Optimizing Dell PowerEdge Configurations for Hadoop

Understanding how to get the most out of Hadoop running on Dell hardware

A Dell technical white paper

July 2013

Michael Pittaro
Principal Architect, Big Data Solutions

Table of contents

July 2013 | Rev 1.0

Introduction

Hadoop is a flexible system, and has many possible applications with varying workloads. This means implementing Hadoop involves multiple decisions about software, tools, operations, servers, and network infrastructure. The Dell™ | Cloudera™ Solution for Apache™ Hadoop® was developed to streamline many of the choices involved in a Hadoop implementation. The solution embodies all the hardware, software, resources, and services needed to deploy and manage Hadoop in a production environment.

Selecting the most suitable hardware platform for Hadoop is a significant part of any implementation. The Dell | Cloudera Solution includes a reference architecture that provides a set of baseline configurations for Hadoop clusters running on Dell™ PowerEdge™ servers. These configurations have been developed jointly with Cloudera, and are based on extensive customer experience with real-world Hadoop production installations. The Dell | Cloudera Reference Architecture provides a general purpose configuration that can handle a large variety of workloads, and can scale from three data nodes to hundreds without changes.

However, the flexibility of Hadoop means that workloads will vary, and one configuration does not fit all potential applications. We regularly get questions about how to modify our published configurations, and the potential implications of those changes. This document provides guidelines for customizing the base configurations in the reference architecture, while remaining within the general guidelines of the reference architecture.

This document does not cover the implementation of heavily customized Hadoop clusters for dedicated workloads. Configuring those systems typically requires benchmark testing of the actual workloads on specific configurations to understand the complete performance characteristics. Also, heavily customized configurations are less flexible, and may not adapt as easily to changing workloads. Allowances for this must be factored into the design of a custom configuration. Dell can provide these customized configurations through a services engagement in collaboration with the Dell Solution Centers and the Enterprise Solution Group responsible for Dell's Big Data offerings.

When modifying the baseline configuration, there are a number of points to keep in mind:

1. Dell fully tests and certifies the actual configurations in the reference architecture. Modifications may not be fully tested as part of the normal Dell | Cloudera Solution release cycle.
2. Dell runs performance benchmarks against the actual configurations in the reference architecture.
3. The reference architecture configurations are readily available in the Dell Solution Centers for Architectural Design Sessions and Proof of Concept engagements. It is Dell's opinion that customized pilot or production-level Hadoop Deployments will require substantially more time to design, implement, and fine-tune.
4. Changing the configuration will impact the performance characteristics compared to the base configuration performance published by Dell.
5. If a production problem arises, Dell will not have the exact configuration immediately available to assist in troubleshooting, and there will be delays in resolving your issue.

1 General configuration

The Dell | Cloudera reference architecture divides the server nodes in a Hadoop cluster into two categories: infrastructure nodes and data nodes. Figure 1 is a high-level architecture diagram that illustrates the functions of the nodes in a minimum size cluster.

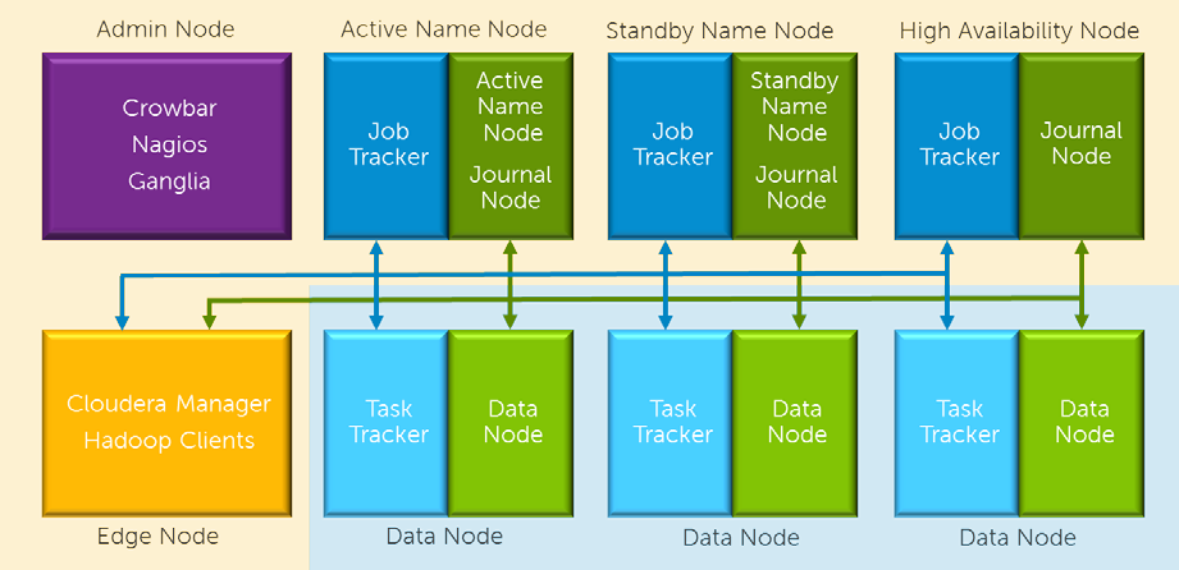

Figure 1 Hadoop node architecture

The infrastructure nodes are the hardware required for the core operations of the cluster. The Dell™ Crowbar software framework's administration node provides deployment, configuration management, and monitoring of the cluster, while the name nodes provide Hadoop Distributed File System (HDFS) directory and Map Reduce job tracking services. Typically, there are three name nodes set up in a high availability configuration. The edge node acts as a gateway to the cluster, and runs the Cloudera Manager server and various Hadoop client tools.

The data nodes are the workhorses of the cluster, and make up the bulk of the nodes in a typical cluster. Figure 1 shows the minimum of three. Clusters are scaled up to larger sizes by adding additional data nodes while using the same core infrastructure nodes.

Because of their different functions, these two types of nodes are configured differently. The configuration for all the data nodes should be consistent to avoid performance variations. All the infrastructure nodes also use consistent configurations. However, sometimes it's necessary to modify the configuration of the edge nodes (for details on this topic, see the "Edge node configuration" section below).

1.1 Cluster sizing

When customizing an implementation, it's best to start with an estimate of the overall cluster sizing characteristics, and then optimize the individual node configurations. There are two common approaches to overall cluster sizing —by storage size or by compute capacity. Since Hadoop scales both compute and storage capacity simultaneously by adding data nodes, the correct approach is to consider both compute and storage in the sizing estimates.

Cluster sizing should take several factors into account:

1. The expected balance of processing and storage
2. The initial storage requirements
3. The amount of new data ingested over time — daily, weekly, or monthly
4. The expected growth rate of newly ingested data
5. The HDFS replication value — typically 3, but variable at the file level
6. The amount of temporary storage for MapReduce — 20% is a good start

Based on these factors, the storage capacity can be calculated. This storage capacity can then be used to calculate both the number of data nodes and the desired storage per node. Depending on the desired storage/processing balance, there are multiple options, from compute heavy to storage heavy. The primary difference between these options is the rate at which compute and storage grow as nodes are added to the cluster.

In practice, there are other physical considerations—such as rack space, power and cooling density, network density, and cost—that also affect the final choice of data node size. The current trend has been toward Hadoop data nodes with approximately 24–36TB of raw storage. This provides a good balance between density and performance, but should be considered a starting point rather than a hard rule.

The remainder of this document provides guidance on optimizing the characteristics of the individual nodes based on the general cluster sizing results.

1.2 Disk drive configuration

1.2.1 Infrastructure node configuration

Infrastructure nodes, including the master and secondary name nodes, store critical metadata for the operation of the Hadoop cluster. They are normally configured as a single RAID 10 file system with a hardware RAID controller for best performance and reliability. We configure approximately 4TB of raw storage on these nodes, resulting in 2TB of usable storage. This is more than enough storage for the OS, and the assorted data files used by the master and secondary name node processes. Adding more disks or using larger disks won't significantly affect cluster performance, and should not be necessary. The drives should be reliable, so we specify enterprise SAS or NL-SAS drives, and these drive types should not be changed. Do not use consumer-grade drives in infrastructure nodes.

1.2.2 Data node configuration

The data nodes perform the bulk of the disk I/O in a Hadoop system. Hadoop MapReduce jobs are broken into tasks, and each task is scheduled on a data node, usually processing data from local disks. I/O is performed sequentially on large blocks of data. Because of the way the Hadoop system parallelizes the I/O processing and HDFS distributes its blocks, there is a high probability that multiple tasks running on a data node will access independent spindles. As a result, the reference architecture aims for a 1 core/1 spindle ratio in the data nodes, with each spindle configured as a separate Linux file system. The reference architecture also specifies the largest drives available (currently 3TB for 3.5-inch drives, and 1TB for 2.5-inch drives), providing a spindle-dense and deep storage configuration.

We use SATA 3Gb/sec, SATA 6Gb/sec, or SAS 6Gb/sec interfaces and controllers, depending on what the particular server platform supports. For drive types, we use enterprise SATA or NL-SAS drives. NL-SAS drives provide the best performance

under heavy I/O, while the SATA drives provide equivalent transfer rates but are slightly slower for multiple concurrent requests. Consumer-grade SATA drives are not recommended for data nodes.

Changing the drive types and controllers is not recommended, since the reference architecture specifies the best performing combinations for each PowerEdge platform.

When changing drive sizes, larger drives have slightly slower seek times. Since Hadoop primarily uses sequential I/O on large blocks of data, differences in seek time have a minimal impact on performance.

The main customization for disk configuration in data nodes is to optimize the spindle/core/storage depth ratio. Hadoop is designed to simultaneously scale compute and storage by adding nodes. A cluster can be built with a spindle-dense, deep storage configuration (the default configuration in the reference architecture) or less deep, less dense storage.

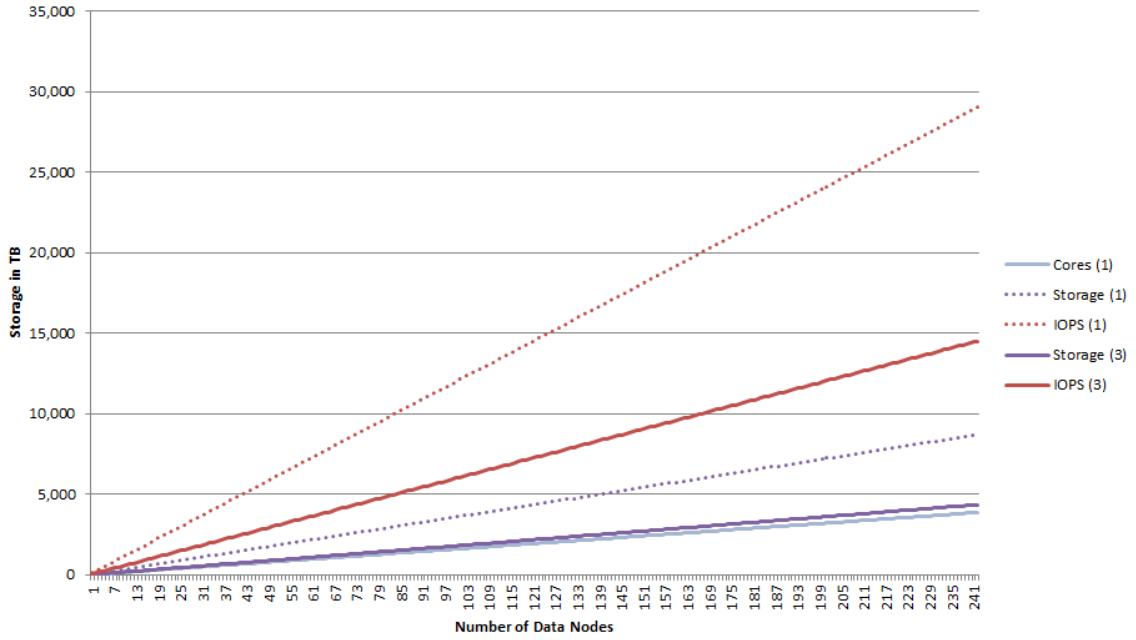

Figure 2 Scaling depth versus width

Figure 2 illustrates the differences in scaling a cluster with data nodes using 12 x 3TB data spindles (legend 1) versus data nodes with 6 x 3TB spindles (legend 3). In both instances, storage, spindles, and compute capacity grow linearly, but at different rates. Using fewer spindles per node results in slower scaling of total I/O capacity (IOPS) compared to compute, and is better for very compute-intensive workloads. Using more spindles per node results in faster scaling of total I/O capacity (IOPS). The same logic applies to total storage per node; nodes with more total storage increase storage at a higher rate

The correct balance depends on the actual workloads and planned growth. If fewer spindles are configured, then future storage growth can be accommodated by adding additional drives instead of additional nodes. However, from a practical point of view, the physical effort involved in adding drives has resulted in a consensus among the Hadoop user community to scale by adding nodes instead of retrofitting existing nodes.

Another factor to consider when changing the spindle/core/storage depth is the amount of active data versus archived data. Some clusters store large amounts of data, but the actively processed data sets are smaller. For these clusters, deep storage may be appropriate even though workload analysis indicates a higher core/storage depth is appropriate.

Changes to disk configuration should also take Hadoop's HDFS replication into account. HDFS stores multiple copies of each HDFS data block for reliability and to improve the potential for parallel processing. The default replication factor is three, and is applied when a file is created. A different replication factor can be specified when a file is created, and the replication factor can also be changed for existing files. The ability to change replication affects the amount of storage required and performance. It also affects the amount of memory required by the name node process.

1.3 Memory configuration

1.3.1 Infrastructure node configuration

The infrastructure nodes are configured with 128GB of memory. The master and secondary name node processes serve their metadata from memory, so optimal performance requires a larger memory configuration. The job tracker also requires additional memory. The larger memory configuration is also useful for the Crowbar administration node.

The name node process uses approximately 1GB of memory per million HDFS blocks. A large number of small files or longer file names increase this usage, so it's just an approximation. HDFS block sizes can vary per file, but 128MB is a common block size. Accounting for a replication factor of 3 (a common default), 24GB of memory is used by the name node process for a 1PB cluster.

The job tracker memory requirements are harder to predict, since they depend on the amount of history the job tracker is configured to maintain, and the number of Hadoop tasks in the individual jobs.

After accounting for name node and job tracker memory requirements and allowing for other processes and operating system overhead, the reference architecture memory configuration will scale comfortably to a cluster with between 4PB and 5PB of available storage without modification.

When configuring memory, it is also important to match DIMM sizes and to populate all processor memory channels for optimal performance. For Dell™ PowerEge™ R720 and PowerEdge™ C8000 servers, there are four channels and two processor sockets, so a minimum of eight DIMMs is recommended. With 8GB DIMMS, that is a practical minimum of 64GB.

Since there are only five infrastructure nodes per cluster—after accounting for the current "sweet spot" for DIMM pricing, planning for future expansion, and optimizing physical memory layout—most configurations will end up between 64GB and 128GB of memory at a good price point.

While it is possible to reduce the infrastructure node memory size to as little as 32GB, this should only be done for smaller clusters during pilot or development work where the cluster will never grow very large, say 15 nodes or 150TB of available storage.

1.3.2 Data node configuration

The data nodes are configured with 64MB of memory. This provides plenty of headroom for a general purpose cluster, including HBase usage.

The actual amount of memory required on a data node depends on the number of tasks allowed to run concurrently in the Hadoop configuration and the amount of memory required in each task. Memory usage per task will vary depending on the actual job characteristics. A good baseline is to assume each task needs between 2GB and 4GB of memory. The base 64MB configuration will support between 16 and 32 concurrent tasks. Since the reference architecture specifies 12 or 16 cores per data node, this is a good general purpose configuration.

If the planned usage of the cluster is known, then it's possible to decrease or increase the amount of memory to match the workloads. This requires characterization of the workloads, their memory usage, and analysis of concurrent workloads.

In general, the main change to make here is to increase memory rather than decrease it.

1.4 Network configuration

The reference architecture supports two options for networking—1GbE or 10GbE. Both configurations use pairs of bonded connections to alternate switches, resulting in 2GbE or 20GbE throughput, with redundancy.

Infrastructure nodes have four network connections (plus a BMC connection). Two of these are used for the cluster production network, and two are used to connect the Crowbar admin and edge nodes to the edge network. These connections are split across two physical networks cards for additional redundancy.

Data nodes have two network connections connected to the cluster production network, plus a baseboard management controller (BMC) connection.

All nodes have an additional 1GbE network connection to facilitate management and PXE booting.

The primary customization is to decide whether to build the cluster with 10GbB or 1GbE networking. If data nodes are configured with large amounts of storage (48TB or more) then 10GbE networking is recommended to minimize the impacts of replication when a node fails. Implementing 10GbE networking also increases Hadoop performance in general, since it provides more bandwidth for data replication when writing data and for transfer between nodes during the shuffle phase of MapReduce processing.

Changing the specified network cards is not recommended, since the cards have been heavily tested in these particular configurations with bonding, and all the possible variations like driver support and firmware have been accounted for.

The reference architecture specifies Dell™ Force10™ S60 and S4810 switches for the 1GbE and 10GbE configurations, respectively. We use these switches for their deep buffers, bonding support, and stacking capabilities. Replacing the switches with other models is possible but may have significant impacts on performance if the alternate switches have significantly different specifications. Repurposing existing switches is a common request, but this can be difficult, and the success of this approach depends on the actual switch models.

The reference architecture specifies fully redundant networking with bonding, and is intended for production clusters. It is possible to eliminate the redundant network and reduce costs. This is only recommended for pilot or test clusters, and will reduce the available bandwidth. This will reduce performance and resilience on the pilot cluster.

1.5 Edge node configuration

The edge node specified in the reference architecture is used as a general purpose server to run various Hadoop "client" applications. Its defining characteristic is connections to both the Hadoop production network and the corporate LAN via the edge network. It is essentially a gateway to the Hadoop cluster. The base configuration matches the other infrastructure nodes, and is a good starting point. However, the configuration of the edge node can vary widely depending on its intended use, so customization of this node is common. The edge node does not run any of the core Hadoop infrastructure processes. However, we recommend installing the Cloudera Manager server and associated databases on this node.

The main consideration for the edge node is to determine its actual usage, and then work toward the correct configuration.

For networking, we specify four network connections. Two of these are for the core cluster production network. The remaining ports can be used to connect to the main or corporate LAN, or to connect to data load or ingest sources.

Memory configuration is entirely dependent on the applications running, and should be adjusted as appropriate.

If the edge node is being used for data ingest, it may be necessary to account for additional disk space for data staging or intermediate files. Expanding the existing RAID 10 configuration is probably the best approach here, but other configurations are also possible. When modifying the amount of storage available, remember to account for the Cloudera Manager databases, which should be on reliable storage.

In some clusters, more than one edge node may be required. For clusters with very high ingest rates, it is common to spread the load across multiple edge nodes. Multiple edge nodes may also be used to host web services or applications for end users.

2 Physical and logistical considerations

Customizations to the reference architecture will affect the power consumption and density of the cluster. Depending on the actual changes and the size of the cluster, these effects may be significant, and need to be accounted for in the site planning and preparation phase. Also, verify that changes that appear good on paper are actually valid configurations for the particular PowerEdge platform, since there are power considerations related to processor, memory, and disk configurations.

When specifying customized configurations, it is important to consider maintenance issues. Uniform configurations make it easier to stock spares or repurpose nodes when failures occur.

3 Crowbar considerations

The guidelines in this document are compatible with clusters deployed and managed through the Dell Crowbar software framework.

However, if changes to server models or controllers are made, they may not be compatible with the reference architecture defaults Crowbar knows about. The following section explains the impacts of server or controller changes, and some possible solutions for custom hardware configurations.

3.1 Using configurations not in the reference architecture

The Dell Crowbar software framework has internal knowledge of PowerEdge servers, disk controllers, and disk configurations that are relevant to the reference architecture configurations. Controller and BIOS information is included in the RAID and BIOS barclamps, while disk layout is included as part of the Cloudera Manager Barclamp.

Changes to server models or controllers are possible, but may not be compatible with the reference architecture defaults Crowbar knows about.

When a node is allocated, the BIOS and RAID configurations on the node are updated based on the settings chosen in the node allocation screen. As long as Crowbar recognizes the server, the BIOS will usually be configured. If Crowbar does not recognize the server/RAID controller combination, the RAID controllers will be left untouched. Community releases of Crowbar handle controller and BIOS configuration differently from Dell Crowbar—they do not make any changes to BIOS or disk controller configurations.

It is possible to create a custom configuration by configuring BIOS and RAID BIOS settings before nodes are allocated in Crowbar. If the BIOS and RAID barclamps are disabled (or a community release is used), then the custom configuration is used instead. However, this requires a manual configuration step for each server.

Disk layout for the operating system is configured when a node is allocated. The operating system is installed on the first drive enumerated at boot. Enumeration order varies depending on the actual hardware configuration, so ensure the OS is installed on the correct volume and the partition is bootable.

Disk layout for the remaining drives is configured when the Cloudera Manager Barclamp is deployed. For data nodes, each drive found is partitioned, formatted, and mounted as an ext3 filesystem. For infrastructure nodes, no additional drives are configured; the barclamp assumes that a large partition was already created at OS install time, on a reliable RAID volume.

4 References

The Dell website publishes detailed memory information and recommendations for PowerEdge servers at http://www.dell.com/poweredge/memory

Hadoop Operations : A Guide for Developers and Administrators [1] is a good reference for general Hadoop configuration guidelines, as well as more detail on the internal operations of Hadoop from an administrator perspective.

Cloudera documentation provides many configuration guidelines. The blog post on hardware recommendations [2] is particularly useful.

Yahoo Developer network has published some information on HDFS and name node scalability recommendations in Scalability of the Hadoop Distributed File System [3]

[1] http://shop.oreilly.com/product/0636920025085.do
[2] http://blog.cloudera.com/blog/2010/03/clouderas-support-team-shares-some-basic-hardware-recommendations/
[3] http://developer.yahoo.com/blogs/hadoop/posts/2010/05/scalability_of_the_hadoop_dist/

To learn more

For more information about the Dell | Cloudera Solution for Apache Hadoop, visit:

www.dell.com/hadoop