[Cloudera Engineering Blog (http://blog.cloudera.com/)](http://blog.cloudera.com/)

Best practices, how-tos, use cases, and internals from Cloudera Engineering and the community

SEARCH

How-to: Select the Right Hardware for Your New Hadoop Cluster

August 28, 2013 (http://blog.cloudera.com/blog/2013/08/how-to-select-the-right-hardware-for-your-new-hadoop-cluster/) | By Kevin O'Dell (http://blog.cloudera.com/?guest-author=Kevin O'Dell) | 11 Comments (http://blog.cloudera.com/blog/2013/08/how-to-select-the-right-hardware-for-your-new-hadoop-cluster/#comments)

Categories: Hadoop (http://blog.cloudera.com/blog/category/hadoop/) Hardware (http://blog.cloudera.com/blog/category/hardware/) How-to (http://blog.cloudera.com/blog/category/how-to/) Performance (http://blog.cloudera.com/blog/category/performance/) Use Case (http://blog.cloudera.com/blog/category/use-case/)

One of the first questions Cloudera customers raise when getting started with Apache Hadoop is how to select appropriate hardware for their new Hadoop clusters.

Although 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. (For example, users with IO-intensive workloads will invest in more spindles per core.)

In this blog post, you'll learn some of the principles of workload evaluation and the critical role it plays in hardware selection. You'll also learn the various factors that Hadoop administrators should take into account during this process.

Marrying Storage with Compute

Over the past decade, IT organizations have standardized on blades and SANs (Storage Area Networks) to satisfy their grid and processing-intensive workloads. While this model makes a lot of sense for a number of standard applications such as web servers, app servers, smaller structured databases, and data movement, the requirements for infrastructure have changed as the amount of data and number of users has grown. Web servers now have caching tiers, databases have gone massively parallel with local disk, and data movement jobs are pushing more data than they can handle locally.

Hardware vendors have created innovative systems to address these requirements including storage blades, SAS (Serial Attached SCSI) switches, external SATA arrays, and larger capacity rack units. However, Hadoop is based on a new approach to storing and processing complex data, with data movement minimized. Instead of relying on a SAN for massive storage and reliability then moving it to a collection of blades for processing, Hadoop handles large data volumes and reliability in the software tier.

Most teams building a Hadoop cluster don't yet know the eventual profile of their workload.

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.

Why Workloads Matter

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

Tweets by @ClouderaEng

Cloudera Engineering
@ClouderaEng

Robust Message Serialization in Apache Kafka Using Apache Avro, Part 3blog.cloudera.com/blog/2018/08/r…

Embed View on Twitter

Categories

Accumulo (http://blog.cloudera.com/blog/category/accumulo/) (1)
Altus (http://blog.cloudera.com/blog/category/altus/) (10)
Analytic Database (http://blog.cloudera.com/blog/category/analytic-database/) (4)
Avro

Tags

analysis (http://blog.cloudera.com/blog/tag/analysis/) analytics (http://blog.cloudera.com/blog/tag/analytics/) apache (http://blog.cloudera.com/blog/tag/apache/) apache hadoop (http://blog.cloudera.com/blog/tag/apache-hadoop/) Apache HBase (http://blog.cloudera.com/blog/tag/apache-hbase/) apache hive (http://blog.cloudera.com/blog/tag/apache-hive/) beta (http://blog.cloudera.com/blog/tag/beta/) Big Data (http://blog.cloudera.com/blog/tag/big-data/) CDH (http://blog.cloudera.com/blog/tag/cdh/) cloudera (http://blog.cloudera.com/blog/tag/cloudera/) Cloudera Manager (http://blog.cloudera.com/blog/tag/cloudera-manager/) Community (http://blog.cloudera.com/blog/tag/community

- Data movement and transformation

Here are several more examples of CPU-bound workloads:

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

Because Cloudera's customers need to thoroughly understand their workloads in order to fully optimize Hadoop hardware, a classic chicken-and-egg problem ensues. Most teams looking to build a Hadoop cluster don't yet know the eventual profile of their workload, and often the first jobs that an organization runs with Hadoop are far different than the jobs that Hadoop is ultimately used for as proficiency increases. Furthermore, some workloads might be bound in unforeseen ways. For example, some theoretical IO-bound workloads might actually be CPU-bound because of a user's choice of compression, or different implementations of an algorithm might change how the MapReduce job is constrained. For these reasons, when the team is unfamiliar with the types of jobs it is going to run, as an initial approach it makes sense to invest in a balanced Hadoop cluster.

The next step would be to benchmark MapReduce jobs running on the balanced cluster to analyze how they're bound. To achieve that goal, it's straightforward to measure live workloads and determine bottlenecks by putting thorough monitoring in place. We recommend installing Cloudera Manager on the Hadoop cluster to provide real-time statistics about CPU, disk, and network load. (Cloudera Manager is an included component of Cloudera Standard and Cloudera Enterprise — in the latter case with enterprise functionality, such as support for rolling upgrades, in place.) With Cloudera Manager installed, Hadoop administrators can then run their MapReduce jobs and check the Cloudera Manager dashboard to see how each machine is performing.

In addition to building out a cluster appropriate for the workload, we encourage customers to work with their hardware vendor to understand the economics of power and cooling. Since Hadoop runs on tens, hundreds, or thousands of nodes, an operations team can save a significant amount of money by investing in power-efficient hardware. Each hardware vendor will be able to provide tools and recommendations for how to monitor power and cooling.

> The first step is to know which hardware your operations team already manages.

Selecting Hardware for Your CDH 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, we advise 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:

/) configuration (http://blog.cloudera.com/blog/tag/configuration/) data (http://blog.cloudera.com/blog/tag/data/) developer (http://blog.cloudera.com/blog/tag/developer/) developers (http://blog.cloudera.com/blog/tag/developers/) development (http://blog.cloudera.com/blog/tag/development/) events (http://blog.cloudera.com/blog/tag/events-2/) Flume (http://blog.cloudera.com/blog/tag/flume/) Guest (http://blog.cloudera.com/blog/tag/guest/) Hadoop (http://blog.cloudera.com/blog/tag/hadoop/) HBase (http://blog.cloudera.com/blog/tag/hbase/) HDFS (http://blog.cloudera.com/blog/tag/hdfs/) Hive (http://blog.cloudera.com/blog/tag/hive/) Hue (http://blog.cloudera.com/blog/tag/hue/) impala (http://blog.cloudera.com/blog/tag/impala-2/) installation (http://blog.cloudera.com/blog/tag/installation/) java (http://blog.cloudera.com/blog/tag/java/) log (http://blog.cloudera.com/blog/tag/log/) logs (http://blog.cloudera.com/blog/tag/logs/) MapReduce (http://blog.cloudera.com/blog/tag/mapreduce/) open source (http://blog.cloudera.com/blog/tag/open-source/) Pig (http://blog.cloudera.com/blog/tag/pig/) platform (http://blog.cloudera.com/blog/tag/platform/) python (http://blog.cloudera.com/blog/tag/python/) questions (http://blog.cloudera.com/blog/tag/questions/) R (http://blog.cloudera.com/blog/tag/r/) release (http://blog.cloudera.com/blog/tag/release/) REST (http://blog.cloudera.com/blog/tag/rest/) Search (http://blog.cloudera.com/blog/tag/search/) security (http://blog.cloudera.com/blog/tag/security/) sql (http://blog.cloudera.com/blog/tag/sql/) Support (http://blog.cloudera.com/blog/tag/support-2/) Testing (http://blog.cloudera.com/blog/tag/testing/) use cases (http://blog.cloudera.com/blog/tag/use-cases/)

Archives

- 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

If you expect your Hadoop cluster to grow beyond 20 machines, we recommend that the initial cluster be configured as if it were to span two racks, where each rack has a top-of-rack 10 GigE switch. As the cluster grows to multiple racks, you will want to add redundant core switches to connect the top-of-rack switches with 40GigE. Having two logical racks gives the operations team a better understanding of the network requirements for intra-rack and cross-rack communication.

> Remember, the Hadoop ecosystem is designed with a parallel environment in mind.

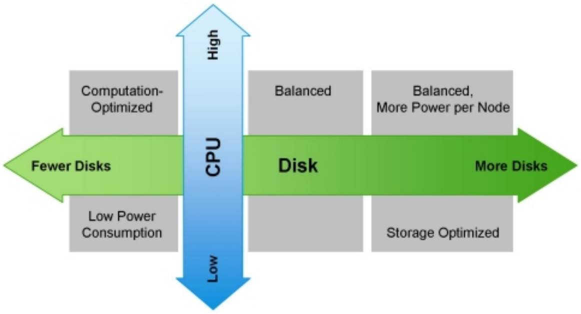
With a Hadoop cluster in place, the team can start identifying workloads and prepare to benchmark those workloads to identify hardware bottlenecks. After some time benchmarking and monitoring, the team will understand how additional machines should be configured. Heterogeneous Hadoop clusters are common, especially as they grow in size and number of use cases – so starting with a set of machines that are not "ideal" for your workload will not be a waste of time. Cloudera Manager offers templates that allow different hardware profiles to be managed in groups, making it simple to manage heterogeneous clusters.

Below is a list of various hardware configurations for different workloads, including our original "balanced" recommendation:

- Light Processing Configuration (1U/machine): Two hex-core CPUs, 24-64GB memory, and 8 disk drives (1TB or 2TB)
- Balanced Compute Configuration (1U/machine): Two hex-core CPUs, 48-128GB memory, and 12 – 16 disk drives (1TB or 2TB) directly attached using the motherboard controller. These are often available as twins with two motherboards and 24 drives in a single 2U cabinet.
- Storage Heavy Configuration (2U/machine): Two hex-core CPUs, 48-96GB memory, and 16-24 disk drives (2TB – 4TB). This configuration will cause high network traffic in case of multiple node/rack failures.
- Compute Intensive Configuration (2U/machine): Two hex-core CPUs, 64-512GB memory, and 4-8 disk drives (1TB or 2TB)

(Note that Cloudera expects to adopt 2×8, 2×10, and 2×12 core configurations as they arrive.)

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.

Beyond MapReduce

Hadoop is far bigger than HDFS and MapReduce; it's an all-encompassing data platform. For that reason, CDH includes many different ecosystem products (and, in fact, is rarely used solely for MapReduce). Additional software components to consider when sizing your cluster include Apache HBase, Cloudera Impala, and Cloudera Search. They should all be run on the DataNode process to maintain data locality.

HBase is a reliable, column-oriented data store that provides consistent, low-latency, random read/write access. Cloudera Search solves the need for full text search on content stored in CDH to simplify access for new types of users, but also open the door for new types of data storage inside Hadoop. Cloudera Search is based on Apache Lucene/Solr Cloud and Apache Tika and extends valuable functionality and flexibility for search through its wider integration with CDH. The Apache-licensed Impala project brings scalable parallel database technology to Hadoop, enabling users to issue low-latency SQL queries to data stored in HDFS and HBase without requiring data movement or transformation.

Focusing on resource management will be your key to success.

HBase users should be aware of heap-size limits due to garbage collector (GC) timeouts. Other JVM column stores also face this issue. Thus, we recommend a maximum of ~16GB heap per Region Server. HBase does not require too many other resources to run on top of Hadoop, but to maintain real-time SLAs you should use schedulers such as fair and capacity along with Linux Cgroups.

Impala uses memory for most of its functionality, consuming up to 80 percent of available RAM resources under default configurations, so we recommend at least 96GB of RAM per node. Users that run Impala alongside MapReduce should consult our recommendations in "Configuring Impala and MapReduce for Multi-tenant Performance." (http://blog.cloudera.com/blog/2013/06/configuring-impala-and-mapreduce-for-multi-tenant-performance/) It is also possible to specify a per-process or per-query memory limit for Impala.

Search is the most interesting component to size. The recommended sizing exercise is to purchase one node, install Solr and Lucene, and load your documents. Once the documents are indexed and searched in the desired manner, scalability comes into play. Keep loading documents until the indexing and query latency exceed necessary values to the project — this will give you a baseline for max documents per node based on available resources and a baseline count of nodes not including and desired replication factor.

Conclusions

Purchasing appropriate hardware for a Hadoop cluster requires benchmarking and careful planning to fully understand the workload. However, Hadoop clusters are commonly heterogeneous and Cloudera recommends deploying initial hardware with balanced specifications when getting started. It is important to remember when using multiple ecosystem components resource usage will vary and focusing on resource management will be your key to success.

We encourage you to chime in about your experience configuring production Hadoop clusters in comments!

Kevin O'Dell is a Systems Engineer at Cloudera.

apache (http://blog.cloudera.com/blog/tag/apache/) apache hadoop (http://blog.cloudera.com/blog/tag/apache-hadoop/) Apache HBase (http://blog.cloudera.com/blog/tag/apache-hbase/) apache zookeeper (http://blog.cloudera.com/blog/tag/apache-zookeeper/) CDH (http://blog.cloudera.com/blog/tag/cdh/) CDH4 (http://blog.cloudera.com/blog/tag/cdh4/) cloud (http://blog.cloudera.com/blog/tag/cloud/) cloudera (http://blog.cloudera.com/blog/tag/cloudera/) Cloudera Customers (http://blog.cloudera.com/blog/tag/cloudera-customers/) cloudera enterprise (http://blog.cloudera.com/blog/tag/cloudera-enterprise/) Cloudera Manager (http://blog.cloudera.com/blog/tag/cloudera-manager/) configuration (http://blog.cloudera.com/blog/tag/configuration/) configuring hadoop (http://blog.cloudera.com/blog/tag/configuring-hadoop/) data (http://blog.cloudera.com/blog/tag/data/) Hadoop (http://blog.cloudera.com/blog/tag/hadoop/) HBase (http://blog.cloudera.com/blog/tag/hbase/) HDFS (http://blog.cloudera.com/blog/tag/hdfs/) impala (http://blog.cloudera.com/blog/tag/impala-2/) java (http://blog.cloudera.com/blog/tag/java/) MapReduce (http://blog.cloudera.com/blog/tag/mapreduce/) monitoring (http://blog.cloudera.com/blog/tag/monitoring/) platform (http://blog.cloudera.com/blog/tag/platform/) questions (http://blog.cloudera.com/blog/tag/questions/) reliability (http://blog.cloudera.com/blog/tag/reliability/) resource management (http://blog.cloudera.com/blog/tag/resource-management/)

11 responses on "How-to: Select the Right Hardware for Your New Hadoop Cluster"

Dean Hiller (http://buffalosw.com)

September 18, 2013 at 1:00 pm

(http://blog.cloudera.com/blog/2013/08/how-to-select-the-right-hardware-for-your-new-hadoop-cluster/#comment-43858)

12TB per node? Will this perform will for hbase type stuff? or is that for purely an analytics point of view in that nothing is slamming the system like in the case of cassandra which is only about 1TB per node?

I suspect if using hbase in a cloud application type environment, 1TB is more realistic for that type of use case compared to the 12TB which is probably an analytical use case of hadoop?

thanks,
Dean

Gaurav

October 3, 2013 at 8:45 pm

(http://blog.cloudera.com/blog/2013/08/how-to-select-the-right-hardware-for-your-new-hadoop-cluster/#comment-43919)

Is it valid for CDH4 too?

Justin Kestelyn (@kestelyn) Post author

October 7, 2013 at 11:27 am

(http://blog.cloudera.com/blog/2013/08/how-to-select-the-right-hardware-for-your-new-hadoop-cluster/#comment-43939)

Gaurav,

Yes of course, valid for CDH 4.x.

---

ved yadav

February 6, 2014 at 10:36 pm

(http://blog.cloudera.com/blog/2013/08/how-to-select-the-right-hardware-for-your-new-hadoop-cluster/#comment-44625)

i want run hadoop system for experimental purpose, for my acadmic reseach. i have core2quad ,4 GB, 320 GB. is this sufficient for making <10 cluster per cluster 15 node.

Jonas Andersson

February 13, 2014 at 5:17 am

(http://blog.cloudera.com/blog/2013/08/how-to-select-the-right-hardware-for-your-new-hadoop-cluster/#comment-44662)

I'm curious to a follow-up article about performance tuning hardware for new hadoop clusters, tools used for measuring stress of clusters and common pitfalls to save myself from reinventing to the wheel all over again. Does anyone have suggestions on resources that proved useful?

Michael

March 19, 2014 at 9:02 am

(http://blog.cloudera.com/blog/2013/08/how-to-select-the-right-hardware-for-your-new-hadoop-cluster/#comment-44914)

Is HP SL4540 3 nodes with15 x 4TB is a good option?

WANG

May 19, 2014 at 10:49 pm

(http://blog.cloudera.com/blog/2013/08/how-to-select-the-right-hardware-for-your-new-hadoop-cluster/#comment-45386)

Will it cause any issue if we have uneven memory configuration among DataNodes? Thank you

---

Anjay (http://NA)
September 20, 2015 at 10:45 am
(http://blog.cloudera.com/blog/2013/08/how-to-select-the-right-hardware-for-your-new-hadoop-cluster/#comment-69798)

12-24 1-4TB hard disks in a JBOD (Just a Bunch Of Disks) configuration

Does it mean that 12 – 24 nodes and per node can have 1-4 TB of size.

Please someone clarify this.

Thanks

Manoj Sundaram
October 26, 2015 at 11:14 pm
(http://blog.cloudera.com/blog/2013/08/how-to-select-the-right-hardware-for-your-new-hadoop-cluster/#comment-72215)

Anjay,
No, it means “each” datanode can have 12 to 24 disks on them and each disk can be 1 TB to 4 TB in size.

Subarahmanyam
October 30, 2015 at 8:02 pm
(http://blog.cloudera.com/blog/2013/08/how-to-select-the-right-hardware-for-your-new-hadoop-cluster/#comment-72540)

I want to do academic research work on big data and deep learning. Presently I am having i7 quad core with 4 GB RAM and 750 GB Hard Disk. I am requesting you to help me what are the minimum Hardware and Software requirements for good speed and economy.

Anilkumar Panda
November 18, 2015 at 5:27 am
(http://blog.cloudera.com/blog/2013/08/how-to-select-the-right-hardware-for-your-new-hadoop-cluster/#comment-72991)

Hi All,
We have a small cluster (c1) with the following spec:
Node Desc No of Nodes Memory/Node Hard Disk/Node Processor Information
Master Node 1 126 GB 710 GB 2 GHz, 4 Physical cores with 8 hyper threads
Slave 3 47 GB 20 TB 2.20 GHz, 4 physical cores with 6 hyper threads

And want to add thses nodes to another cluster (c2) . c2 configurations are :
Node Description No of Nodes Memory/Node Hard Disk/Node Processor Information
Master 2 256 GB 16TB (Intel Xeon Processor E5-2680 v3 12C 2.5GHz 30MB Cache 2133MHz 120W)*2
Slave 6 256 GB 56TB (Intel Xeon Processor E5-2680 v3 12C 2.5GHz 30MB Cache 2133MHz 120W) *2
Edge 1 512 GB 20 TB (Intel Xeon Processor E5-2680 v3 12C 2.5GHz 30MB Cache 2133MHz 120W)*2

Will the old nodes affect the performance on the c2 cluster? We generally process analytical type of workloads
.

‹ How Cloudera Ensures HBase Client API Compatibility in CDH (http://blog.cloudera.com/blog/2013/08/how-cloudera-ensures-hbase-client-api-compatibility-in-cdh/)

Distributed Systems Get Simpler with Apache Helix › (http://blog.cloudera.com/blog/2013/09/distributed-systems-get-simpler-with-apache-helix/)

Partner (https://www.cloudera.com/partners.html)
Developers (https://www.cloudera.com/developers.html)
Community (http://community.cloudera.com/)
Resources (https://www.cloudera.com/resources.html)
Documentation (https://www.cloudera.com/documentation.html)
Career (http://jobs.jobvite.com/cloudera/)

(https://www.linkedin.com/company/cloudera)

(https://www.facebook.com/cloudera)

(https://twitter.com/cloudera)

(https://www.cloudera.com/contact-us.html)

Contact (https://www.cloudera.com/contact-us.html)

United States: +1 888 789 1488 (tel:18887891488)
Outside the US: +1 650 362 0488 (tel:16503620488)

Terms & Conditions (https://www.cloudera.com/legal/terms-and-conditions.html)
Privacy Policy and Data Policy (https://www.cloudera.com/legal/policies.html)