Ready Solutions for Data Analytics

Hortonworks Hadoop 3.0

Architecture Guide

December 2019

H17561.1

Abstract

This reference architecture guide describes the architectural recommendations for Hortonworks Data Platform (HDP) 3.0 software on Dell EMC PowerEdge servers and Dell EMC PowerSwitch switches.

Dell EMC Solutions

Dell EMC
Hopkinton, Massachusetts 01748-9103
1-508-435-1000 In North America 1-866-464-7381
www.DellEMC.com

CONTENTS

FIGURES

TABLES

CHAPTER 1

Executive Summary

This chapter presents the following topics:

Document purpose

This document describes the Dell EMC server hardware and networking configuration that is recommended for running the Hortonworks Data Platform (HDP).

It also includes recommendations for the location of HDP core services and components. The installation of additional components and services is flexible and depends on the applications and workloads.

For additional and relevant information, see the Dell EMC Ready Architectures for Hadoop web page.

Note: Cloudera, Inc., as a result of a merger transaction, is now the parent company of Hortonworks, Inc.

Audience

This document is for customers and system architects who require information about configuring Hadoop clusters in their information technology environment for Big Data analytics.

Solution overview

Every business is now a data business. Data is your organization's future and its most valuable asset. Hortonworks Data Platform (HDP) helps enterprises transform their business by unlocking the full potential of Big Data.

Hortonworks Connected Data Platforms deliver an open-architected solution to manage both data-in-motion and data-at-rest, empowering organizations to deliver actionable intelligence to end users through modern data applications. Hortonworks DataFlow (HDF) manages data-in-motion by securely acquiring and transporting data to HDP. HDP manages data-at-rest for all types of data with enterprise-grade governance, security, and operations.

As a leader in open data platforms and modern data applications, Hortonworks focuses on driving innovation through open source communities into core Apache Hadoop as well as associated projects including Apache NiFi, Apache Hive, and Apache Spark. Hortonworks and Dell EMC provide the expertise, training, and services that enable customers to unlock transformational value for their organizations across any line of business.

Solution use case summary

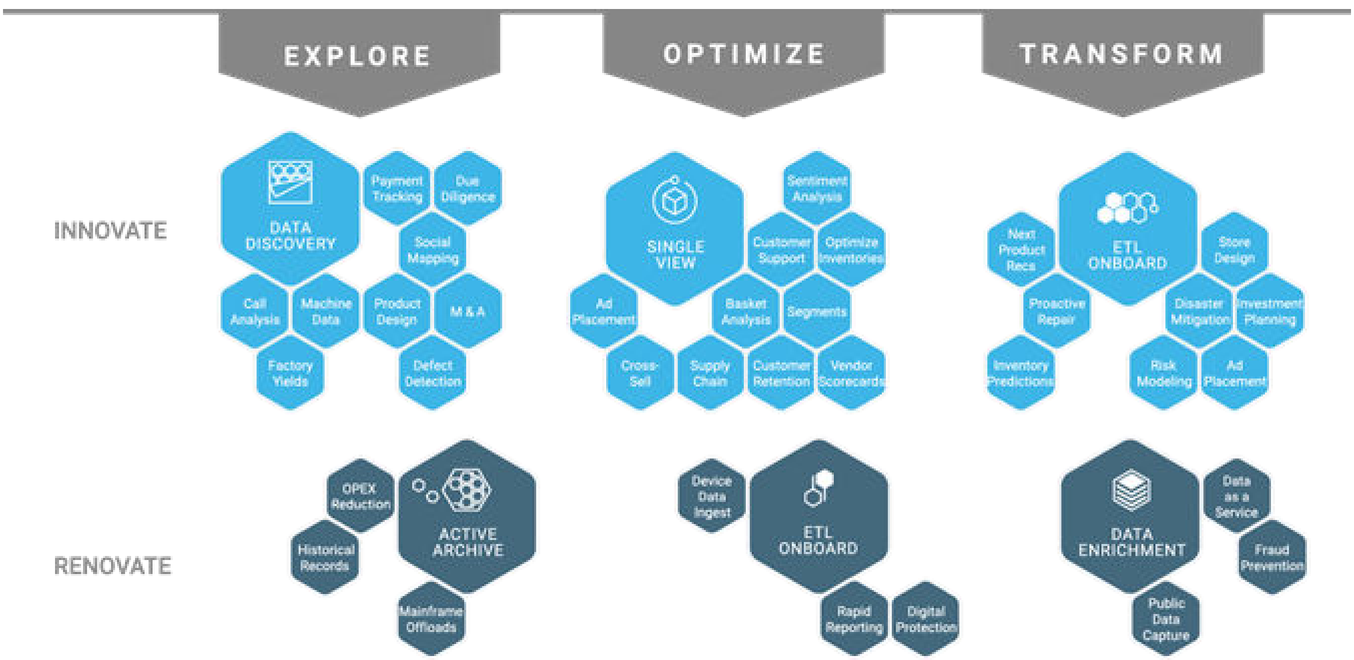
The Hortonworks Data Platform (HDP) helps customers create actionable intelligence to transform their businesses.

Whether it is 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
- Enterprise Data Warehouse (EDW) optimization

The following figure illustrates the HDP use cases.

Figure 1 Hortonworks Data Platform use cases



CHAPTER 2

Ready Architecture Components

This chapter presents the following topics:

Solution components

This Ready Architecture combines the Hortonworks Data Platform (HDP) with Dell EMC PowerEdge servers and Dell EMC PowerSwitch switches to implement a complete Hadoop platform.

Dell EMC PowerEdge rack servers

This Ready Architecture uses Dell EMC's latest server solutions.

Feature summary

- Up to two Intel Xeon scalable processors (SPs), with up to 28 cores per processor
- 24 DDR4 DIMM slots
- Support for RDIMM/LRDIMM, with speeds of up to 2,666 MT/s
- IPMI 2.0 compliance
- iDRAC9 with Lifecycle Controller
- Quick Sync 2 wireless module
- Titanium and platinum efficiency power supplies

Highlights

- Highly optimized air flow design that enables exceptional configuration flexibility and industry-leading energy efficiency
- Out of band management architecture that facilitates rapid bare metal deployment and remediation regardless of operating system state
- Embedded SupportAssist that reduces troubleshooting and downtime with embedded diagnostics and automated case creation

Automated productivity

- Up to 4 times performance improvement in common management tasks with the new iDRAC9 dual-core ARM processor (compared to iDRAC8)
- Use of the same next-generation of embedded automation to standardize BIOS and secure boot configuration, firmware updates, server asset inventory, health monitoring, and power/reset control across all Dell EMC PowerEdge servers
- Embedded proactive automated support that resolves issues up to 90 percent faster

Comprehensive security

- Fully signed firmware updates in which embedded trust only allows authenticated code to run
- Security lock-down that protects your server configuration and firmware (BIOS, iDRAC, and RAID) from malicious changes
- Secure instant erase for HDDs, SSDs, and NVMs
- A more secure, unique default password
- Redfish, a new REST-based management API, that is more secure and scalable than legacy IPMI

Dell EMC PowerEdge R640 server

The Dell EMC PowerEdge R640 server is a dense, general purpose, scale-out compute node.

Dell EMC PowerEdge R640 is an ideal choice for dense scale-out data center computing and storage in a 1U/2S platform. It enables optimization of application performance, price performance, or performance per watt per unit of rack space in most data center environments.

The following figure shows the server.

Figure 2 Dell EMC PowerEdge R640 server 10 x 2.5 in. chassis

Dell EMC PowerEdge R740xd server

The Dell EMC PowerEdge R740xd server is a highly configurable software-defined storage server.

The Dell EMC PowerEdge R740xd is the ideal platform for uncompromising storage performance and data set processing in a 2U/2S form factor. It provides excellent storage performance and density for applications such as software-defined storage. The Dell EMC PowerEdge R740xd is designed with the versatility that is demanded by cloud service providers, Hadoop and Big Data users, and for colocation hosting.

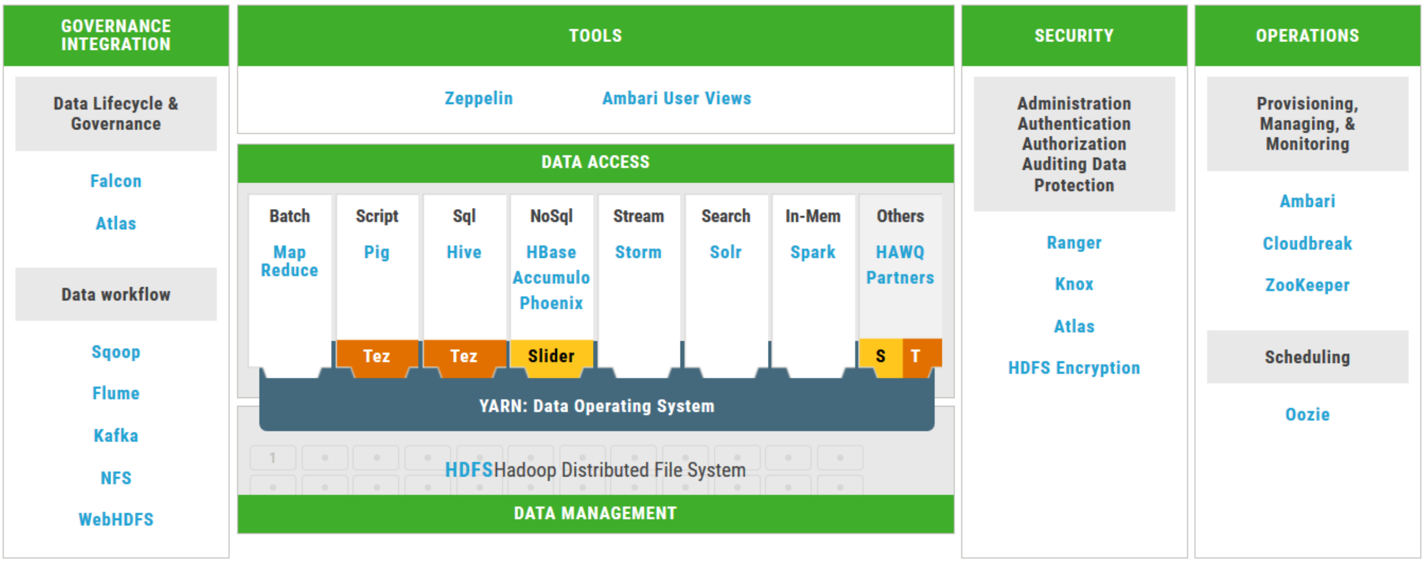
The following figure shows the server.

Figure 3 Dell EMC PowerEdge R740xd server 3.5 in. chassis

Hortonworks Data Platform

Hortonworks Data Platform (HDP) is a secure, enterprise-ready, open source Hadoop distribution that is based on a centralized architecture, as shown in the following figure.

Figure 4 Hortonworks Data Platform components



HDP addresses a range of data-at-rest use cases, powers real-time customer applications, and delivers robust analytics that accelerate decision making and innovation.

The following sections describe the HDP components:

- Data management
- Data access
- HDP Operations
- Security and governance

Data management

The foundational components of HDP are Yet Another Resource Negotiator (YARN) and Hadoop Distributed File System (HDFS).

While HDFS provides the scalable, fault-tolerant, cost-efficient storage for a Hadoop-powered Big Data lake, YARN provides the centralized architecture that enables organizations to process multiple workloads simultaneously. YARN also provides the resource management and pluggable architecture to enable a wide variety of data access methods.

Data access

With YARN at its architectural center, HDP provides a range of processing engines that allow users to interact simultaneously with data in multiple ways.

YARN enables a range of access methods to coexist in the same cluster against shared datasets to avoid unnecessary and costly data silos. HDP enables multiple data processing engines that range from interactive SQL, real-time streaming, data science, and batch processing. These processing engines leverage data that is stored in a single platform, unlocking an entirely new approach to analytics.

HDP Operations

HDP Operations enables IT organizations to bring Hadoop online quickly by taking the guesswork out of manual processes and replacing them with automated preconfigured best practices, guided configurations, and full operation control.

Because of the rapid emergence of Hadoop, many users lack an optimal way to provision and operate the environment, leading them to waste time on inefficient troubleshooting, monitoring, and configuration. HDP Operations makes it easy to operate distributed multi-user, multi-tenant, and multiple data access engines and helps manage HDP clusters at scale through an integrated web UI or single pane of glass.

Apache Ambari is an open source management platform for provisioning, managing, monitoring, and securing Hadoop clusters. Ambari removes the manual and often error-prone tasks that are associated with operating Hadoop. It also provides the necessary integration points to fit seamlessly into the enterprise and enables the IT operator to focus on delivering world-class service and support for their HDP consumers.

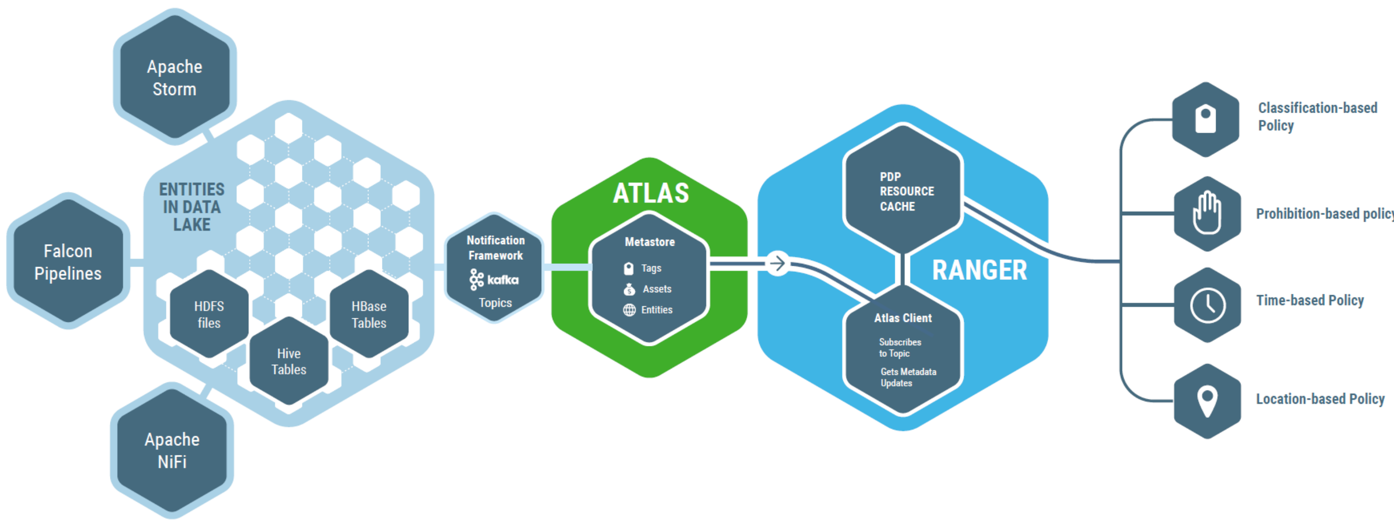
Security and governance

Hortonworks created Data Governance Initiative (DGI), a consortium of cross-industry leaders, to address the need for an open source governance solution to manage data classification, lineage, security, and data life cycle management.

Apache Atlas, created as part of the DGI, empowers organizations to apply consistent data classification across the data ecosystem. Apache Ranger provides centralized security administration for Hadoop.

By integrating Atlas with Ranger, Hortonworks empowers enterprises to institute dynamic access policies at runtime that proactively prevent violations from occurring. The Atlas and Ranger integration represents a paradigm shift for Big Data governance and security. By integrating Atlas with Ranger, enterprises can now implement dynamic classification-based security policies. Ranger's centralized platform empowers data administrators to define a security policy that is based on Atlas metadata tags or attributes, and then apply this policy in real time to the entire hierarchy of data assets including databases, tables, and columns.

Figure 5 Hortonworks Data Platform next generation security



Security

Hadoop-powered Big Data lakes can provide a robust foundation for a new generation of analytics and insight. However, it is important to secure the data before launching or expanding a Hadoop initiative.

By ensuring that data protection and governance are built into their Big Data environment, enterprises can exploit the full value of advanced analytics without exposing their businesses to new risks. Hortonworks and Dell EMC understand the importance of security and governance for every business. To ensure effective protection for customers, Hortonworks uses a holistic approach that is based on five pillars:

- Administration
- Authentication and perimeter security
- Authorization
- Audit
- Data protection

For each of these areas, Hortonworks provides differentiated capabilities so that customers achieve the highest possible level of protection. As a result, Big Data does not have to incur considerable risks and companies can put it to work without compromising data assets.

Governance

As organizations pursue Hadoop initiatives to capture new opportunities for data-driven insight, data governance requirements can pose a key challenge. The management of information to identify its value and enable effective control, security, and compliance for customer and enterprise data is a core requirement for both traditional and Big Data architectures.

As part of its commitment to drive enterprise readiness for Hadoop, Hortonworks established the Data Governance Initiative (DGI) in collaboration with Aetna, Merck, Target, SAS, Schlumberger, and a global financial institution. The charter of this initiative was to introduce a common metadata-powered approach to data governance into the open source community, and to establish a framework with the flexibility to be applied across industries.

Hortonworks Data Platform for teams

Successful deployment of Hadoop in any organization depends on using existing skill sets and resources to adopt the Big Data architecture.

HDP provides valuable tools and capabilities for every persona on your Big Data team:

- Data scientists
- Business analysts
- Developers
- Hadoop operators

Data scientists

Data science is an interdisciplinary field that involves scientific exploration of data to uncover meaning or insight, and the construction of software systems to use such insights in a business or social context. A data scientist uses a combination of econometrics, machine learning, statistics, visualization, and computer science to extract valuable business insights hidden in data and builds operational systems to deliver that value.

Apache Spark, part of HDP, plays an important role in data science. Data scientists commonly use machine learning, a set of techniques and algorithms that can learn from data. These algorithms are often iterative, and Spark's ability to cache the data in memory greatly speeds up data processing, making it an ideal processing engine for implementing such algorithms.

Business analysts

Business analysts are also tasked with extracting value from data. They use business intelligence, analytics, visualization, and dashboard tools for this purpose. They create reports, visualizations, dashboards, scorecards, and key performance indicators (KPIs) to identify trends and patterns in data.

Hadoop makes it technically and economically viable for business analysts to gain actionable insights from data that in the past was too sizable and expensive to retain and analyze. HDP provides business analysts with quick access to vast amounts of data by using SQL on Hadoop interfaces provided by Hive, Spark, SQL, and Phoenix.

Through these interfaces, business analysts can use their favorite Business Intelligence (BI) and business analytics tools to create reports, visualizations, dashboards, and scorecards to make more effective insight-driven decisions.

Developers

Developers are responsible for building Big Data applications in, on, and around Hadoop. They expect a vibrant and powerful set of tools, frameworks, and interfaces to simplify this task. Developers focus on delivering value through applications and do not want to be concerned with the mechanical details of integration with Hadoop.

HDP provides a rich set of tools for application developers, such as Spark and Apache Zeppelin. HDP also features a set of native APIs that simplify development:

- WebHDFS—Provides a REST interface to write, manipulate, and delete HDFS files

- WebHCAT—Provides a critical point of integration to access metadata and schema for Hadoop data

Hadoop operators

Hadoop operators are responsible for smooth operation, support for business use cases, high uptime, and security of the Hadoop environment.

They need full visibility into the cluster health to make it easier to manage, provision, and troubleshoot Hadoop clusters. By using Ambari, operators realize the following benefits to streamline Hadoop operations:

- More flexible upgrades—Ambari enables faster cluster upgrades by automating both maintenance and feature releases while the cluster is down.
- Simplified security operations—Service configurations for Apache Ranger provide a continuation of the new user experience. Optional storage of Kerberos credentials and customizable security settings simplifies administration and provides robust security.
- Improved troubleshooting—For easier and faster troubleshooting, Ambari provides a customizable metric widget graph display and the ability to export metrics to identify and respond to problems quickly.

CHAPTER 3

Solution Architecture Overview

This chapter presents the following topics:

Cluster architecture

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

Node architecture

The Hortonworks Data Platform (HDP) consists of many Hadoop components that address 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 Ready Architecture, we classify physical nodes into roles, and then map services to these roles. The assignment of services and roles to physical nodes is quite flexible, based on cluster workload. The following table shows the physical node classification.

Table 1 Cluster physical node roles

Physical node role	Requirement	Server hardware configuration
Administration Node	Optional	Custom
Active NameNode	Required	Infrastructure
Standby NameNode	Required	Infrastructure
HA Node	Required	Infrastructure
Edge Node	Required (at least one node)	Infrastructure or custom Edge
Worker Node 1 - N	Required	Data

Node definitions

The following list provides node definitions for this solution.

- Administration Node—Provides cluster deployment and management capabilities. The Administration Node is optional in cluster deployments, depending on whether existing provisioning, monitoring, and management infrastructure is used.
- Active NameNode—Runs all the services that manage the HDFS data storage and YARN resource management. It is sometimes called the “master name node.” There are four primary services running on the Active NameNode:
  - YARN Resource Manager—Supports cluster resource management, including MapReduce jobs
  - NameNode—Supports HDFS data storage
  - Journal Manager—Supports high availability
  - ZooKeeper—Supports coordination
- Standby NameNode—When quorum-based HA mode is used, runs the standby namenode process, a second journal manager, and an optional standby resource manager. This node also runs the Spark History Server and a second ZooKeeper service.
- HA Node—Provides the third journal node for HA. The Active NameNodes and Standby NameNodes provide the first and second journal nodes. It also runs a third ZooKeeper service.

The operational databases that are required for Ambari and additional metastores are on the HA node.

- Edge Node—Provides an interface between the data and processing capacity that is available in the Hadoop cluster and a user of that capacity. An Edge Node has an additional connection to the Edge Network and is sometimes called a “gateway node.” At least one Edge Node is required.
- Worker Node—Runs all the services that are required to store blocks of data on the local hard drives and run processing tasks against that data. A minimum of five Worker Nodes are required. Larger clusters are scaled primarily by adding Worker Nodes. The primary services running on the Worker Nodes are:
  - DataNode daemon (to support HDFS data storage)
  - NodeManager daemon (to support YARN job execution)

  Other services such as HBase also run on Worker Nodes. Spark jobs run on the Worker Nodes. However, there is no persistent service that is associated with Spark jobs.

Core Hortonworks Data Platform services

The following table lists the core Hortonworks Data Platform (HDP) services.

Table 2 HDP services

Service	Function	Master	Worker
HDFS	Hadoop distributed filesystem	Active NameNode, Standby NameNode	Worker 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
Ranger	Security administration	Ranger Gateway	N/A
Ambari	Hadoop cluster management	Ambari Server	Ambari Agent

Node locations

The following table describes the node locations and functions of the cluster services.

Table 3 Service locations by node

Physical node	Software function
Administration Node	Systems Management Services
First Edge Node	• Hadoop Clients • Ambari • Ranger Gateway
Active NameNode	• NameNode

Table 3 Service locations by node (continued)

Physical node	Software function
	• YARN Resource Manager • App Timeline Server • History Server • ZooKeeper • Metrics Collector • Oozie Server
Standby NameNode	• Yum Repositories • Standby NameNode (SNameNode) • HMaster • Hive Metastore • HiveServer2 • Spark2 History Server • ZooKeeper • Activity Explorer
HA Node	• Grafana • Activity Analyzer • Operational Databases (PostgreSQL) • ZooKeeper • Infra Solr Instance • Grafana • Activity Analyzer • Operational Databases (PostgreSQL)
Worker Node(N)	• DataNode • YARN NodeManager • HBase RegionServer • Tez Client • DataNode • YARN NodeManager • HBase RegionServer • Tez Client

High Availability

This Ready Architecture implements high availability (HA) at multiple levels through a combination of hardware redundancy and software support.

Hadoop redundancy

The Hadoop distributed file system implements redundant storage for data resiliency and is aware of node and rack locality.

Data is replicated across multiple nodes and across racks. This replication provides multiple copies of data for reliability if there are disk or node failures. It can also increase performance. The number of replicas defaults to three and can easily be changed. Hadoop automatically maintains replicas when a node fails—the network provides enough bandwidth to handle replication traffic as well as production traffic.

Note: The Hadoop job parallelism model can scale to larger and smaller numbers of nodes, enabling jobs to run when parts of the cluster are offline.

Network redundancy

Optionally, the production network can use bonded connections to pairs of switches in each pod, with the switch pairs configured using Virtual link Trunking (VLT). This configuration provides increased bandwidth capacity, and enables operation at reduced capacity in the event of a network port, network cable, or switch failure.

Note: If you choose to use bonded connections, adjust the recommended network and rack configurations accordingly.

HDFS highly available NameNodes

This Ready Architecture implements high availability (HA) for the Hadoop Distributed File System (HDFS) directory through a quorum mechanism that replicates critical namenode data across multiple physical nodes. Production clusters normally implement namenode HA.

In quorum-based HA, typically, two namenode processes run on two physical servers. At any point in time, one of the NameNodes is in an Active state and the other is in a Standby state. The Active NameNode is responsible for all client operations in the cluster, while the Standby NameNode acts as a slave, maintaining enough state to provide a fast failover if necessary.

For the Standby NameNode to keep its state synchronized with the Active NameNode in this implementation, both nodes communicate with a group of separate daemons called JournalNodes. When the Active NameNode performs any namespace modification, it durably logs a record of the modification to a majority of these JournalNodes.

The Standby NameNode can read the edits from the JournalNodes and constantly monitors them for changes to the edit log. As the Standby NameNode detects the edits, it applies them to its own namespace. If a failover occurs, the Standby NameNode ensures that it has read all of the edits from the JournalNodes before promoting itself to the Active state. This action ensures that the namespace state is fully synchronized before a failover occurs.

To provide a fast failover, it is necessary that the Standby NameNode has up-to-date information about the location of blocks in the cluster. Therefore, the Worker Nodes are configured with the location of both NameNodes, and they send block location information and heartbeats to both NameNodes.

Because edit log modifications must be written to a majority of JournalNodes, there must be an odd number of (and at least three) JournalNode daemons. The JournalNode daemons run on the Active NameNode, Standby NameNode, and HA Node in this architecture.

YARN resource manager high availability

This Ready Architecture supports high availability (HA) for the Hadoop YARN resource manager.

Without resource manager HA, currently running jobs to fail when a Hadoop resource manager fails. When resource manager HA is enabled, jobs can continue running if a resource manager fails.

On failover, the applications can resume from their last check-pointed state. For example, completed map tasks in a MapReduce job are not rerun on a subsequent attempt. This action enables events such as machine crashes or planned maintenance to be handled without any significant performance effect on running applications.

An Active/Standby pair of resource managers implements resource manager HA. On startup, each resource manager is in the standby state, which means that the process is started, but the state is not loaded. When transitioning to the active state, the resource manager loads the internal state from the designated state store and starts all the internal services. The stimulus to transition-to-active comes from either the administrator or through the integrated failover controller when automatic failover is enabled.

Note: Resource manager HA is not always implemented in production clusters.

Database server high availability

This Ready Architecture supports high availability (HA) for the operational databases.

The database server that is used for both the Ambari operational and metadata databases stores its data on a RAID 10 partition to provide redundancy in the event of a drive failure.

Note: Our default installation uses a single PostgreSQL instance. Therefore, there is a single point of failure. You can implement database server HA by using one or more additional PostgreSQL instances on other nodes in the cluster or by using an external database server.

Rack server hardware configurations

This Ready Architecture supports the Dell EMC PowerEdge R640 and Dell EMC PowerEdge R740xd servers using configurations for the following:

- Infrastructure Nodes on page 27
- Worker Nodes on page 27
- Edge Nodes on page 28

For more information about configuration, refer to Appendix A, which provides the recommended rack layout for Dell EMC PowerEdge R740xd clusters.

Infrastructure Nodes

Infrastructure Nodes host the critical cluster services. The configuration is optimized to reduce downtime and provide high performance.

The following table lists the recommended configuration.

Table 4 Hardware configuration: Dell EMC PowerEdge R640 Infrastructure Nodes

Machine function	Infrastructure Nodes
Platform	Dell EMC PowerEdge R640 server
Chassis	2.5 in. chassis with up to 10 hard drives and 2 PCIe slots
Processor	Dual Intel Xeon Gold 6234 3.3 GHz (8 Core) 24.75 M cache
RAM	192 GB (12 x 16 GB 2933 MT/s)
Network Daughter Card	Mellanox ConnectX-4 Lx Dual Port 25 GbE DA/SFP rNDC
Boot configuration	From PERC controller
Storage controller	Dell EMC PERC H740P RAID Controller, 8 GB NV Cache, Minicard
Disk - spindles	8 x 1 TB 7.2K RPM NLSAS 12 Gbps
Disk - SSD	2 x 480 GB SSD SAS Mix Use 12 Gbps
Drive configuration	Combination of RAID 1, RAID 10, and dedicated drives

Note: Consult your Dell EMC account representative before changing the recommended disk sizes.

Worker Nodes

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

The following table shows a 2U chassis option using large form-factor (LFF) 3.5 in. drives for data. This option provides dense storage capability with high performance compute and solid state storage for fast caching of temporary data.

Table 5 Hardware configurations: Dell EMC PowerEdge R740xd Worker Nodes

Component	Details
Platform	Dell EMC PowerEdge R740xd server
Chassis	Chassis with up to 12 x 3.5 in. HDD, 4 x 3.5 in. HDDs on MP and 4 x 2.5 in. HDDs on Flex Bay
Processor	Dual Intel Xeon Gold 6240 2.6 GHz, 18 Core, 25 M Cache
RAM (minimum)	384 GB (12 x 32 GB 2933 MT/s)
Network Daughter Card	Mellanox ConnectX-4 Lx Dual Port 25 GbE DA/SFP rNDC
Boot configuration	BOSS controller card + with 2 M.2 Sticks 240 GB
Storage controller	Dell EMC PERC HBA330 RAID Controller, 12 Gb Minicard
Disk - spindles	16 x 4 TB 7.2 K RPM SATA 6 Gb/ps 512n 3.5 in. hot-plug hard drive
Disk - SSD	4 x 480 GB SSD SAS mixed-use 12 Gb/ps
Drive configuration	RAID 1 - OS JBOD - data drives

Edge Nodes

Edge Nodes are the primary interface through which the data moves in and out of the cluster. They are also used to run applications that access the cluster. Because of the wide variation in applications, Edge Node configurations can vary significantly. The main characteristic of Edge Nodes is a connection to the Cluster Data network and additional network connections for external access.

A common baseline choice for Edge Node configuration uses the same configuration as an infrastructure node, as shown in the following table.

Table 6 Hardware configurations: Dell EMC PowerEdge R640 Edge Nodes

Component	Details
Platform	Dell EMC PowerEdge R640 server
Chassis	2.5 in. chassis with up to 10 hard drives and 2 PCIe slots
Processor	Dual Intel Xeon Gold 6234 3.3 GHz (8 Core) 24.75 M cache
RAM	192 GB (12 x 16 GB 2933 MT/s)
Network Daughter Card	Mellanox ConnectX-4 Lx Dual Port 25 GbE DA/SFP rNDC
Boot configuration	From PERC controller
Storage controller	Dell EMC PERC H740P RAID Controller, 8 GB NV Cache, Minicard
Disk - spindles	8 x 1 TB 7.2K RPM NLSAS 12 Gbps
Disk - SSD	2 x 480 GB SSD SAS Mix Use 12 Gbps
Drive configuration	Combination of RAID 1, RAID 10, and dedicated drives.

Storage sizing notes

For drive capacities greater than 4 TB or node storage density over 48 TB, special consideration is required for Hadoop Distributed File System (HDFS) setup. Configurations of this size approach the limit of Hadoop per-node storage capacity.

At a minimum, the HDFS block size must be no less than 128 MB and can be as large as 1,024 MB. Because the number of files, blocks per file, compression, and reserved space factor into the calculations, the configuration requires an analysis of the intended cluster usage and data.

When sizing nodes, per-node density also has an impact on cluster performance in the event of node failure. The bandwidth that is required to replicate the lost data impacts overall performance, the time that is required to finish the recovery is lengthy, and data is under-replicated and at risk during the recovery

Note: Never configure a single Worker Node with more than 100 TB of storage.

Your Dell EMC representative can assist you with these estimates and calculations.

Network architecture

The cluster network is designed 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 that is based on 25 GbE networking technologies. It uses Dell EMC PowerSwitch S5248-ON switches for the leaves and Dell EMC PowerSwitch S5232F-ON switches for the spine.

IPv4 is used for the network layer. At this time, the architecture does not support the use of IPv6 for network connectivity.

Cluster networks

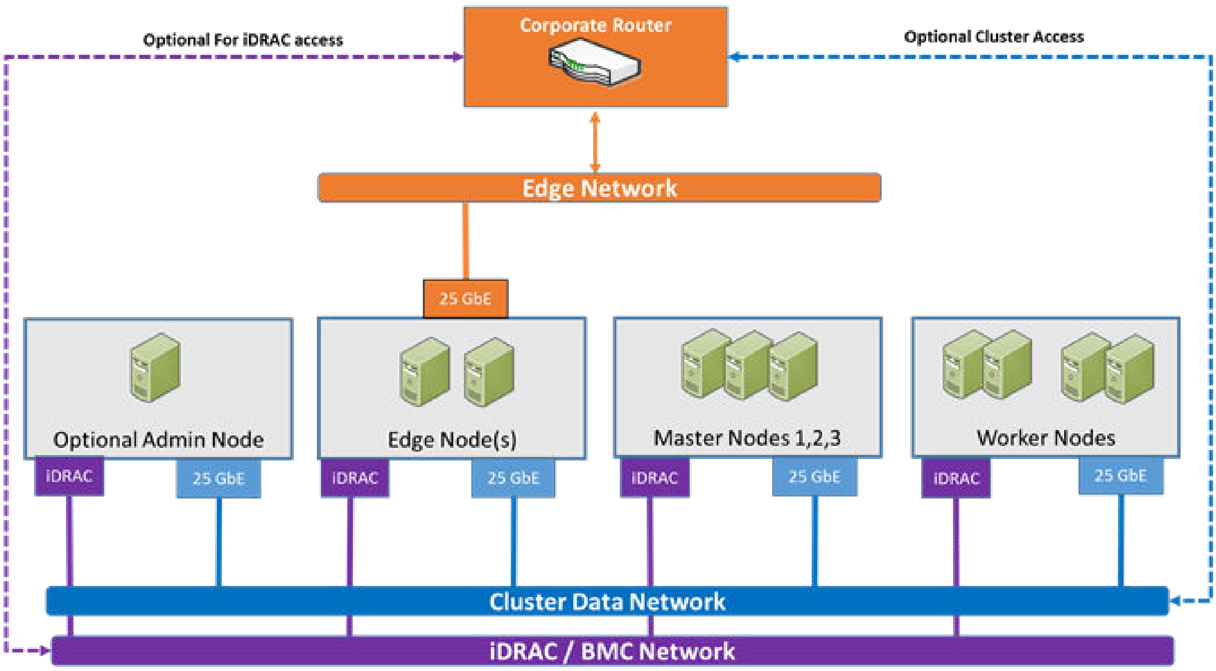
The following table lists the distinct networks that are used in the cluster.

Table 7 Cluster networks

Logical network	Connection	Switch
Cluster Data network	25 GbE	Top-of-rack (ToR) (pod) switches and aggregation switches
BMC network	1 GbE	Dedicated switch per rack
Edge network	25 GbE	Direct to Edge network or through a pod or aggregation switch

Each network uses a separate VLAN and dedicated components when possible. The following figure shows the logical organization of the network.

Figure 6 Hadoop 25 GbE network connections



Network definitions

The following table defines the networks for this Ready Architecture.

Table 8 Network definitions

Network	Description	Available services
Cluster Data network	The Data network carries the bulk of the traffic in the cluster. This network is aggregated in each pod and pods are aggregated into the cluster switch.	The Hortonworks Data Platform services are available on this network. Note: The Hortonworks Data Platform services do not support multihoming and are only accessible on the Cluster Data Network.
iDRAC/BMC network	The BMC network connects the BMC or iDRAC ports and the out of band management ports of the switches and is used for hardware provisioning and management. It is aggregated into a management switch in each rack.	This network provides access to the BMC and iDRAC functionality on the servers. It also provides access to the management ports of the cluster switches.
Edge network	The Edge network provides connectivity from the Edge Nodes to an existing premises network, either directly, or through the pod or cluster aggregation switches.	SSH access is available on this network. Other application services might be configured and available.

You can adapt connectivity between the cluster and existing network infrastructure to specific installations. Common scenarios are:

- The Cluster Data network is isolated from any existing network and access to the cluster is by using the edge network only.
- The Cluster Data network is exposed to an existing network. In this scenario, the Edge network is either not used or is used for application access or ingest processing.

Physical network components

The physical networks of this Ready Architecture consist of the following components:

- Server node connections on page 31
- Network fabric on page 32
  - 25 GbE pod switches on page 33
  - 25 GbE Layer 2 Dell EMC PowerSwitch S5232F-ON cluster aggregation on page 35
- iDRAC Management Network on page 35

Network integration information is presented in:

- Core network integration on page 35
- Layer 2 and Layer 3 separation on page 35

All equipment is listed in:

- Network equipment summary: 25 GbE configurations on page 36

Server node connections

Server connections to the network switches for the Data network use Ethernet technology. Connections to the network use 25 GbE, which is recommended for new Dell EMC PowerEdge R640 and Dell EMC PowerEdge R740xd server deployments.

Edge Nodes have an additional available network connection. This connection facilitates high-performance cluster access between applications running on those nodes and the optional Edge network.

Server connections to the BMC network use a single connection from the iDRAC port to a S3048-ON management switch in each rack, as shown in the following figures.

Figure 7 Dell EMC PowerEdge R640 Infrastructure Node network ports

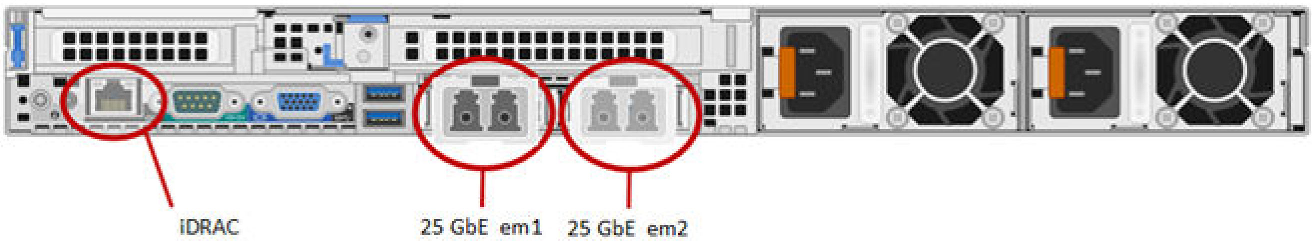
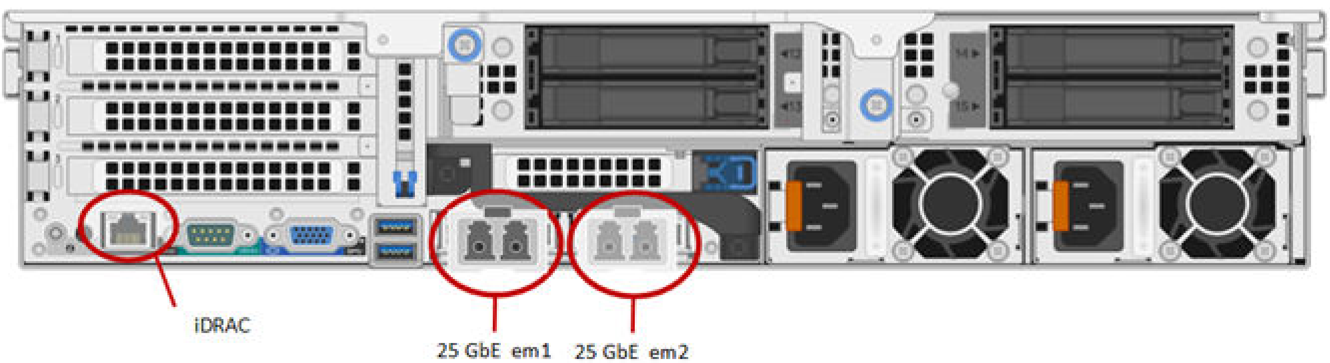


Figure 8 Dell EMC PowerEdge R740xd Worker Node network ports



The following table shows the mapping of individual interfaces to networks and bonds.

Table 9 Network/bond/interface cross reference

Server Platform	Network	Interface	Bond
Dell EMC PowerEdge R740xd	Cluster Data	em1	none
Dell EMC PowerEdge R640	Cluster Data	em1	none
Dell EMC PowerEdge R640	Edge	em2	none

Network fabric

We recommend 25 GbE for new deployments of Dell EMC PowerEdge R740xd and Dell EMC PowerEdge R640 servers.

Clusters larger than a single pod require an aggregation layer. The aggregation layer can be implemented at either Layer 2 (L2) or Layer 3 (L3). The choice depends on the initial size and planned scaling.

Layer 2 aggregation provides lower cost and medium scalability, and can support approximately 250 nodes.

Layer 3 aggregation is recommended for:

- Larger initial deployments of over 250 nodes
- Deployments where extreme scale-out is planned to about 1500 nodes
- Instances where the cluster must be colocated with other applications in a different rack

The actual scalability depends on the switches used and the oversubscription ratio.

The following sections describe the fabric:

- 25 GbE pod switches
- 25 GbE Layer 2 Dell EMC PowerSwitch S5232F-ON cluster aggregation on page 35

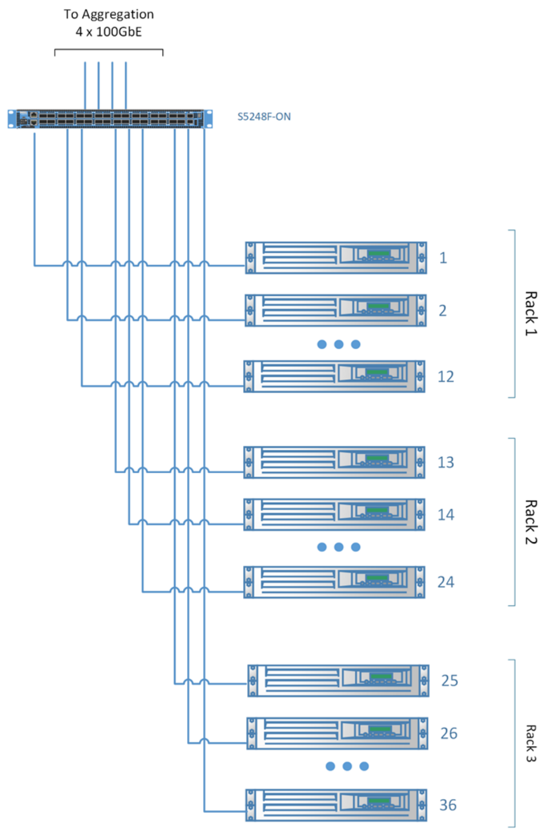
25 GbE pod switches

Each pod uses a Dell EMC PowerSwitch S5248-ON switch as the first layer switch. The pod switches are often referred to as top-of-rack (ToR) switches, although this architecture splits a physical rack from a logical pod.

The Dell EMC PowerSwitch S5248-ON switch is a 1RU spine switch that is optimized for high performance, ultra-low latency data center requirements. The Dell EMC PowerSwitch S5248-ON provides six ports of 100 GbE (QSFP28) and 48 ports of 25 GbE (QSFP+).

The following figure shows the single pod network configuration with a Dell EMC PowerSwitch S5248-ON switch aggregating the pod traffic.

Figure 9 25 GbE single pod networking equipment



For a single pod, the pod switch can act as the aggregation layer for the entire cluster. Multiple-pod clusters require a cluster aggregation layer.

In this Ready Architecture, each pod is managed as a separate entity from a switching perspective and the individual pod switches connect only to the aggregation switch.

25 GbE cluster aggregation switches

For clusters consisting of more than one pod, this Ready Architecture uses the Dell EMC PowerSwitch S5232F-ON switch for an aggregation switch.

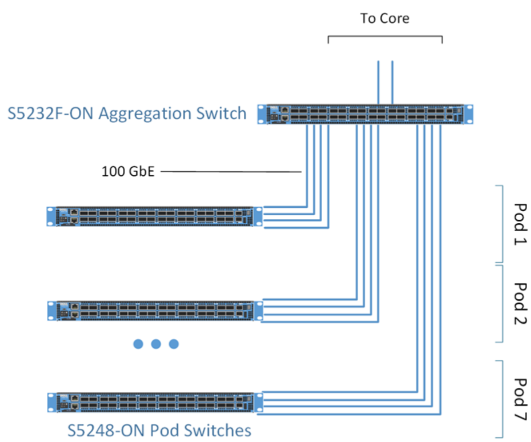
The Dell EMC PowerSwitch S5232F-ON is a 100 GbE 1RU spine switch that is optimized for high-performance, ultra-low latency data center requirements. It provides 32 ports of 100 GbE (QSFP28).

The Dell EMC PowerSwitch S5232F-ON can be used for both Layer 2 and Layer 3 implementations.

25 GbE Layer 2 Dell EMC PowerSwitch S5232F-ON cluster aggregation

The following figure illustrates the configuration for a multiple-pod cluster using the S5232F-ON switch for cluster aggregation switch with Layer 2 networking.

Figure 10 Dell EMC PowerSwitch S5232F-ON multiple-pod networking equipment



The uplink from each S5248-ON pod switch to the aggregation layer uses four 100 GbE interfaces in a bonded configuration, providing a collective bandwidth of 400 Gb from each pod.

iDRAC Management Network

The iDRAC (or BMC) network is a separate network that is provided for cluster management.

The iDRAC management ports are aggregated into a per-rack Dell EMC PowerSwitch S3048-ON switch with a dedicated VLAN. This aggregate provides a dedicated iDRAC/BMC network for hardware provisioning and management. Switch management ports are also connected to this network.

The management switches can be connected to the core or connected to a dedicated management network if out of band management is required.

Core network integration

The aggregation layer functions as the network core for the cluster.

In most instances, the cluster connects to a larger core in the enterprise, as shown in Figure 10 on page 35. When you use the Dell EMC PowerSwitch S5232F-ON switch, four 100 GbE ports are reserved at the aggregation level for connection to the core. Connection details are site-specific and must be determined as part of the deployment planning.

Layer 2 and Layer 3 separation

The Layer 2 and Layer 3 boundaries are separated at either the pod or the aggregation layer. Either option is equally viable. This Ready Architecture is based on Layer 2 for switching in the cluster.

Network equipment summary: 25 GbE configurations

The following tables summarize the required cluster networking equipment.

Table 10 Per rack network equipment

Component	Quantity
Total racks	1 (12 nodes nominal)
Management switch	1 x Dell EMC PowerSwitch S3048-ON
Switch interconnect cables	1 x 1 GbE cables (to next rack management switch)

Table 11 Per pod network equipment

Component	Quantity
Total racks	3 (36 Nodes)
Top-of-rack switches	1 x Dell EMC PowerSwitch S5248-ON
Pod uplink cables (to aggregate switch)	4 x 100 Gb QSFP+ cables

Table 12 Per cluster aggregation network switches for multiple pods

Component	Quantity
Total pods	8
Aggregation layer switches	1 x Dell EMC PowerSwitch S5232F-ON

The following table summarizes the number of cables that are needed for a cluster.

Table 13 Per node network cables required

Description	1 GbE cables required	25 GbE connections with QSFP+ required
Master Nodes	1 x number of nodes	1 x number of nodes
Edge Nodes	1 x number of nodes	2 x number of nodes
Worker Nodes	1 x number of nodes	1 x number of nodes

CHAPTER 4

References

This chapter presents the following topics:

Hortonworks partnership and certification

Note: Cloudera, Inc., as a result of a merger transaction, is now the parent company of Hortonworks, Inc.

Cloudera is a key contributor to the Apache Hadoop project. The Cloudera Distribution for Apache Hadoop (CDH) is a highly scalable open source platform for high-volume data management and analytics. CDH integrates with existing enterprise IT infrastructure, enabling data engineers and data scientists to quickly and easily develop and deploy Hadoop applications in a cost-efficient manner.

Dell EMC is a Platinum member of the Hortonworks IHV Program. Platinum membership is the highest level of partnership and indicates Dell EMC's ongoing commitment to Hortonworks and our customers.

The Dell EMC infrastructure in this guide is Hortonworks-certified.

Dell EMC Customer Solution Centers

Our global network of dedicated Dell EMC Customer Solution Centers are trusted environments where world class IT experts collaborate with customers and prospects to share best practices, facilitate in-depth discussions of effective business strategies using briefings, workshops, or proof of concept (PoCs), and help businesses become more successful and competitive. Dell EMC Customer Solution Centers reduce the risks associated with new technology investments and can help improve speed of implementation.

Technical support

The following table lists where to obtain technical support for the various components of this Ready Architecture.

Table 14 Ready Architecture support matrix

Category	Component	Version	Available support
Operating system	Red Hat Enterprise Linux Server	7.5	Red Hat Linux support
Operating system	CentOS	7.5	Dell EMC hardware support
Java Virtual Machine	Sun Oracle JVM	Java 8 (1.8.0_112)	N/A
Hadoop	Hortonworks Data Platform	3.0.1	Hortonworks support
Hadoop	Ambari	2.7.1.0	Hortonworks support

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

APPENDIX A

Dell EMC PowerEdge R740xd Worker Nodes Physical Rack Configuration

This appendix contains suggested rack layouts for single-rack, single-pod, and multiple-pod installations. Rack layouts vary depending on power, cooling, and loading constraints.

PowerEdge R740xd Worker Nodes single-rack configuration

Table 15 Single-rack configuration: Worker Nodes

RU	RACK1
42	R1 - Switch 1: Dell EMC PowerSwitch S5248-ON
41	Cable management
40	Cable management
39	Cable management
38	R1 - Dell EMC PowerSwitch S3048-ON iDRAC Management switch
37	Cable management
36	Cable management
35 29	Empty
28 27	Edge01: Dell EMC PowerEdge R640
26 25	Active NameNode: Dell EMC PowerEdge R640
24 23	Standby NameNode: Dell EMC PowerEdge R640
22 21	HA Node: Dell EMC PowerEdge R640
20 19	Empty
18 17	Empty
16 15	R1 - Chassis08: Dell EMC PowerEdge R740xd
14 13	R1 - Chassis07: Dell EMC PowerEdge R740xd
12 11	R1 - Chassis06: Dell EMC PowerEdge R740xd
10 9	R1 - Chassis05: Dell EMC PowerEdge R740xd
8	R1 - Chassis04: Dell EMC PowerEdge R740xd

Table 15 Single-rack configuration: Worker Nodes (continued)

RU	RACK1
7	
6 5	R1 - Chassis03: Dell EMC PowerEdge R740xd
4 3	R1 - Chassis02: Dell EMC PowerEdge R740xd
2 1	R1 - Chassis01: Dell EMC PowerEdge R740xd

PowerEdge R740xd Worker Nodes initial rack configuration

Table 16 Initial pod rack configuration: Dell EMC PowerEdge R740xd Worker Nodes

RU	RACK1	RACK2	RACK3
42	Empty	R2 - Switch 1: Dell EMC PowerSwitch S5248-ON	Empty
41	Empty	Empty	Empty
40	Cable management	Cable management	Cable management
39	Cable management	Cable management	Cable management
38	R1 - Dell EMC PowerSwitch S3048-ON iDRAC Management switch	R2 - Dell EMC PowerSwitch S3048-ON iDRAC Management switch	R3 - Dell EMC PowerSwitch S3048-ON iDRAC Management switch
37	Cable management	Cable management	Cable management
36	Cable management	Cable management	Cable management
35	Active NameNode: Dell EMC PowerEdge R640	Edge01: Dell EMC PowerEdge R640	R3 - Switch 1: Dell EMC PowerSwitch S5232F-ON
34	Empty	Empty	Empty
33 32	Empty	Standby NameNode: Dell EMC PowerEdge R640	HA Node: Dell EMC PowerEdge R640
31 21	Empty	Empty	Empty
20 19	R1 - Chassis10: Dell EMC PowerEdge R740xd	R2 - Chassis10: Dell EMC PowerEdge R740xd	R3 - Chassis10: Dell EMC PowerEdge R740xd
18 17	R1 - Chassis09: Dell EMC PowerEdge R740xd	R2 - Chassis09: Dell EMC PowerEdge R740xd	R3 - Chassis09: Dell EMC PowerEdge R740xd

Table 16 Initial pod rack configuration: Dell EMC PowerEdge R740xd Worker Nodes (continued)

RU	RACK1	RACK2	RACK3
16 15	R1 - Chassis08: Dell EMC PowerEdge R740xd	R2 - Chassis08: Dell EMC PowerEdge R740xd	R3 - Chassis08: Dell EMC PowerEdge R740xd
14 13	R1 - Chassis07: Dell EMC PowerEdge R740xd	R2 - Chassis07: Dell EMC PowerEdge R740xd	R3 - Chassis07: Dell EMC PowerEdge R740xd
12 11	R1 - Chassis06: Dell EMC PowerEdge R740xd	R2 - Chassis06: Dell EMC PowerEdge R740xd	R3 - Chassis06: Dell EMC PowerEdge R740xd
10 9	R1 - Chassis05: Dell EMC PowerEdge R740xd	R2 - Chassis05: Dell EMC PowerEdge R740xd	R3 - Chassis05: Dell EMC PowerEdge R740xd
8 7	R1 - Chassis04: Dell EMC PowerEdge R740xd	R2 - Chassis04: Dell EMC PowerEdge R740xd	R3 - Chassis04: Dell EMC PowerEdge R740xd
6 5	R1 - Chassis03: Dell EMC PowerEdge R740xd	R2 - Chassis03: Dell EMC PowerEdge R740xd	R3 - Chassis03: Dell EMC PowerEdge R740xd
4 3	R1 - Chassis02: Dell EMC PowerEdge R740xd	R2 - Chassis02: Dell EMC PowerEdge R740xd	R3 - Chassis02: Dell EMC PowerEdge R740xd
2 1	R1 - Chassis01: Dell EMC PowerEdge R740xd	R2 - Chassis01: Dell EMC PowerEdge R740xd	R3 - Chassis01: Dell EMC PowerEdge R740xd

PowerEdge R740xd Worker Nodes additional pod rack configuration

Table 17 Additional pod rack configuration: Dell EMC PowerEdge R740xd Worker Nodes

RU	RACK1	RACK2	RACK3
42	Empty	R2 - Switch 1: Dell EMC PowerSwitch S5248-ON	Empty
41	Empty	Empty	Empty
40	Cable Management	Cable management	Cable management
39	Cable Management	Cable management	Cable management
38	R1 - Dell EMC PowerSwitch S3048-ON iDRAC Management switch	R2 - Dell EMC PowerSwitch S3048-ON iDRAC Management switch	R3 - Dell EMC PowerSwitch S3048-ON iDRAC Management switch
37	Cable management	Cable management	Cable management
36	Cable management	Cable management	Cable management
35	Empty	Empty	Empty

Table 17 Additional pod rack configuration: Dell EMC PowerEdge R740xd Worker Nodes (continued)

RU	RACK1	RACK2	RACK3
25			
24 23	R1 - Chassis12: Dell EMC PowerEdge R740xd	R2 - Chassis12: Dell EMC PowerEdge R740xd	R1 - Chassis12: Dell EMC PowerEdge R740xd
22 21	R1 - Chassis11: Dell EMC PowerEdge R740xd	R2 - Chassis11: Dell EMC PowerEdge R740xd	R3 - Chassis11: Dell EMC PowerEdge R740xd
20 19	R1 - Chassis10: Dell EMC PowerEdge R740xd	R2 - Chassis10: Dell EMC PowerEdge R740xd	R3 - Chassis10: Dell EMC PowerEdge R740xd
18 17	R1 - Chassis09: Dell EMC PowerEdge R740xd	R2 - Chassis09: Dell EMC PowerEdge R740xd	R3 - Chassis09: Dell EMC PowerEdge R740xd
16 15	R1 - Chassis08: Dell EMC PowerEdge R740xd	R2 - Chassis08: Dell EMC PowerEdge R740xd	R3 - Chassis08: Dell EMC PowerEdge R740xd
14 13	R1 - Chassis07: Dell EMC PowerEdge R740xd	R2 - Chassis07: Dell EMC PowerEdge R740xd	R3 - Chassis07: Dell EMC PowerEdge R740xd
12 11	R1 - Chassis06: Dell EMC PowerEdge R740xd	R2 - Chassis06: Dell EMC PowerEdge R740xd	R3 - Chassis06: Dell EMC PowerEdge R740xd
10 9	R1 - Chassis05: Dell EMC PowerEdge R740xd	R2 - Chassis05: Dell EMC PowerEdge R740xd	R3 - Chassis05: Dell EMC PowerEdge R740xd
8 7	R1 - Chassis04: Dell EMC PowerEdge R740xd	R2 - Chassis04: Dell EMC PowerEdge R740xd	R3 - Chassis04: Dell EMC PowerEdge R740xd
6 5	R1 - Chassis03: Dell EMC PowerEdge R740xd	R2 - Chassis03: Dell EMC PowerEdge R740xd	R3 - Chassis03: Dell EMC PowerEdge R740xd
4 3	R1 - Chassis02: Dell EMC PowerEdge R740xd	R2 - Chassis02: Dell EMC PowerEdge R740xd	R3 - Chassis02: Dell EMC PowerEdge R740xd
2 1	R1 - Chassis01: Dell EMC PowerEdge R740xd	R2 - Chassis01: Dell EMC PowerEdge R740xd	R3 - Chassis01: Dell EMC PowerEdge R740xd

Dell EMC PowerEdge R740xd Worker Nodes Physical Rack Configuration

APPENDIX B

Tested Component Versions

This appendix describes the versions of software and firmware that are used during validation of this Ready Architecture:

Software versions

Table 18 Software versions

Component	Version
Operating system	Red Hat Enterprise Linux 7.5
Hortonworks Data Platform	3.0.1.0-187
HDFS	3.1.1
YARN	3.1.1
MapReduce2	3.1.1
Tez	0.9.1
Hive	3.1.0
Zookeeper	3.4.6
Spark2	2.3.1

Network switch firmware versions

Table 19 Network switch firmware versions

Component	Version
Dell EMC S5248-ON	9.12(1.0)
Dell EMC S3048-ON	9.11(2.4)

Dell EMC PowerEdge R640 firmware versions

Table 20 Dell EMC PowerEdge R640 firmware versions

Component	Version
BIOS	1.6.12
iDRAC with LC	3.21.26.22
Mellanox ConnectX-4 LX 25 GbE SFP Rack NDC	14.21.30.12
Driver for operating system deployment	18.10.17
Dell 12 Gb expander firmware	2.25
PERC H740P	50.5.0.1750
CPLD	1.0.2

Dell EMC PowerEdge R740xd firmware versions

Table 21 Dell EMC PowerEdge R740xd firmware versions

Component	Version
BIOS	1.6.12
iDRAC with LC	3.21.26.22
Mellanox ConnectX-4 LX 25 GbE SFP Rack NDC	14.21.30.12
Driver for operating system deployment	18.10.17
Dell 12 Gb expander firmware	2.25
HBA 330 mini	16.17.00.03
CPLD	1.0.6
Nonexpander storage backplane	4.26

GLOSSARY

A

API Application Programming Interface

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

B

BIOS Basic input/output system

BMC Baseboard Management Controller

BMP Bare Metal Provisioning

C

CDH The Cloudera distribution including Apache Hadoop

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

CMC Chassis Management Controller

CRM Customer Relationship Management

D

DBMS Database Management System

DGI Data Governance Initiative

DTK Dell EMC OpenManage Deployment Toolkit

E

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

ERP Enterprise Resource Planning

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.

F

FQDN A Fully Qualified Domain Name (FQDN) is the portion of an Internet Uniform Resource Locator that fully identifies the server to which an Internet request is addressed. The FQDN includes the second-level domain name, such as `dell.com`, and any other levels as required.

H

HBA Host Bus Adapter

HDF Hortonworks DataFlow

HDFS Hadoop Distributed File System

HDP Hortonworks Data Platform

HVE Hadoop Virtualization Extensions

I

iDRAC Integrated Dell Remote Access with Lifecycle Controller

IPMI Intelligent Platform Management Interface

J

JBOD Just a Bunch of Disks

JDBC Java Database Connectivity

JDK Java Development Kit

K

KPI Key Performance Indicator

L

LACP Link Aggregation Control Protocol

LAG Link Aggregation Group

LOM Local Area Network on Motherboard

M

MTU A maximum transmission unit is the largest size packet or frame, in octets, that can be sent over a packet/frame-based computer network.

N

NIC Network Interface Card

NTP Network Time Protocol

NVM Node Version Manager

O

OS Operating system

OS-HCTK A configuration utility with sample scripts and configuration files that is used to automate the setup and configuration of BIOS and RAID settings for Dell EMC servers in OpenStack and Hadoop open source software solutions.

P

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

Q

QSFP Quad Small Form-factor Pluggable

R

RAID Redundant Array of Independent Disks

REST Representational State Transfer

RPM Red Hat Package Manager

RSTP Rapid Spanning Tree Protocol

RTO Recovery Time Objective

RU A Rack Unit measures 1.75 inches, or 44.45 mm, in a 19-inch or 23-inch electronic equipment rack frame.

S

SIEM Security Information and Event Management

SLA Service Level Agreement

SSD Solid-state Drive (or Solid-state Disk)

T

THP Transparent Huge Pages

ToR Top-of-Rack Switch/Router

U

UID A code identifying each user on a Unix or Unix-like computer system

V

VLT Virtual Link Trunking

VRRP Virtual Router Redundancy Protocol

Y

YARN Yet Another Resource Negotiator

INDEX