conversations around Data Node memory amounts. With the rise in Spark (and other in-memory storage and processing engines) we are seeing customer push this to 512GB. Using more than 512 GB of RAM is not recommended.

Edge Node Configurations

Edge nodes are the primary interface through which data traverses in and out of the cluster. As such, they can vary significantly depending on use-case to use-case. The main characteristic of edge nodes is a connection to the cluster data network, and additional connections for external access.

A basic edge node configuration can use the same configuration as a master node, as shown in *Table 5: Server Hardware Configuration - Master Nodes* on page 18. This is a good choice for an initial development or small production cluster. An alternative edge node profile is shown in *Table 11: Staging Edge Node Hardware* on page 21. This configuration is optimized for a larger amount of local storage, and would typically be used in ETL scenarios where data is staged before moving in or out of the cluster.

Table 11: Staging Edge Node Hardware

Component	Hardware Option
Platform	Dell EMC PowerEdge R730xd (12-Drive Option with Flex Bay)
Processor	2x Intel Xeon E5-2650 v4 2.2 GHz (12-Core)
RAM (min)	256 GB
Network Daughter Card	Intel X520 Dual-port 10GbE + I350 Dual-port 1GbE (LACP Bonded)
Add-in PCI-E Network Card	Intel X520 Dual-port 10GbE Server Adapter (LACP Bonded)
Disk (Hot-Plug)	12x 4TB 7.2K RPM SAS 12Gbps (Data) – RAID 5 or RAID 10
Disk (Flex Bay)	2x 600GB 10K RPM SAS 12Gbps (OS) – RAID 1 (Mirror)
Storage Controller	Dell PowerEdge RAID Controller (PERC) H730

Sizing Guidelines

Dell EMC recognizes that use cases for Hadoop range from small development clusters all the way through large multi petabyte production installations. It is a good idea to leverage the Dell EMC Customer Solution Centers subject matter experts to help determine you exact needs.

Node Count Recommendations

As a starting point, three cluster configurations can be defined for typical use:

- **Proof of Concept Cluster** – This is a minimum size cluster, targeted at proof of concept projects. The performance of this cluster will not demonstrate the highly distributed nature of HDFS but it is large enough to demonstrate parallel execution.
- **Minimum Development Cluster** – This is a good starting point for development efforts after a proof of concept is completed. This cluster provides the resiliency that is expected in today's production IT world and additional scalability.
- **Minimum Production Cluster** – The minimum 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 12: Recommended Cluster Sizes

Node Type	Proof of Concept Cluster	Minimum Development Cluster	Minimum Production Cluster
NameNode	2	2	2
HA Node	1	1	1
Edge Node(s)	0	1	1
Admin Node	0	1	1
Data Nodes	5	7	10
1 GbE Switches (Dell Networking S3048-ON)	1	1	1
10 GbE Switches (S4048-ON)	1	2	2
Rack Units	18U	27U	33U across two racks

Note: In the proof of concept cluster, the HA node also serves as both the Ambari Admin and Edge Node.

Note: The 1GbE switches are used for access to the Dell Remote Access Controllers (iDRAC) for out-of-band management.

Note: The 10GbE switches we recommend have some important attributes such as non-blocking backplanes and dedicated per-port packet buffers.

Appendix A

References

Topics:

- *To Learn More*

Additional information can be obtained at *http://www.dell.com/hadoop*.

If you need additional services or implementation help, please contact your Dell EMC sales representative.

To Learn More

For more information on the Dell EMC Hortonworks Hadoop Solution, visit *http://www.dell.com/hadoop*.