HP Reference Architectures

HP Verified Reference Architecture for running HBase on HP BDRA

Configuration guide for running HBase on HP Big Data Reference Architecture

Table of contents

Click here to verify the latest version of this document

Executive summary

This white paper outlines the benefits of running HBase on HP Big Data Reference Architecture (BDRA) and recommends the best practice configurations to evaluate, balance and monitor HBase.

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 (Hadoop, NoSQLs, MPP DBMSs, etc.) with optimal hardware, to support rigid workload partitioning for departmental requirements or simply as a byproduct of multi-generational hardware. This often leads to data duplication and movement of large amounts of data between systems to accomplish an organization's business requirements. We find some customers 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 customer challenges, HP engineers challenged the conventional wisdom that compute should always be co-located with data. We used Hadoop/HBase with HP BDRA to build and test a system that will speedily perform big data capture, management and analysis to meet the business needs of our customers.

HP BDRA is the most modern and flexible design today to improve access to big data and deploy big data solutions rapidly in response to the ever-changing requirements in a Hadoop ecosystem.

Hadoop enables effective distributed processing and management of large data sets. HBase is a distributed database that runs on top of HDFS (Hadoop Distributed File System) used by Hadoop clusters. HDFS is great for reading and writing large chunks of data all at once but random Read/Writes are a weakness. HBase is a speedy data store that makes up for what Hadoop lacks in particular situations to meet all of your service level requirements. Combining Hadoop and HBase means that Hadoop handles distributed management of big data and simple aggregation while HBase simultaneously performs random searches of data.

Clearly, big data solutions are evolving from a simple model where each application was deployed on a dedicated cluster of identical nodes. By integrating the significant changes that have occurred in fabrics, storage, container-based resource management and workload-optimized servers, HP has created the next-generation, data center-friendly, big data cluster. In this paper, we recommend that teaming HBase with HP BDRA makes a perfect combination for big data analysis solutions as balancing the company's infrastructure is optimized and made significantly more efficient when HBase runs on HP BDRA.

This white paper discusses how to evaluate and balance HBase performance using Yahoo Cloud Serving Benchmark (YCSB) to get maximum performance. Performance numbers are presented from these three scenarios:

- Multiple workloads and distributions while running with a dataset completely in memory
- Multiple workloads and distributions while running with a dataset that fits on SSD
- Multiple workloads and distributions while running with a dataset that does not fit on SSD

Our test findings were significant and impressive. They constitute the highest YCSB HBase performance numbers that have been achieved to date.

This guide also describes the best way to configure Ganglia in order to monitor HBase.

Target audience: This document is intended for HP customers who are investigating the deployment of big data solutions, or those who already have big data deployments in operation and are looking for a modern architecture to consolidate current and future solutions while offering the widest support for big data tools.

Document purpose: The purpose of this document is to provide guidance on how to configure HBase to run on HP BDRA.

This white paper describes testing performed by HP in the spring of 2015.

Introduction

Today's data centers are typically based on a converged infrastructure featuring a number of servers and server blades accessing shared storage over a storage array network (SAN). In such environments, the need to build dedicated servers to run particular applications has disappeared. However, with the explosive growth of big data, the Hadoop platform is becoming widely adopted as the standard solution for the distributed storage and processing of big data. Since it is often cost-prohibitive to deploy shared storage that is fast enough for big data, organizations are often forced to build a dedicated Hadoop cluster at the edge of the data center; then another; then perhaps an HBase cluster, a Spark cluster, and so on. Each cluster utilizes different technologies and different storage types; each typically has three copies of big data.

The current paradigm for big data products like Hadoop, Cassandra, and Spark insists that performance in a Hadoop cluster is optimal when work is taken to the data, resulting in nodes featuring Direct-Attached Storage (DAS), where compute and storage resources are co-located on the same node. In this environment, however, data sharing between clusters is constrained by network bandwidth, while individual cluster growth is limited by the necessity to repartition data and distribute it to new disks.

Thus, many of the benefits of a traditional converged architecture are lost in today's Hadoop cluster implementations. You cannot scale compute and storage resources separately; and it is no longer practical for multiple systems to share the same data.

HP BDRA is a reference architecture that is based on hundreds of man-hours of research and testing by HP engineers. To allow customers to deploy solutions based on this architecture, HP offers detailed Bills of Materials (BOMs) based on a proof-of-concept. To facilitate installation, HP has developed a broad range of Intellectual Property (IP) that allows HP BDRA solutions to be implemented by HP or, jointly, with partners and customers. Additional assistance is available from HP Technical Services.

It is very important to understand that the HP Big Data Reference Architecture is a highly optimized configuration built using servers offered by HP – the HP ProLiant SL4540 for the high density storage layer, and HP Moonshot for the high density computational layer. It is also the result of a great deal of testing and optimization done by our HP engineers which resulted in the right set of software, drivers, firmware and hardware to yield extremely high density and performance. Simply deploying Hadoop onto a collection of traditional servers in an asymmetric fashion will not yield the kind of benefits that we see with our reference architecture. In order to simplify the build for customers, HP will provide the exact bill of materials in this document to allow a customer to purchase their cluster. Then, our Technical Services Consulting organization will provide a service that will install our prebuilt operating system images and verify all firmware and versions are correctly installed then run a suite of tests that verify that the configuration is performing optimally. Once this has been done, the customer is free to do a fairly standard HBase/BDRA installation on the cluster using the recommended guidelines in this document.

Benefits of the HP BDRA solution

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

- Elasticity – HP BDRA is designed for flexibility. Compute nodes can be allocated very flexibly without redistributing data; for example, you can allocate nodes by time-of-day or even for a single job. You are no longer committed to yesterday's CPU/storage ratios, leading to much more flexibility in design and cost. Moreover, with HP BDRA, you only need to grow your system where needed.
- Consolidation – HP 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. You can consolidate the various pools of data currently being used in big data projects into a single, central repository.
- 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 your requirements, you 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 HP 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; time-to-solution is typically faster.

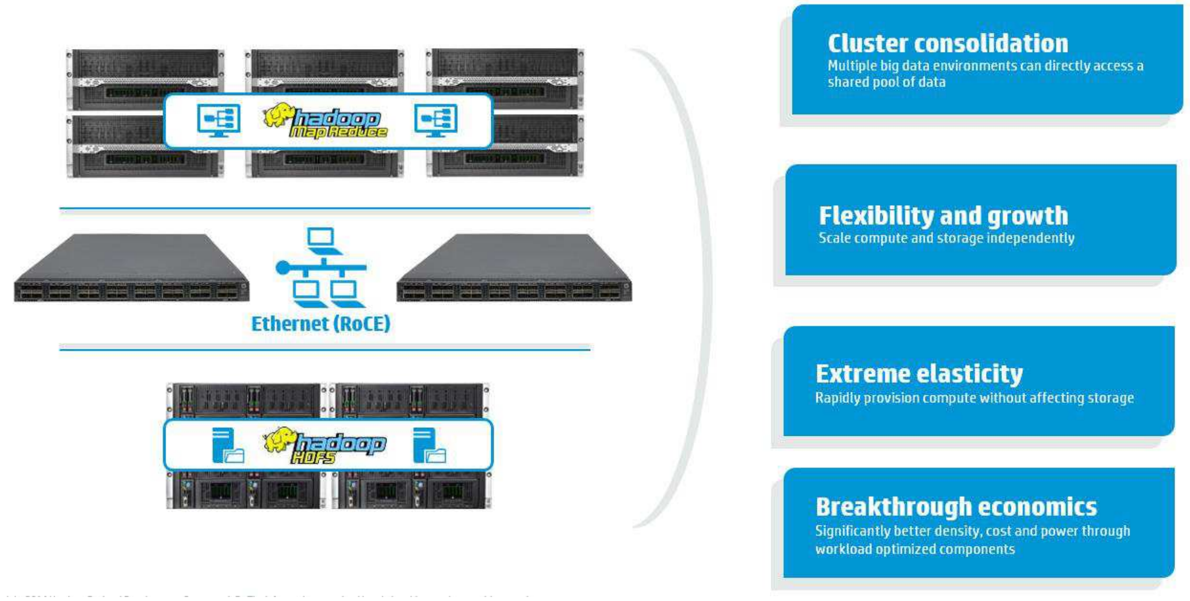
Solution overview

HP has created HP BDRA, a reference architecture for big data, which is shown in Figure 1. This dense, converged solution can facilitate the consolidation of multiple pools of data, allowing Hadoop, Vertica, Spark and other big data technologies to share a common pool of data. The flexibility to adapt to future workloads has been built into the design.

This converged design features an asymmetric cluster where compute and storage resources are deployed on separate tiers. Storage is direct-attached; specialized SAN technologies are not used. Workloads and storage can be directed to optimized nodes. Interconnects are standard Ethernet; protocols between compute and storage are native Hadoop, such as HDFS and HBase.

HP BDRA serves as a proof of concept and has been benchmarked to demonstrate substantially improved price/performance – along with significantly increased density – compared with a traditional Hadoop architecture. For example, HP has demonstrated that storage nodes in HP BDRA actually perform better now that they have been decoupled and are dedicated to running HDFS, without any Java or MapReduce overhead. Moreover, because modern Ethernet fabrics are capable of delivering more bandwidth than a server's storage subsystem, network traffic between tiers does not create a bottleneck. Indeed, testing indicated that read I/Os increased by as much as 30% in an HP BDRA configuration compared with a conventional Hadoop cluster.

Figure 1. HP Big Data Reference Architecture, changing the economics of work distribution in big data


hadoop MapReduce
Ethernet (RoCE)
hadoop HDFS
Cluster consolidation
Multiple big data environments can directly access a shared pool of data
Flexibility and growth
Scale compute and storage independently
Extreme elasticity
Rapidly provision compute without affecting storage
Breakthrough economics
Significantly better density, cost and power through workload optimized components

The HP BDRA design is anchored by the following HP technologies:

- Storage nodes – HP ProLiant SL4540 Gen8 2 Node Servers (SL4540) make up the HDFS storage layer, providing a single repository for big data.
- Compute nodes – HP Moonshot System cartridges deliver a scalable, high-density layer for compute tasks and provide a framework for workload-optimization.

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 co-located with storage, Hadoop does 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 you scale within a rack, overall scalability is not affected.

With compute and storage de-coupled, you can again enjoy many of the advantages of a traditional converged system. For example, you can scale compute and storage independently, simply by adding compute nodes or storage nodes. Testing carried out by HP indicates that most workloads respond almost linearly to additional compute resources. For more information on HP BDRA, please refer to http://h20195.www2.hp.com/V2/GetDocument.aspx?docname=4AA5-6141ENW

Solution components

HP recommends the components described below, which were utilized in the proof-of-concept. This section lists the equipment used to configure, monitor and evaluate HBase on HP BDRA. This solution is comprised of the following components.

Hardware

- HP Moonshot 1500 Chassis fully populated with 45 HP Moonshot m710 server cartridges
- HP SL4540 servers configured with 3TB drives
- HP ProLiant DL360p Gen8 servers
- HP 5930 ToR switches
- HP 5900 iLO switch
- HP 642 1200mm Shock Intelligent rack and power supplies

Software

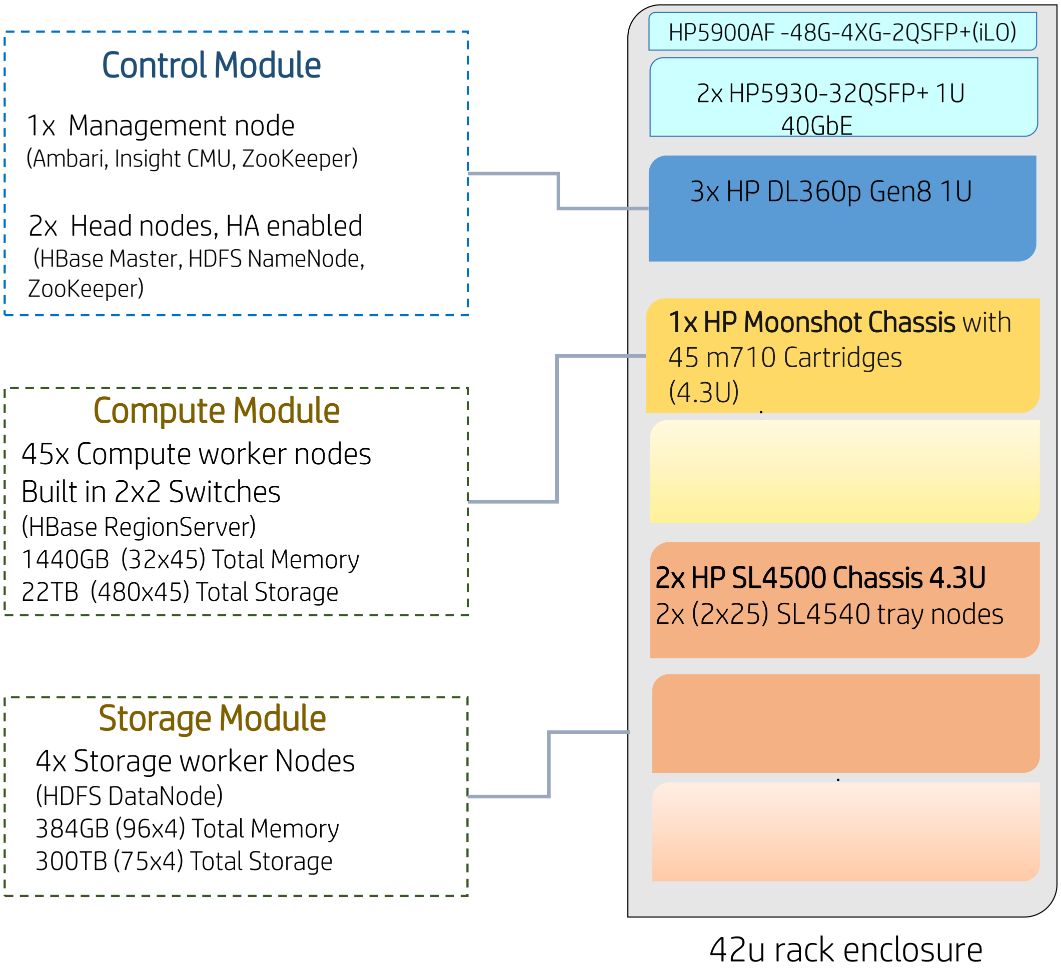
- Red Hat® Enterprise Linux® (RHEL) 6.5
- Hortonworks Data Platform (HDP) 2.2
- HP Insight Cluster Management Utility (Insight CMU)

More detailed component information can be found in the Bill of Materials section.

Key point
Hortonworks HDP 2.2 and RHEL 6.5 where used in this particular testing scenario; however, the same results can be expected when using other Hadoop distributions such as Cloudera or MapR.

Figure 2 provides a basic conceptual overview of the components involved in the test.

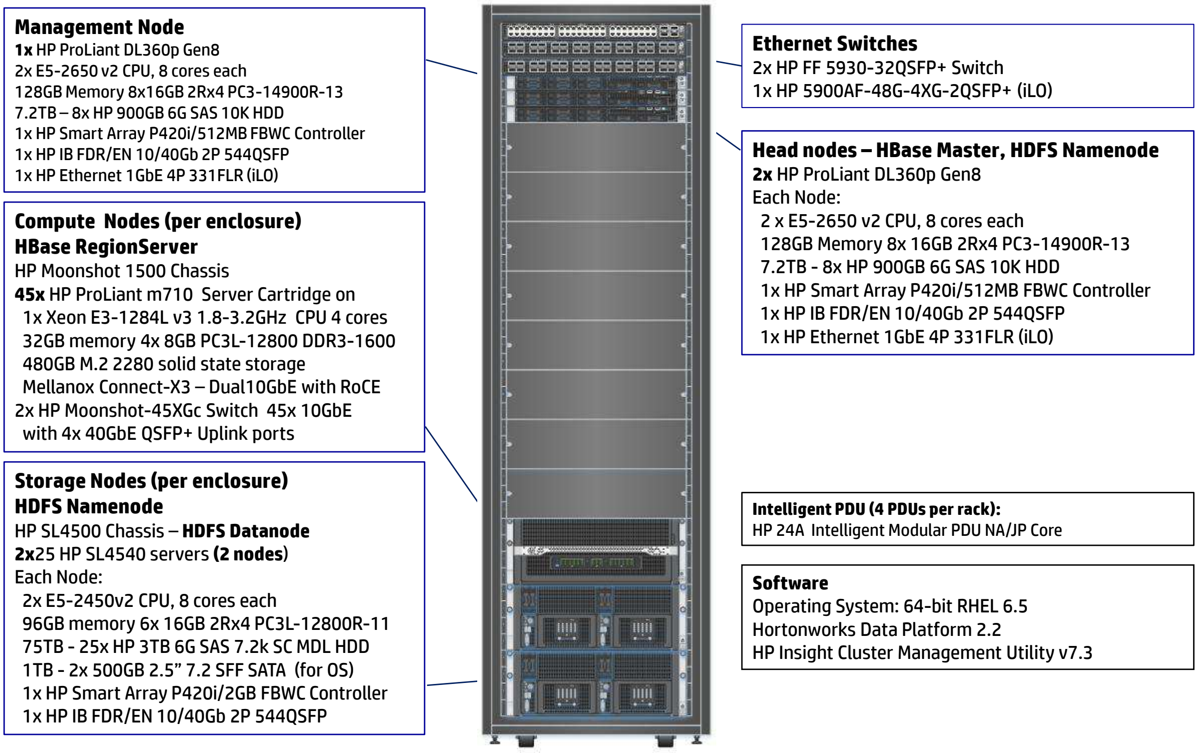
Figure 2. BDRA configuration used in test


Control Module
1x Management node
(Ambari, Insight CMU, ZooKeeper)
2x Head nodes, HA enabled
(HBase Master, HDFS NameNode, ZooKeeper)
Compute Module
45x Compute worker nodes
Built in 2x2 Switches
(HBase RegionServer)
1440GB (32x45) Total Memory
22TB (480x45) Total Storage
Storage Module
4x Storage worker Nodes
(HDFS DataNode)
384GB (96x4) Total Memory
300TB (75x4) Total Storage
HP5900AF -48G-4XG-2QSFP+(iLO)
2x HP5930-32QSFP+ 1U 40GbE
3x HP DL360p Gen8 1U
1x HP Moonshot Chassis with 45 m710 Cartridges (4.3U)
2x HP SL4500 Chassis 4.3U
2x (2x25) SL4540 tray nodes
42u rack enclosure

Configuration

Figure 3 shows the HP BDRA configuration, with 45 worker nodes and four storage nodes housed in a single 42U rack.

Figure 3. HP BDRA configuration, with detailed lists of components



Management Node
1x HP ProLiant DL360p Gen8
2x E5-2650 v2 CPU, 8 cores each
128GB Memory 8x16GB 2Rx4 PC3-14900R-13
7.2TB – 8x HP 900GB 6G SAS 10K HDD
1x HP Smart Array P420i/512MB FBWC Controller
1x HP IB FDR/EN 10/40Gb 2P 544QSFP
1x HP Ethernet 1GbE 4P 331FLR (iLO)

Compute Nodes (per enclosure)
HBase RegionServer
HP Moonshot 1500 Chassis
45x HP ProLiant m710 Server Cartridge on
1x Xeon E3-1284L v3 1.8-3.2GHz CPU 4 cores
32GB memory 4x 8GB PC3L-12800 DDR3-1600
480GB M.2 2280 solid state storage
Mellanox Connect-X3 – Dual10GbE with RoCE
2x HP Moonshot-45XGc Switch 45x 10GbE
with 4x 40GbE QSFP+ Uplink ports

Storage Nodes (per enclosure)
HDFS Namenode
HP SL4500 Chassis – **HDFS Datanode**
2x25 HP SL4540 servers **(2 nodes)**
Each Node:
2x E5-2450v2 CPU, 8 cores each
96GB memory 6x 16GB 2Rx4 PC3L-12800R-11
75TB - 25x HP 3TB 6G SAS 7.2k SC MDL HDD
1TB - 2x 500GB 2.5" 7.2 SFF SATA (for OS)
1x HP Smart Array P420i/2GB FBWC Controller
1x HP IB FDR/EN 10/40Gb 2P 544QSFP

Ethernet Switches
2x HP FF 5930-32QSFP+ Switch
1x HP 5900AF-48G-4XG-2QSFP+ (iLO)

Head nodes – HBase Master, HDFS Namenode
2x HP ProLiant DL360p Gen8
Each Node:
2 x E5-2650 v2 CPU, 8 cores each
128GB Memory 8x 16GB 2Rx4 PC3-14900R-13
7.2TB - 8x HP 900GB 6G SAS 10K HDD
1x HP Smart Array P420i/512MB FBWC Controller
1x HP IB FDR/EN 10/40Gb 2P 544QSFP
1x HP Ethernet 1GbE 4P 331FLR (iLO)

Intelligent PDU (4 PDUs per rack):
HP 24A Intelligent Modular PDU NA/JP Core

Software
Operating System: 64-bit RHEL 6.5
Hortonworks Data Platform 2.2
HP Insight Cluster Management Utility v7.3

The following nodes are used in the base HP BDRA configuration:

- Compute nodes – The HP BDRA configuration features one HP Moonshot 1500 Chassis containing a total of 45 m710 cartridge servers.
- Storage nodes – There are two SL4500 Chassis, each with two SL4540 servers, for a total of four storage nodes. Each node is configured with 25 disks; each typically runs HDFS DataNode.
- Management/head nodes – Three HP ProLiant DL360p Gen8 servers are configured as management/head nodes with the following functionality:
  - Management node, with Ambari, HP Insight CMU and Hadoop ZooKeeper
  - Head node, with HBase Master, HDFS NameNode and ZooKeeper
  - Head node, with secondary HDFS NameNode and ZooKeeper

Key point
HP recommends starting with a minimum of one full Moonshot 1500 chassis, with 45 compute nodes on each. To provide high availability, you should start with a minimum of three SL4500 chassis, with a total of six storage nodes in order to provide the redundancy that comes with the default replication factor of three.

Recommended components

For a detailed description of the solution components and the basic configuration of HP BDRA, please refer to: http://h20195.www2.hp.com/V2/GetDocument.aspx?docname=4AA5-6141ENW

Capacity and sizing

Workload description

The configuration used for testing the HBase/HP BDRA project consisted of BDRA using one HP Moonshot chassis fully populated with 45 m710 cartridges and two HP SL4500 chassis with four HP SL4540 servers (2 x 25 nodes) configured with 3TB drives. Figure 3 shows the basic HP BDRA configuration. Please refer to the Bill of materials in this paper for part numbers and descriptions.

The test bed employed 16 Yahoo Cloud Serving Benchmark (YCSB) client driver machines as auxiliary servers used to run the benchmark. HP ProLiant DL360p Gen8 servers were used as YCSB client machines to simulate Read/Write operations during the test.

YCSB was used to provide a framework and common set of workloads for evaluating the performance of different "key-value" and cloud-serving stores. YCSB is an extensible workload generator. So you can define new and different workloads not covered by the core workload and also examine system aspects, or application scenarios, not adequately covered by the core workload. YCSB is extensible to support benchmarking HBase and other databases.

The testing used was that of YCSB benchmark client configuration to evaluate and balance HBase performance. Performance numbers are presented below in the *Analysis and recommendations* section for the following scenarios:

- Multiple workloads and distributions while running with a dataset completely in memory
- Multiple workloads and distributions while running with a dataset that fits on SSD
- Multiple workloads and distributions while running with a dataset that does not fit on SSD

Analysis and recommendations

HBase results

We used two basic scenarios in our test – WorkloadC consisted of 100% random reads while WorkloadA consisted of 50% reads and 50% writes. For all test cases, the HBase region servers where running on the Moonshot servers only.

Testing was done on a table with 450,000,000 rows, using YCSB default row and field sizes. The amount of physical memory on the Moonshot servers (HBase RegionServer) and the size of HBase blockcache was altered in order to simulate datasets that fit in memory, on SSD and on HDD.

Tables 1a and 1b show the results of running HBase with BDRA, using Uniform, Zipfian and HotSpot distributions. Figures 4 and 5 show the results of WorkloadC and WorkloadA. The results are expressed in operations per second (ops).

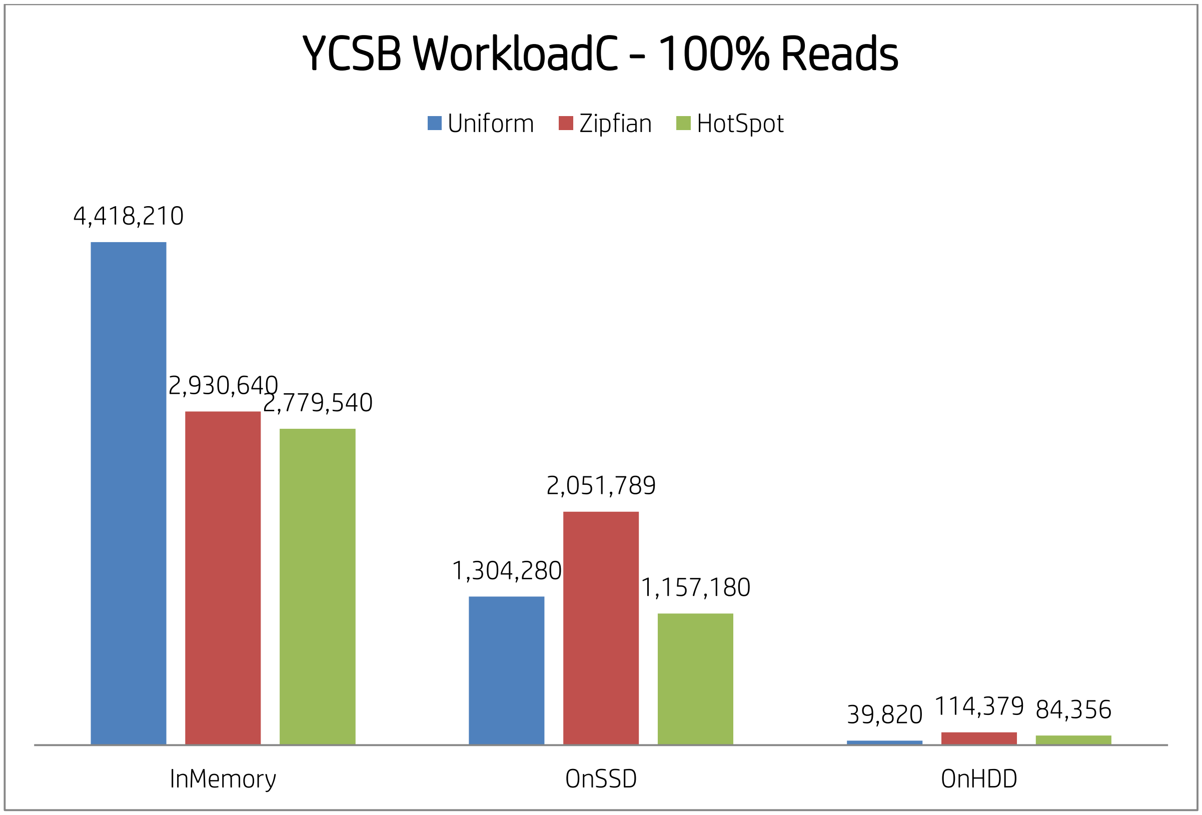
Table 1a. WorkloadC (Operations per second)

YCSB	Use Case	Distribution	HP BDRA
WorkloadC	In Memory	Uniform Zipfian HotSpot	4,418,210 2.930,640 2,779,540
WorkloadC	On SSD	Uniform Zipfian HotSpot	1,304,280 2,051,789 1,157,180
WorkloadC	On HDD	Uniform Zipfian HotSpot	39,820 114,379 84,356

Table 1b. WorkloadA (Operations per second)

YCSB	Use Case	Distribution	HP BDRA
WorkloadA	In Memory	Uniform Zipfian HotSpot	1,198,280 1,069,500 1,013,930
WorkloadA	On SSD	Uniform Zipfian HotSpot	718,930 989,243 830,938
WorkloadA	On HDD	Uniform Zipfian HotSpot	79,836 197,567 129,330

Figure 4. YCSB WorkloadC – 100% reads



WorkloadC

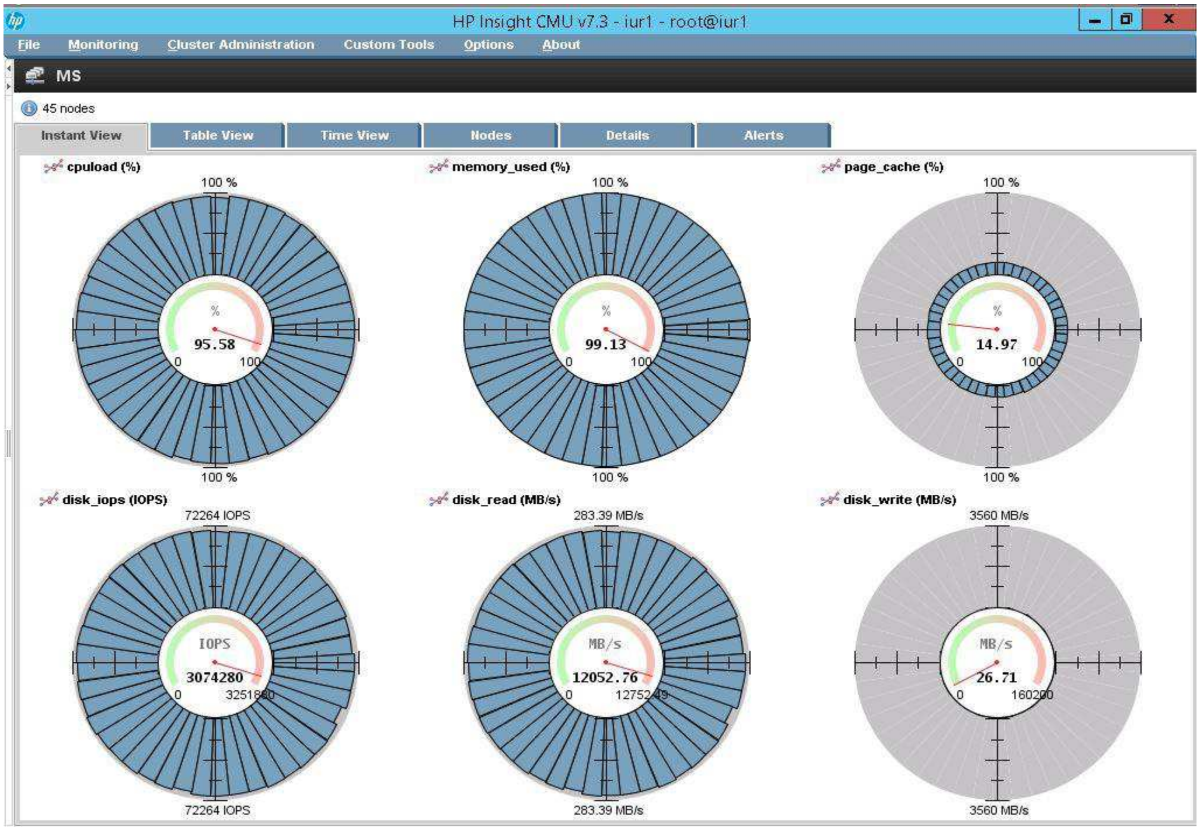
The following description presents the results of WorkloadC, which is a 100% random read workload.

When the working dataset fits completely into system memory, we obtain the greatest performance as expected. Our tested system has 1440GB memory, from which about 1TB can be allocated directly to HBase cache. Additional memory can be added by using multiple Moonshot chassis. Note that by using a uniform access distribution on a properly tuned system, we measured 4.4 million operations per second (ops). This number is significantly better than any other YCSB WorkloadC number published to date. When using a non-uniform access distribution like Zipfian or HotSpot, we unbalance the system because some servers will receive more requests compared to others since they host the "hot" data; which is why the non-uniform access distribution numbers are lower in our test case.

When the working dataset doesn't fit completely into system memory, but still fits into the SSD cache, we notice slower performance compared to what is in memory in the 1.5x-3.5x range. Our tested system has 22TB of SSD cache, from which about 20TB can be allocated directly to HBase cache, significantly more than the amount of memory (~20x more). Using a uniform access distribution, we measured 1.3 million operations per second (ops). This number is directly linked to the random performance of the SSD used and the CPU load generated by this I/O workload. Each YCSB read generates two I/Os to the SSD, one to read the actual data and one to read the checksum for the data. Our system was able to serve 3-3.5 million IOPS from SSDs. Please refer to Figure 5 for more information. When using a Zipfian access distribution, we obtained better performance since this distribution has a smaller size "hot" data and was able to cache it in memory.

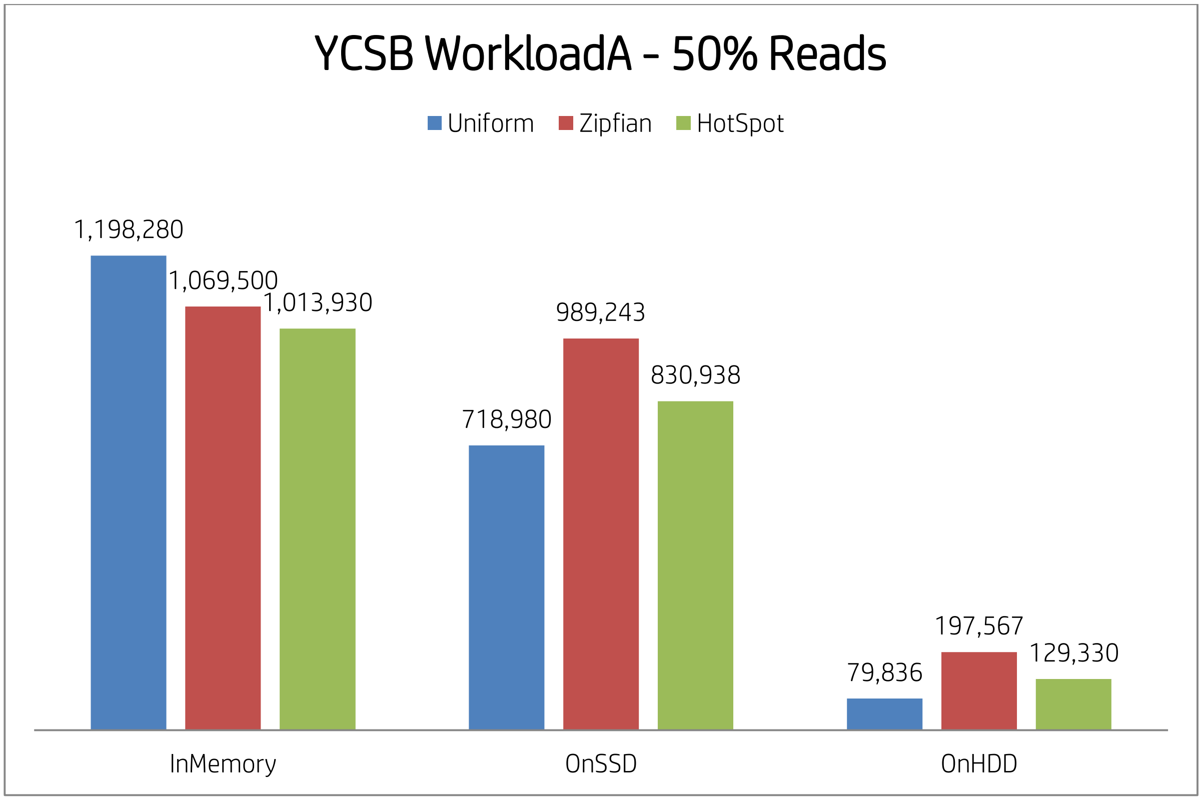
When the working dataset doesn't fit completely into the SSD cache, we notice significantly slower performance, in the 100x range. Using a uniform access distribution, we measured 39,820 operations per second (ops). This number is directly linked to the random performance of the HDD used. When using non-uniform access distributions, such as Zipfian and HotSpot, we obtained better performance since some of the "hot" data was able to fit into the SSD cache.

Figure 5. HP Insight CMU



WorkloadA

Figure 6. YCSB WorkloadA – 50% reads



WorkloadA
The following description presents the results of WorkloadA, which is a 50% random read workload.

With this workload, the performance delta between running with a dataset that fits into system memory compared to running on SSD cache is much smaller (1.1x-1.6x), especially for non-uniform access distributions (Zipfian and HotSpot). It's important to note that in this test scenario HBase client-side write caching was disabled and all writes went directly to the storage nodes. The intention was to simulate a real world use case where the potential for data loss from HBase client-side write caching would not be acceptable.

When the working dataset doesn't fit completely into the SSD cache, we notice significantly slower performance (5x-15x), but not as significant as with WorkloadC. These numbers are limited by the random read performance of the rotational disks, the same as for WorkloadC.

Recommendations

While running with a big dataset that didn't fit into the SSD cache or the system memory, the HBase performance was significantly slower. We highly recommend sizing the HP BDRA so that your dataset can fit into SSD cache. Since HP BDRA allows scaling compute resources, which hosted memory and SSD in this case, independently from storage resources, it's easy and cost effective to add more Moonshot servers. Keep in mind that each Moonshot chassis with 45 m710 cartridges offers 1TB of HBase memory cache and 20TB of HBase SSD cache.

Other recommendations

Expanding the base configuration

As needed, you can add compute and/or storage nodes to a base HP BDRA configuration. The minimum HP BDRA configuration with one Moonshot 1500 chassis and three SL4500 chassis can expand to include a mix of eight chassis in a single rack. In this way, a single-rack solution can form the basis of a much larger, multi-rack solution. For more information, please refer to the HP BDRA at http://h20195.www2.hp.com/V2/GetDocument.aspx?docname=4AA5-6141ENW

HP BDRA cluster sizing

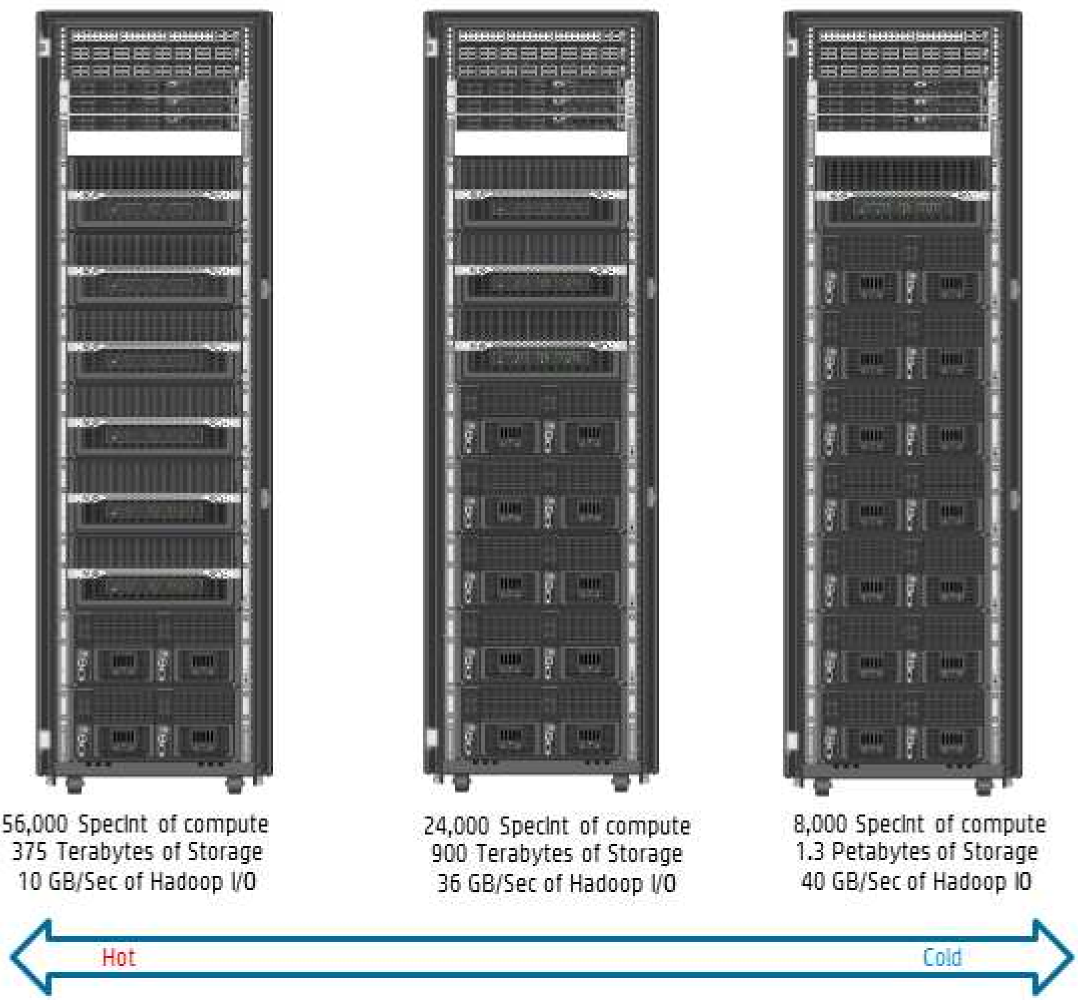
HP BDRA consists of two tiers of servers that deliver storage and compute resources for a modern Hadoop cluster; tiers are interconnected by high-performance HP networking components. De-coupling storage and compute resources allows each tier to be scaled independently, providing maximum flexibility in the design, procurement, and growth of the cluster, as shown in Figure 7.

Three different HP BDRA single-rack configurations are shown side-by-side. The racks can be characterized as follows:

- Left rack – This is a hot configuration, where compute resources are relatively high and storage capability is lower.
- Middle rack – This configuration has a balanced compute/storage mix.
- Right rack – This is a cold configuration, where storage resources are relatively high and compute capability is lower.

Thus, the flexibility of HP BDRA allows you to create a solution that meets your particular needs.

Figure 7. Three different storage solutions



Best practices for the solution

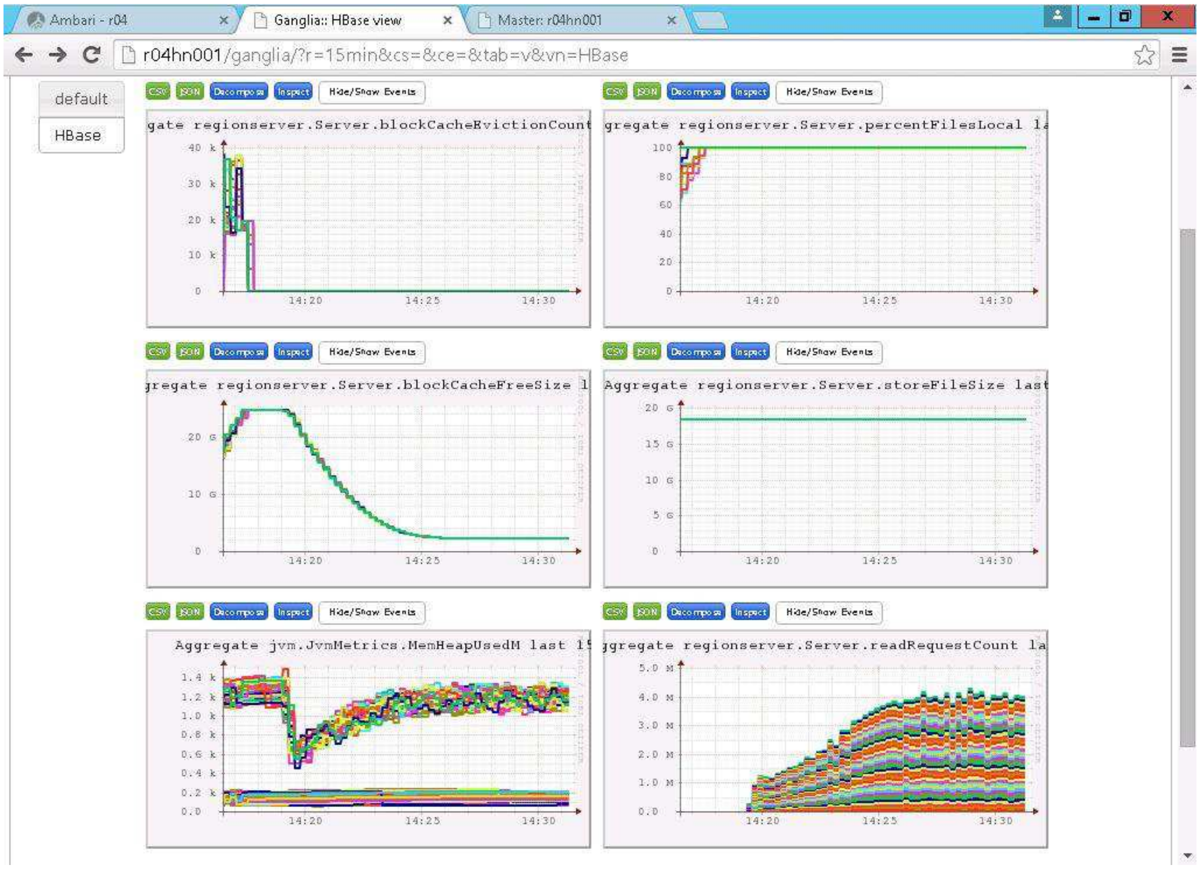
Tuning parameters

In order to properly tune and balance an HBase system we need a way to monitor its internals. HBase exposes metrics that can be viewed with a variety of tools. For our case we chose to use Ganglia for monitoring since it came standard (default) with Hortonworks HDP 2.2. Figure 8 shows a custom Ganglia view we used for HBase monitoring.

Note
Starting with Hortonworks HDP 2.2.4.2 Ganglia was replaced with Ambari Metrics as the default Hadoop monitoring software.

Ganglia

Figure 8. Ganglia view for HBase monitoring


Ambari - r04
Ganglia:: HBase view
Master: r04hn001
r04hn001/ganglia/?r=15min&cs=&ce=&tab=v&vn=HBase
default
HBase
CSV
JSON
Decompose
Inspect
Hide/Show Events
gate regionserver.Server.blockCacheEvictionCount
40 k
30 k
20 k
10 k
0
14:20
14:25
14:30
gregate regionserver.Server.percentFilesLocal la
100
80
60
40
20
0
14:20
14:25
14:30
gregate regionserver.Server.blockCacheFreeSize l
20 G
10 G
0
14:20
14:25
14:30
Aggregate regionserver.Server.storeFileSize last
20 G
15 G
10 G
5 G
0
14:20
14:25
14:30
Aggregate jvm.JvmMetrics.MemHeapUsedM last 15
1.4 k
1.2 k
1.0 k
0.8 k
0.6 k
0.4 k
0.2 k
0.0
14:20
14:25
14:30
ggregate regionserver.Server.readRequestCount la
5.0 M
4.0 M
3.0 M
2.0 M
1.0 M
0.0
14:20
14:25
14:30

Ganglia metrics

The following HBase metrics are recommended to monitor with Ganglia:

- regionserver.Server.blockCacheEvictedCount – Number of blocks that had to be evicted from the block cache due to heap size constraints. If this stays at 0 it means all your data fits completely into HBase blockcache.
- regionserver.Server.percentFilesLocal – Percentage of store file data that can be read from the local DataNode. With BDRA this is always 0 since the HDFS data is remote. For non BDRA architectures this should be 100 to reflect the fact that the regions server has local access to HDFS data.
- regionserver.Server.storeFileSize – Aggregate size of the store files on disk. Make sure this is similar on all region servers in order to properly balance the HBase load.
- regionserver.Server.blockCacheExpressHitPercent – The percentage of time that requests with the cache turned on hit the cache. The intent is to have this close to 100; less than 100 means the hot data cannot be entirely fit into blockcache.

- regionserver.Server.blockCacheFreeSize – Number of bytes that are free in the blockcache. Tells you how much of the cache is used. It is a good indicator if your data is "warmed" by moving it into cache.
- regionserver.Server.readRequestCount – The number of read requests received. This is used to monitor HBase read activity.
- regionserver.Server.writeRequestCount – The number of write requests received. This is used to monitor HBase write activity.
- regionserver.Server.flushQueueLength – Current depth of the memstore flush queue. If increasing, we are falling behind with clearing memstores out to HDFS

We did encounter some issues using Ganglia on HDP 2.2 out of the box. Please refer to Appendix A for common troubleshooting and fixes that produced the results we needed in an operational test monitoring system.

YCSB changes

The YCSB benchmark tool from Yahoo that we used for the test could not be used off the shelf and needed to be modified. We made the following changes to YCSB benchmark during testing.

The HBase configuration parameter hbase.client.write.buffer was set to 0 in order to disable client write caching. This action was necessary so that we could obtain realistic write performance numbers.

We added 0 left padding to records so that HBase sort order would match the YCSB sort order. This action was necessary so that we could properly split the HBase regions.

We added support for YCSB insertstart and insertcount parameters to run the workloads. This action was necessary to help with debugging and for the proper use of non-uniform distributions.

HBase tuning

The methods below were used to obtain the best YCSB performance numbers when tuning HBase.

- During our testing, we noticed significant performance variations when the HBase cluster was not properly load balanced. The whole system runs at the speed of the slowest node. In a traditional HBase deployment, it is important to make sure the HBase regions are properly balanced across Hadoop nodes and that HDFS data is local to each HBase region server. With HP BDRA, the HDFS data is always remote, so from that perspective all HBase region servers are always balanced. For example, when you load data into HBase you have to make sure region servers do not go down while loading. If any region server goes down during load, the data won't be evenly spread across nodes. Also, if you run the HDFS balancer at any point after loading data into HBase, it will disrupt HDFS locality which has a significant performance impact in a traditional Hadoop deployment. None of these behaviors are noticed using HP BDRA.
- Use a small HBase slop factor such as 0.02 for the HBase balancer in order to properly balance the regions across all nodes. Note that HBase balancer works by number of regions per server, not their size or number of requests (hbase.regions.slop).
- Stop Hadoop services, such as flume, spark, etc., that are not in use. It is easier to troubleshoot issues and tune performance as some of the Java metrics cannot distinguish between various Hadoop services.
- Make sure the SSDs have readahead disabled by running "`blockdev --setra 0 /dev/sda`" on all Moonshot servers
- Make sure you presplit HBase regions correctly. We recommend one region per core. An example presplit for 1 billion rows and 360 regions is the following:

  ```
  create 'usertable', {NAME => 'f', BLOCKSIZE => 4096},{ SPLITS => (1..360-1).map
  {|i| "user#{sprintf('%010d',i*(1000000000)/360)}"}, MAX_FILESIZE => 4*1024**3}
  ```
- Use the smaller HBase blocksizes that are closer to request size; 4k works best for default YCSB settings. Be aware that a smaller HBase blocksize uses much more HBase blockcache memory. For example a 4k HBase blocksize uses approximately 30% more HBase blockcache memory compared to a 64k HBase blocksize. If you alter HBase blocksize for a table you need to `major_compact` the table in order for the new HBase blocksize to take effect.
- Another way to balance the system is to apply `move/merge/split` regions manually.
- Do not use compression as it will use less disk space, but significantly more memory and extra CPU.
- Use HBase offheap bucket cache instead of heap cache for memory testing. This way you avoid Java GC activity

YCSB example run

```
java -Djava.library.path=/usr/hdp/2.2.0.0-2041/hadoop/lib/native/ -cp
/usr/hdp/2.2.0.0-2041/hbase/conf/:/usr/hdp/2.2.0.0-
2041/hbase/lib/*:/usr/hdp/2.2.0.0-2041/hadoop/*:/usr/hdp/2.2.0.0-
2041/hadoop/lib/*:/usr/hdp/2.2.0.0-2041/hadoop-hdfs/*:/usr/hdp/2.2.0.0-
2041/hadoop-hdfs/lib/*:/ycsb-0.1.4/hbase-binding/lib/hbase-binding-
0.1.4.jar:/ycsb-0.1.4/core/lib/core-0.1.4.jar com.yahoo.ycsb.Client -db
com.yahoo.ycsb.db.HBaseClient -p columnfamily=f -P /ycsb-0.1.4/workloads/workloadc
-p recordcount=1000000000 -p operationcount=720000000 -p maxexecutiontime=1200 -p
table=usertable -p requestdistribution=uniform -p insertorder=ordered -p
insertstart=50000000 -p insertcount=50000000 -threads 300 -s -t
```

Deployment guidance

Deployment for testing involved tuning the HP BDRA system. For more information on tuning, please refer to the *Best practices for solution* section of this white paper. For more information and guidance on deployment, please refer to the white paper on HP Big Data Reference Architecture at
http://h20195.www2.hp.com/V2/GetDocument.aspx?docname=4AA5-6141ENW

Bill of materials

Bills of Materials (BOMs) provided here are based on the tested configuration for a single-rack HP BDRA solution featuring the following key components:

45 compute nodes: HP Moonshot m710

Four storage nodes: HP SL4540 Gen8

One management node: HP DL360p Gen8

Two head nodes: HP DL360p Gen8

Two ToR switches: HP 5930-32QSFP+

One HP iLO switch: HP 5900AF-48G-4XG-2QSFP+

Hortonworks Data Platform: HDP 2.2

Note
Part numbers are based on the part numbers that were used for testing and subject to change. The bill of materials does not include complete support options. If you have questions regarding ordering, please consult with your HP representative.

Compute nodes

Table 2 shows a BOM for a single HP Moonshot chassis with 45 cartridges for the compute nodes, as featured in the tested configuration.

Table 2. Bill of materials for HP Moonshot Chassis with 45 cartridges

Qty	Part Number	Description
1	755371-B21	HP Moonshot 1500 Chassis
4	684532-B21	HP 1500W Ht Plg Pwr Supply Kit
2	704654-B21	HP Moonshot-45XGc Switch Kit
2	704652-B21	HP Moonshot 4QSFP Uplink Kit
45	755860-B21	HP ProLiant m710 Server Cartridge
45	765483-B21	HP Moonshot 480GB M.2 2280 FIO Kit
1	681254-B21	HP 4.3U Rail Kit
45	C6N36A	HP Insight Control ML/DL/BL Bundle E-LTU
45	663201-B21	HP Insight Control ML/DL/BL FIO Bndl Lic (optional if E-LTU is not available)
45	G3J28AAE	RHEL Svr 2 Sckt/2 Gst 1yr 24x7 E-LTU

Storage

Table 3 provides a BOM for one SL4500 chassis with two SL4540 servers for the storage nodes. The tested solution featured two chassis and four servers.

Table 3. Bill of materials HP SL4500 chassis and SL4540 servers

Qty	Part Number	Description
1	663600-B22	HP 2xSL4500 Chassis
4	512327-B21	HP 750W CS Gold Ht Plg Pwr Supply Kit
1	681254-B21	HP 4.3U Rail Kit
1	681260-B21	HP 0.66U Spacer Blank Kit
2	664644-B22	HP 2xSL4540 Gen8 Tray Node Svr
2	740695-L21	HP SL4540 Gen8 E5 2450v2 FIO Kit
2	740695-B21	HP SL4540 Gen8 E5-2450v2 Kit
12	713985-B21	HP 16GB 2Rx4 PC3L-12800R-11 Kit
2	631681-B21	HP 2GB FBWC for P-Series Smart Array
4	655708-B21	HP 500GB 6G SATA 7.2k 2.5in SC MDL HDD
50	652766-B21	HP 3TB 6G SAS 7.2K 3.5in SC MDL HDD
2	692276-B21	HP Smart Array P420i Mezz Ctrllr FIO Kit
2	682632-B21	HP SL4500 Storage Mezz to PCIe Opt Kit
2	668943-B21	HP 12in Super Cap for Smart Array
2	G3J28AAE	RHEL Svr 2 Sckt/2 Gst 1yr 24x7 E-LTU
2	C6N27A	HP Insight Control Lic
2	649281-B21	HP IB FDR/EN 10/40Gb 2P 544QSFP Adptr
4	498385-B23	HP 3M 4X DDR/QDR QSFP IB Cu Cable

Management and head nodes

Table 4 shows a BOM for a single HP ProLiant DL360p Gen8 server used for the management and head nodes. The tested solution featured total of three servers.

Table 4. Bill of materials for HP ProLiant DL360p Gen8

Qty	Part Number	Description
1	654081-B21	HP DL360p Gen8 8-SFF CTO Server
1	712726-L21	HP DL360p Gen8 E5-2650v2SDHS FIO Kit
1	712726-B21	HP DL360p Gen8 E5-2650v2SDHS Kit
8	708641-B21	HP 16GB 2Rx4 PC3-14900R-13 Kit
8	652589-B21	HP 900GB 6G SAS 10K 2.5in SC ENT HDD
1	661069-B21	HP 512MB FBWC for P-Series Smart Array
1	649281-B21	HP IB FDR/EN 10/40Gb 2P 544QSFP Adptr
1	684208-B21	HP Ethernet 1GbE 4P 331FLR FIO Adptr
1	663201-B21	HP 1U SFF BB Gen8 Rail Kit
2	655874-B21	HP QSFP/SFP+ Adaptor Kit
2	656362-B21	HP 460W CS Plat PL Ht Plg Pwr Supply Kit
1	C6N36A	HP Insight Control ML/DL/BL FIO Bndl Lic
1	G3J28AAE	RHEL Svr 2 Sckt/2 Gst 1yr 24x7 E-LTU
2	SG508A	HP C13 - C14 WW 250V 10Amp IPD 1.37m 1pc Jumper Cord
3	498385-B23	HP 3M 4X DDR/QDR QSFP IB Cu Cable

Networking

Table 5 provides a BOM for two HP 5930 ToR switches and one HP 5900 iLO switch, as featured in the tested configuration for networking.

Table 5. Bill of materials for HP 5930 and HP 5900 switches

Qty	Part Number	Description
2	JG726A	HP FF 5930-32QSFP+ Switch
4	JG553A	HP X712 Bck(pwr)-Frt(prt) HV Fan Tray
4	JC680A	HP A58x0AF 650W AC Power Supply
4	JC680A#B2B	HP Power PDU Cable
1	JG510A	HP 5900AF-48G-4XG-2QSFP+ Switch
2	JC680A	HP A58x0AF 650W AC Power Supply
2	JC682A	HP 58xDAF Bck(pwr)-Frt(ports) Fan tray
4	JG330A	QSFP+ 4SFP+ 3m DAC cable
20	JG327A	HP X240 40G QSFP + QSFP 3m DAC cable
10	JG326A	HP X240 40G QSFP+ QSFP+ 1m DAC Cable
12	AF595A	HP 3.0M, Blue, CAT6 STP, Cable Data

Other hardware

Quantities listed in Table 6 are based on a full rack.

Table 6. Bill of materials for one HP 642 1200mm Shock Intelligent rack with four PDUs and other hardware

Qty	Part Number	Description
1	BW908A	HP 642 1200mm Shock Intelligent Rack
1	BW908A 001	HP Factory Express Base Racking Service
1	BW946A	HP 42U Location Discovery Kit
1	BW930A	HP Air Flow Optimization Kit
1	BW930A B01	Include with complete system
1	BW909A	HP 42U 1200mm Side Panel Kit
1	BW891A	HP Rack Grounding Kit
4	AF520A	HP Intelligent Mod PDU 24a Na/Jpn Core
6	AF547A	HP 5xC13 Intlgnt PDU Ext Bars G2 Kit

Software

Tables 7 and 8 show the BOMs for HP Insight CMU and Hortonworks HDP 2.2.

Table 7. Bill of materials for HP Insight CM with three-year subscription options

Qty	Part Number	Description
1	BD476AAE	HP Insight CMU 3yr 24x7 Flex E-LTU
1	BD477A	HP Insight CMU Media

Hortonworks software

Options listed in Table 8 are based on a single node. While HP is a certified reseller of Hortonworks software subscriptions, all application support (level-one through level-three) is provided by Hortonworks.

Table 8. Bill of materials for Hortonworks HDP 2.2

Qty	Part Number	Description
5	F5Z52A	Hortonworks Data Platform Enterprise 4 Nodes or 50TB Raw Storage 1 year 24x7 Support LTU

Summary

This white paper provided guidance on how to configure HBase to run on HP BDRA and provided the results of testing in each of these three scenarios:

- Multiple workloads and distributions while running with a dataset completely in memory
- Multiple workloads and distributions while running with a dataset that fits on SSD
- Multiple workloads and distributions while running with a dataset that does not fit on SSD

Overall, the results were outstanding and impressive. Not only were the performance numbers significant and record-breaking in terms of speed, but they also represented the highest YCSB HBase performance numbers achieved to date. Teaming HBase with HP BDRA would appear to make a perfect combination for big data analysis.

For a company looking to optimize support for different types of workloads and grow clusters where needed while minimizing TCO, the HBase/HP BDRA combination provides the best solutions.

Implementing a proof-of-concept

As a matter of best practice for all deployments, HP 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 HP Services representative (http://www8.hp.com/us/en/business-services/it-services/it-services.html) or your HP partner.

Appendix A: Ganglia troubleshooting

The following issues may occur when monitoring with Ganglia software. The error and fix for each issue are shown below.

500 Internal server error

The Ganglia web frontend displays a "500 internal server" error a few minutes after Ganglia service starts. The error is displayed only with IE; Google Chrome presents a blank page or the error number is not displayed. We found that this issue is related to PHP memory limit. The workaround is to increase memory_limit in php.ini to 512MB.

Ganglia server very slow

Rrdcached keeps the Ganglia server disk I/O at 100% the majority of the time. The HBase metrics for regions are enabled by default and this issue generates 99.5% of the traffic; however, these region metrics are generally not useful in day-to-day operations. We found that in order to filter properly, the settings on every region server need to be edited. The correct file to edit is: `/var/lib/ambari-agent/cache/stacks/HDP/2.0.6/services/HBASE/package/templates/hadoop-metrics2-hbase.properties-GANGLIA-RS.j2`.

No Ganglia charts

The Ganglia charts do not appear, instead a broken picture icon is displayed. We found that restarting Ganglia usually fixes this issue.

No matching metrics error

Although gstat shows that metrics are received from HDP* clusters, charts are obtained with error, "no matching metrics detected," including those for basic metrics such as cpu_system. We found that ensuring the metrics filtering is typo-free and that the server selection for chart generation is correct will address this issue.

Ganglia configuration changes not working

The location of the true Ganglia configuration file is `/var/www/html/ganglia/conf_default.php`, not in `./var/www/html/ganglia/ganglia-web/debian/conf.php`

HBase metrics not shown

HDP metrics name are in this format: `regionserver.Server.readRequestCount`, not in `hbase.regionserver.readRequestCount`.

For more information

HP Big Data Reference Architecture:
http://h20195.www2.hp.com/V2/GetDocument.aspx?docname=4AA5-6141ENW

Hortonworks-specific HP BDRA:
http://h20195.www2.hp.com/V2/GetDocument.aspx?docname=4AA5-6136ENW

Cloudera-specific HP BDRA:
http://h20195.www2.hp.com/V2/GetDocument.aspx?docname=4AA5-6137ENW

MapR-specific HP BDRA:
http://h20195.www2.hp.com/V2/GetDocument.aspx?docname=4AA5-7447ENW

HP Technology Consulting Services: hp.com/services/bigdata

HP Deployment Services: http://www8.hp.com/us/en/business-services/it-services.html?compURI=1078339

To help us improve our documents, please provide feedback at hp.com/solutions/feedback.

Sign up for updates
hp.com/go/getupdated

© Copyright 2015 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice. The only warranties for HP products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HP shall not be liable for technical or editorial errors or omissions contained herein.

Java is a registered trademark of Oracle and/or its affiliates. Red Hat is a registered trademark of Red Hat, Inc. in the United States and other countries. Linux is the registered trademark of Linus Torvalds in the U.S. and other countries.

4AA5-8757ENW, May 2015