HPE Reference Configuration for MapR on HPE Elastic Platform for Big Data Analytics (EPA)

HPE EPA Traditional cluster design with HPE ProLiant DL380 Gen10 for MapR 6.x

Reference Architecture

Contents

Executive summary

Hewlett Packard Enterprise and MapR allow you to derive new business insights from all of your data by providing a platform to store, manage, and process data at scale. This Reference Configuration provides several performance optimized configurations for deploying MapR clusters on Hewlett Packard Enterprise infrastructure that provide a significant reduction in complexity and a recognized increase in value and performance.

The configurations described in this Reference Configuration are based on the MapR 6.x release and the HPE Elastic Platform for Big Data Analytics (EPA) infrastructure; and they highlight solutions based on HPE EPA Traditional systems. These configurations have been designed and developed by Hewlett Packard Enterprise to provide the highest levels of computational performance for MapR.

This Reference Configuration (RC) describes deployment options for the MapR 6.x using the HPE Elastic Platform for Big Data Analytics - modular building blocks of compute and storage optimized for modern workloads. This RC also provides suggested configurations that highlight the benefits of a building block approach to address the diverse processing and storage requirements typical of modern Big Data platforms.

The Hewlett Packard Enterprise software, HPE ProLiant DL380 Gen10 servers, and the HPE networking switches, and all of their respective configurations, that are recommended in this RC have been carefully tested with a variety of I/O, CPU, network, and memory bound workloads. The configurations included provide optimum MapReduce, YARN, Spark, Hive, and HBase computational performance, resulting in a significant performance increase at an optimal cost. The HPE EPA solutions provide excellent performance and availability, with integrated software, services, infrastructure, and management – all delivered as one proven configuration, described in more detail at hpe.com/info/hadoop . The HPE Reference Library provides a comprehensive list of technical articles on Big Data, http://h17007.www1.hpe.com/us/en/enterprise/reference-architecture/info-library/index.aspx?workload=big_data.

Target audience: This document is intended for decision makers, system and solution architects, system administrators and experienced users who are interested in reducing the time to design and purchase an HPE and MapR solution. An intermediate knowledge of Apache Hadoop and scale out infrastructure is recommended. Those already possessing expert knowledge about these topics may proceed directly to the Solution components section.

Document purpose: The purpose of this document is to describe a Reference Configuration, highlighting recognizable benefits to technical audiences and providing guidance for end users on selecting the right configuration for building their Hadoop cluster needs.

This white paper describes testing performed in April 2018.

HPE Pointnext Services

In order to simplify the build for customers, HPE provides a bill of materials in this document to allow customers to purchase this complete solution. HPE recommends that customers purchase the option of services from HPE Pointnext, as detailed in Appendix B: HPE Pointnext value-added services and support, to install and configure the operating system, verify if all firmware and versions are installed correctly, and run a suite of tests that verify that the configuration is performing optimally. Once this has been done, the customer can perform a standard MapR Converged Data Platform installation using the recommended guidelines in this document.

MapR Converged Enterprise Edition overview

The MapR Converged Data Platform delivers distributed processing, real-time analytics, and enterprise-grade requirements across cloud and on-premises environments, while leveraging the significant ongoing development in open source technologies including Spark and Hadoop. Converge-X Data Fabric powers the shared services of the Converged Data Platform including high availability, unified security, multi-tenancy, disaster recovery, global namespace, data management, automation, global event streaming and real-time data access.

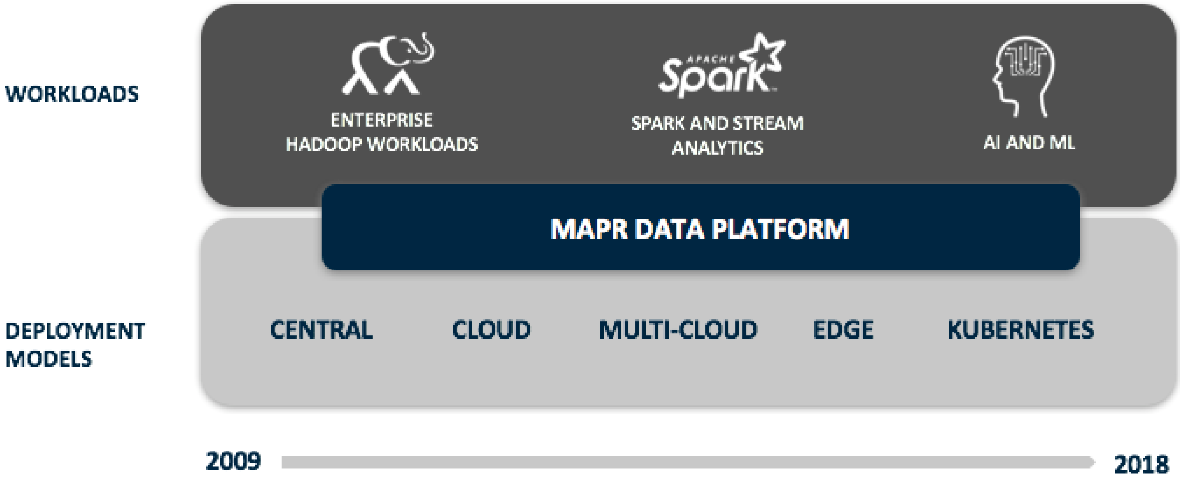


Figure 1. MapR Converged Data Platform – Architected To Evolve Quickly With Industry Trends

MapR 6.x is providing new MapR Control System (MCS), a beautiful and unified management solution for efficiently administering all data in the MapR Converged Data Platform and underlying cluster infrastructure. Key aspects of MapR 6.x include best-in-class database management system for building global data-intensive applications, Data science offering to fit the needs of all data teams and automated platform health and security capabilities.

MapR-XD Cloud-Scale Data Store provides exabyte-scale data store for building intelligent applications with the MapR Converged Data Platform. MapR-XD includes all the functionality you need to manage large amounts of conventional data.

MapR-DB is a high performance NoSQL ("Not Only SQL") database management system built into the MapR Converged Data Platform. It is a highly scalable multi-model database that brings together operations and analytics, and real-time streaming and database workloads to enable a broader set of next-generation data-intensive applications in organizations. Starting with MapR 6.x, we recommend using SSDs as data drives to customers who want to use the secondary indexes feature in MapR-DB.

MapR-ES is the first big data-scale streaming system built into a converged data platform, and the only big data streaming system to support global event replication reliably at IoT scale. The MapR Converged Data Platform allows you to quickly and easily build breakthrough, reliable, real-time applications by providing:

- **Single cluster** for streams, file storage, database, and analytics.
- **Persistence** of streaming data, providing direct data access to batch and interactive frameworks, eliminating data movement.
- **Unified security** framework for data-in-motion and data-at-rest, with authentication, authorization, and encryption.
- **Utility-grade reliability** with self-healing and no single point-of-failure architecture.

Solution overview

These configurations are based on the MapR 6.x specifically and HPE EPA Traditional systems which includes the HPE ProLiant DL380 Gen10 server platform.

HPE EPA Traditional solution

The HPE EPA Traditional solution infrastructure blueprints are composed of four blocks: storage/compute blocks, control blocks, network blocks and rack blocks. Listed below are the blocks and model in a HPE EPA Traditional solution:

Table 1. HPE EPA Traditional standard solution components

Blocks	Model
Control Block	HPE ProLiant DL360 Gen10
Compute/Storage Block	HPE ProLiant DL380 Gen10
Network Block	HPE FlexFabric 5940 48XGT 6QSFP28 switch
Rack Block	1200mm or 1075mm

Note

Use one 150GB SATA RI M.2 DS SSD for OS disk in SATA AHCI mode and not software RAID. If software RAID is used, the two 150GB M.2 SSD disks are managed by HPE Smart Array S100i SR Gen10 SW RAID controller using in-distro open-source software to create a two-disk RAID1 boot volume. Software RAID can require a significant amount of the server's resources and harm performance. For more info: http://h20564.www2.hpe.com/hpsc/doc/public/display?docId=a00017763en_us.

If RAID1 for OS drives are required we recommend configuring the HPE ProLiant DL380 Gen10 2SFF rear drives for OS.

For detailed information, please refer: HPE Reference Configuration for Elastic Platform for Big Data Analytics.

Solution components and configuration guide

Single-rack Reference Configuration

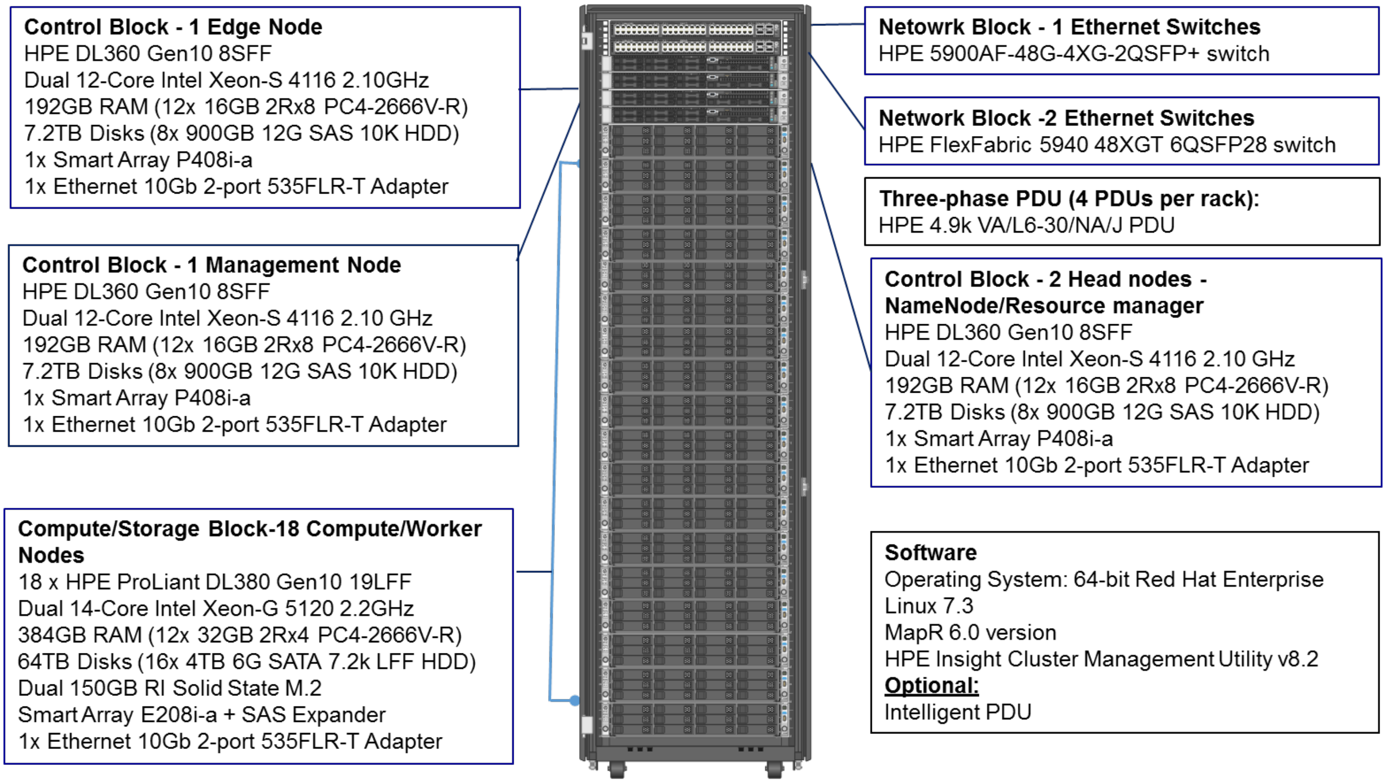
The single-rack Hadoop Reference Configuration (RC) is designed to perform well as a single-rack cluster design but also form the basis for a much larger multi-rack design. When moving from the single-rack to multi-rack design, one can simply add racks to the cluster without having to change any components within the single-rack. This RC reflects the following:

- **Single-rack network block**
  - The HPE FlexFabric 5940 48XGT 6QSFP28 switch is a high density ToR switch available as a 1RU 48-port 10GbE. This switch can be used for high-density 10GbE ToR with 100GbE/40GbE/25GbE/10GbE spine/ToR connectivity. 100GbE ports may be split into four 25GbE ports and can also support 40GbE which can be split into four by 10GbE for a total of 80 25/10GbE ports. The HPE FlexFabric 5940 48XGT 6QSFP28 switch includes six 100GbE uplinks which can be used to connect the switches in the rack into the desired network or to the 100GbE HPE FlexFabric 5950 32QSFP28 aggregation switch. Keep in mind that if IRF bonding is used, it requires 2x 100GbE ports per switch, which would leave 4x 100GbE ports on each HPE FlexFabric 5940 48XGT 6QSFP28 switch for uplinks.
- **Power and cooling**
  - In planning for large clusters, it is important to properly manage power redundancy and distribution. To ensure the servers and racks have adequate power redundancy we recommend that each server have a backup power supply, and each rack have at least two Power Distribution Units (PDUs). There is an additional cost associated with procuring redundant power supplies.

Reference Configuration for standard HPE EPA Traditional with HPE ProLiant DL380 Gen10 balanced block

Refer to Figure 2 for a rack-level view of the single-rack Reference Configuration for this solution with HPE ProLiant DL380 balanced block.

For more information on configuration for the Balanced Optimization solution refer to HPE Reference Configuration for Elastic Platform for Big Data Analytics.



Control Block - 1 Edge Node
HPE DL360 Gen10 8SFF
Dual 12-Core Intel Xeon-S 4116 2.10GHz
192GB RAM (12x 16GB 2Rx8 PC4-2666V-R)
7.2TB Disks (8x 900GB 12G SAS 10K HDD)
1x Smart Array P408i-a
1x Ethernet 10Gb 2-port 535FLR-T Adapter

Control Block - 1 Management Node
HPE DL360 Gen10 8SFF
Dual 12-Core Intel Xeon-S 4116 2.10 GHz
192GB RAM (12x 16GB 2Rx8 PC4-2666V-R)
7.2TB Disks (8x 900GB 12G SAS 10K HDD)
1x Smart Array P408i-a
1x Ethernet 10Gb 2-port 535FLR-T Adapter

Compute/Storage Block-18 Compute/Worker Nodes
18 x HPE ProLiant DL380 Gen10 19LFF
Dual 14-Core Intel Xeon-G 5120 2.2GHz
384GB RAM (12x 32GB 2Rx4 PC4-2666V-R)
64TB Disks (16x 4TB 6G SATA 7.2k LFF HDD)
Dual 150GB RI Solid State M.2
Smart Array E208i-a + SAS Expander
1x Ethernet 10Gb 2-port 535FLR-T Adapter

Netowrk Block - 1 Ethernet Switches
HPE 5900AF-48G-4XG-2QSFP+ switch

Network Block -2 Ethernet Switches
HPE FlexFabric 5940 48XGT 6QSFP28 switch

Three-phase PDU (4 PDUs per rack):
HPE 4.9k VA/L6-30/NA/J PDU

Control Block - 2 Head nodes - NameNode/Resource manager
HPE DL360 Gen10 8SFF
Dual 12-Core Intel Xeon-S 4116 2.10 GHz
192GB RAM (12x 16GB 2Rx8 PC4-2666V-R)
7.2TB Disks (8x 900GB 12G SAS 10K HDD)
1x Smart Array P408i-a
1x Ethernet 10Gb 2-port 535FLR-T Adapter

Software
Operating System: 64-bit Red Hat Enterprise Linux 7.3
MapR 6.0 version
HPE Insight Cluster Management Utility v8.2
Optional:
Intelligent PDU

Figure 2. Single-rack MapR Reference Configuration – Rack-level view

For multi-rack architecture, please refer to HPE Reference Configuration for Elastic Platform for Big Data Analytics.

Best practice

For each server, Hewlett Packard Enterprise recommends that each power supply is connected to a different PDU than the other power supply on the same server. Furthermore, the PDUs in the rack can each be connected to a separate data center power line to protect the infrastructure from a data center power line failure.

Additionally, distributing the server power supply connections evenly to the in-rack PDUs, as well as distributing the PDU connections evenly to the data center power lines, ensures an even power distribution in the data center and avoids overloading any single data center power line. When designing a cluster, check the maximum power and cooling that the data center can supply to each rack and ensure that the rack does not require more power and cooling than is available.

Pre-deployment considerations

The operating system and the network are key factors you need to consider prior to designing and deploying a MapR cluster. The following subsections articulate the design decisions in creating the baseline configurations for the Reference Configurations.

Operating system

MapR 6.x supports 64-bit operating systems, visit https://maprdocs.mapr.com/home/InteropMatrix/r_os_matrix_6.x.html for the minimum requirements. In this RC, we have tested with Red Hat Enterprise Linux® (RHEL) 7.3.

Key point

Hewlett Packard Enterprise recommends all HPE ProLiant systems be upgraded to the latest BIOS and firmware versions before installing the OS. HPE Service Pack for ProLiant (SPP) is a comprehensive systems software and firmware update solution, which is delivered as a single ISO image. The minimum SPP version recommended is 2017.10.1. The latest version of SPP is available at: http://h17007.www1.hpe.com/us/en/enterprise/servers/products/service_pack/spp/index.aspx

Computations

Employing Hyper-Threading increases effective core count, potentially allowing the YARN ResourceManager to assign more cores as needed.

Storage capacity

The number of disks and their corresponding storage capacity determines the total amount of the storage capacity for your cluster.

Redundancy

Hadoop ensures that a certain number of block copies are consistently available. This number is configurable in the block replication factor setting, which is typically set to three. If a Hadoop worker node goes down, Hadoop will replicate the blocks that had been on that server onto other servers in the cluster to maintain the consistency of the number of block copies. For example, if the NIC (Network Interface Card) on a server with 16TB of block data fails, 16TB of block data will be replicated between other servers in the cluster to ensure the appropriate number of replicas exist. Furthermore, the failure of a non-redundant ToR (Top of Rack) switch will generate even more replication traffic. Hadoop provides data throttling capability in the event of a node/disk failure so as to not overload the network.

I/O performance

The more disks you have, the less likely it is that you will have multiple tasks accessing a given disk at the same time. This avoids queued I/O requests and incurring the resulting I/O performance degradation.

Disk configuration

For management nodes, storage reliability is important and SAS drives are recommended. For worker nodes, one has the choice of SAS or SATA and as with any component there is a cost/performance tradeoff. Specific details around disk and RAID configurations will be provided in the server tuning section in Appendix A.

Network

Configuring a single ToR switch per rack introduces a single point of failure for each rack. In a multi-rack system such a failure will result in a very long replication recovery time as Hadoop rebalances storage; and, in a single-rack system such a failure could bring down the whole cluster. Consequently, configuring two ToR switches per rack is recommended for all production configurations as it provides an additional measure of redundancy. This can be further improved by configuring link aggregation between the switches. The most desirable way to configure link aggregation is by bonding the two physical NICs on each server. Port1 wired to the first ToR switch and Port2 wired to the second ToR switch, with the two switches IRF bonded. When done properly, this allows the bandwidth of both links to be used. If either of the switches fail, the servers will still have full network functionality, but with the performance of only a single link. Not all switches have the ability to do link aggregation from individual servers to multiple switches; however, the HPE FlexFabric 5940 48XGT 6QSFP28+ switch supports this through HPE Intelligent Resilient Fabric (IRF) technology. In addition, switch failures can be further mitigated by incorporating dual power supplies for the switches.

Hadoop is rack-aware and tries to limit the amount of network traffic between racks. The bandwidth and latency provided by two bonded 10 Gigabit Ethernet (GbE) connections from the worker nodes to the ToR switch is more than adequate for most Hadoop configurations.

A more detailed white paper for Hadoop Networking best practices is available at, http://h20195.www2.hpe.com/V2/GetDocument.aspx?docname=a00004216enw.

For sizing the cluster use the HPE EPA Sizing tool, available at, https://www.hpe.com/h20195/v2/GetDocument.aspx?docname=a00005868enw.

High Availability considerations

The following are some of the High Availability (HA) features considered in this Reference Configuration:

- **CLDB HA –** The configurations in this white paper utilize quorum-based journaling high-availability feature. For this feature, servers should have similar I/O subsystems and server profiles so that each (Container Location Database) CLDB server can potentially take the role of another. Another reason to have similar configurations is to ensure that ZooKeeper's quorum algorithm is not affected by a machine in the quorum that cannot make a decision as fast as its quorum peers.
- **ResourceManager HA –** To make a YARN cluster highly available (similar to JobTracker HA in MR1), the underlying architecture of an Active/Standby pair is configured, hence the completed tasks of in-flight MapReduce jobs are not re-run on recovery after the ResourceManager is restarted or failed over. One ResourceManager is Active and one or more ResourceManagers are in standby mode waiting to take over should anything happen to the Active ResourceManager. MCS provides a simple wizard to enable HA for YARN ResourceManager.
- **OS availability and reliability –** For the reliability of the server, the OS disk is configured in a RAID1 configuration thus preventing failure of the system from OS hard disk failures.
- **Network reliability –** The Reference Configuration uses the standard HPE EPA Traditional network block with two HPE FlexFabric 5940 48XGT 6QSFP28 switches for redundancy, resiliency and scalability through using Intelligent Resilient Fabric (IRF) bonding. We recommend using redundant power supplies.
- **Power supply** – To ensure the servers and racks have adequate power redundancy we recommend that each server have a backup power supply, and each rack have at least two Power Distribution Units (PDUs).

Software components for control blocks

The control block is made up of three HPE ProLiant DL360 Gen10 servers, with an optional fourth server acting as an edge or gateway node depending on the customer enterprise network requirements.

Management node

The management node hosts the applications that submit jobs to the Hadoop cluster. We recommend that you install with the software components shown in Table 2.

Table 2. Management node basic software components

Software	Description
Red Hat Enterprise Linux 7.3	Recommended Operating System
HPE Insight CMU 8.2	Infrastructure Deployment, Management, and Monitoring
Oracle JDK 1.8	Java Development Kit
MCS	MapR Control System
ZooKeeper	Cluster coordination service

Best practice

For performance critical Hadoop clusters, Hewlett Packard Enterprise and MapR recommend two disks in RAID1 for OS, two disks in RAID1 for FS metadata, one JBOD or RAID0 for ZooKeeper, one JBOD or RAID0 for QJN (Quorum Journal Node), and the rest for database.

Head nodes

The head node servers contain the following software components with HA feature enabled. See the following link for more information on installing and configuring the service, https://maprdocs.mapr.com/home/install.html.

Table 3 shows the head node servers base software components.

Table 3. Head node server base software components

Software	Description
Red Hat Enterprise Linux 7.3	Recommended Operating System
Oracle JDK 1.8	Java Development Kit
ResourceManager	YARN ResourceManager
CLDB	Maintains the locations of services, containers and other cluster information
HiveServer2	Hive Service to run SQL-like ad hoc queries
ZooKeeper	Cluster coordination service
Fileserver	Disk Storage for MapR
HBaseMaster	The HBase Master for the Hadoop cluster (Only if running HBase)
Warden	Manage, Monitor and report on the services on each node
Job History Server	Job History for ResourceManager

Edge nodes

The edge node hosts the client configurations that submit jobs to the Hadoop cluster, but this optional control block depending on the customer enterprise network requirements. We recommend that you install the following software components shown in Table 4.

Table 4. Edge node basic software components

Software	Description
Red Hat Enterprise Linux 7.3	Recommended Operating System
Oracle JDK 1.8	Java Development Kit
Gateway Services	Hadoop Gateway Services (FS, YARN, MapReduce, HBase, and others)

Software components for compute/storage blocks

The worker nodes run the DataNode, NodeManager and YARN container processes and thus storage capacity and compute performance are important factors.

Balanced block software components

Table 5 lists the worker node software components. See the following link for more information on installing and configuring the NodeManager and DataNode, https://maprdocs.mapr.com/home/install.html.

Table 5. Compute/storage node base software components

Software	Description
Red Hat Enterprise Linux 7.3	Recommended Operating System
Oracle JDK 1.8	Java Development Kit
NodeManager	The NodeManager process for MR2/YARN
Fileserver	Disk Storage for MapR

For Hardware Configuration guidelines, please refer to the HPE EPA Traditional Standard block HPE Reference Configuration for Elastic Platform for Big Data Analytics.

Capacity and sizing

Hadoop cluster storage sizing requires careful planning and identifying the current and future storage and compute needs. Use the following as general guidelines for data inventory:

- Sources of data
- Frequency of data
- Raw storage
- Processed FS storage
- Replication factor
- Default compression turned on
- Space for intermediate files

Best Practices and tuning guidelines

YARN configuration

For configuring YARN, update the default values of the following attributes with ones that reflect the cores and memory available on a worker node.

- yarn.nodemanager.resource.memory-mb - Defines the memory available to processing Yarn containers on the node in MB.
- yarn.nodemanager.resource.cpu-vcores - Defines the number of CPUs available to process YARN containers on the node.

While configuring YARN for MapReduce jobs make sure that the following attributes have been specified with sufficient vcores and memory. They represent resource allocation attributes for map and reduce containers. Note that the optimum values for these attributes depend on the nature of workload/use case.

- mapreduce.map.memory.mb - Defines the container size for map tasks in MB.
- mapreduce.reduce.memory.mb - Defines the container size for reduce tasks in MB.

In mapred-site.xml, the following properties define the number of disks a MapReduce task requires. On nodes, this value should be 0 so that the number of MapReduce tasks spawned does not depend on the number of disks present on the node.

- mapreduce.mapr.disk - Defines the number of disks a map task requires. Set to 0 to disable checking the value.

- mapreduce.reduce.disk - Defines the number of disks that a reduce task requires. Set to 0 to disable checking the value.

Similarly, specify the appropriate size for map and reduce task heap sizes using the following attributes:

- mapreduce.map.java.opts - Java options for map tasks.
- mapreduce.reduce.java.opts - Java options for reduce tasks.

MapR Best Practice
MapR-XD is configured using 4 disks per storage pool so that all 16 disks are utilized for MapR-XD.

Syntax: /opt/mapr/server/disksetup <options><disk list file>

Setting up the Storage Pool specified in the file /tmp/disks.txt: /opt/mapr/server/disksetup -F -W 4 /tmp/disks.txt

Options: -F: Forces formatting of all specified disks

-W: Specifies the number of disks per storage pool.

Note
Add equal number of disks to each storage pool to attain better performance.

For more information on configuring storage node refer to, https://maprdocs.mapr.com/home/AdvancedInstallation/InstallingMapRSoftware-config-node-storage.html.

Isolating CLDB nodes
In a large cluster (100 nodes or more) create CLDB-only nodes to ensure high performance. This configuration also provides additional control over the placement of the CLDB data, for load balancing, fault tolerance, or high availability (HA). Setting up CLDB-only nodes involves restricting the CLDB volume to its own topology and making sure all other volumes are on a separate topology.

Isolating ZooKeeper Nodes
For large clusters (100 nodes or more), isolate ZooKeeper on nodes that do not perform any other function. Isolating ZooKeeper enables the node to perform its functions without competing for resources with other processes. Installing a ZooKeeper-only node is similar to any typical node installation, but with a specific subset of packages.

Warning
Do not install the Fileserver package on an isolated ZooKeeper node in order to prevent MapR from using this node for data storage.

Refer to MapR best practices, for more information: http://maprdocs.mapr.com/home/ReferenceGuide/BestPractices.html

Best practice
These servers have 6 memory channels per processor and in order to get optimal performance all channels should be used, which means 12 DIMMs (or 24 DIMMs). For details on the HPE Server Memory Options Population Rules visit, http://www.hpe.com/docs/memory-population-rules.

HPE Sizer for the Elastic Platform for Big Data Analytics

HPE has developed the HPE Sizer for the Elastic Platform for Big Data Analytics to assist customers with proper sizing of these environments. Based on design requirements, the sizer will provide a suggested bill of materials (BOM) and metrics data for an HPE EPA Traditional cluster which can be modified further to meet customer requirements.

To download the HPE Sizer for the Elastic Platform for Big Data Analytics, visit hpe.com/info/sizers.

HPE Insight Cluster Management Utility

The HPE Insight Cluster Management Utility (Insight CMU) is a collection of tools used to manage and monitor a large group of nodes, specifically High performance Computing and large Linux Clusters such as Big data environments. Insight CMU helps manage, install, and monitor the nodes of your cluster from a single interface. A simple graphical interface enables an “at-a-glance” real-time or 3D historical view of the entire cluster for both infrastructure and application (including Hadoop) metrics, provides frictionless scalable remote management and analysis, and allows rapid provisioning of software to all nodes of the system.

Best practice

HPE recommends using HPE Insight CMU for all Hadoop clusters. HPE Insight CMU allows one to easily correlate Hadoop metrics with cluster infrastructure metrics, such as CPU Utilization, Network Transmit/Receive, Memory Utilization and I/O Read/Write. This allows characterization of Hadoop workloads and optimization of the system thereby improving the performance of the Hadoop cluster. HPE Insight CMU Time View Metric Visualizations will help you understand, based on your workloads, whether your cluster needs more memory, a faster network or processors with faster clock speeds. In addition, HPE Insight CMU also greatly simplifies the deployment of Hadoop, with its ability to create a golden Image from a node and then deploy that image to up to 4000 nodes. HPE Insight CMU is able to deploy 800 nodes in 30 minutes.

HPE Insight CMU is highly flexible and customizable, offers both GUI and CLI interfaces supports for Ansible, and can be used to deploy a range of software environments, from simple compute farms to highly customized, application-specific configurations. HPE Insight CMU is available for HPE ProLiant and HPE BladeSystem servers, and is supported on a variety of Linux operating systems, including Red Hat Enterprise Linux, SUSE Linux Enterprise Server, CentOS, and Ubuntu. HPE Insight CMU also includes options for monitoring graphical processing units (GPUs) and for installing GPU drivers and software. Figures 3 and 4 shows the real-time view of the HPE Insight CMU.

HPE Insight CMU is free for managing up to 32 nodes.

HPE Insight CMU can be configured to support High Availability with an active-passive cluster. For more information, see hpe.com/info/cmu.

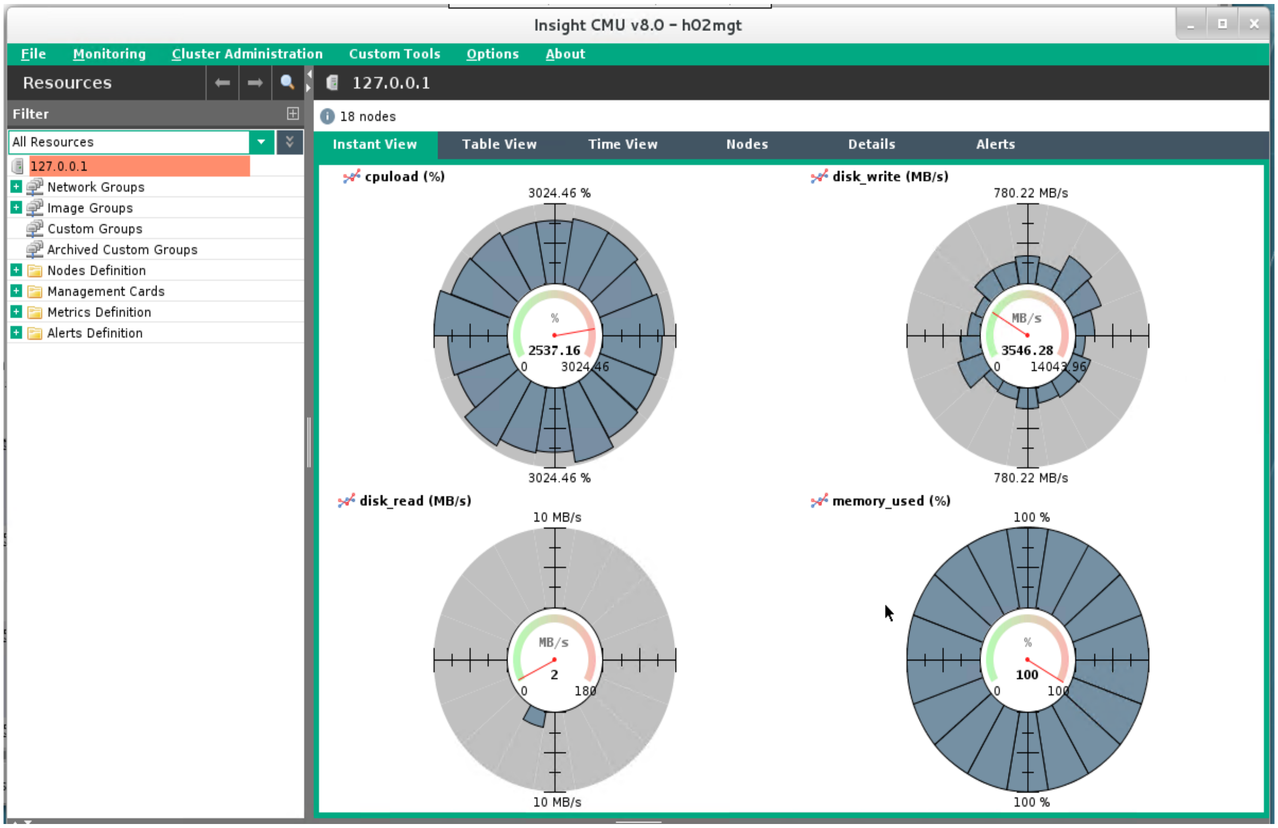
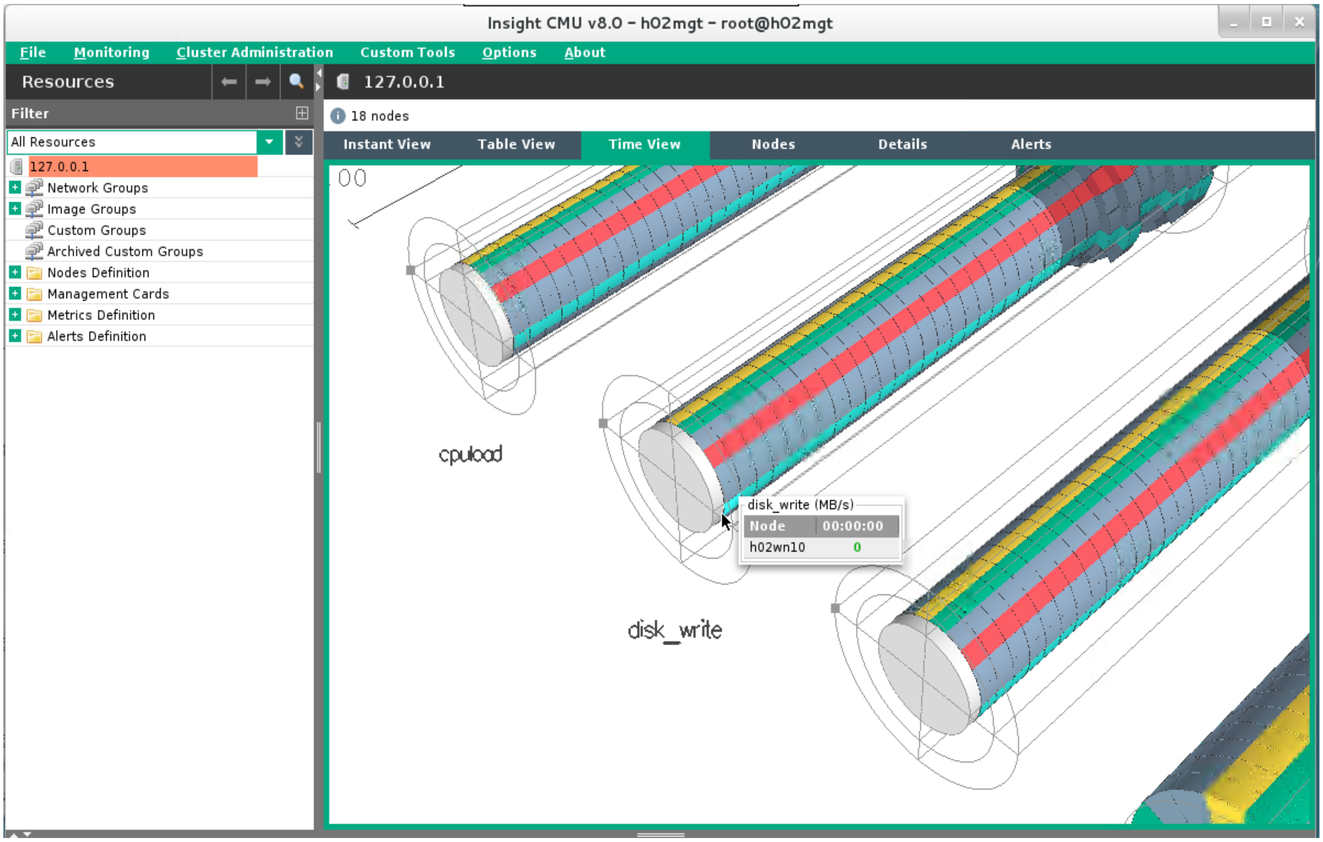


Figure 3. HPE Insight CMU Interface – real-time view

Figure 4 shows the Time View of the HPE Insight CMU.



disk_write (MB/s)

Node	00:00:00
h02wn10	0

Figure 4. HPE Insight CMU Interface – Time view

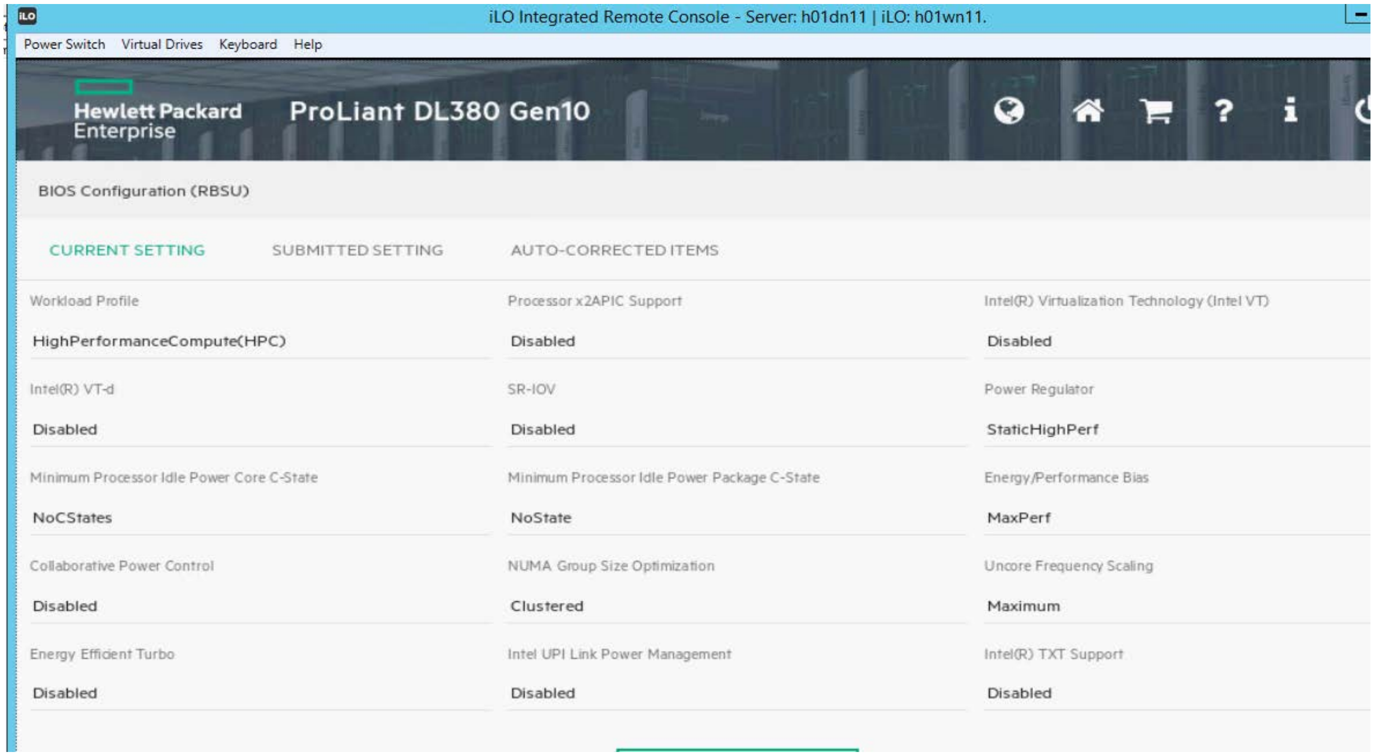
Summary

Hewlett Packard Enterprise and MapR allow one to derive new business insights from Big Data by providing a platform to store, manage and process data at scale. However, designing and ordering Hadoop clusters can be both complex and time consuming. This white paper provides several Reference Configurations for deploying clusters of varying sizes with the MapR 6.x on HPE infrastructure and management software. These configurations leverage HPE balanced building blocks of servers, storage and networking, along with integrated management software and bundled support. In addition, this white paper has been created to assist in the rapid design and deployment of MapR on HPE infrastructure for clusters of various sizes.

Appendix A: Hadoop cluster tuning/optimization

Gen10 Server BIOS Configuration

HPE recommends changing the default BIOS setting to the following using Workload Profile HighPerformanceCompute (HPC) on all HPE ProLiant servers hosting Hadoop to ensure highest performance.


iLO Integrated Remote Console - Server: h01dn11 | iLO: h01wn11.
Power Switch
Virtual Drives
Keyboard
Help
Hewlett Packard Enterprise
ProLiant DL380 Gen10
BIOS Configuration (RBSU)
CURRENT SETTING
SUBMITTED SETTING
AUTO-CORRECTED ITEMS

Workload Profile HighPerformanceCompute(HPC)	Processor x2APIC Support Disabled	Intel(R) Virtualization Technology (Intel VT) Disabled
Intel(R) VT-d Disabled	SR-IOV Disabled	Power Regulator StaticHighPerf
Minimum Processor Idle Power Core C-State NoCStates	Minimum Processor Idle Power Package C-State NoState	Energy/Performance Bias MaxPerf
Collaborative Power Control Disabled	NUMA Group Size Optimization Clustered	Uncore Frequency Scaling Maximum
Energy Efficient Turbo Disabled	Intel UPI Link Power Management Disabled	Intel(R) TXT Support Disabled

Figure 5. HPE ProLiant DL380 Gen10 BIOS workload profile

The following are the BIOS setting recommended for HPE ProLiant DL380 Gen10.

Table 6. HPE ProLiant DL380 Gen10 BIOS settings

Parameter	Description	Settings
Boot_Mode	Recommended Operating System	UEFI_Mode
UEFI_Optimized_Boot	UEFI Optimized Boot	Enabled
Workload_Profile	Sets power and performance settings for application workloads.	High_Performance_Compute
Processor_x2APIC_Support	Enables or disables x2APIC support.	Disabled
Intel_VT	Controls whether a Virtual Machine Manager (VMM) supporting Virtualization Technology can use hardware capabilities provided by UEFI Intel processors	Disabled
Intel_VT-d	Enables or disables Intel Virtualization Technology for Directed I/O (VT-d) on a Virtual Machine Manager (VMM).	Disabled
SR-IOV	Enables or disables the BIOS to allocate more PCI resources to PCIe devices.	Disabled
Power_Regulator	Sets the power regulator mode.	Static_High_Performance_Mode
Minimum_Processor_Idle_Power_Core_C-State	Sets the lowest processor idle power state (C-State).	No_C-States
Minimum_Processor_Idle_Power_Package_C-State	Sets the lowest processor idle power state (C-State).	No_Package_State
Energy/Performance_Bias	To optimize the processor's performance and power usage.	Maximum_Performance
Collaborative_Power_Control	Enables or disables collaborative power control for operating systems that support the Processor Clocking Control (PCC) interface.	Disabled
Intel_Hyper-Threading	Enables or disables the logical processor cores on processors supporting Intel Hyperthreading technology.	Enabled
Intel_Turbo_Boost_Technology	Enables or disables Intel Turbo Boost Technology to control whether the processor transitions to a higher frequency than the processor's rated speed if the processor has available power and is within temperature.	Enabled
Energy_Efficient_Turbo	Controls whether the processor uses an energy efficient based policy.	Disabled
Maximum_Memory_Bus_Frequency	Configures the system to run memory at a lower maximum speed than that supported by the installed processor and DIMM configuration.	Auto
Channel_Interleaving	Enables or disables a higher level of memory interleaving.	Enabled
Intel_UPI_Link_Power_Management	To place the Ultra Path Interconnect (UPI) links into a low power state when the links are not being used.	Disabled
Intel_TXT_support	Enables or disable Intel TXT (Trusted Execution Technology) support for servers with Intel processors.	Disabled
Embedded_SATA_Configuration	Sets the mode for the embedded SATA controller	Enable_AHCI_Support
NUMA_Group_Size_Optimization	The number of logical processors in a NUMA (Non-Uniform Memory Access) node.	Clustered
Uncore_Frequency_Scaling	Controls the frequency scaling of the processor's internal buses (the uncore).	Maximum

Server tuning

Below are some general guidelines for tuning the server OS and the storage controller for a typical Hadoop proof-of-concept (POC). Please note that these parameters are recommended for YARN workloads which are most prevalent in Hadoop environments. Please note that there is no silver bullet performance tuning. Modifications will be needed for other types of workloads.

- OS tuning

  As a general recommendation, update to the latest patch level available to improve stability and optimize performance. The recommended Linux file system is ext4, 64 bit OS:

  – Enable defaults, `nodiratime,noatime (/etc/fstab)`

  – Do not use logical volume management (LVM)

  – Tune OS block readahead to 8K (/etc/rc/local):

  ```
  blockdev --setra 8192 <storage device>
  ```

  – Decrease kernel swappiness to minimum 1:

  ```
  Set sysctl vm.swappiness=1 in /etc/sysctl.conf
  ```

  – Tune ulimits for number of open files to a high number:

  Example: in /etc/security/limits.conf:

  ```
  soft nofile 65536
  hard nofile 65536
  Set nproc = 65536
  ```

  Add it to end of (/etc/security/limits.conf)

  – Set IO scheduler policy to deadline on all the data drives:

  ```
  echo deadline > /sys/block/<device>/queue/scheduler
  ```

  – For persistency across boot, append the following to kernel boot line in /etc/grub.conf:

  ```
  elevator=deadline
  ```

  – Configure network bonding on two 10GbE server ports, for 20GbE throughput.

  – Ensure forward and reverse DNS is working properly.

  – Install and configure ntp to ensure clocks on each node are in sync to the management node.

  – Setting tuned profile network-latency for the server. Profile for low latency network tuning It additionally disables transparent hugepages, NUMA balancing and tunes several other network related sysctl parameters:

  ```
  tuned-adm profile network-latency
  ```

  – For good performance improvements, disable transparent huge page compaction:

  ```
  echo never > /sys/kernel/mm/transparent_hugepage/enabled
  ```

  – Disable SELinux on RHEL 7 by editing /etc/selinux/config and setting SELINUX=disabled

- HPE Smart Array E208i-a/ P408i-a/ P816i-a

  – Configure each Hadoop data drive as a separate RAID0 array with stripe size of 1024KB

  ```
  ssacli ctrl slot=<slot number> ld <ld number> modify ss=1024
  ```

  - Set power mode be set to maxperformance:

    ```
    ssacli ctrl slot=0 modify powermode=maxperformance
    ```

  - For data drivers we recommend enabling drive write cache:

    ```
    ssacli ctrl slot=0  modify dwc=enable
    ```

  - Turn Off “Array Acceleration” / ”Caching” for all data drives

    Example:

    ```
        ctrl slot=<slot number> ld all modify arrayaccelerator=disable ← disable arrayaccelerator on
        all logical drives on 1st ctrlr

        ctrl slot=<slot number> ld 1 modify arrayaccelerator=enable ← enable arrayaccelerator on the
        OS logical drive on 1st ctrlr
    ```

- Oracle Java

  ```
      java.net.preferIPv4Stack set to true
  ```

- Patch common security vulnerabilities

Check Red Hat Enterprise Linux and SUSE security bulletins for more information.

Appendix B: HPE Pointnext value-added services and support

In order to help customers jump-start their Big Data solution development, HPE Pointnext offers flexible, value-added services, including Factory Express and Big Data Consulting services which can accommodate and end-to-end customer experience.

HPE Pointnext Factory Express Services

Factory-integration services are available for customers seeking a streamlined deployment experience. With the purchase of Factory Express services, your cluster will arrive racked and cabled, with software installed and configured per an agreed upon custom statement of work, for the easiest deployment possible. HPE Factory Express Level 4 Service (HA454A1) is the recommended Factory Integration service for Big Data covering hardware and software integration, as well as end-to-end delivery project management. Please engage HPE Pointnext Factory Express for details and quoting assistance. For more information and assistance on Factory Integration services, you can go to: https://www.hpe.com/us/en/services/factory-express.html

Or contact:

- AMS: easy.solutions.americas@hpe.com
- APJ: ap.fe-engagement@hpe.com
- EMEA: sol_eng_support@hpe.com

HPE Pointnext Big Data Consulting – Reference Configuration Implementation Service for Hadoop

With the HPE Reference Configuration Implementation Service for Hadoop, experienced HPE Big Data consultants install, configure, deploy, and test your Hadoop environment based on the HPE Reference Configuration for Hadoop. HPE will implement a Hadoop design: naming, hardware, networking, software, administration, backup and operating procedures and work with you to configure the environment according to your goals and needs. HPE will also conduct an acceptance test to validate and prove that the system is operating as defined in the Reference Configuration.

HPE Pointnext Advisory, Transform and Manage - Big Data Consulting Services

HPE Pointnext Big Data Consulting Services cover the spectrum of services to advise, transform, and manage your Hadoop environment, helping you to reshape your IT infrastructure to corral increasing volumes of bytes – from e-mails, social media, and website downloads – and convert them into beneficial information. Our Big Data solutions encompass strategy, design, implementation, protection and compliance. We deliver these solutions in three steps.

1. Big Data Architecture Strategy and Roadmap: We'll define the functionalities and capabilities needed to align your IT with your Big Data initiatives. Through transformation workshops and roadmap services, you'll learn to capture, consolidate, manage and protect business-aligned information, including structured, semi-structured and unstructured data.
2. Big Data System Infrastructure: HPE experts will design and implement a high-performance, integrated platform to support a strategic architecture for Big Data. Choose from design and implementation services, Reference Configuration implementations and integration services. Your flexible, scalable infrastructure will support Big Data variety, consolidation, analysis, share and search on HPE platforms.
3. Big Data Protection: Ensure availability, security and compliance of Big Data systems. Our consultants can help you safeguard your data, achieve regulatory compliance and lifecycle protection across your Big Data landscape, as well as improve your backup and continuity measures.

For additional information, visit: hpe.com/us/en/services/consulting/big-data.html

Hewlett Packard Enterprise Support options

HPE offers a variety of support levels to meet your needs:

- **HPE Datacenter Care** - HPE Datacenter Care provides a more personalized, customized approach for large, complex environments, with one solution for reactive, proactive, and multi-vendor support needs.
- **HPE Support Plus 24** - For a higher return on your server and storage technology, our combined reactive support service delivers integrated onsite hardware/software support services available 24x7x365, including access to HPE technical resources, 4-hour response onsite hardware support and software updates.
- **HPE Proactive Care** - HPE Proactive Care begins with providing all of the benefits of proactive monitoring and reporting along with rapid reactive care. You also receive enhanced reactive support, through access to HPE's expert reactive support specialists. You can customize your reactive support level by selecting either 6 hour call-to-repair or 24x7 with 4 hour onsite response. You may also choose DMR (Defective Media Retention) option.
- **HPE Proactive Care with the HPE Personalized Support Option** - Adding the Personalized Support Option for HPE Proactive Care is highly recommended. The Personalized Support option builds on the benefits of HPE Proactive Care Service, providing you an assigned Account Support Manager who knows your environment and delivers support planning, regular reviews, and technical and operational advice specific to your environment. These proactive services will be coordinated with Microsoft's proactive services that come with Microsoft® Premier Mission Critical, if applicable.
- **HPE Proactive Select** - And to address your ongoing/changing needs, HPE recommends adding Proactive Select credits to provide tailored support options from a wide menu of services, designed to help you optimize capacity, performance, and management of your environment. These credits may also be used for assistance in implementing updates for the solution. As your needs change over time you flexibly choose the specific services best suited to address your current IT challenges.
- **Other offerings** - In addition, Hewlett Packard Enterprise highly recommends HPE Education Services (for customer training and education) and additional Pointnext, as well as in-depth installation or implementation services as may be needed.

Resources and additional links

MapR, https://mapr.com/

MapR 6.x, https://maprdocs.mapr.com/home/

HPE Solutions for Apache Hadoop, hpe.com/info/hadoop

Hadoop and Vertica, hpe.com/info/vertica

HPE Insight Cluster Management Utility (CMU), hpe.com/info/cmu

HPE FlexFabric 5900 switch series, hpe.com/networking/5900

HPE FlexFabric 5940 switch series, hpe.com/us/en/product-catalog/networking/networking-switches/pip.hpe-flexfabric-5940-switch-series.1009148840.html

HPE FlexFabric 5950 switch series, hpe.com/us/en/product-catalog/networking/networking-switches/pip.hpe-flexfabric-5950-switch-series.1008901775.html

HPE ProLiant servers, hpe.com/info/proliant

HPE Networking, hpe.com/networking

HPE Services, hpe.com/services

Red Hat, redhat.com

HPE EPA Sizing tool: HPE EPA Sizing Tool

HPE Education Services: http://h10076.www1.hpe.com/ww/en/training/portfolio/bigdata.html

To help us improve our documents, please provide feedback at hpe.com/contact/feedback.

About MapR

Headquartered in San Jose, Calif., MapR provides the industry's only Converged Data Platform that enables customers to harness the power of Big Data by combining analytics in real-time to operational applications to improve business outcomes. With MapR, enterprises have an unparalleled data management platform for undertaking digital transformation initiatives to achieve competitive edge. World-class companies have realized more than five times their return on investment using MapR. Amazon, Cisco, Google, Microsoft, SAP and other leading businesses are part of the global MapR partner ecosystem. For more information, visit MapR.com.

Sign up for updates

Hewlett Packard
Enterprise

a00048919enw, June 2018