

Extract Insights on a Scalable and Security-Enabled Data Platform from Cloudera

In collaboration with:



Tech Note by

Todd Mottershead

Todd.mottershead@dell.com

Seamus Jones

Seamus.jones@dell.com

Amandeep Raina

Amandeep.Raina@intel.com

Krzysztof Cieplucha

krzysztof.cieplucha@intel.com

Summary

This joint paper outlines a brief discussion on the key hardware considerations when configuring a successful Cloudera deployment and recommends configurations based on the most recent 15th Generation PowerEdge Server portfolio offerings.

Enterprises today are modernizing their IT infrastructures to store, distribute, and analyze enormous amounts of data. Many are migrating to container-based environments to take advantage of their inherent agility and scalability. But some data must be kept on premises, especially in industries like healthcare and finance.

Cloudera® Data Platform (CDP) Private Cloud is a scalable data platform that allows data to be managed across its lifecycle—from ingestion to analysis—without leaving the data center. It comprises two products: Cloudera Private Cloud Base (the on-premises portion built on Dell EMC™ PowerEdge™ servers) and Cloudera Private Cloud Experiences. By disaggregating compute and storage, CDP Private Cloud allows each product to be scaled independently according to an organization's needs. And for organizations with containerized applications deployed on Kubernetes®, CDP Private Cloud brings both agility and predictable performance to analytics applications.

Key Considerations

- **Data throughput.** CDP Private Cloud on Dell EMC PowerEdge servers is built on high-performing Intel® architecture. Intel Ethernet network controllers, adapters and accessories enable agility in the data center and support high throughput.¹ Unlike many other point solutions, CDP Private Cloud is an end-to-end platform for data, from collecting and engineering to reporting and using AI capabilities.
- **Balanced system configuration.** CDP Private Cloud can handle multiple varying workloads, including analytics and machine learning (ML). Its capabilities are supported by generation-over-generation improvements in underlying Intel technologies that offer more cores and higher memory capacity.
- **Data latency.** As data grows and needs to be accessed across the cluster, data-access response times are critical, especially for real-time analytics applications. CDP Private Cloud takes a tiered memory and storage approach using DRAM, Intel® Optane™ persistent memory (PMem), and Intel Optane Solid State Drives (SSDs) for low-latency response times.

Available Configurations

Cloudera® Data Platform (CDP) Private Cloud Base Cluster				
	Edge Node (1 Node) + Master Nodes (Minimum of Three Nodes Required)	Worker Nodes for Use with External Storage System (Minimum of Three Nodes Required)	Worker Nodes with Local All-Flash Storage (Minimum of Three Nodes Required)	Worker Nodes with Local HDDs (Minimum of Three Nodes Required)
Functions	Edge node: Apache Hadoop® clients, Cloudera Manager Master nodes: NameNode, Resource Manager, Apache ZooKeeper™	DataNode, NodeManager, CDP DC (YARN) workloads		
Platform	Dell EMC™ PowerEdge™ R650 server supporting 10 x 2.5" Serial-Attached SCSI (SAS)/Serial ATA (SATA)/NVM Express® (NVMe®) drives	Dell R650 supporting 10 x 2.5" SAS/SATA/NVMe drives	Dell R750 chassis with up to 24 x 2.5" drives (16 x SAS/SATA + 8 x NVMe)	Dell R750 chassis with up to 12 x 3.5" drives and 2 x 2.5" rear storage (NVMe)
CPU	2 x Intel® Xeon® Gold 6326 processor (16 cores at 2.9 GHz) or better	2 x Intel® Xeon® Gold 6348 processor (28 cores at 2.6 GHz)		
DRAM	256 GB (16 x 16 GB DDR4-3200)	Standard configuration: 512 GB (16 x 32 GB DDR4-3200) Large memory configuration: 256 GB (16 x 16 GB DDR4-3200)		
Persistent Memory ⁱⁱ	None	Standard configuration: <i>None</i> Large memory configuration: 1 TB (8 x 128 GB Intel® Optane™ PMem 200 series)		
Boot device	From Dell™ PowerEdge RAID Controller (PERC)	Dell EMC™ Boot Optimized Server Storage (BOSS)-S2 with 2 x 480 GB M.2 SATA (RAID1)		
Storage adapter	Dell PERC H755	Dell PERC HBA355i		
Storage (NVMe)	1 x 1.6 TB Intel SSD P5600 (PCIe Gen4, mixed use)	1 x 3.2 TB Intel SSD P5600 (PCIe Gen4, mixed use, 3 DWPD)		
Storage (SATA SSD)	7 x 3.84TB Intel SSD S4510 SATA	Not required (use external storage system instead).	12 x (up to 16 x) 3.84TB Intel SSD S4510 (SATA, read intensive, 1 DWPD)	SAS/SATA: 12 x (up to 16 x) 4 TB 7.2K RPM NLSAS 12 Gbps 512n 3.5" hot-plug HDD
Storage (Ozone metadata), optional ^{iii, iv}	1 x 400GB Intel Optane P5800X (PCIe Gen4) or 375GB Intel Optane SSD 4800X	None	None	None

The information in this publication is provided as is. Dell Inc. makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Use, copying, and distribution of any software described in this publication requires an applicable software license.

Copyright © 2021 Dell Inc. or its subsidiaries. All Rights Reserved. Dell, EMC, PowerEdge and other trademarks are trademarks of Dell Inc. or its subsidiaries. Other trademarks may be the property of their respective owners.

Dell Inc. believes the information in this document is accurate as of its publication date. The information is subject to change without notice.

Network interface controller (NIC)	Intel® Ethernet Network Adapter E810-XXVDA2 for OCP3 (dual-port 25-gigabit Ethernet [GbE])			
Additional NIC ^v	None	Intel E810-XXVDA2 PCIe adapter (dual-port 25GbE)	None	None

Note: For storage-only configuration (HDFS/Ozone), customers can still choose traditional high-density storage nodes with high-capacity rotational HDDs based on the PowerEdge R740xd2 platform, although external storage systems like Dell EMC PowerScale or ECS are recommended. Customers should be aware that using large capacity HDDs increases the time of background scans (bit-rot detection) and block report generation for HDFS and significantly increases recovery time after full node failure. Also, using nodes with more than 100 TB of storage is not recommended by Cloudera. Source: <https://blog.cloudera.com/disk-and-datanode-size-in-hdfs/>. For more information and specifications, please contact a Dell representative.

CDP Private Cloud Experiences (Red Hat® OpenShift® Kubernetes®) Cluster			
	Container Services Administration Host	Master Nodes (Three Nodes Required)	Worker Nodes (10 Nodes or More)
Functions	OpenShift administration services	OpenShift services, Kubernetes services	Kubernetes operators, Cloudera® Data Platform (CDP) Private Cloud workload pods
Platform	Dell EMC™ PowerEdge™ R650 server supporting 10 x 2.5" SAS/SATA/NVMe® drives		Dell R650 supporting 10 NVMe drives (direct connection/no PERC)
CPU	2 x Intel® Xeon® Gold 6326 processor (16 cores at 2.9 GHz) or better		2 x Intel® Xeon® Gold 6348 processor (28 cores at 2.6 GHz) or better
DRAM	128 GB (16 x 8 GB DDR4-3200)		Standard configuration: 512 GB (16 x 32 GB DDR4-3200) Large memory configuration: 256 GB (16 x 16 GB DDR4-3200)
Persistent memory ^{vi}	None	None	Standard configuration: None Large memory configuration: 2 TB (16 x 128 GB Intel® Optane™ PMem 200 series)
Boot device	Dell EMC™ BOSS-S2 with 2 x 480 GB Intel® SSD S4510 M.2 SATA (RAID1)		
Storage adapter	Not required for all-NVMe configuration.		
Storage (NVMe)	1 x 1.6 TB Intel SSD P5600 NVMe	1 x 3.2 TB Intel SSD P5600 NVMe	1 x 7.68 TB Intel SSD P5500 NVMe
NIC	Intel Ethernet Network Adapter E810-XXVDA2 for OCP3 (dual-port 25 GbE)		
Additional NIC	Intel Ethernet Network Adapter E810-XXV (dual-port 25 GbE)		

Learn More

Contact your [Dell](#) or Intel account team for a customized quote [1-877-289+-3355](tel:1-877-289-3355)

Read the blog: [Maximizing performance of Apache Kudu block cache with Intel Optane DCPMM](#)

ⁱ For workloads requiring particularly high network bandwidth, customers might use an Intel Ethernet Network Adapter E810-CQDA2 with PCIe (dual-port 100 GbE) and 100 GbE top-of-rack (ToR) switches.

ⁱⁱ Larger Intel Optane PMem modules (256 GB or 512 GB) can be used for up to 8 TB/nodes. When using persistent memory in Memory Mode, the DRAM to PMem ratio needs to be between 1:4 and 1:16.

ⁱⁱⁱ Not required for the Edge Node or if not using Ozone.

^{iv} Intel Optane P5800X is recommended, but previous-generation P4800X can be used instead if P5800X is not yet available.

^v Additional NIC is recommended for connectivity to an external storage system via a dedicated storage network.

^{vi} [repeat endnote 2] Larger Intel Optane PMem modules (256 GB or 512 GB) can be used for up to 8 TB/nodes. When using persistent memory in Memory Mode, the DRAM to PMem ratio needs to be between 1:4 and 1:16.