Hewlett Packard Enterprise

HPE Big Data Reference Architecture for Hortonworks Data Platform 2.3

Using HPE Apollo 2000 and Apollo 4200 Servers

Hortonworks®

Technical white paper

Contents

Executive summary

This white paper describes a big data solution deploying the Hortonworks Data Platform (HDP) on the Hewlett Packard Enterprise Big Data Reference Architecture (BDRA). The paper also describes the HPE Apollo 2000 compute system and the Apollo 4200 storage nodes as key components of the solution and highlights recognizable benefits of using Hortonworks Data Platform in the HPE BDRA. In addition to simplifying the procurement process, this paper also provides guidelines for configuring HDP once the system has been deployed.

There is an ever-growing need for a scalable, modern architecture for the consolidation, storage, access, and processing of big data analytics. Big data solutions using Hadoop are evolving from a simple model where each application was deployed on a dedicated cluster of identical nodes into a complex model where applications are deployed on a cluster of asymmetric nodes. Analytics and processing engines themselves have grown from MapReduce to a broader set now including Spark, Hive and Tez. By integrating the significant advances that have occurred in fabrics, storage, container-based resource management, and workload-optimized servers since the inception of the Hadoop architecture in 2005, our BDRA provides a cost-effective and flexible solution to optimize computing infrastructure in response to these ever-changing requirements in the Hadoop ecosystem. This implementation focuses on Apache YARN, Hive and Tez. These technologies and how they are configured are described in detail in this reference architecture.

The combination of HDP 2.3 running Hive with Tez on YARN on HPE BDRA provides the ideal blend of speed, flexibility, scalability and optimization of operations needed by today's enterprises: a high performance data warehouse infrastructure built on top of Hadoop that provides data summarization, query and analysis, supports analysis of large datasets stored in HDFS, and provides a SQL-like interface. See the section Benefits of the HPE BDRA solution in this paper for more on the advantages of this solution.

Target audience: This paper is intended for decision makers, system and solution architects, system administrators and experienced users that are interested in reducing design time or simplifying the purchase of a big data architecture containing both HPE and Hortonworks components. An intermediate knowledge of Apache Hadoop and scale-out infrastructure is recommended.

Document purpose: The purpose of this document is to describe the optimal way to configure Hortonworks Data Platform services including HDFS, YARN, Hive, and Tez on the BDRA solution.

Collaborative testing between HPE and Hortonworks Engineering for this reference architecture was performed in October 2015.

Introduction

As companies grow their big data implementations, they often find themselves deploying multiple clusters to support their needs. This could be to support different big data environments (MapReduce, Spark, NoSQL DBs, MPP DBMSs, etc.), to support rigid workload partitioning for departmental requirements or simply as a byproduct of multi-generational hardware. These multiple clusters often lead to data duplication and movement of large amounts of data between systems to accomplish an organization's business requirements. Many enterprise customers are searching for a way to recapture some of the traditional benefits of a converged infrastructure, such as the ability to more easily share data between different applications running on different platforms, the ability to scale compute and storage separately and the ability to rapidly provision new servers without repartitioning data to achieve optimal performance.

To address these needs, HPE engineers challenged the traditional Hadoop architecture which always co-locates compute elements in a server with data. While this approach works, the real power of Hadoop is that tasks run against specific slices of data without the need for coordination. Consequently, the distributed lock management and distributed cache management found in older parallel database designs is no longer required.

In fact, in a modern server, there is often more network bandwidth available to ship data off the server than there is bandwidth to disk. The HPE BDRA for HDP is deployed as an asymmetric cluster with some nodes dedicated to YARN compute resources and others dedicated to Hadoop Distributed File System (HDFS). Hadoop still works the same way – tasks still have complete ownership of their portion of the data and functions are still being shipped to the data – but the computation is executed on a node optimized for this task and the file system operations are executed on a node that is optimized for that work. Of particular interest is that this approach can actually perform better than a traditional Hadoop cluster for several reasons. For more information on this architecture and the benefits that it provides please see the overview master document at http://h20195.www2.hpe.com/V2/GetDocument.aspx?docname=4AA5-6141ENW.

Solution overview

The HPE BDRA is a highly optimized configuration built using unique servers offered by HPE – the Apollo 4200 Gen9 for the high density storage layer and the Apollo 2000 System with ProLiant XL170r nodes for the high density computational layer. This configuration is the result of a great deal of testing and optimization done by HPE engineers resulting in the right set of software, drivers, firmware and hardware to yield extremely high density and performance. As shown in Figure 1, this architecture is changing the economics of work distribution in big data.

Simply deploying Hadoop onto a collection of traditional servers in an asymmetric fashion will not yield the kind of benefits that is seen with HPE BDRA. In order to simplify the build for customers, HPE provides the exact bill of materials in this document to allow a customer to purchase this complete solution. HPE recommends that customers purchase the option in which HPE Technical Services Consulting will install the prebuilt operating system images, verify all firmware and versions are correctly installed, 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 Hortonworks installation using the recommended guidelines in this document.

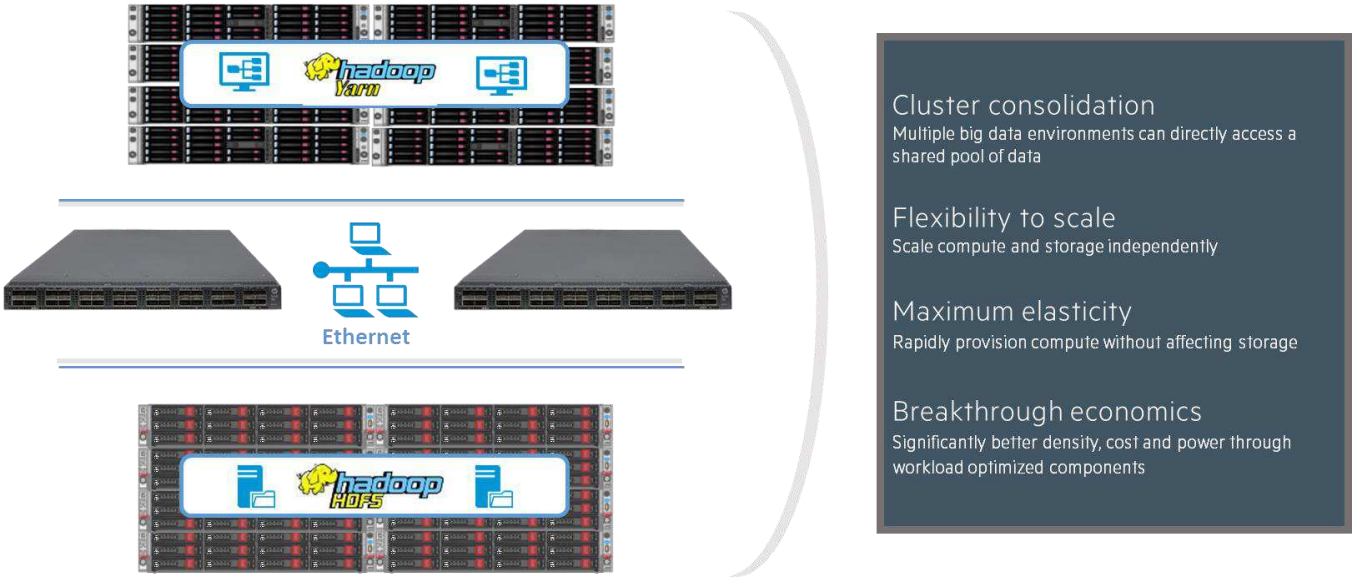

Figure 1. HPE BDRA, changing the economics of work distribution in big data

The HPE BDRA design is anchored by the following HPE technologies.

Storage nodes

Apollo 4200 Gen9 Servers make up the HDFS storage layer, providing a single repository for big data.

Compute nodes

Apollo 2000 System nodes deliver a scalable, high-density layer for compute tasks and provide a framework for workload-optimization with four ProLiant XL170r nodes on a single chassis.

High-speed networking separates compute nodes and storage nodes, creating an asymmetric architecture that allows each tier to be scaled individually; there is no commitment to a particular CPU/storage ratio. Since big data is no longer collocated with storage, Hadoop does not need to achieve node locality. However, rack locality works in exactly the same way as in a traditional converged infrastructure; that is, as long as one scales within a rack, overall scalability is not affected.

With compute and storage decoupled, many of the advantages of a traditional converged system can be enjoyed again. For example, compute and storage can be scaled independently simply by adding compute nodes or storage nodes. Testing carried out by HPE indicates that most workloads respond almost linearly to additional compute resources.

Hadoop YARN

YARN is a key feature of the latest generation of Hadoop and of the BDRA. It decouples MapReduce resource management and scheduling capabilities from the data processing components, allowing Hadoop to support more varied processing approaches and a broader array of applications.

New to YARN is a concept of labels for groupings of compute nodes. Jobs submitted through YARN can now be flagged to perform their work on a particular set of nodes when the appropriate label name is included with the job. Thus, an organization can now create groups of compute resources that are designed, built, or optimized for particular types of computational work, allowing for jobs to be passed out to groups of hardware that are more geared to the type of computation work. In addition, the use of labels allows for isolating compute resources that may be required to perform a high-priority job, for example, ensuring that sufficient resources are available at any given time.

Apache Hive

Hive is a data warehouse infrastructure built on top of Hadoop that provides data summarization, query and analysis. Apache Hive supports analysis of large datasets stored in HDFS. It provides a SQL-like language called HiveQL with schema on read and transparently converts queries to MapReduce and Apache Tez jobs.

Apache Tez

Tez is a new application framework that is a more flexible, powerful successor of the MapReduce framework. The application allows for a complex directed-acyclic-graph (DAG) of tasks for processing data. It is currently built atop Hadoop YARN. By allowing projects like Hive and Pig to run a complex DAG of tasks, Tez is used to process the same data in a single job that previously took MapReduce multiple jobs to accomplish.

Hortonworks Data Platform

HDP is a platform for multi-workload data processing. As shown in Figure 2, the platform can utilize a range of processing methods – from batch to interactive and real-time – all supported by solutions for governance, integration, security, and operations. HDP integrates with and augments solutions like HPE BDRA, allowing you to maximize the value of big data. HDP enables Open Enterprise Hadoop, a full suite of essential Hadoop capabilities in the following functional areas: data management, data access, data governance and integration, security, and operations.

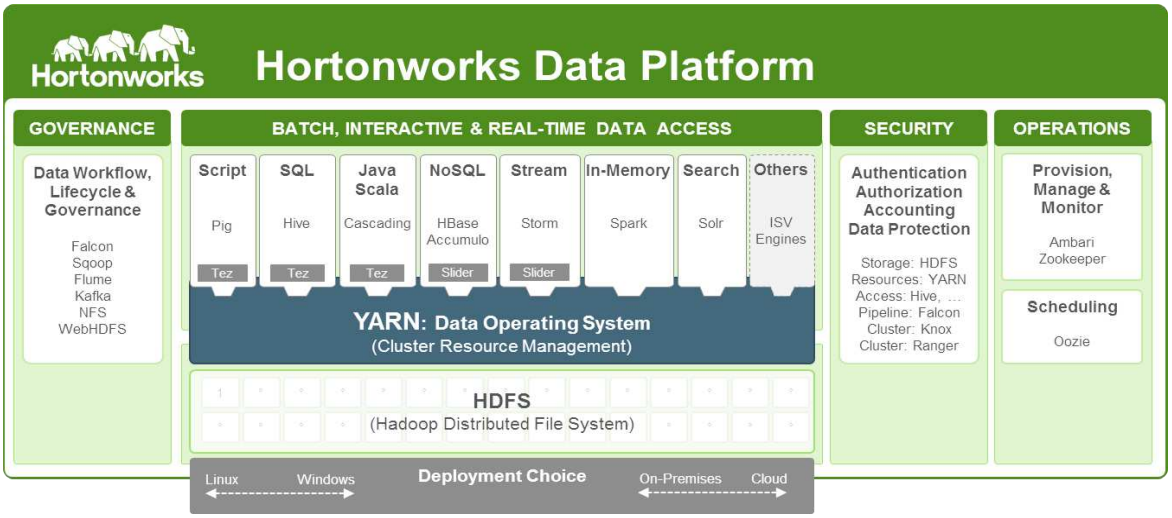

Figure 2. HDP blueprint for Enterprise Hadoop

Key highlights of HDP 2.3 include the following:

- Dramatic improvement in user experience
- Enhanced security
- Workload Management advancements
- Apache Hive improvements, including:
  - Batch and interactive queries via Apache Tez
  - Cost-based optimizer powered by Apache Calcite
  - Increased performance through Vectorized Map Joins
  - SQL: Union, Interval Types, CURRENT_TIMESTAMP, CURRENT_DATE
  - Usability through Configurations, Hive and Tez Views
- High-performance ETL via Pig and Tez
- Stream processing via Apache Storm and Apache Kafka
- Search via Apache Solr

For more information on HDP, refer to hortonworks.com/hdp.

Benefits of the HPE BDRA solution

While the most obvious benefits of the HPE BDRA solution center on density and price/performance, other benefits include:

- **Elasticity** – HPE BDRA is designed for flexibility. Compute nodes can be allocated very flexibly without redistributing data; for example, nodes can be allocated by time-of-day or even for a single job. The organization is no longer committed to yesterday's CPU/storage ratios, leading to much more flexibility in design and cost. Moreover, with HPE BDRA, the system is only grown where needed.
- **Consolidation** – HPE BDRA is based on HDFS, which has enough performance and can scale to large enough capacities to be the single source for big data within any organization. Various pools of data currently being used in big data projects can be consolidated into a single, central repository. YARN-compliant workloads access the big data directly via HDFS; other workloads can access the same data via appropriate connectors.
- **Workload-optimization** – There is no one go-to software for big data; instead there is a federation of data management tools. After selecting the appropriate tool to meet organizational requirements, run the job using the compute nodes that are best suited for the workload, such as low-power cartridges or compute-intense cartridges.
- **Enhanced capacity management** – Compute nodes can be provisioned on the fly, while storage nodes now constitute a smaller subset of the cluster and, as such, are less costly to overprovision. In addition, managing a single data repository rather than multiple different clusters reduces overall management costs.
- **Faster time-to-solution** – Processing big data typically requires the use of multiple data management tools. When these tools are deployed on conventional Hadoop clusters with dedicated – often fragmented – copies of the data, time-to-solution can be lengthy. With HPE BDRA, data is unfragmented and consolidated in a single data lake, allowing different tools to access the same data. Thus, more time can be spent on analysis, less on shipping data; therefore time-to-solution is typically faster.

Solution components

Figure 3 provides a basic conceptual diagram of HPE BDRA.

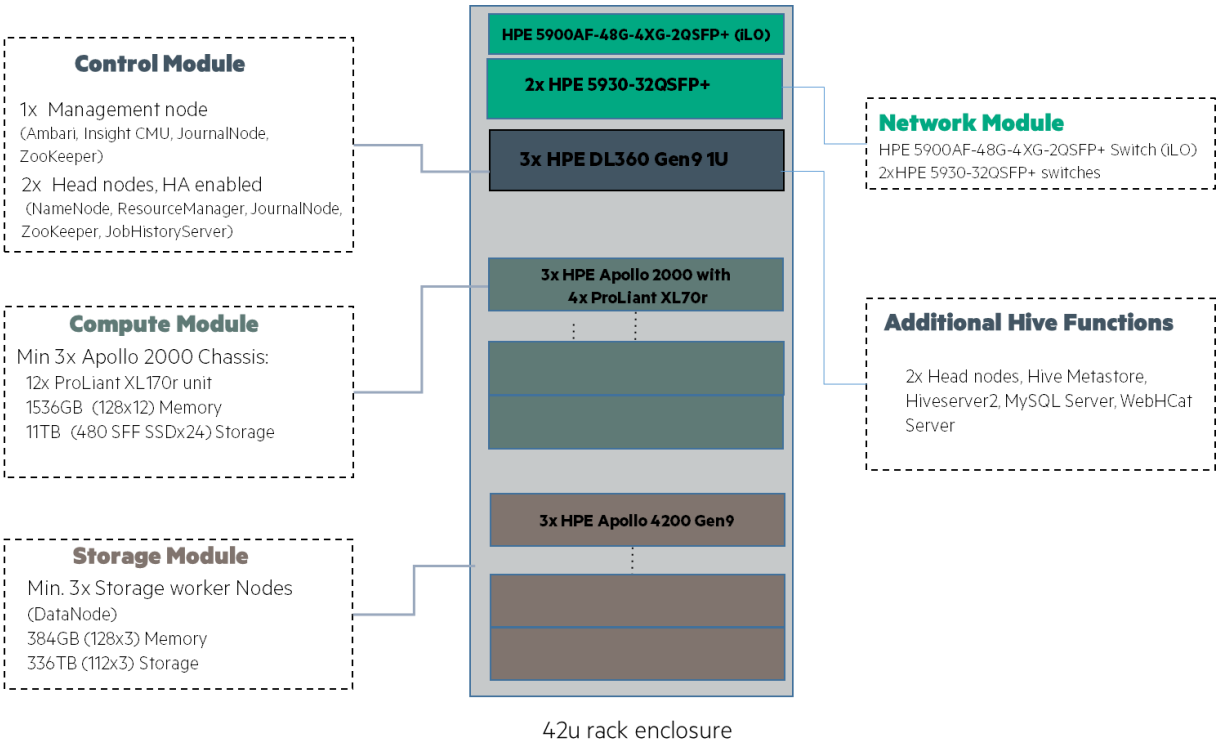

Figure 3. HPE BDRA conceptual diagram

For full BOM listings of products selected for the proof of concept, refer to the Bill of materials section of this white paper.

Minimum configuration

Figure 4 shows the minimum recommended HPE BDRA configuration, with 12 worker nodes and 3 storage nodes housed in a single 42U rack.

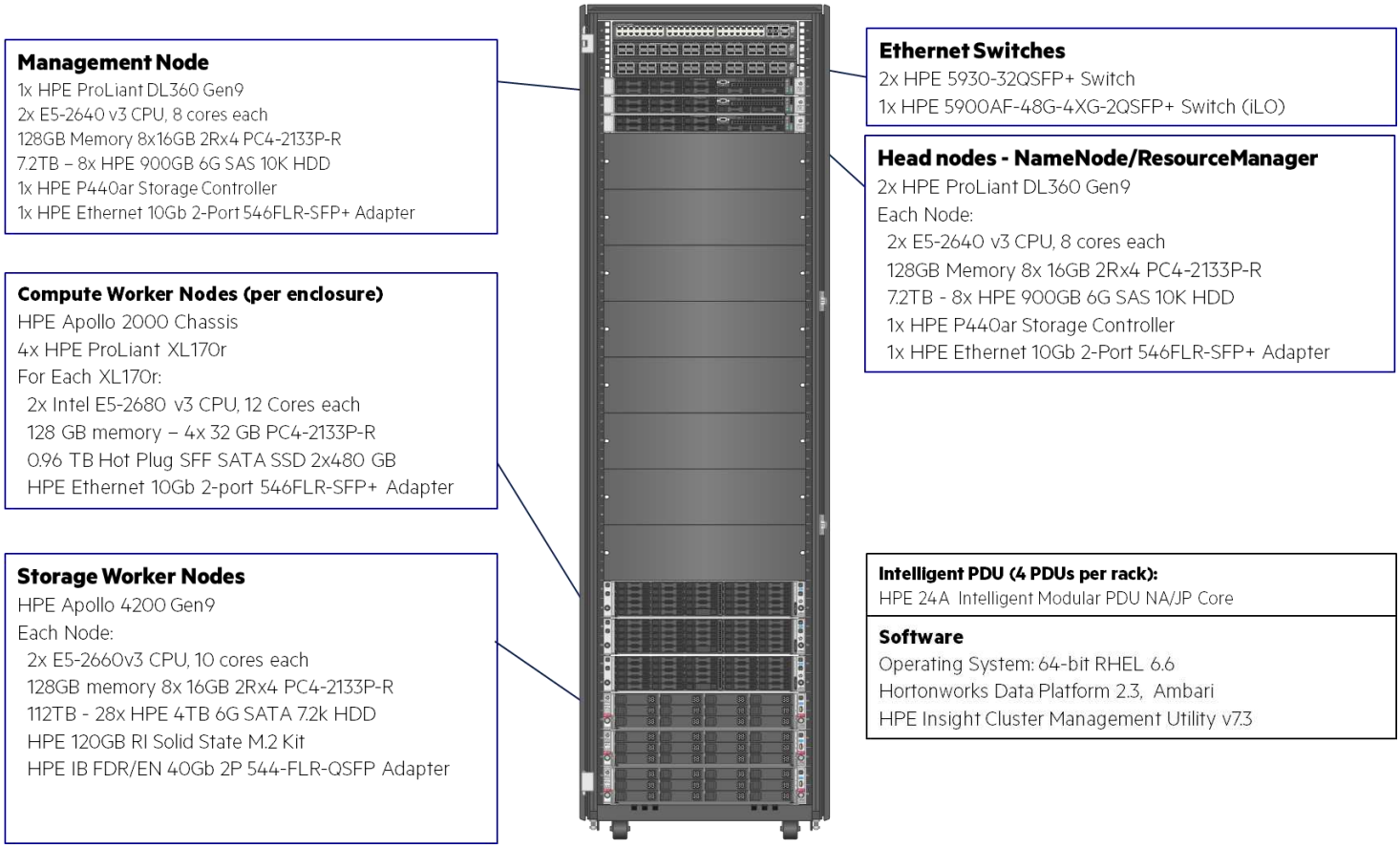

Figure 4. Base HPE BDRA configuration, with detailed lists of components

Best practice

HPE recommends starting with three Apollo 2000 chassis consisting of 12 ProLiant XL170r nodes with 128 GB of memory for each node, combined with three Apollo 4200 Gen9 Servers as storage nodes. For Hive/Tez installations, consider increasing the memory to 256 GB in the compute nodes.

The following nodes are used in the base HPE BDRA configuration.

Compute nodes

The base BDRA configuration features three Apollo 2000 Chassis containing a total of 12 ProLiant XL170r Gen9 hot-pluggable server nodes as shown in Figures 5 and 6.

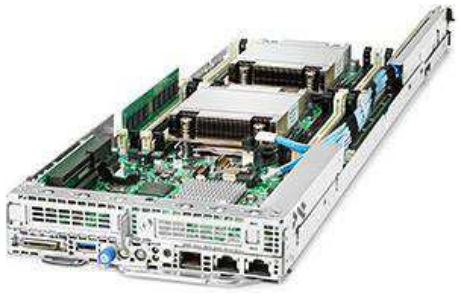

Figure 5. HPE ProLiant XL170r Gen9

Figure 6. HPE Apollo 2000 chassis (back and front)

The HPE Apollo 2000 System is a dense solution with four independent HPE ProLiant XL170r Gen9 hot-pluggable server nodes in a standard 2U chassis. Each ProLiant XL170r Gen9 Server node is serviced individually without impacting the operation of other nodes sharing the same chassis to provide increased server uptime. Each server node harnesses the performance of 2133 MHz memory (16 DIMM slots per node) and dual Intel® Xeon® E5-2600 v3 processors in a very efficient solution that shares both power and cooling infrastructure. Other features of the ProLiant XL170r Gen9 server include:

- Support for high-performance memory (DDR4) and Intel Xeon E5-2600 v3 processor up to 18C, 145W
- Additional PCIe riser options for flexible and balanced I/O configurations
- FlexibleLOM feature for additional network expansion options
- Support for dual M.2 drives

For more information on Apollo 2000 Chassis, visit http://www8.hp.com/us/en/products/proliant-servers/product-detail.html?oid=7832023.

For more information on the ProLiant XL170r Gen9 server, visit hpe.com/servers/xl170r

Each of these compute nodes typically runs YARN NodeManager.

Note

Alternate compute nodes: HPE ProLiant m710p server cartridges (Intel Xeon E3-1284L v4 2.9GHz-3.8GHz) housed on HPE Moonshot 1500 chassis provides users with an alternate choice of compute servers to use in the HPE BDRA configuration for denser compute nodes. The ProLiant m710p cartridges come with Intel's 4th Generation Broadwell chips. For BOM information, refer to Appendix A: Table A-4.

Storage nodes

There are three Apollo 4200 Gen9 servers. Each server is configured with 28 LFF disks; each typically runs HDFS DataNode. The Apollo 4200 Gen9 storage server is shown in Figure 7.

Figure 7. HPE Apollo 4200 Gen9 server

The HPE Apollo 4200 Gen9 Server offers revolutionary storage density for data intensive workloads such as Apache Hive on Hadoop HDFS. The Apollo 4200 allows you to save valuable data center space through its unique density optimized 2U form factor which holds up to 28 LFF disks and with capacity for up to 224 TB per server. It has the ability to grow your big data solutions with an infrastructure that is ready to scale. Another benefit is that the Apollo 4200 fits easily into standard racks with a depth of 32-inches per server – no special racks are required.

For more detailed information, visit http://www8.hp.com/us/en/products/proliant-servers/product-detail.html?oid=8261831.

The storage controllers in the Apollo 4200 and ProLiant XL170r support the HPE Secure Encryption. HPE Secure Encryption is a Smart Array controller-based data encryption solution. It provides encryption for data at rest, an important component for complying with government regulations which have data privacy requirements, such as HIPAA and Sarbanes-Oxley.

Management/head nodes

Three HPE ProLiant DL360 Gen9 servers are configured as management/head nodes with the following functionality:

- Management node with Ambari, HPE Insight Cluster Management Utility (Insight CMU), ZooKeeper, and JournalNode
- Head node 1 with active ResourceManager/standby NameNode, ZooKeeper, JobHistoryServer and JournalNode. Additional functions for Hive: Hiveserver2 and MySQL Server
- Head node 2 with active NameNode/standby ResourceManager, ZooKeeper, JobHistoryServer and JournalNode. Additional functions for Hive: Hive Metastore and WebHCat Server.

Power and cooling

When planning large clusters, it is important to properly manage power redundancy and distribution. To ensure servers and racks have adequate power redundancy, HPE recommends that each Apollo 2000 chassis, each Apollo 4200 Gen9 server, and each ProLiant DL360 Gen9 server should have a backup power supply and each rack should have at least two Power Distribution Units (PDUs).

There is additional cost associated with procuring redundant power supplies; however, the need for redundancy is less critical in larger clusters where the inherent redundancy within HDP ensures there would be less impact in the event of a failure.

Best practice

For each chassis, Apollo 4200 Gen9 and ProLiant DL360 Gen9 server, HPE recommends connecting each of the device's power supplies to a different PDU. Furthermore, PDUs should each be connected to a separate data center power line to protect the infrastructure from a line failure.

Distributing server power supply connections evenly to the PDUs while also distributing PDU connections evenly to data center power lines ensures an even power distribution in the data center and avoids overloading any single 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.

Networking

As shown in Figure 8, two IRF-bonded HPE 5930-32QSFP+ switches are specified in each rack for high performance and redundancy. Each provides six 40GbE uplinks that can be used to connect to the desired network or, in a multi-rack configuration, to another pair of 5930-32QSFP+ switches that are used for aggregation. The other 22 ports are available for Hadoop nodes.

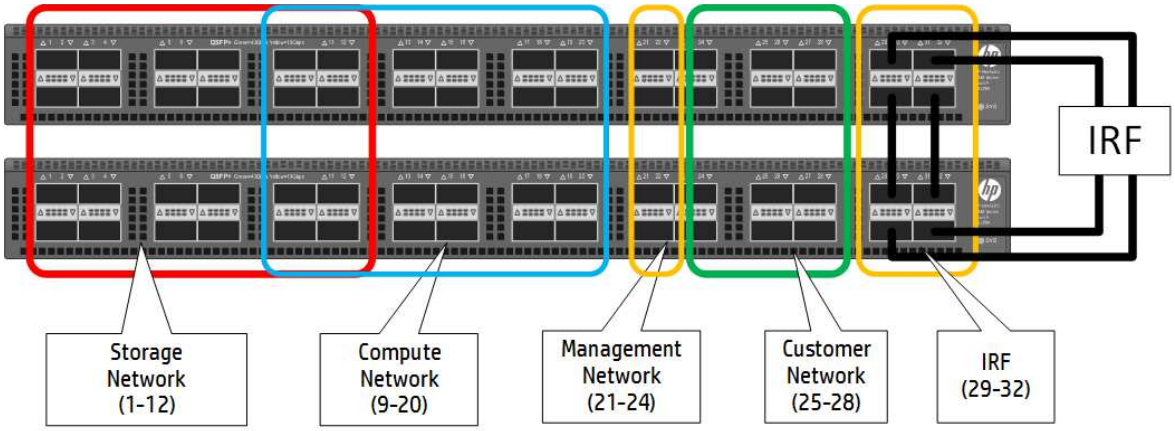

Figure 8. HPE 5930-32QSFP+ switch port allocation based upon node type

Note

IRF-bonding requires four 40GbE ports per switch, leaving six 40GbE ports on each switch for uplinks. QSFP+ 4x10G SFP+ DAC cables are used to connect Compute and Management nodes to the QSFP+ ports on the HPE 5930 switches. Therefore a single QSFP+ port on the 5930 switch handles 4 compute node connections.

iLO network

A single HPE 5900 switch is used exclusively to provide connectivity to HPE Integrated Lights-Out (iLO) management ports, which run at or below 1GbE. The iLO network is used for system provisioning and maintenance.

Cluster isolation and access configuration

It is important to isolate the Hadoop cluster so that external network traffic does not affect the performance of the cluster. In addition, isolation allows the Hadoop cluster to be managed independently from its users, ensuring that the cluster administrator is the only person able to make changes to the cluster configuration.

Thus, HPE recommends deploying ResourceManager, NameNode, and Worker nodes on their own private Hadoop cluster subnet.

Key point

Once a Hadoop cluster is isolated, the users still need a way to access the cluster and submit jobs to it. To achieve this, HPE recommends multi-homing the management node so that it can participate in both the Hadoop cluster subnet and a subnet belonging to users.

Ambari is a web application that runs on the management node, allowing users to manage and configure the Hadoop cluster and view the status of jobs without being on the same subnet – provided that the management node is multi-homed. Furthermore, this approach allows users to shell into the management node and run Apache Pig or Apache Hive command line interfaces and submit jobs to the cluster in that way.

Staging data

After the Hadoop cluster has been isolated on its own private network, you must determine how to access HDFS in order to ingest data. The HDFS client must be able to reach every Hadoop DataNode in the cluster in order to stream blocks of data on to the HDFS.

HPE BDRA provides the following options for staging data:

- **Using the management node** – The already multi-homed management server is used to stage data. HPE recommends configuring this server with eight disks to provide a sufficient amount of disk capacity to provide a staging area for ingesting data into the Hadoop cluster from another subnet.
- **Using the ToR switches** – Use the remaining 40GbE ports on the ToR switches for multi-homed systems outside the Hadoop cluster to move data into the cluster.
- **Using WebHDFS** – WebHDFS provides HTTP access to securely read and write data to and from HDFS. For more information on WebHDFS, refer to http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.2.4/bk_hdfs_admin_tools/content/ch11.html.

Note

The benefit of using dual-homed management node(s) to isolate in-cluster Hadoop traffic from the ETL traffic flowing to the cluster may be debated. This enhances security; however, it may result in ETL performance/connectivity issues, since relatively few nodes are capable of ingesting data. For example, you may wish to initiate Sqoop tasks on the compute nodes to ingest data from an external RDBMS in order to maximize the ingest rate. However, this approach requires worker nodes to be exposed to the external network to parallelize data ingestion, which is less secure.

Consider the options before committing to an optimal network design for your particular environment. HPE recommends dual-homing the entire cluster, allowing every node to ingest data.

Capacity and sizing

Expanding the base configuration

As needed, compute and/or storage nodes can be added to a base HPE BDRA configuration. The minimum HPE BDRA configuration (refer to Figure 4) can expand to include a total of eighteen mixed chassis in a single rack. There are a number of options for configuring a single-rack HPE BDRA solution, ranging from hot (with a large number of compute nodes and minimal storage) to cold (with a large number of storage nodes and minimal compute).

Multi-rack configuration

The single-rack BDRA configuration is designed to perform well as a standalone solution but also form the basis of a much larger multi-rack solution, as shown in Figure 9. When moving from a single rack to a multi-rack solution, you simply add racks without having to change any components within the base rack.

A multi-rack solution assumes the base rack is already in place and extends its scalability. For example, the base rack already provides sufficient management services for a large scale-out system.

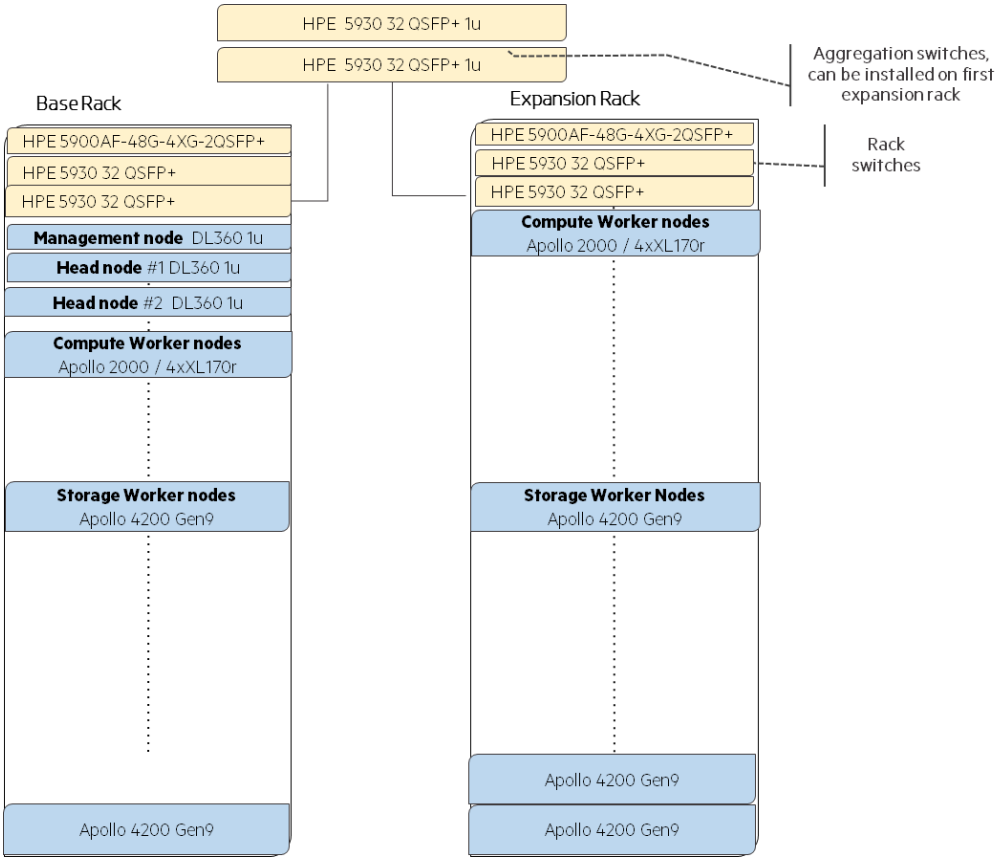

Figure 9. Multi-rack HPE BDRA, extending the capabilities of a single rack

Note

While much of the architecture for the multi-rack solution was borrowed from the single-rack design, the architecture suggested here for multi-rack solutions is based on previous iterations of Hadoop testing on the ProLiant DL380e platform rather than Apollo servers. It is provided here as a general guideline for designing multi-rack Hadoop clusters.

Extending networking in a multi-rack configuration

For performance and redundancy, two HPE 5930-32QSFP+ ToR switches are specified per expansion rack. The 5930-32QSFP+ switch includes up to six 40GbE uplinks that can be used to connect these switches to the desired network via a pair of 5930-32QSFP+ aggregation switches.

Guidelines for calculating storage needs

Hadoop cluster storage sizing requires careful planning, based on the identification of current and future storage and compute needs. The following are general considerations for data inventory:

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

To calculate your storage needs, you should identify the number of TB of data needed per day, week, month, and year; and then add the ingestion rates of all data sources.

It makes sense to identify storage requirements for the short-, medium-, and long-term.

Another important consideration is data retention – both size and duration. Which data must you keep? For how long?

In addition, consider maximum fill-rate and file system format space requirements on hard drives when estimating storage size.

Configuration guidance

The instructions provided below assume that the Hadoop cluster has already been created on an HPE BDRA solution. They are intended to assist in optimizing the setup for the various HDP services on this reference architecture.

The Hive and Tez configuration settings were derived through testing using a set of queries of varying complexity from a hive database schema consisting of multiple dimension and fact tables. It is recommended that the settings shown below be used as a starting point for future test scenarios.

Setting up HDP

Compute node components

Management and head nodes contain the software shown in Tables 1 and 2.

Table 1. Management node base software components

SOFTWARE	DESCRIPTION
Red Hat® Enterprise Linux® (RHEL) 6.6	Recommended operating system
Oracle Java Development Kit	JDK
Ambari Server	Hortonworks HDP management
Hadoop Client components	Used to initiate the various Hadoop jobs.
ZooKeeper	Hadoop consensus management utility
Insight CMU	Cluster Management Utility

Table 2. Head node base software components

SOFTWARE	DESCRIPTION
Red Hat Enterprise Linux (RHEL) 6.6	Recommended operating system
Oracle Java Development Kit	JDK
Ambari Agent	Hortonworks HDP management
NameNode	HDFS name node Service
SNameNode	Secondary HDFS name node service
ResourceManager	YARN resource management service
HiveServer2	Apache Hive server
Hive Metastore	Stores metadata for Hive tables and partitions
MySQL Server	Database
WebHcat Server	web API for HCatalog and related Hadoop components
ZooKeeper	Hadoop consensus management utility

Compute node components

Compute nodes contain the software shown in Table 3.

Table 3. Compute node base software components

SOFTWARE	DESCRIPTION
Red Hat Enterprise Linux (RHEL) 6.6	Recommended operating system
Oracle Java Development Kit	JDK
Ambari Agent	Hortonworks HDP management
NodeManager	NodeManager process for MR2/YARN

Storage node components

Storage nodes contain the software shown in Table 4.

Table 4. Storage node base software components

SOFTWARE	DESCRIPTION
RHEL 6.6	Recommended operating system
Oracle Java Development it	JDK
Ambari Agent	Hortonworks HDP management
DataNode	DataNode process for HDFS

Configuring HDFS

Make the following changes to the HDFS configuration:

- Increase the dfs.blocksize value to allow more data to be processed by each map task, thus reducing the total number of mappers and NameNode memory consumption.

  `dfs.block.size, dfs.blocksize 536870912 (default 134217728)`

- Increase the dfs.namenode.handler.count value to better manage multiple HDFS operations from multiple clients.

  `dfs.namenode.handler.count 180 (default 30)`

- Increase numbers of retries for network-intensive environments. Make the following changes in the HDFS Custom HDFS site:

  `dfs.client.block.write.locateFollowingBlock.retries 35`

Configuring MapReduce2 and YARN

The changes described below optimize the compute nodes for YARN and MapReduce workloads. For a cluster of 12 ProLiant XL170r compute nodes with two E5-2680 v3 processors each, these changes will allow the YARN scheduler to schedule 574 maps for the cluster, which averages out to about 48 maps per node, and 286 reduces for the cluster which averages out to about 24 reduces per node. You can use the following formula to calculate the number of maps and reduces you should use for MapReduce/YARN workloads.

`Maps: (Number of vcores per Compute Node * 2 * Number of Compute Nodes) - 2`

`Reduces: (Number of vcores per Compute Node * Number of Compute Nodes) - 2`

Note

The YARN application master requires the equivalent of 2 map or reduce slots, which is indicated in the above formulas.

MapReduce2

HPE recommends the following changes in the MapReduce2 Service.

- Specify the amount of memory for a job's map task:

  `Default virtual memory for a job's map-task (maximum container size / vcores / 2)`

- Specify the amount of memory for a job's reduce task:

  `Default virtual memory for a job's reduce-task (maximum container size / vcores)`

- Enable compression of Map Containers:

  `mapreduce.map.output.compress=true`

- Specify Snappy Codec to be used for compressing Map Containers: This parameter is added to Custom mapred-site:

  `mapreduce.map.output.compress.codec org.apache.hadoop.io.compress.SnappyCodec`

- Specify that the MapReduce file output be compressed:

  `mapreduce.output.fileoutputformat.compress=true`

YARN

Make the following changes in the YARN Service.

Capacity Scheduler

- Specify the amount of memory for all YARN containers on a node, and the maximum amount of memory a YARN container can use. Typically, this is the total memory of the node less about 5% for the OS to use.

  ```
  yarn.nodemanager.resource.memory.mb=Amount of memory dedicated to YARN. (0.95% Total Memory of a node).
  yarn.scheduler.maximum-allocation-mb=Amount of memory dedicated to YARN. (0.95% Total Memory of a node).
  ```

- Specify the number of cores available on compute nodes:

  ```
  yarn.nodemanager.resource.cpu-vcores=Number of processor cores on a node.
  ```

- The node-locality-delay specifies how many scheduling intervals to let pass attempting to find a node local slot to run on prior to searching for a rack local slot. This setting is very important for small jobs that do not have a large number of maps or reduces as it will better utilize the Compute Nodes. We highly recommend this value be set to 1.

  ```
  yarn.scheduler.capacity.node-locality-delay=1
  ```

Configuring Hive

The following parameters can be configured through Ambari, a direct modification of `hive-site.xml`, or at the time of executing particular queries; they allow for the use of some base settings globally and also for the ability to override the settings for specific queries that may require special tuning.

- Enables the cost based optimizer:

  ```
  hive.cbo.enable=true
  ```

- Enables tez as the driving engine:

  ```
  hive.execution.engine=tez
  ```

- Enables auto gather stats for the columns during inserts/updates for better calculations in cost based optimizer:

  ```
  hive.stats.fetch.column.stats=true
  hive.compute.query.using.stats=true
  ```

- Allows vectorized execution:

  ```
  hive.vectorized.execution.enabled=true
  ```

- Enable auto joining:

  ```
  hive.auto.convert.join.noconditionaltask=true
  hive.auto.convert.sortmerge.join=true
  hive.enforce.sortmergebucketmapjoin=true
  ```

- Turn off speculative execution for reducers:

  ```
  hive.mapred.reduce.tasks.speculative.execution=false
  ```

- Specific tunings for test database (10TB database):

  ```
  Set hive.auto.convert.join.noconditionaltask.size to the sum of sizes of tables that can be converted into hash maps that fit in memory (n-1 tables in the join).
  hive.auto.convert.join.noconditionaltask.size=32000000
  ```

Configuring Tez

Similar to tuning of Hive itself, the following Tez settings can be made system wide (`tez-site.xml`) through Ambari, or at the command line when executing queries through HiveQL.

- Tez application manager resources settings. (Most queries will execute with 4096 MB memory but as the queries get more complex this may need to be raised to 6144 or even 8192.)

  ```
  tez.am.resource.memory.mb=4096
  ```

  ```
  tez.am.container.reuse.locality.delay-allocation-millis=1
  ```

- Set the container size for Tez jobs to override the default map/reduce settings:

  ```
  hive.tez.container.size=(Maximum Container Memory / vcores / 2)
  ```

Configuring compression

HPE recommends using compression for map outputs since it reduces file size on disk and speeds up disk and network I/Os. Using Snappy compressions, set the following MapReduce parameters to compress job output files:

```
mapreduce.output.fileoutputformat.compress=true
```

```
mapreduce.map.output.compress.codec=org.apache.hadoop.io.compress.SnappyCodec
```

```
mapreduce.map.output.compress=true
```

Compression for Hive

Set the following Hive parameters to compress the Hive output files using Snappy compression:

```
hive.exec.compress.output=true
```

```
hive.exec.orc.default.compress=SNAPPY
```

Hive/Tez concurrency

For multiple concurrent short running queries, HPE recommends modifying the capacity scheduler and the Hive/Tez configuration parameters to represent the number of queries running. The capacity scheduler changes would involve the creation of additional queues and limit the resources used by those queues so that the cluster isn't completely taken over by a single query. The hive setting changes would modify the following parameters:

```
hive.server2.tez.default.queues=<comma separated list of queues to run on>
```

```
hive.server2.tez.sessions.per.default.queue=<#concurrent queries/#of queues>
```

Use the Hive concurrency settings shown in Table 5 as a guideline for the number of queues and number of sessions per queue.

Table 5. Hive concurrency settings

NUMBER OF CONCURRENT QUERIES	NUMBER OF QUEUES	NUMBER OF SESSIONS PER QUEUE
Less than 10	2-5	1 or 2
10	5	2
15	5	3
20	5	4

For example, assume that typically there are four concurrent short running queries:

Capacity Scheduler changes (additional queues)

```
yarn.scheduler.capacity.root.queues=hive1,hive2
```

```
yarn.scheduler.capacity.root.hive1.capacity=50
```

```
yarn.scheduler.capacity.root.hive2.capacity=50
```

Hive parameter changes (assigning YARN capacity scheduler queues to hive and the number of sessions per queue)

```
hive.tez.server2.defualt.queues=hive1,hive2
```

```
hive.tez.sessions.per.default.queue=2
```

This configuration would allow for four concurrent queries with no query being able to use more than 50% of the total capacity of the cluster.

Bill of materials

The BOMs outlined in Tables 6-13 below are based on the recommended configuration for a single rack reference architecture with the following key components:

- One management node
- Two head nodes
- Twelve compute nodes (refer to Appendix A – Alternate compute node components)
- Three storage nodes (refer to Appendix B – Alternate storage node components)
- Two ToR switches
- One HPE iLO switch
- Hortonworks Data Platform (HDP 2.3 was tested)

The following BOMs contain electronic license to use (E-LTU) parts. Electronic software license delivery is now available in most countries. HPE recommends purchasing electronic products over physical products (when available) for faster delivery and for the convenience of not tracking and managing confidential paper licenses. For more information, please contact your reseller or an HPE representative.

Note
Part numbers listed are current as of the time of publication and subject to change. The bill of materials does not include complete support options or other rack and power requirements. If you have questions regarding ordering, please consult with your HPE Reseller or HPE Sales Representative for more details. hpe.com/us/en/services/consulting.html

Management node and head nodes

Important
Table 6 provides a BOM for one HPE ProLiant DL360 Gen9 server. The minimum recommended solution and the tested solution both feature one management server and two head nodes, requiring a total of three ProLiant DL360 Gen9 servers.

Table 6. BOM for a single HPE ProLiant DL360 Gen9 server

QTY	PART NUMBER	DESCRIPTION
1	755258-B21	HPE DL360 Gen9 8SFF CTO Server
1	755386-L21	HPE DL360 Gen9 E5-2640v3 FIO Kit
1	755386-B21	HPE DL360 Gen9 E5-2640v3 Kit
8	726719-B21	HPE 16GB 2Rx4 PC4-2133P-R Kit
8	652589-B21	HPE 900GB 6G SAS 10K 2.5in SC ENT HDD
1	749974-B21	HPE Smart Array P440ar/2G FIO Controller
1	779799-B21	HPE Ethernet 10Gb 2P 546FLR-SFP+ Adptr
1	734807-B21	HPE 1U SFF Easy Install Rail Kit
2	720478-B21	HPE 500W FS Plat Ht Plg Pwr Supply Kit
1	764646-B21	HPE DL360 Gen9 Serial Cable
1	764636-B21	HPE DL360 Gen9 SFF Sys Insght Dsply Kit
2	AF595A	HPE 3.0M, Blue, CAT6 STP, Cable Data

Compute nodes

Note

Table 7 provides a BOM for one Apollo 2000 chassis with four ProLiant XL170r servers. The minimum recommended solution features three Apollo chassis with a total of 12 ProLiant XL170r servers.

Table 7. BOM for a single Apollo 2000 with four ProLiant XL170r servers

QTY	PART NUMBER	DESCRIPTION
1	798153-B21	HPE Apollo r2600 24SFF CTO Chassis
2	800059-B21	HPE Apollo 2000 FAN-module Kit
4	798155-B21	HPE ProLiant XL170r Gen9 CTO Svr
4	793028-L21	HPE XL1x0r Gen9 E5-2680v3 FIO Kit
4	793028-B21	HPE XL1x0r Gen9 E5-2680v3 Kit
16	728629-B21	HPE 32GB 2Rx4 PC4-2133P-R Kit
4	779799-B21	HPE Ethernet 10Gb 2-port 546FLR-SFP+ Adapter
8	832414-B21	HPE 480GB 6Gb SATA 2.5in MU-2 SC SSD
4	798178-B21	HPE XL1x0r Gen9 LP PCIex16 L Riser Kit
4	798180-B21	HPE XL170r FLOM x8 R Riser Kit
4	798192-B21	HPE XL170r/190r Dedicated NIC IM Board Kit
4	800060-B21	HPE XL170r Mini-SAS B140 Cbl Kit
2	720620-B21	HPE 1400W FS Plat Pl Ht Plg Pwr Spply Kit
1	740713-B21	HPE t2500 Strap Shipping Bracket
1	611428-B21	HPE DL2000 Hardware Rail Kit
8	AF595A	HPE 3.0M, Blue, CAT6 STP, Cable Data
2	JG330A	HPE X240 QSFP+ 4x10G SFP+ 3m DAC Cable (2 per 798153-B21)

Storage nodes

Important

Table 8 provides a BOM for one Apollo 4200 Gen9 server. The minimum recommended solution features three Apollo 4200 Gen9 servers. (See Appendix B for the BOM for alternate components for storage nodes.)

Table 8. BOM for a single Apollo 4200 Gen9 server

QTY	PART NUMBER	DESCRIPTION
1	808027-B21	HPE Apollo 4200 Gen9 24LFF CTO Svr
1	803311-L21	HPE Apollo 4200 Gen9 E5-2660v3 FIO Kit
1	803311-B21	HPE Apollo 4200 Gen9 E5-2660v3 Kit
8	726719-B21	HPE 16GB 2Rx4 PC4-2133P-R Kit
1	806563-B21	HPE Apollo 4200 Gen9 LFF Rear HDD Cage Kit
28	797265-B21	HPE 4TB 6G SATA 7.2k 3.5in MDL LP HDD
1	813546-B21	HPE SAS Controller Mode for Rear Storage
1	764285-B21	HPE IB FDR/EN 40Gb 2P 544_FLR-QSFP Adapter
2	720479-B21	HPE 800W FS Plat Hot Plug Power Supply Kit
1	806565-B21	HPE Apollo 4200 Gen9 IM Card Kit
1	788028-B21	HPE 120GB RI Solid State M.2 Kit
1	806562-B21	HPE Apollo 4200 Gen9 Redundant Fan Kit
1	822731-B21	HPE 2U Shelf-Mount Adjustable Rail Kit
2	JG327A	HPE X240 40G QSFP+ QSFP+ 3m DAC Cable
2	AF595A	HPE 3.0M, Blue, CAT6 STP, Cable Data

Networking

Table 9 provides a BOM for two ToR switches and one iLO switch, as featured in the tested configuration.

Table 9. BOM for two 5930 switches (ToR) and one 5900 switch (HPE iLO)

QTY	PART NUMBER	DESCRIPTION
2	JG726A	HPE FF 5930-32QSFP+ Switch
4	JG553A	HPE X712 Bck(pwr)-Frt(prt) HV Fan Tray
4	JC680A	HPE A58x0AF 650W AC Power Supply
1	JG510A	HPE 5900AF-48G-4XG-2QSFP+ Switch
2	JC680A	HPE A58x0AF 650W AC Power Supply
2	JC682A	HPE 58x0AF Bck(pwr)-Frt(ports) Fan Tray
2	JG329A	HPE X240 QSFP+ 4x10G SFP+ 1m DAC Cable
4	JG326A	HPE X240 40G QSFP+ QSFP+ 1m DAC Cable

Other hardware

Important
Quantities listed in Table 10 are based on a rack with three switches, 12 compute nodes, and three storage nodes.

Table 10. BOM for a single rack with four PDUs

QTY	PART NUMBER	DESCRIPTION
1	BW904A	HPE 42U 600x1075mm Enterprise Shock Rack
1	BW946A	HPE 42U Location Discovery Kit
1	BW930A	HPE Air Flow Optimization Kit
1	TK817A	HPE CS Rack Side Panel 1075mm Kit
1	TK816A	HPE CS Rack Light Kit
1	TK815A	HPE CS Rack Door Branding Kit
1	BW891A	HPE Rack Grounding Kit
4	AF520A	HPE Intelligent Mod PDU 24a Na/Jpn Core
6	AF547A	HPE 5xC13 IntIgnt PDU Ext Bars G2 Kit
2	C7536A	HPE Ethernet 14ft CAT5e RJ45 M/M Cable

Recommended service components

Table 11 provides a BOM for three recommended service components for Factory Express Build, TS Consulting and HPE BDRA.

Table 11. Recommended service components

QTY	PART NUMBER	DESCRIPTION
--	HA454A1	HPE Factory Express Level 4 Service (recommended)
1	H8E04A1	HPE Hadoop Custom Consulting Service (recommended)
1	P6L57A	HPE Big Data Reference Architecture

Software options

Important
Quantities listed in Table 12 may vary. Quantities below are based on a rack with three management/head nodes, 12 ProLiant XL170r compute nodes and three Apollo 4200 storage nodes.

Table 12. BOM for software options

QTY	PART NUMBER	DESCRIPTION
18	E6U59ABE	HPE iLO Adv incl 1yr TS U E-LTU
18	QL803BAE	HPE Insight CMU 1yr 24x7 Flex E-LTU
18	G3J29AAE	RHEL Svr 2 Sckt/2 Gst 1yr 9x5 E-LTU
3	C9A82AAE	HPE Secure Encryption per Server Entitlement

Hortonworks software

Important
Table 13 provides the BOM for the Hortonworks license and support. One Hortonworks license covers a single Apollo 2000 chassis of compute nodes.

For the Moonshot compute option, one license covers up to 16 Moonshot m710p cartridges. Therefore, three of these licenses will cover a fully populated Moonshot 1500 chassis. While HPE is a certified reseller of Hortonworks software subscriptions, all application support (level-one through level-three) is provided by Hortonworks.

Table 13. BOM for Hortonworks software

QTY	PART NUMBER	DESCRIPTION
--	F5Z52A	Hortonworks Data Platform Enterprise 4 Nodes or 50TB Raw Storage 1 year 24x7 Support LTU

Implementing a proof-of-concept

As a matter of best practice for all deployments, HPE recommends implementing a proof-of-concept using a test environment that matches as closely as possible the planned production environment. In this way, appropriate performance and scalability characterizations can be obtained. For help with a proof-of-concept, contact an HPE Services representative (hpe.com/us/en/services/consulting.html) or your HPE partner.

Summary

HPE BDRA is a modern, flexible architecture for big data solutions that improves overall access to information and time-to-solution, while providing the inherent flexibility to support the rapidly changing requirements of big data applications. Our BDRA leverages HPE's innovative big data building blocks of servers, storage, and networking, along with integrated management software and bundled support.

The combination of HDP 2.3 running Hive with Tez on YARN on HPE BDRA provides the ideal blend of speed, flexibility, scalability and optimization of operations needed by today's enterprises: a high performance data warehouse infrastructure built on top of Hadoop that provides data summarization, query and analysis, supports analysis of large datasets stored in HDFS, and provides a SQL-like interface

HPE BDRA and HDP allow an organization to derive new business insights from big data by providing a platform to store, manage, and process data at scale. However, the design, procurement and deployment of a Hadoop cluster can be both complex and time consuming. Thus, this white paper outlined reference architectures for heterogeneous clusters of varying sizes with HDP 2.3 running Hive on Tez on HPE infrastructure and management software. Guidelines for optimizing HDP software settings were provided.

Appendix A: Alternate compute node components

This appendix provides BOMs for alternate processors, memory, and disk drives for the ProLiant XL170r servers used as compute nodes, and also includes the BOM for optional HPE Moonshot cartridge nodes.

Table A-1. BOM Alternate processors - Apollo 2000 - ProLiant XL170r

QTY	PART NUMBER	DESCRIPTION
1	793028-L21	HPE XL1x0r Gen9 E5-2680v3 FIO Kit (12C, 2.5 GHz)
1	793028-B21	HPE XL1x0r Gen9 E5-2680v3 Kit
1	793024-L21	HPE XL1x0r Gen9 E5-2660v3 FIO Kit (10C, 2.6 GHz)
1	793024-B21	HPE XL1x0r Gen9 E5-2660v3 Kit
1	793020-L21	HPE XL1x0r Gen9 E5-2640v3 FIO Kit (8C, 2.6 GHz)
1	793020-B21	HPE XL1x0r Gen9 E5-2640v3 Kit

Table A-2. Alternate memory - Apollo 2000 - ProLiant XL170r

QTY PER NODE	PART NUMBER	DESCRIPTION
4/8/16	728629-B21	HPE 32GB 2Rx4 PC4-2133P-R Kit
8/16	726719-B21	HPE 16GB 2Rx4 PC4-2133P-R Kit
16	759934-B21	HPE 8GB 2Rx8 PC4-2133P-R Kit

Table A-3. Alternate disk drives - Apollo 2000 - ProLiant XL170r

QTY PER NODE	PART NUMBER	DESCRIPTION
2/4/6	804625-B21	HPE 800GB 6G SATA MU-2 SFF SC SSD
2/4/6	817011-B21	HPE 1.92TB 6G SATA MU-3 SFF SC SSD
2/4/6	757339-B21	HPE 1.6TB 6G SATA VE 2.5in SC EV SSD

Table A-4. BOM for a single HPE Moonshot chassis with 45 m710p server cartridges

QTY	PART NUMBER	DESCRIPTION
1	755371-B21	HPE Moonshot 1500 Chassis
4	684532-B21	HPE 1500W Ht Plg Pwr Supply Kit
2	704654-B21	HPE Moonshot-45XGc Switch Kit
2	704652-B21	HPE Moonshot 4QSFP Uplink Kit
45	808915-B21	HPE ProLiant m710p Server Cartridge
45	765483-B21	HPE Moonshot 480G SATA VE M.2 2280FIO Kit
1	681254-B21	HPE 4.3U Rail Kit
1	681260-B21	HPE 0.66U Spacer Blank Kit
6	JG327A	HPE X240 40G QSFP+ QSFP+ 3m DAC Cable
1	JG326A	HPE X240 40G QSFP+ QSFP+ 1m DAC Cable
1	755371-B21	HPE Moonshot 1500 Chassis
3	AF595A	HPE 3.0M, Blue, CAT6 STP, Cable Data

Appendix B: Alternate storage node components

This appendix provides BOMs for alternate processors, memory, and disk drives for the Apollo 4200 servers used as storage nodes.

Table B-1. BOM alternate processors - Apollo 4200

QTY	PART NUMBER	DESCRIPTION
1	821791-L21	HPE Apollo 4200 Gen9 E5-2697v3 FIO Kit (14C, 2.6GHz)
1	821791-B21	HPE Apollo 4200 Gen9 E5-2697v3 Kit
1	803314-L21	HPE Apollo 4200 Gen9 E5-2680v3 FIO Kit (12C, 2.5 GHz)
1	803314-B21	HPE Apollo 4200 Gen9 E5-2680v3 Kit
1	803308-L21	HPE Apollo 4200 Gen9 E5-2640v3 FIO Kit (8C, 2.6 GHz)
1	803308-B21	HPE Apollo 4200 Gen9 E5-2640v3 Kit

Table B-2. Alternate memory - Apollo 4200

QTY PER NODE	PART NUMBER	DESCRIPTION
4/8/16	728629-B21	HPE 32GB 2Rx4 PC4-2133P-R Kit
4/8/16	726717-B21	HPE 4GB 1Rx8 PC4-2133P-R Kit
4/8/16	759934-B21	HPE 8GB 2Rx8 PC4-2133P-R Kit

Table B-3. Alternate disk drives - Apollo 4200

QTY PER NODE	PART NUMBER	DESCRIPTION
Up to 28	797269-B21	HPE 6TB 6G SATA 7.2K 3.5in LP MDL HDD
Up to 28	797271-B21	HPE 3TB 6G SATA 7.2k 3.5in MDL LP HDD
Up to 28	805334-B21	HPE 8TB 6G SATA 7.2k 3.5in MDL LP HDD

Appendix C: HPE value added services and support

In order to help you jump-start your Hadoop solution development, HPE offers a range of big data services, which are outlined in this appendix.

Factory Express Services
Factory-integration services are available for customers seeking a streamlined deployment experience. With the purchase of Factory Express services, your Hadoop cluster will arrive racked and cabled, with software installed and configured per an agreed-upon custom statement of work, for the easiest deployment possible. You should contact HPE Technical Services for more information and for assistance with a quote.

Technical Services Consulting – Reference Architecture Implementation Service for Hadoop (Hortonworks)
With HPE Reference Architecture Implementation Service for Hadoop, HPE can install, configure, deploy, and test a Hadoop cluster that is based on HPE BDRA. Experienced consultants implement all the details of the original Hadoop design: naming, hardware, networking, software, administration, backup, disaster recovery, and operating procedures. Where options exist, or the best choice is not clear, HPE works with you to configure the environment to meet your goals and needs. HPE also conducts an acceptance test to validate that the system is operating to your satisfaction.

Technical Services Consulting – Big Data Services
HPE Big Data Services can help you reshape your IT infrastructure to corral increasing volumes of data – from e-mails, social media, and website downloads – and convert them into beneficial information. These services encompass strategy, design, implementation, protection, and compliance. Delivery is in the following three steps:

1. **Architecture strategy:** HPE defines 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. **System infrastructure:** HPE designs and implements a high-performance, integrated platform to support a strategic architecture for big data. Choose from design and implementation services, reference architecture implementations, and integration services. Your flexible, scalable infrastructure will support big data variety, consolidation, analysis, share, and search on HPE platforms.
3. **Data protection:** Ensure the availability, security, and compliance of your big data systems. HPE can help you safeguard your data and achieve regulatory compliance and lifecycle protection across your big data landscape, while also enhancing your approach to backup and business continuity.

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

HPE Support options
HPE offers a variety of support levels to meet your needs. More information is provided below.

HPE Support Plus 24
HPE can provide integrated onsite hardware/software support services, available 24x7x365, including access to HPE technical resources, four-hour response onsite hardware support and software updates.

HPE Proactive Care
HPE Proactive Care provides all of the benefits of proactive monitoring and reporting, along with rapid reactive support through HPE's expert reactive support specialists. You can customize your reactive support level by selecting either six-hour call-to-repair or 24x7 with four-hour onsite response.

HPE Proactive Care helps prevent problems, resolve problems faster, and improve productivity. Through analysis, reports, and update recommendations, you are able to identify and address IT problems before they can cause performance issues or outages.

HPE Proactive Care with the HPE Personalized Support Option
Adding the Personalized Support Option for HPE Proactive Care is highly recommended. This option builds on the benefits of HPE Proactive Care Service, providing you an assigned Account Support Manager who knows your environment and can deliver support planning, regular reviews, and technical and operational advice specific to your environment.

HPE Proactive Select

To address your ongoing/changing needs, HPE recommends adding Proactive Select credits to provide tailored support options from a wide menu of services that can help you optimize the capacity, performance, and management of your environment. These credits may also be used for assistance in implementing solution updates. As your needs change over time, you have the flexibility to choose the services best suited to address your current challenges.

HPE Datacenter Care

HPE Datacenter Care provides a more personalized, customized approach for large, complex environments, providing a single solution for reactive, proactive, and multi-vendor support needs. You may also choose the Defective Media Retention (DMR) option.

Other offerings

HPE highly recommends HPE Education Services (customer training and education) and additional Technical Services, as well as in-depth installation or implementation services when needed. HPE Analytics & Data Management services provide a range of Business Intelligence (BI) modernization services and include advisory and platform services to implement and manage Hadoop solutions.

More information

For additional information, visit:

- HPE Education Services: http://h10076.www1.hp.com/education/bigdata.htm
- HPE Technology Consulting Services: hpe.com/us/en/services/consulting/big-data.html
- HPE Deployment Services: hpe.com/services/deployment

Resources and additional links

Hortonworks: hortonworks.com

Hortonworks partner site: hortonworks.com/partner/hp/

Hortonworks HDP 2.3: http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.2/bk_Sys_Admin_Guides/content/index.html

HPE Solutions for Apache Hadoop: hpe.com/us/en/solutions/big-data.html

HPE Insight Cluster Management Utility (Insight CMU): http://www8.hp.com/us/en/products/server-software/product-detail.html?oid=3296361

HPE 5930 Switch Series: hpe.com/networking/5930

HPE ProLiant servers: hpe.com/servers/proliant

HPE Apollo Systems: hpe.com/us/en/servers/apollo.html

QuickSpecs for Apollo 2000 System: http://h20195.www2.hpe.com/V2/GetDocument.aspx?docname=c04542554

QuickSpecs for HPE ProLiant XL170r Gen9 Server: http://h20195.www2.hpe.com/V2/GetDocument.aspx?docname=c04545616

HPE Moonshot system: hpe.com/info/moonshot

HPE Software: hpe.com/software

HPE Networking: hpe.com/us/en/networking.html

HPE Integrated Lights-Out (HPE iLO): hpe.com/info/ilo

HPE Product Bulletin (QuickSpecs): hpe.com/info/qs

HPE Services: hpe.com/services

HPE Support and Drivers: hpe.com/support

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

About Hortonworks

Hortonworks develops, distributes and supports the only 100% open source Apache Hadoop data platform. Our team comprises the largest contingent of builders and architects within the Hadoop ecosystem who represent and lead the broader enterprise requirements within these communities. The Hortonworks Data Platform provides an open platform that deeply integrates with existing IT investments and upon which enterprises can build and deploy Hadoop-based applications. Hortonworks has deep relationships with the key strategic data center partners that enable our customers to unlock the broadest opportunities from Hadoop. For more information, visit hortonworks.com.

Sign up for updates

★ Rate this document

4AA5-6136ENW, November 2015, Rev. 3

Hewlett Packard Enterprise