Hardware and Software Requirements

**HARDWARE REQUIREMENTS**

Hadoop is designed to run on industry-standard hardware, recommending an ideal cluster configuration is not as easy as delivering a list of hardware specifications. Selecting hardware that provides the best balance of performance and economy for a given workload requires testing and validation.

## STORAGE WITH COMPUTE

Hadoop distributes data across a cluster of balanced machines and uses replication to ensure data reliability and fault tolerance. Because data is distributed on machines with compute power, processing can be sent directly to the machines storing the data. Since each machine in a Hadoop cluster stores as well as processes data, those machines need to be configured to satisfy both data storage and processing requirements.

## WORKLOADS

In nearly all cases, a MapReduce job will either encounter a bottleneck reading data from disk or from the network (known as an IO-bound job) or in processing data (CPU-bound). An example of an IO-bound job is sorting, which requires very little processing (simple comparisons) and a lot of reading and writing to disk. An example of a CPU-bound job is classification, where some input data is processed in very complex ways to determine ontology.

Here are several more examples of IO-bound workloads:

* Indexing
* Grouping
* Data importing and exporting
* Data movement and transformation

Here are several more examples of CPU-bound workloads:

* Clustering/Classification
* Complex text mining
* Natural-language processing
* Feature extraction

## SELECTING HARDWARE FOR CLUSTER

The first step in choosing a machine configuration is to understand the type of hardware your operations team already manages. Operations teams often have opinions or hard requirements about new machine purchases, and will prefer to work with hardware with which they’re already familiar. Hadoop is not the only system that benefits from efficiencies of scale. Again, as a general suggestion, if the cluster is new or you can’t accurately predict your ultimate workload, it is advised that you use balanced hardware.

There are four types of roles in a basic Hadoop cluster: NameNode (and Standby NameNode), JobTracker, TaskTracker, and DataNode. (A node is a machine performing a particular task.) Most machines in your cluster will perform two of these roles, functioning as both DataNode (for data storage) and TaskTracker (for data processing).

Here are the recommended specifications for DataNode/TaskTrackers in a balanced Hadoop cluster:

* 12-24 1-4TB hard disks in a JBOD (Just a Bunch Of Disks) configuration
* 2 quad-/hex-/octo-core CPUs, running at least 2-2.5GHz
* 64-512GB of RAM
* Bonded Gigabit Ethernet or 10Gigabit Ethernet (the more storage density, the higher the network throughput needed)

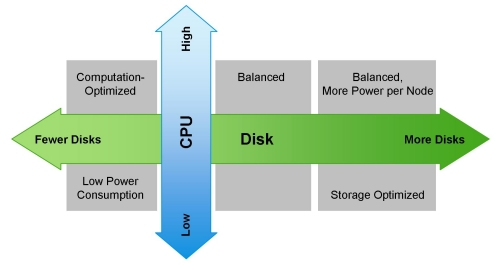
The NameNode role is responsible for coordinating data storage on the cluster, and the JobTracker for coordinating data processing. (The Standby NameNode should not be co-located on the NameNode machine for clusters and will run on hardware identical to that of the NameNode.) Cloudera recommends that customers purchase enterprise-class machines for running the NameNode and JobTracker, with redundant power and enterprise-grade disks in RAID 1 or 10 configurations.

The NameNode will also require RAM directly proportional to the number of data blocks in the cluster. A good rule of thumb is to assume 1GB of NameNode memory for every 1 million blocks stored in the distributed file system. With 100 DataNodes in a cluster, 64GB of RAM on the NameNode provides plenty of room to grow the cluster. We also recommend having HA configured on both the NameNode and JobTracker, features that have been available in the CDH4 line for some time.

Here are the recommended specifications for NameNode/JobTracker/Standby NameNode nodes. The drive count will fluctuate depending on the amount of redundancy:

* 4–6 1TB hard disks in a JBOD configuration (1 for the OS, 2 for the FS image [RAID 1], 1 for Apache ZooKeeper, and 1 for Journal node)
* 2 quad-/hex-/octo-core CPUs, running at least 2-2.5GHz
* 64-128GB of RAM
* Bonded Gigabit Ethernet or 10Gigabit Ethernet

The following diagram shows how a machine should be configured according to workload:



# OTHER CONSIDERATIONS

It is important to remember that the Hadoop ecosystem is designed with a parallel environment in mind. When purchasing processors, we do not recommended getting the highest GHz chips, which draw high watts (130+). This will cause two problems: higher consumption of power and greater heat expulsion. The mid-range models tend to offer the best bang for the buck in terms of GHz, price, and core count.

When we encounter applications that produce large amounts of intermediate data — outputting data on the same order as the amount read in — we recommend two ports on a single Ethernet card or two channel-bonded Ethernet cards to provide 2 Gbps per machine. Bonded 2Gbps is tolerable for up to about 12TB of data per nodes. Once you move above 12TB, you will want to move to bonded 4Gbps(4x1Gbps). Alternatively, for customers that have already moved to 10 Gigabit Ethernet or Infiniband, these solutions can be used to address network-bound workloads. Confirm that your operating system and BIOS are compatible if you’re considering switching to 10 Gigabit Ethernet.

When computing memory requirements, remember that Java uses up to 10 percent of it for managing the virtual machine. We recommend configuring Hadoop to use strict heap size restrictions in order to avoid memory swapping to disk. Swapping greatly impacts MapReduce job performance and can be avoided by configuring machines with more RAM, as well as setting appropriate kernel settings on most Linux distributions.

It is also important to optimize RAM for the memory channel width. For example, when using dual-channel memory, each machine should be configured with pairs of DIMMs. With triple-channel memory each machine should have triplets of DIMMs. Similarly, quad-channel DIMM should be in groups of four.

**SOFTWARE REQUIREMENTS**

# SUPPORTEED OPERATING SYSTEMS FOR CLOUDERA MANAGER

Cloudera Manager supports a range of operating systems including:

* Red Hat-compatible systems
* Red Hat Enterprise Linux 5.7 and CentOS 5.7, 64-bit
* Red Hat Enterprise Linux 6.2 and 6.4, and CentOS 6.2 and 6.4, 64-bit
* Oracle Enterprise Linux 5.6 with Unbreakable Enterprise Kernel, 64-bit

* SLES systems
* SUSE Linux Enterprise Server 11, 64-bit. Service Pack 1 or later is required. Also, the SUSE Linux Enterprise Software Development Kit 11 SP1 is required on cluster hosts running the Cloudera Manager Agents (not required on the Cloudera Manager Server host);

* Debian systems
* Debian 6.0 (Squeeze), 64-bit
* Ubuntu systems
* Ubuntu 10.04 (Lucid Lynx), 64-bit
* Ubuntu 12.04 (Precise Pangolin), 64-bit

# SUPPORTED BROWSERS FOR CLOUDERA MANAGER ADMIN CONSOLE

The Cloudera Manager Admin console, which you use to configure, manage, and monitor CDH, supports the following browsers:

* Firefox 11 or later
* Google Chrome
* Internet Explorer 9
* Safari 5 or later

# RESOURCES

Cloudera Manager requires sufficient:

* Disk space. A minimum of:
* 5 GB on the partition hosting /var.
* 500 MB on the partition hosting /usr.
* RAM. 4GB is appropriate for most cases, and is required when using Oracle databases. 2GB may be sufficient for non-Oracle deployments involving fewer than 100 hosts.