Cloudera Data Platform (CDP) Private Cloud Base on IBM Power10 and IBM Elastic Storage System

Best practices for deployment

Table of contents

Introduction

Cloudera Data Platform (CDP) Private Cloud Base is the on-premises version of Cloudera Data Platform deployed on IBM® Power10 processor-based virtualized servers.

New with this Power10 blueprint is the option to use a **traditional data cluster** with S1024 servers and internal NVMe storage as well as the **elastic deployment topology** using S1022 and Elastic Storage System.

CDP Private Cloud Base consists of a variety of components such as Apache Hadoop Distributed File System (HDFS), Apache Hive, Apache HBase, and Apache Spark, along with many other components to establish an intelligent data fabric.

You can select any combination of these services to create clusters that address your business requirements. Although Cloudera is available in the public cloud, many enterprise customers with large volumes of data prefer on-premises and hybrid solutions.

Challenge

Traditional Big Data solutions are implemented in data clusters.

This limits elasticity and results in a large number of assets with **low utilization** to host three copies of data.

Solution

Decouple data from compute to create a **modular and practitioner focused** architecture.

Planning your deployment

The primary step in planning your deployment is to understand user requirements. It is important to know which CDP Private Cloud Base capabilities must be used, who will use them, and what size of data will be used. In addition, this information along with the users' performance expectations help in determining the size and configuration of each server, and the number of servers that are needed for an initial CDP Private Cloud Base deployment.

Cloudera provides a deployment guide for CDP Private Cloud Base on Linux®. The instructions in this white paper help during the planning phase to understand the CDP

Software

- CDP Private Cloud Base 7.1.8

Hardware

- IBM Power S1022 and Power S1024 servers
- IBM Elastic Storage System (ESS) 3500

Network

- 100 GbE data network
- 1 Gb admin network
- 10/25 Gb access/data ingest network

Private Cloud Base deployment modes and other software and environmental requirements.

You can find the latest guide at: https://docs.cloudera.com/cdp-private-cloud-base/latest/index.html

Beyond this, IBM Power® offers several additional flexible deployment options for CDP Private Cloud Base.

The Cloudera Support Matrix: https://supportmatrix.cloudera.com/ details the CDP versions and dependencies compatible with the IBM Power Processor based offerings.

While the deployment options provide examples to select server configuration for the CDP Private Cloud Base workload, clients can contact their IBM and Cloudera account team for assistance in sizing their workload.

This ensures that the hardware environment is properly sized for the client-specific CDP Private Cloud Base workload.

IBM Elastic Storage System deployment options

Moving from a static data cluster to a data services architecture requires **separated storage and compute resources**. Disaggregated compute or storage with IBM Elastic Storage® System (ESS) reduces the traditional HDFS 3-way data replication overhead of HFDS data by up to 85%[1] using IBM Spectrum® Scale Native RAID.

[1] HDFS, with a default replication factor of 3, have a 200% capacity overhead. ESS Native Raid (8+2) has typical 30% capacity overhead. This means that 85% less capacity overhead is required with the IBM Power and ESS solution.

A typical ESS system provides data throughput exceeding local NVMe storage implementations.

Shuffle and sort performance is also typically higher on spectrum scale file systems remotely accessed from an ESS, compared to local solid-state drive (SSD).

In addition, the ESS provides multi-protocol data access. This results in greatly **simplified data ingest** scenarios where data might get ingested directly over NFS or POSIX without being copied into HDFS from a local file system.

ESS 3500

Adds 12% increased throughput compared to earlier blueprint with ESS3200.

Provides the ability to mix NVMe and serial-attached SCSI (SAS) storage modules for an optimum balance of performance and capacity.

Using AFM/DR, the ESS system can copy data to a **secondary location**. This greatly simplifies DR scenarios and avoids installing additional software components in your elastic data topology cluster for the same.

Therefore, the IBM Power and ESS elastic deployment topology solution provides both **performance advantages** as well as potential **cost savings** from the reduced raw data requirements and simplified DR scenarios.

The most suitable deployment model of Power and Spectrum Scale for CDP is following the HDFS Transparency DataNode colocation architecture. Cluster Export Services (CES) nodes provide HDFS NameNode services and the transparency connector between HDFS and Spectrum Scale is installed on all nodes requiring HDFS access.

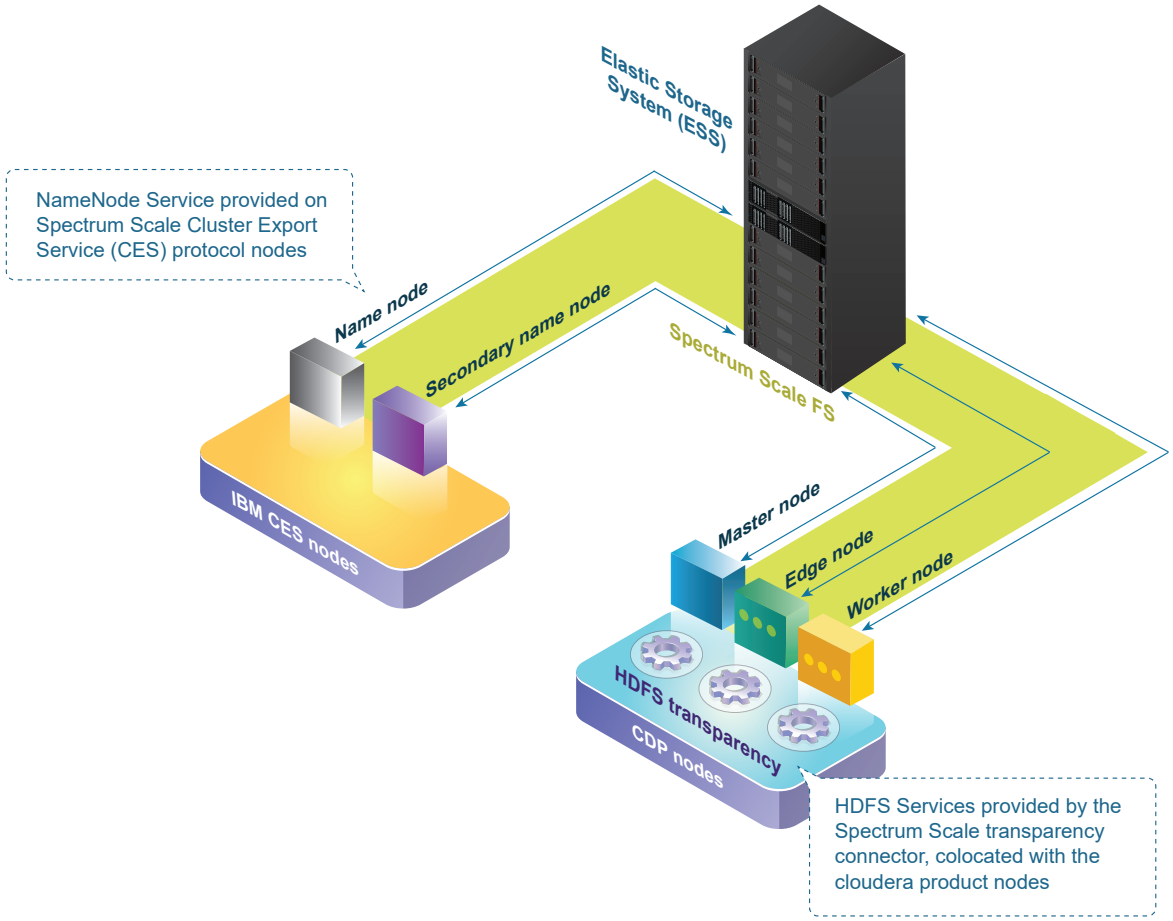

Figure 1. DataNode colocation architecture with CDP on Power10 with ESS

IBM Power10 deployment options

The IBM Power portfolio of servers enables flexible deployment options for running CDP Private Cloud Base. IBM recommends the IBM Power® S1022 and S1024 servers for CDP Private Cloud Base deployment.

The IBM Power servers provides performance, virtualization, reliability, availability and delivers twice the throughput of Intel® processor-based offerings and is highly economical for elastic data services deployments.

IBM PowerVM® allows for virtualizing the IBM Power Systems server without performance penalties traditionally associated with software-based hypervisors, and this is due to the enablement of single root I/O virtualization (SR-IOV) and dedicated I/O -virtualization options.

Naming

Cloudera component naming is aligned with Cloudera reference architecture: CDP Private Cloud Base Reference Architecture

Spectrum Scale HDFS consideration

The NameNode and Secondary NameNode functionality is provided by the Spectrum Scale Cluster Export Services (CES) and are not deployed on the CDP Master nodes. (Reference Figure 1).

Cloudera recommends deploying up to four machine types into a production environment:

Master nodes - Master nodes host most of the management functions and some of the storage functions. Hadoop master daemons such as ZooKeeper, HBase Master, JobHistory Server, and Spark History Server are hosted on the master nodes.

There should be a **minimum of three master nodes in a production environment**. Three master nodes are required to provide basic high availability (HA) capability. As the number of worker nodes increase, the number of master nodes typically increases to provide the additional capacity to manage the larger number of worker nodes. The published role assignment provides some guidance from Cloudera on appropriate master nodes that typically have somewhat lower hardware demands than worker nodes.

Master nodes can be configured on the same hardware as the worker nodes in the cluster and allow the servers for each node type to be interchangeable. Processor, memory, and storage capacity can be individually assigned to each type based on actual requirements.

Worker nodes - A worker node in an IBM Power server with ESS deployment mainly contain the functions to run applications that are parts of jobs. Job execution is typically distributed across multiple worker nodes to provide parallel execution. There are typically three or more (often many more) worker nodes in a cluster.

Worker nodes are usually the most common node type in a cluster, accounting for perhaps 80% (or more) of the nodes. DataNode, YARN NodeManager, and HBase RegionServer are hosted on worker nodes.

Overall cluster performance and behavior is strongly influenced by the design of the worker nodes. Thus, the design and configuration of the worker nodes should be considered early in the design process, with significant attention to the requirements of the deployment. Worker nodes are frequently optimized for performance of the network and storage functions and for performance when running applications. This leads to the following recommendations:

- High speed dedicated or SR-IOV virtualized 100 GbE data network connections
- Larger memory capacity (256 GB or more per node is common)
- Persistent memory for the operating system and applications, deployed on NVMe rather than spinning disks or SATA SSDs
- Virtual CPU capacity provided that typically corresponds to 50% of a Power server for each of two or four virtual machines (VMs) per server

The CDP Private Cloud Base and the HDFS architecture tolerate significant failures within the collection of worker nodes. Thus, worker node components can typically be chosen, which optimize performance and capacity versus resilience.

Utility nodes - Cloudera Manager and the Cloudera Management Services are deployed on utility nodes. It can also host a PostgreSQL (or another supported) database instance, which is used by Cloudera Manager, Hive, Ranger and other Hadoop-related components. Care needs to be given to the design in order to protect this critical data in case of any failure.

Edge nodes – An edge node provides a control point for user access, and it provides a dedicated capacity to handle data import and export. Edge nodes contain all client-facing configurations and services, including gateway configurations for HDFS, YARN, Hive, and HBase. The edge node is also a good place for Hue, Oozie, HiveServer2. HiveServer2 serves as a gateway to external applications, such as the business intelligence (BI) tools. *Edge nodes are also known as* ***Gateway nodes***.

The IBM Power and ESS solution requires two or more additional types of nodes to be deployed for managing the solution. They include:

System management node - The system management node is a server hosting the software that accomplishes the provisioning and management of the infrastructure. The system management node is not visible or used by the CDP Private Cloud Base cluster. It is used exclusively for infrastructure and cluster-level purposes.

Cluster Export Services (CES) nodes - Cluster Export Services (CES) provides highly available file and object services to a Cloudera cluster by using Network File System (NFS), Hadoop Distributed File System (HDFS), Object, and Server Message Block (SMB) protocols. *A minimum of two CES HDFS nodes are required for production environments and minimum of at least one for non-production or PoC setups*.

Active File Management nodes (DR, optional) - Active File Management (AFM) enables

sharing of data across clusters, even if the networks are unreliable or have high latency. By using AFM, you can build a common namespace across locations, and automate the flow of file data. You can duplicate data for disaster recovery purposes without suffering from WAN latencies.

Elastic deployment topology (more than 100 TB)

To gain the benefits of the IBM Power and ESS technology stack, an elastic deployment topology is recommended. Figure 2 shows a cluster deployed across three nodes. Each node is connected to two switches, two 100 Gbps connections for data network and one 1 Gbps for management network. 10 Gbps data ingestion network (not shown) is attached to an existing customer network.

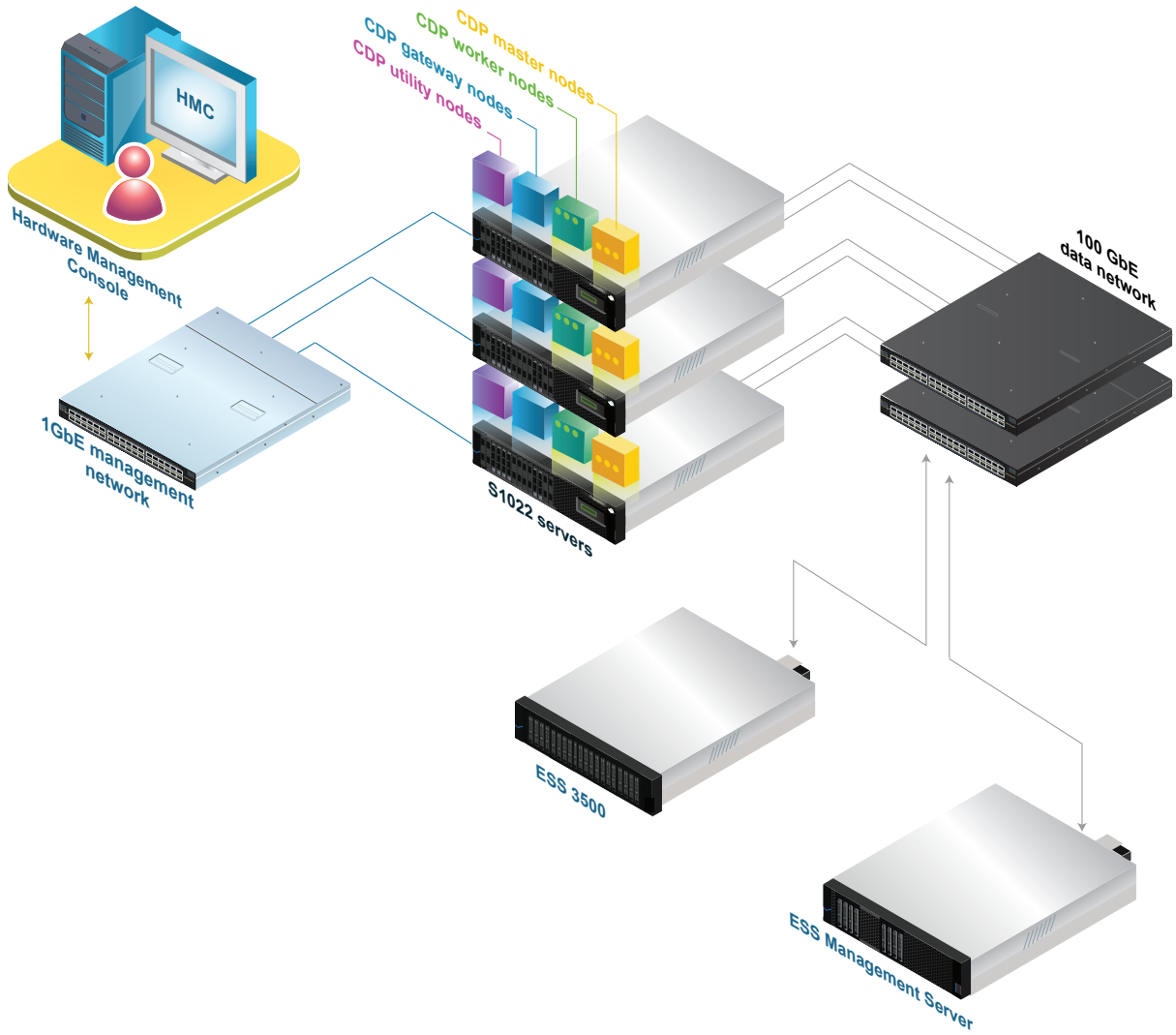

Figure 2. Elastic and scalable CDP deployment architecture using S1022 and ESS

In a production environment, it is highly recommended to use four physical links for data and six more physical links for management (when running four VMs per physical Power S1022 server). This provides a redundant path that is used to provide resilience for these networks.

Elastic S1022 and ESS configuration

This section describes a reference design for an elastic deployment topology solution. It is an example of a system design that complies with the architecture explained in the earlier sections.

This reference design is intended as a reference only. Any specific design, with appropriately sized components that are suitable for a specific deployment, requires additional review and sizing that is appropriate for the intended use.

	Hardware Management Console	1GbE Switch	100GbE Swtich	S1022	Elastic Storage System	ESS Mgmt Server
Server model	7063-CR2	8831-T48	8831-00M	9105-22A	5141-FN2	5105-22E
Small cluster	1	2	2 - 8	3 - 20	1-2	1
Medium cluster	2	4	8 - 16	20 - 40	2-4	2
Sockets	1			2		1
CPU Cores	6			24-40		20
Memory	64 GB			512 - 4096 GB	46 - 737 raw TB	128 GB
Storage	2x 1.8TB SAS			8x 1.6TB NVMe	12-24x NVMe	2x 1.8TB SAS
Network* - 1 GbE	Internal (4 ports)	48 ports	2 ports	6 (2 HMC+ 4 (4xEC2T+EB48))	Internal (2 ports)	Internal 2 ports+AJZQ 4 port
Cables* - 1 GbE	3 (2 HMC + 1 OS)	48	2 BMC	6 (2 HMC+ 4 OS)	2 BMC	4 (2 BMC + 2 OS)
Network - 10 GbE**				4 (2xEC2T+EB46)		
Cables - 10 GbE**				4 SR Fibers to customer DC		
Network* - 100 GbE**			32	2x EC75 2-ports (4 ports)	2xAJP1 2-ports (4 ports)	1x AJP1 2-ports (2 ports)
Cables* - 100 GbE**			16 cables (DACs)	4 cables (DACs)	4 cables (DACs)	2 cables (DACs)
Operating System	HMC Firmware			RHEL 8.6 ppc64le	Embedded RHEL for ESS	Embedded RHEL for ESS

* The 1 GbE network infrastructure hosts the following logical networks: campus, management, provisioning and service networks
** The 10 GbE network infrastructure hosts the data ingestion network.
*** The 100 GbE network infrastructure hosts the data network.

Figure 3. Hardware and OS configuration for a small to medium production cluster

	Hardware Management Console	1GbE Switch	100GbE Swtich	S1022	Elastic Storage System	ESS Mgmt Server
Server model	7063-CR2	8831-T48	8831-00M	9105-22A	5141-FN2	5105-22E
PoC Cluster	1	1	1	2-4	1	1
Sockets	1			2		1
CPU Cores	6			24 - 40		20
Memory	64 GB			512 GB	46 - 737 raw TB	128 GB
Storage	2x 1.8TB SAS			8x 800GB NVMe	12-24x NVMe	2x 1.8TB SAS
Network* - 1 GbE	Internal (4 ports)	48 ports	2 ports	6 (2 HMC+ 4 (4xEC2T+EB48))	Internal (2 ports)	Internal 2 ports+AJZQ 4 port
Cables* - 1 GbE	3 (2 HMC + 1 OS)	48	2 BMC	6 (2 HMC+ 4 OS)	2 BMC	4 (2 BMC + 2 OS)
Network - 10 GbE**				4 (2xEC2T+EB46)		
Cables - 10 GbE**				4 SR Fibers to customer DC		
Network* - 100 GbE**			32	2x EC75 2-ports (4 ports)	2x AJP1 2-ports (4 ports)	1x AJP1 2-ports (2 ports)
Cables* - 100 GbE**			16 cables (DACs)	4 cables (DACs)	4 cables (DACs)	2 cables (DACs)
Operating System	HMC Firmware			RHEL 8.6 ppc64le	Embedded RHEL for ESS	Embedded RHEL for ESS

* The 1 GbE network infrastructure hosts the following logical networks: campus, management, provisioning and service networks
** The 10 GbE network infrastructure hosts the data ingestion network.
*** The 100 GbE network infrastructure hosts the data network.

Figure 4. Hardware and OS configuration for a PoC cluster

	System management node	Master nodes	Utility nodes	Gateway nodes	Worker Nodes	Cluster Export Services*
Server model	S1022 VM	S1022 VM	S1022 VM	S1022 VM	S1022 VM	S1022 VM
Small cluster	1	3	1	1	3 - 20	2 x 128GB RAM
Medium cluster	1	3	2	2	20 - 80	2 x 256GB RAM
Large cluster	1	3	7	4	80 - 200	2 x 256GB RAM
Extra large cluster	1	5	7	4	200 - 1000	2 x 512GB RAM
Cores (vCPU)	2 - 10	5 - 12	5 - 12	5 - 12	5 - 20	8 - 12
Memory	32 GB	256 GB	256 GB	256 GB	256 GB	(see above)
Operating System	RHEL 8.6	RHEL 8.6	RHEL 8.6	RHEL 8.6	RHEL 8.6	RHEL 8.6

*Cluster Export Services only required in an Elastic cluster architecture

Figure 5. CDP runtime distribution and deployment architecture

For more details, visit Runtime Cluster Hosts and Role Assignments.

Power10 data cluster deployment topology (less than 100 TB)

A new option with the Power S1024 server is the traditional data cluster deployment topology. This avoids the high-speed network and ESS for a MVP or non-scalable scenarios. The Power S1024 server can accommodate 16x6.4 TB NVMe persistent storage modules. Each NVMe persistent storage module is **independently assigned to a VM**, which leaves a large number of high-capacity modules available for hosting worker nodes in the same physical server together with master nodes and gateway nodes.

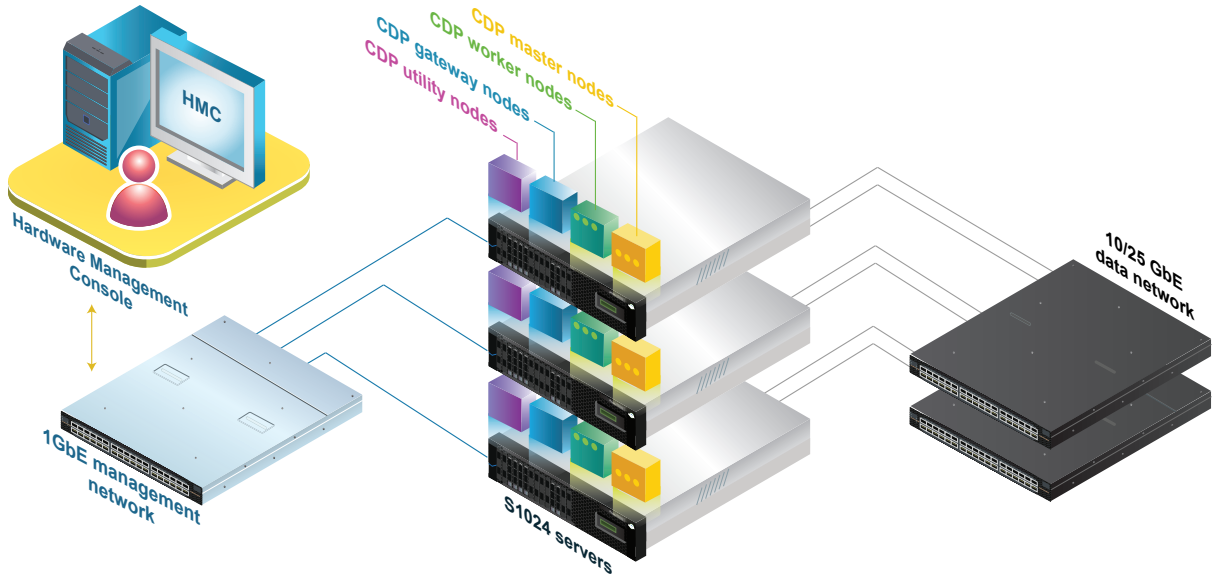

Figure 6. Data sluster deployment architecture using three S1024 servers

Starting with Power10 and the Power S1024 server, it becomes feasible to deploy a solution (having less than 100 TB usable capacity) according to Figure 6: Data Cluster Deployment Architecture.

Using three Power S1024 servers, each DataNode holds a full data replica. Therefore, inter-server communication is not expected to require high bandwidth. This eliminates the 100 GbE network requirement and the ability to deploy the data network on regular 10/25 Gb ports provided in the data center.

Data cluster production configuration

This section describes a reference design for a data cluster. A data cluster has a much lower initial acquisition price point, sacrificing the flexibility and modular approach of the elastic design. It is therefore ideal to showcase the value of the Cloudera components in production as well as non-production use cases below 100 TB data capacity, where there is no DR requirement. Going beyond four physical servers is not recommended.

	Hardware Management Console	1GbE Switch	S1024
Server model	7063-CR2	8831-T48	9105-42A
Small cluster	1	1	3-4
Sockets	1		2
CPU Cores	6		24-48
Memory	64 GB		512 - 8192 GB
Storage	2x 1.8TB SAS		16x 6.4TB NVMe
Network* - 1 GbE	Internal (4 ports)	48 ports	6 (2 HMC+ 4 (4xEC2U+EB48))
Cables* - 1 GbE	3 (2 HMC + 1 OS)	48	6 (2 HMC+ 4 OS)
Network - 10 GbE**			4 (2xEC2U+EB46)
Cables - 10 GbE**			4 SR Fibers to customer DC
Operating System	HMC Firmware		RHEL 8.6 ppc64le

* The 1 GbE network infrastructure hosts the following logical networks: management, provisioning and service networks.
** The 10 GbE network infrastructure hosts the data ingestion and cluster interconnect network.

Figure 7. Data cluster production configuration

Cluster sizing

Sizing a system for CDP Private Cloud Base is a significant and complex topic. Sizing is relevant in several dimensions. Factors such as throughput, response time, ingest rate, high availability, disaster recovery and so on may be relevant.

A complete treatment of the design and sizing process for CDP Private Cloud Base on IBM Power is beyond the scope of this paper.

Note that consulting is required to properly size a cluster for a client deployment. The Cloudera sizing team can be engaged to provide the IBM team a HDFS size based on an agreed erasure coding factor (and agreed free space amount) and peak sustained I/O rate to help estimate the cluster sizing.

However, guidance for one simple process (sizing for storage capacity) is provided in the following section for reference. Refer to the following process for a data capacity driven sizing on Power and ESS:

1. Gather client requirements.
    - Usable storage capacity needed (HDFS content)
    - Usage modes and cases expected
    - Number of users expected
    - Data ingest rate expected
    - Number of landscapes (Production/Non-production)
    - Networking preferences or policies
    - Availability requirements
2. Choose the ESS model and quantities based on IBM Storage Modeller or FOS DE analysis of expected workload and usable capacity required (RawStorageCapacity – see Formula 1).
3. Calculate the amount of storage per worker node based on expected workload density.
4. Calculate the total number of worker node VMs to support the data density (Formula 2).
5. Choose the number of master nodes as a function of the number of worker nodes. If necessary, increase the master node count to handle any additional demand expected on master nodes.
6. Choose the number of edge nodes as a function of the number of users and expected data ingest rate.
7. Choose the logical network topology, typically based on client networking preferences or policies.
8. Choose the network switches and network redundancy preferred for each network class.

9. Confirm or adjust the configuration for each node type – beginning with the reference configuration for each node type for the chosen cluster type.
10. Confirm or adjust the network link capacities and switch capacities as appropriate.
11. Confirm or adjust the node counts to meet availability requirements.
12. Confirm or adjust any selection based on growth expectation or initial headroom required.

```
RawStorageCapacity = ((UsableStorageCapacity * IntermediateDataUplift ) /
CompressionRatio ) + Freespace
```

Formula 1. Raw storage capacity calculation

```
NumberOfWorkerNodes = RawStorageCapacity / StoragePerWorkerNode
```

Formula 2. Number of worker nodes calculation

Multirack considerations

Configurations that require more than one rack introduce some additional factors that must be considered. Most of these considerations are the same as those that apply to other Multi Rack cluster deployments. These considerations include:

- Providing additional physical infrastructure for the additional nodes and switches (for example racks, power distribution unit, and so on)
- Scaling and designing the network appropriately for the total number of nodes (including cable lengths)
- Distributing master nodes and edge nodes across racks to improve availability

The first item is largely a matter of choosing the number of nodes per rack, choosing where to place the switches, and configuring sufficient power for the components in the rack.

The second item is beyond the scope of this paper, and network design consulting is recommended for any configuration that exceeds more than two racks.

Third is somewhat specific to CDP Private Cloud Base deployment, the master and edge nodes should be spread across the different racks.

Cluster configuration best practices

This section describes some of the best practices for deploying CDP Private Cloud Base.

Implementation

Before implementation ensure that you first read the entire section of the published implementation roadmap, because there are deviations from CDP Private Cloud Base installation documentation when IBM Spectrum Scale is integrated.

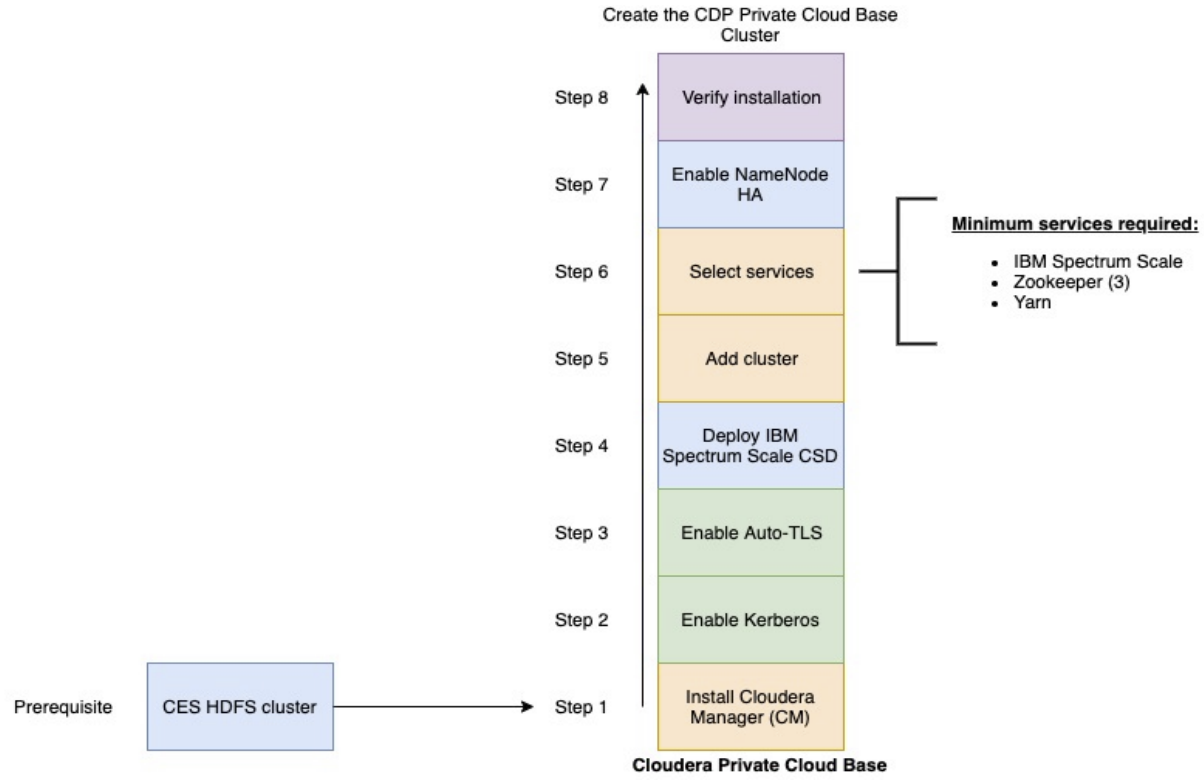

Figure 8. Overview of CDP Private Cloud Base Installation roadmap with Power and ESS

ZooKeeper

ZooKeeper is sensitive to disk latency. While it uses only a modest amount of resources, having ZooKeeper swap out or wait for a disk operation can result in that ZooKeeper node being considered *dead* by its quorum peers. For this reason, Cloudera does not recommend deploying ZooKeeper on worker nodes where loads are unpredictable and are prone to spikes. It is acceptable to deploy Zookeeper on master nodes where load is more uniform and predictable (or on any node where it can have unobstructed access to data).

HDFS

NameNode HA

Because the compute and storage architecture are decoupled, the server-side administration of NameNode HA is managed by the IBM Spectrum Scale CES protocol.

Please reference https://www.ibm.com/docs/en/spectrum-scale-bda?topic=configuring-enabling-namenode-ha

The following are the two required steps for the HA enablement process:

- **Server side**: NameNode HA can be enabled in the CES HDFS cluster during the installation and deployment using the IBM Spectrum Scale installation toolkit. if NameNode HA is not enabled on your CES HDFS cluster, follow Change CES HDFS NON-HA cluster into CES HDFS HA cluster to enable it.
- **Client side**: Enable the NameNode HA for the IBM Spectrum Scale service in the Cloudera Manager by enabling the NameNode HA for the CDP Private Cloud Base cluster.

When a NameNode failover event occurs in the IBM Spectrum Scale CES HDFS cluster, HDFS clients and Hadoop workloads running on the CDP Private Cloud Base cluster connect to the HA environment.

Block size

HDFS stores files in blocks that are distributed over the cluster. A block is typically stored contiguously on disks to provide high read throughput. The choice of block size influences how long these high- throughput read operations run for, and over how many nodes a file is distributed.

When reading many blocks of a single file, a too low block size spends more overall time in slow disk seek, and a large block size has reduced parallelism. Data processing that is I/O heavy benefits from larger block sizes, and data processing that is processor heavy benefits from smaller block sizes.

The default block size provided by Cloudera Manager is 128 MB. The block size can also be specified by an HDFS client on a per-file basis.

YARN

The YARN service manages Apache Hadoop MapReduce and Spark tasks. Applications run in YARN containers and use Linux c-groups for resource management and process isolation. The *Cloudera Installation and Upgrade* manual has a section on YARN tuning guidance.

Summary

Deploying CDP Private Cloud Base on Power10 and ESS allows for an elastic and modular architecture that enables your organization to build a data services architecture. It is an enterprise analytics and data management solution that requires a properly planned infrastructure to meet user requirements.

The underlying infrastructure matters to ensure that performance expectations and SLA requirements are met. Rely on Power10 and ESS configurations details shared in this technical white paper to ensure that an optimized infrastructure deployment is achieved.

Contact your IBM or Cloudera sales representatives for any questions or assistance with selecting the right IBM Power10 deployment and configuration for your needs.

Get more information

To learn more about CDP Private Cloud Base on IBM Power Systems, contact your IBM representative or IBM Business Partner.

About the author

Fredrik Lundholm is an IT Infrastructure Architect in the IBM Technology - Server Solutions Technical Sales team, covering the IBM server platforms in the IBM Global Markets organization. Fredrik has more than 15 years of experience in the IBM server and storage ecosystem. You can reach Fredrik at fredrikl@ae.ibm.com or https://www.linkedin.com/in/fredrik-lundholm-497a68

IBM Corporation New Orchard Road Armonk, NY 10504

Produced in the
United States of America January 2023