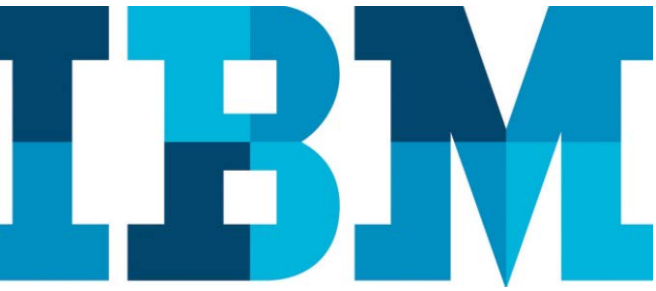
October 2020

Cloudera Data Platform (CDP) Private Cloud Base on IBM Power Systems

Best practices for deployment

Table of contents



Cloudera Data Platform (CDP) Private Cloud Base is the on-premises version of Cloudera Data Platform. CDP Private Cloud Base offers best of Cloudera Enterprise Data Hub (CDH) and Hortonworks Data Platform (HDP) Enterprise along with new features and enhancements across the stack. CDP Private Cloud Base consists of a variety of components such as Apache Hadoop Distributed File System (HDFS), Apache Hive 3, Apache HBase, and Apache Spark, along with many other components for specialized workloads. You can select any combination of these services to create clusters that address your business requirements and workloads.

This white paper helps in Cloudera Data Platform Private Cloud Base planning and deployment on IBM Power Systems.

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. These are the fundamentals to determine the optimal infrastructure environment to support the users. 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 would be helpful during the planning phase to understand the CDP 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/7.1.3/index.html

Beyond this, IBM® Power Systems™ offer several additional flexible deployment options for CDP Private Cloud Base. The next section outlines these options. While the deployment options provide examples to select server configuration for the CDP Private Cloud Base workload, clients can contact their IBM account team for assistance in sizing their workload from the sizing team. This ensures that the hardware environment is properly sized for the client-specific CDP Private Cloud Base workload.

IBM POWER9 deployment options

The IBM Power Systems portfolio of servers enables flexible deployment options for running CDP Private Cloud Base. The IBM portfolio offers ultra-flexible systems with the highest reliability*. IBM recommends the IBM Power® System IC922 server for CDP Private Cloud Base deployment. Power IC922 is one of the finest servers designed to provide a compute-intensive and low-latency infrastructure. The IBM POWER9™ processor-based Power IC922 server provides advanced interconnects (for example, Peripheral Component Interconnect Express (PCIe) Gen4 and OpenCAPI) to support faster data throughput and decreased latency. Customers can also make use of IBM PowerVM® based virtualized systems such as the IBM Power System S922 server for CDP Private Cloud Base deployment. Power S922 provides performance, virtualization, reliability, availability and delivers twice the throughput of Intel® processor-based offerings and superior economics for scale-out deployments.

CDP Private Cloud Base clusters are generally deployed for proof of concept (PoC) or production deployment. A cluster can have up to six types of machines based on the need:

* Global Server Hardware, Server OS Reliability Survey

Master nodes - Master nodes host most of the management functions and some of the storage functions. Hadoop master daemons such as NameNode, Secondary NameNode, 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 role assignment provides some guidance from Cloudera on appropriate master nodes typically have somewhat lower hardware demands than worker nodes. Master codes can be configured with the same hardware as the worker nodes if it is required to have one node configuration in the cluster and allow the servers for each node type to be interchangeable. However, processor, memory, and network configurations can be the same or somewhat less than what is configured for the worker nodes.

Worker nodes - A worker node serves two primary roles: First, each worker node contains some physical storage which is used for HDFS, and it hosts some storage functions that allow it to manage this storage as part of HDFS. These storage functions communicate and cooperate to form the distributed file system across the collection of worker nodes in the cluster. Second, a worker node is also used by the management functions to run applications that are parts of jobs. Job execution is typically distributed across multiple worker nodes to provide parallel execution of the job. There are typically three or more (often many more) worker nodes in a cluster. Three worker nodes provide the ability to directly support the common HDFS replication factor of three. Worker nodes are usually the most common node type in a cluster, accounting for perhaps 80% to 90% (or more) of the nodes in the cluster. HDFS, 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 storage functions and for performance when running applications. This commonly leads to the following recommendations:

- Higher CPU core counts and clock rates, often the maximum offered by the server model chosen
- Larger memory sizes (128 GB or more per node is common)
- Data nodes should have as many storage drives as possible for HDFS. Ten or more drives per node is common. To obtain the required storage capacity per node, it is preferred to use more drives with smaller capacity versus fewer drives with larger capacity
- High performance storage controllers are preferred, but significant Redundant Array of Independent Disks (RAID) capability is not required as the HDFS storage is typically configured as just a bunch of disks (JBOD)
- Significant network bandwidth to the data network (25 GbE per node or better is common)

Worker nodes generally need not be configured for high availability characteristics. 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. Every worker node is typically configured with the same hardware.

Utility nodes - Cloudera Manager and the Cloudera Management Services are deployed on utility nodes. It can also host a MariaDB (or another supported) database instance, which is used by Cloudera Manager, Hive, Ranger and other Hadoop-related projects.

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.

Machine learning/Deep learning nodes - One or more machine learning or deep learning nodes can be added to the cluster to support the running of machine learning or deep learning workloads or workloads that use GPU capabilities for acceleration.

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.

Minimum production configuration

This section describes a reference design for this solution. It is an example of a system design that complies with the architecture explained in the earlier section. This reference design is a *minimum production* configuration as it is designed and sized with a minimum set of elements that would be generally appropriate for consideration as a minimum starting point for a production deployment.

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.

	System management node	Master node	Utility node	Gateway node	Worker nodes
Server model	2U IC922	2U IC922	2U IC922	2U IC922	2U IC922
Small cluster	1	3	1	1	3 - 20
Medium cluster	1	3	2	1	20 - 80
Large cluster	1	3	8	1	80 - 200
Extra large cluster	1	5	8	2	200 - 1000
Sockets	1	2	2	2	2
Cores	12	40	40	40	40
Memory	32 GB	256 GB	256 GB	256 GB	256 GB
Storage backplane (front)	1	3	3	3	3
Storage - HDD (front)	4x 2.4TB HDD	4x 2.4TB HDD	8x 2.4TB HDD	8x 2.4TB HDD	22x 2.4TB HDD
Storage - SSD (front)					
OS Storage - HDD (front)		2x 2.4TB HDD	2x 2.4TB HDD	2x 2.4TB HDD	2x 2.4TB HDD
Storage controller	1x Broadcom 9300-8i	1x Broadcom MegaRAID 9361-8i 1X Broadcom 9305-16i	1x Broadcom MegaRAID 9361-8i 1X Broadcom 9305-16i	1x Broadcom MegaRAID 9361-8i 1X Broadcom 9305-16i	1x Broadcom MegaRAID 9361-8i 1X Broadcom 9305-16i
Network* - 1 GbE	Internal (2 ports OS)	Internal (2 ports OS)	Internal (2 ports OS)	Internal (2 ports OS)	Internal (2 ports OS)
Cables* - 1 GbE	3 (2 OS + 1 BMC)	3 (2 OS + 1 BMC)	3 (2 OS + 1 BMC)	3 (2 OS + 1 BMC)	3 (2 OS + 1 BMC)
Network - 25 GbE**	1x 2-port (2 ports)	1x 2-port (2 ports)	2x 2-port (4 ports)	2x 2-port (4 ports)	1x 2-port (2 ports)
Cables - 25 GbE**	2 cables (DACs)	2 cables (DACs)	4 cables (DACs)	4 cables (DACs)	2 cables (DACs)
Operating System	RHEL 7.6_Alt, 7.7 ppc64le	RHEL 7.6_Alt, 7.7 ppc64le	RHEL 7.6_Alt, 7.7 ppc64le	RHEL 7.6_Alt, 7.7 ppc64le	RHEL 7.6_Alt, 7.7 ppc64le

* The 1 GbE network infrastructure hosts the following logical networks: campus, management, provisioning and service networks
** The 25 GbE network infrastructure hosts the data network.

Figure 1. Hardware and OS configuration reference for production cluster

	System management node	Master node	Utility/Gateway node	Worker node
Server model	2U IC922	2U IC922	2U IC922	2U IC922
Number of servers	1	1	1	3-10
Sockets	1	2	2	2
Cores	12	40	40	40
Memory	32GB	256GB	256GB	256GB
Storage backplane (front)	1	3	3	3
Storage - HDD (front)	4x 2.4TB HDD	4x 2.4TB HDD	8x 2.4TB HDD	8x 2.4TB HDD
Storage - SSD (front)				
OS storage - HDD (front)				2x 2.4TB HDD
Storage controller	1x Broadcom 9300-8i	1x Broadcom 9305-16i 1x Broadcom 9300-8i	1x Broadcom 9305-16i 1x Broadcom 9300-8i	1x Broadcom 9305-16i 1x Broadcom 9300-8i
Network* - 1 GbE	Internal (2 ports OS)	Internal (2 ports OS)	Internal (2 ports OS)	Internal (2 ports OS)
Cables* - 1 GbE	3 (2 OS + 1 BMC)	3 (2 OS + 1 BMC)	3 (2 OS + 1 BMC)	3 (2 OS + 1 BMC)
Network - 25 GbE**	1x 2-port (2 ports)	1x 2-port (2 ports)	2x 2-port (4 ports)	1x 2-port (2 ports)
Cables - 25 GbE**	2 cables (DACs)	2 cables (DACs)	4 cables (DACs)	2 cables (DACs)
Operating System	RHEL 7.6_Alt, 7.7 ppc64le	RHEL 7.6_Alt, 7.7 ppc64le	RHEL 7.6_Alt, 7.7 ppc64le	RHEL 7.6_Alt, 7.7 ppc64le

* The 1 GbE network infrastructure hosts the following logical networks: campus, management, provisioning and service networks.
** The 25 GbE network infrastructure hosts the data network.

Figure 2. Hardware and OS configuration reference for PoC cluster

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

Deployment topology

Figure 3 shows a simple cluster deployed across several nodes. Each node is connected to two switches, one 25 Gbps connection for data network and one 1 Gbps for management network. In a production environment, it is highly recommended to use two physical links for data and two more physical links for management. This provides a redundant path that is used to provide resilience for these networks. The connection between each server and the switches for the data network can be configured for link aggregation using Link Aggregation Control Protocol (LACP) on the server and on the switch. This can offer increased bandwidth (up to 50 Gbps) between the nodes.

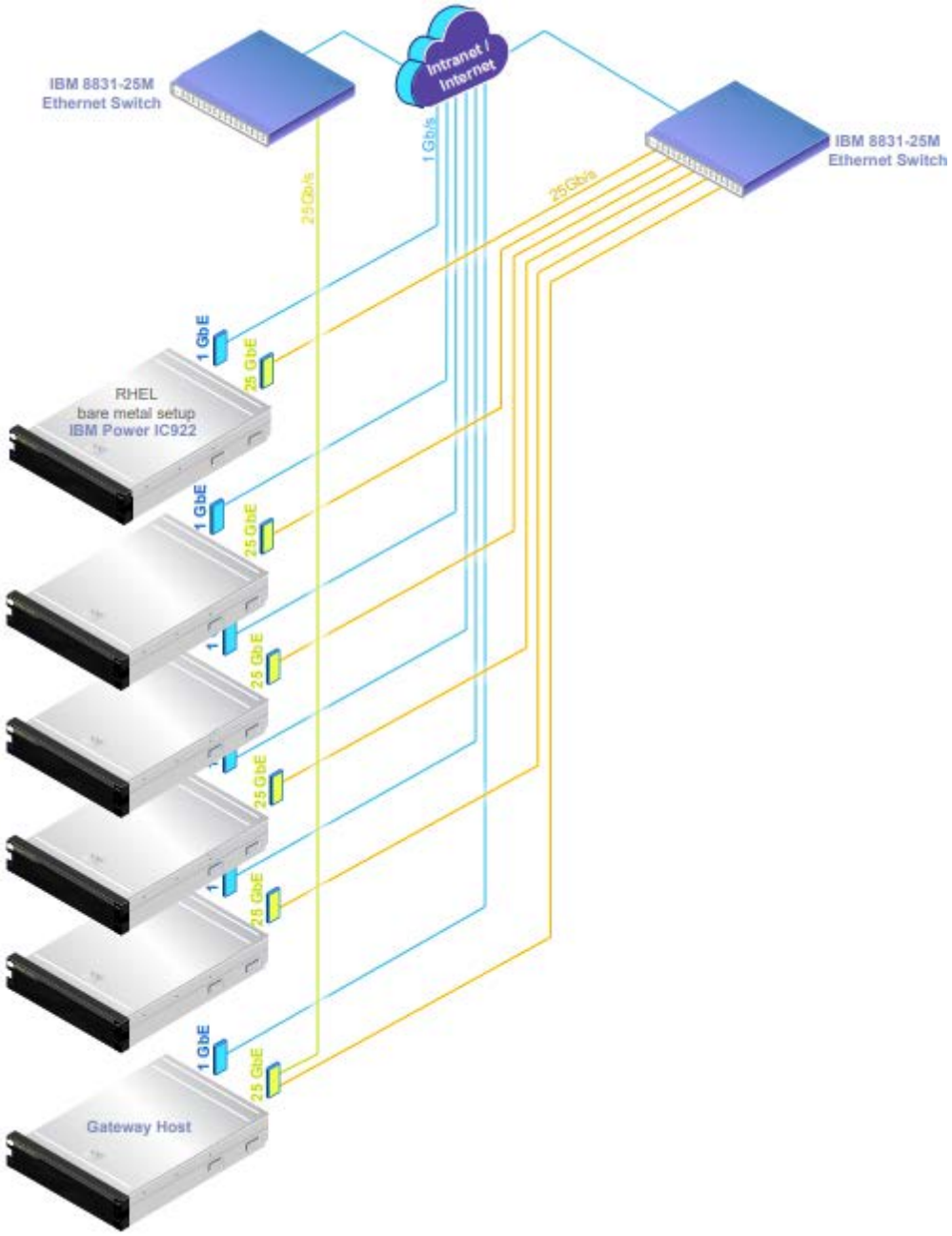


Figure 3. CDP Private Cloud Base deployment topology

Cluster sizing

Sizing a system for CDP Private Cloud Base is a significant and complex topic. Sizing is relevant in a number of dimensions. For example, in terms of requirements, factors such as throughput, response time, ingest rate, and so on may be relevant. In terms of the system design, parameters, such as number of each node type, processor cores total and per node, memory total and per node, HDFS storage capacity total and per node, network adapter ports and bandwidth, and so on must be chosen.

A complete treatment of the design and sizing process for CDP Private Cloud Base on IBM Power is beyond the scope of this reference architecture. Note that some skilled consulting is required to properly size a cluster for a client deployment. However, some guidance for one common process (sizing for storage capacity) is provided in the following section for reference.

Refer to the following process for a data capacity driven sizing:

1. Gather client requirements.
    - Usable storage capacity (represented as UsableStorageCapacity in Formula 1) needed (HDFS content)
    - Usage modes and cases expected
    - Storage calculation input parameters (DataReplicationFactor, IntermediateDataUplift, CompressionRatio, FreeSpace - see Formula 1)
    - Addition demand on master nodes expected – especially database operations
    - Number of users expected
    - Data ingest rate expected
    - Networking preferences or policies
    - Availability requirements
2. Choose the cluster type (for example, Balanced, Performance, or Storage Dense) as a function of the usage modes or cases expected.
3. Choose the drive size and number of drives per server for worker nodes (typically fully populate all drive bays on worker nodes).
4. Calculate the amount of raw storage per worker node (StoragePerWorkerNode).
5. Calculate the total system raw storage capacity needed (RawStorageCapacity - see Formula 1).
6. Calculate the total number of worker nodes (see Formula 2).
7. 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.
8. Choose the number of edge nodes as a function of the number of users and expected data ingest rate.
9. Choose the logical network topology, typically based on client networking preferences or policies.
10. Choose the network switches and network redundancy preferred for each network class.
11. Confirm or adjust the configuration for each node type – beginning with the reference configuration for each node type for the chosen cluster type.
12. Confirm or adjust the network link capacities and switch capacities as appropriate.
13. Confirm or adjust the node counts to meet availability requirements.
14. Confirm or adjust any selection based on growth expectation or initial headroom required.

RawStorageCapacity = ((UsableStorageCapacity * DataReplicationFactor + 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 multitrack 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
- 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 reference architecture, and network design consulting is recommended for any configuration that exceeds more than a very few racks. Third is somewhat specific to CDP Private Cloud Base deployment, the master and edge nodes should be spread across the different racks.

Design variations

The following variations are included as part of this reference design. Each of these variations brings with it some trade-offs that may be non-obvious or difficult to quantify. If any of these variations are applied, care should be taken to ensure that the resulting behavior and characteristics of the system meet the requirements of the deployment.

- Serial Advanced Technology Attachment (SATA) disk drives may be used for any of the HDDs for a node type. This may be done for any or all the ode types in the cluster. However, if done, this substitution is typically most appropriate to apply first to the worker nodes and finally to the master nodes. This variation trades some performance and reliability, availability, and serviceability (RAS) characteristics for lower price.
- CPU for a node type is assumed to be two sockets. It may be reduced to as low as 12 cores on Power IC922 (based on POWER9 processors) by using one socket. This variation trades performance for lower price. If a single socket processor option is chosen, note that other features of the server may not be available or other capacities (for example, maximum memory) may be reduced.
- Memory for a node type may be increased up to 2 TB. 2 TB is the maximum memory available for the server models in this reference design. This variation may improve performance, and it typically increases price.

- Memory for a node type may be reduced to 128 GB. 128 GB is recommended as the minimum memory for worker, master, and edge nodes. This variation typically lowers price, and it may reduce performance.
- HDD sizes may be increased up to 3.8 TB per drive. This variation increases the total storage capacity with a reduction in performance likely when compared to the same capacity spread over a larger number of smaller drives.
- HDD sizes may be decreased to 960 GB per drive. This variation reduces the total storage capacity with an increase in performance likely when compared to the same capacity spread over a fewer number of smaller drives.
- If additional network bandwidth is needed for data network, 100 GbE connections can be used instead of the 25 GbE connections. A 100 GB EDR switch is recommended for this network.

Cluster configuration best practices

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

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 disk). ZooKeeper's dataDir and dataLogDir can be also be configured to use separate disks to help minimize any I/O related issues.

HDFS

Java heap sizes

The NameNode memory should be increased over time as HDFS has more files and blocks stored. Cloudera Manager can monitor and alert on memory usage. A rough estimate is that the NameNode needs 1 GB of memory for every 1 million files. Setting the heap size too large when it is not needed leads to inefficient Java™ garbage collection, which might result in an erratic behavior that is hard to diagnose. NameNode and Standby NameNode heap sizes must always be the same and must be adjusted together.

NameNode metadata locations

When a quorum-based high availability HDFS configuration is used, JournalNodes handle the storage of metadata write operations. The NameNode daemons require a local location to store metadata. Cloudera recommends that only a single directory be used if the underlying disks are configured as RAID, or two directories on different disks if the disks are mounted as JBOD.

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.

Replication
Bottlenecks can occur on a small number of nodes when only small subsets of files on HDFS are being heavily accessed. Increasing the replication factor of the files so that their blocks are replicated over more nodes can alleviate this. This is done at the expense of storage capacity on the cluster. This can be set on individual files, or recursively on directories with the `-R` parameter, by using the Hadoop shell command, `hadoop fs -setrep`. By default, the replication factor is 3.

Erasure coding
Erasure coding (EC) is an alternative to the 3x replication scheme. See Data Durability for details on how EC works.

YARN

The YARN service manages 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

CDP Private Cloud Base is an enterprise analytics and data management solution that requires a properly planned infrastructure to meet user requirements. The underlying infrastructure matters to ensure performance expectations and SLA requirements are met. Key POWER9 configurations should be considered 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 POWER9 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 authors

Santoshkumar Hiremath is a technical architect for Cloudera on IBM Power platform in the IBM Cognitive Systems organization. Santosh has more than 13 years of experience in the Hadoop and Cloud ecosystem. You can reach Santosh at santoshh@in.ibm.com or https://www.linkedin.com/in/santoshkumarhiremath-7005701a/

Grace Liu is a principal offering manager for Cloudera and Linux Infrastructure in the IBM Cognitive Systems organization. Grace has more than 17 years of experience with IBM Power Systems. You can reach Grace at gliu@us.ibm.com or https://www.linkedin.com/in/grace-l-7663453/

IBM Systems
3039 Cornwallis Road
RTP, NC 27709

Produced in the United States of America