Multi-node Hadoop Cluster Using Cloudera

CSE 487

Cloud Computing

Summer 172

Department of Computer Science and Engineering

United International University

Submitted by:

1. Asif Ahmed - aahmed141068@bscse.uiu.ac.bd
2. Md. Younus Bipul - mbipul141075@bscse.uiu.ac.bd
3. Syed Md. Imran - simran141086@bscse.uiu.ac.bd
4. Niger Sultana Tahniat - ntahniat141088@bscse.uiu.ac.bd
5. Toha Khan Mozlish - tmozlish141089@bscse.uiu.ac.bd

Multi-node Cluster

Multi node or Fully Distributed Cluster is a typical hadoop cluster which follows a master-slave architecture. It will basically comprise of one master machine (running the NameNode and TaskTracker daemon) and one or more slave machines (running the DataNode and TaskTracker daemon). The default replication factor for a multi node cluster is 3. It is basically used for full stack development of hadoop application and projects.

Hadoop

Apache Hadoop is an open-source software framework used for distributed storage and processing of dataset of big data using the MapReduce programming model. It consists of computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common occurrences and should be automatically handled by the framework.

The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part which is a MapReduce programming model. Hadoop splits files into large blocks and distributes them across nodes in a cluster. It then transfers packaged code into nodes to process the data in parallel. This approach takes advantage of data locality, where nodes manipulate the data they have access to. This allows the dataset to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are distributed via high-speed networking

Hive

Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data summarization, query, and analysis Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop. Traditional SQL queries must be implemented in the MapReduce Java API to execute SQL applications and queries over distributed data. Hive provides the necessary SQL abstraction to integrate SQL-like queries (HiveQL) into the underlying Java without the need to implement queries in the low-level Java API. Since most data warehousing applications work with SQL-based querying languages, Hive aids portability of SQL-based applications to Hadoop.[1] While initially developed by Facebook, Apache Hive is used and developed by other companies such as Netflix and the Financial Industry Regulatory Authority (FINRA). Amazon maintains a software fork of Apache Hive included in Amazon Elastic MapReduce on Amazon Web Services

VirtualBox

A VirtualBox or VB is a software virtualization package that installs on an operating system as an application. VirtualBox allows additional operating systems to be installed on it, as a Guest OS, and run in a virtual environment. In 2010, VirtualBox was the most popular virtualization software application. Supported operating systems include Windows XP, Windows

Vista, Windows 7, macOS X, Linux, Solaris, and OpenSolaris. The current version is 5.1.28 . It is a type-2 hypervisor that sits on an host OS and can run multiple gues OS in it.

Cloudera Manager

Coudera manager is a software that makes it easy to manage Hadoop deployments of any scale in production. Quickly deploy, configure and monitor your cluster through an intuitive UI-complete with roling upgrades, backups and disaster recovery and customizable alerting. Cloudera manager is available as integrated and supported part of Cloudera Enterprise. The current version is Cloudera Manager 5.12.1 .

CDH 5.x.x Requirements

- **Operating System : Ubuntu :**
  CDH 5.3.x runs on both Ubuntu Trusty (14.04)

- **In VM, bridged network**

- **Internet Protocol& Access :**
  Protocol: IPv4
  Internet access to allow the wizard to install software packages or parcels from *archive.cloudera.com*

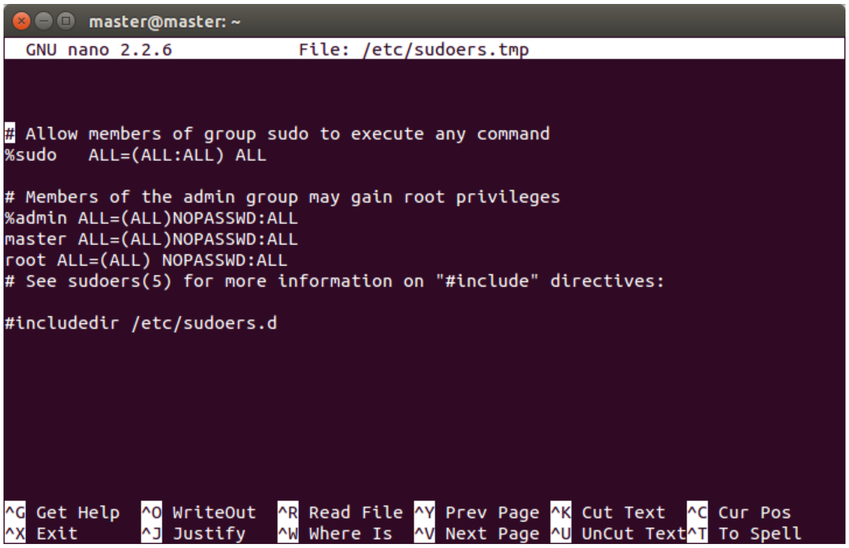
  In ubuntu, go to terminal and run:

  $ sudo su

1. Passwordless sudo priviledge

   $ sudo visudo

   add this line -
   %<username> ALL=(ALL) NOPASSWD:ALL



```
  GNU nano 2.2.6              File: /etc/sudoers.tmp



# Allow members of group sudo to execute any command
%sudo   ALL=(ALL:ALL) ALL

# Members of the admin group may gain root privileges
%admin ALL=(ALL)NOPASSWD:ALL
master ALL=(ALL)NOPASSWD:ALL
root ALL=(ALL) NOPASSWD:ALL
# See sudoers(5) for more information on "#include" directives:

#includedir /etc/sudoers.d
```

^G Get Help   ^O WriteOut   ^R Read File ^Y Prev Page ^K Cut Text   ^C Cur Pos
^X Exit   ^J Justify   ^W Where Is   ^V Next Page ^U UnCut Text ^T To Spell

2. disable ipv6

   Check if IPv6 is disabled

```
$ cat /proc/sys/net/ipv6/conf/all/disable_ipv6
```

   Note : 0 means it's enabled and 1 is disabled.

   To disable IPv6

```
$ sudo su -
$ nano  /etc/sysctl.conf
```

Add these lines to sysctl.conf file

```
#disable ipv6
net.ipv6.conf.all.disable_ipv6 = 1
net.ipv6.conf.default.disable_ipv6 = 1
net.ipv6.conf.lo.disable_ipv6 = 1
```

Save sysctl.conf file with new config and Reboot your system

3. fqdn server

   in each node,

```
$ ifconfig
```

   note down ip address (inet add)

```
$ hostname
```

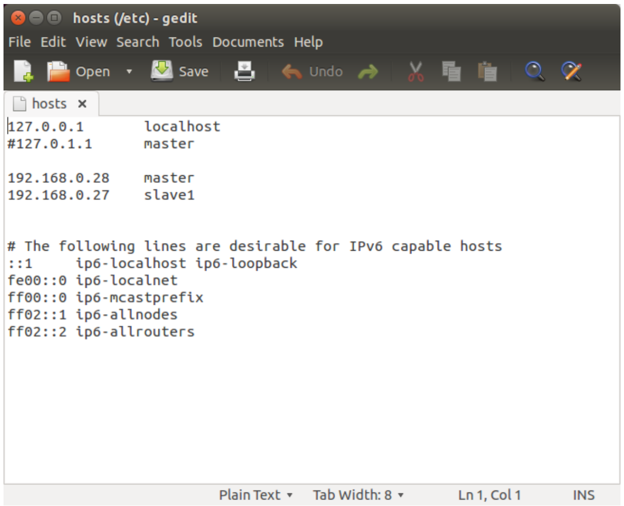
   (and hostname)

   then in each node

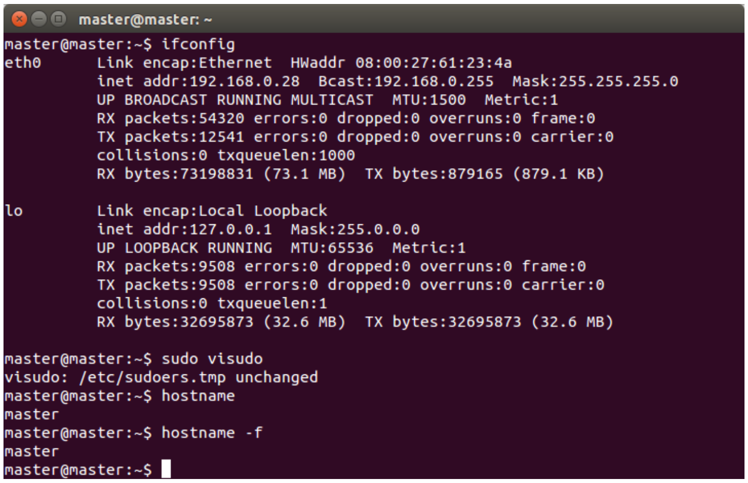
```
$ sudo gedit /etc/hosts
```

   add –

```
ipaddress_of_current_node        hostname_of_current_node

ipaddress_of_other_node          hostname_of_other_node
```



```
127.0.0.1       localhost
#127.0.1.1      master

192.168.0.28    master
192.168.0.27    slave1


# The following lines are desirable for IPv6 capable hosts
::1     ip6-localhost ip6-loopback
fe00::0 ip6-localnet
ff00::0 ip6-mcastprefix
ff02::1 ip6-allnodes
ff02::2 ip6-allrouters
```



```
master@master:~$ ifconfig
eth0      Link encap:Ethernet  HWaddr 08:00:27:61:23:4a
          inet addr:192.168.0.28  Bcast:192.168.0.255  Mask:255.255.255.0
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:54320 errors:0 dropped:0 overruns:0 frame:0
          TX packets:12541 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:73198831 (73.1 MB)  TX bytes:879165 (879.1 KB)

lo        Link encap:Local Loopback
          inet addr:127.0.0.1  Mask:255.0.0.0
          UP LOOPBACK RUNNING  MTU:65536  Metric:1
          RX packets:9508 errors:0 dropped:0 overruns:0 frame:0
          TX packets:9508 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1
          RX bytes:32695873 (32.6 MB)  TX bytes:32695873 (32.6 MB)

master@master:~$ sudo visudo
visudo: /etc/sudoers.tmp unchanged
master@master:~$ hostname
master
master@master:~$ hostname -f
master
master@master:~$ 
```

---

then do ping on each other

$ ping <hostname>

4. create ssh connection:

   in each node

```
$sudo apt-get install openssh-client

$sudo apt-get install openssh-server
```

Configuring passwordless SSH.

We need to configure SSH access to localhost for the user

```
    $ sudo gedit /etc/ssh/sshd_config
```

Note : Set *PubkeyAuthentication* to *Yes*.

```
    $ sudo /etc/init.d/ssh reload
```

To generate SSH key

```
    $ ssh-keygen
    $ ssh-add
    $ sudo cat .ssh/id_pub.rsa >> .ssh/sauthorized_keys
```

in master node or namenode :

```
$ ssh-copy-id –i datanode_hostname@datanode_ip_add
```

do it for all datanodes with namenodes also

Now from namenode, check ssh connection with datanodes

```
$ ssh hostname@ipaddress (of datanodes)
```

5. add repository:

Path to repository address –

https://www.cloudera.com/documentation/enterprise/5-8-x/topics/cm_ig_install_path_b.html

In terminal

```
$ sudo add-apt-repository "deb [arch=amd64] http://archive.cloudera.com/cm5/ubuntu/trusty/amd64/cm trusty-cm5 contrib"
```

```
$ sudo add-apt-repository "deb-src http://archive.cloudera.com/cm5/ubuntu/trusty/amd64/cm trusty-cm5 contrib"
```

```
$ apt-get update
```

```
$ sudo apt-get install oracle-j2sdk1.7
```

Now go to cloudera download page

https://www.cloudera.com/downloads/manager/5-12-1.html

```
$ wget https://archive.cloudera.com/cm5/installer/latest/cloudera-manager-installer.bin
$ chmod u+x cloudera-manager-installer.bin
$ sudo ./cloudera-manager-installer.bin
```

6. Deploy cdh with cloudera manager :

   in web browser, go to

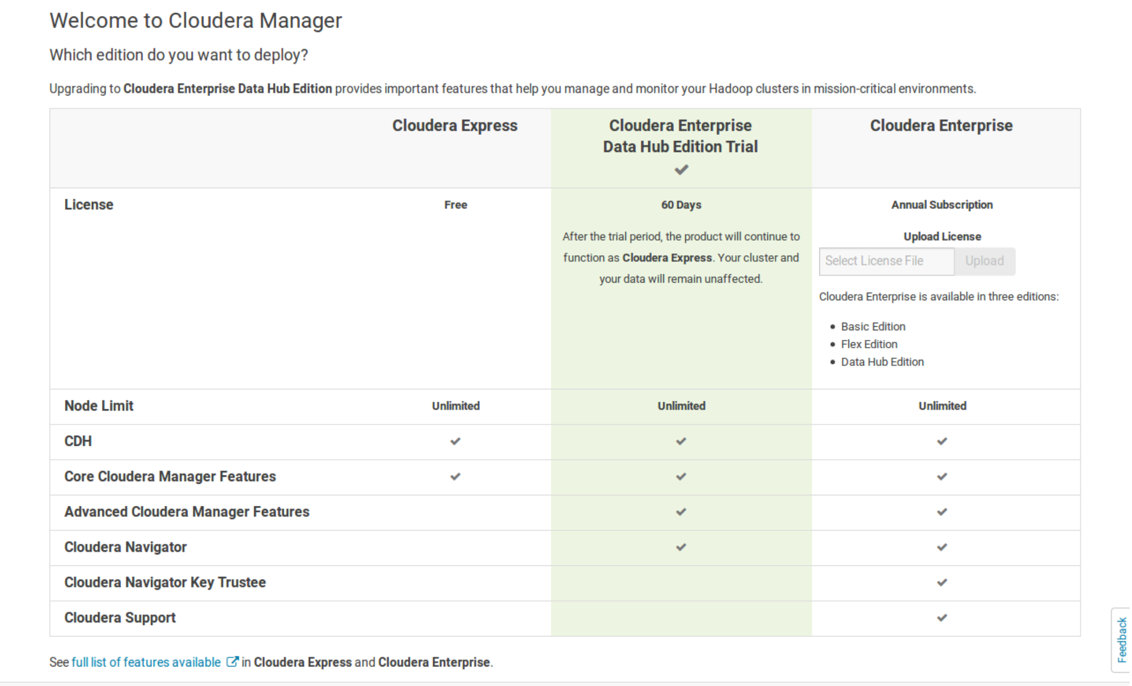
   localhost:7180

   login credential:

   user : admin

   password: admin

7. Setting up Cluster:’



Welcome to Cloudera Manager

Which edition do you want to deploy?

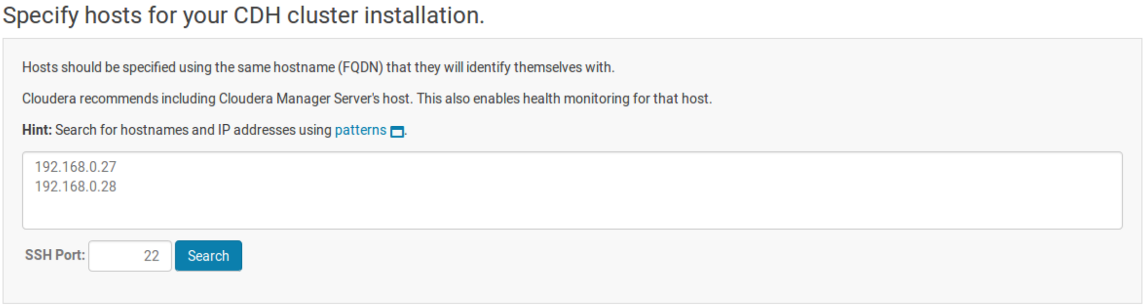
Upgrading to **Cloudera Enterprise Data Hub Edition** provides important features that help you manage and monitor your Hadoop clusters in mission-critical environments.

	Cloudera Express	Cloudera Enterprise Data Hub Edition Trial ✓	Cloudera Enterprise
License	Free	60 Days After the trial period, the product will continue to function as **Cloudera Express**. Your cluster and your data will remain unaffected.	Annual Subscription Upload License Select License File Upload Cloudera Enterprise is available in three editions: • Basic Edition • Flex Edition • Data Hub Edition
Node Limit	Unlimited	Unlimited	Unlimited
CDH	✓	✓	✓
Core Cloudera Manager Features	✓	✓	✓
Advanced Cloudera Manager Features		✓	✓
Cloudera Navigator		✓	✓
Cloudera Navigator Key Trustee			✓
Cloudera Support			✓

See full list of features available in **Cloudera Express** and **Cloudera Enterprise**.

Select Cloudera Express.

Search nodes via ip



Specify hosts for your CDH cluster installation.

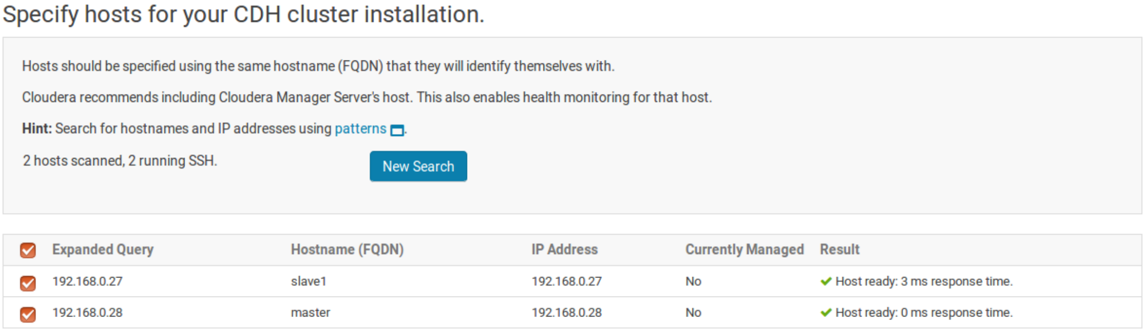
Hosts should be specified using the same hostname (FQDN) that they will identify themselves with.

Cloudera recommends including Cloudera Manager Server's host. This also enables health monitoring for that host.

Hint: Search for hostnames and IP addresses using patterns.

```
192.168.0.27
192.168.0.28
```

SSH Port: 22



Specify hosts for your CDH cluster installation.

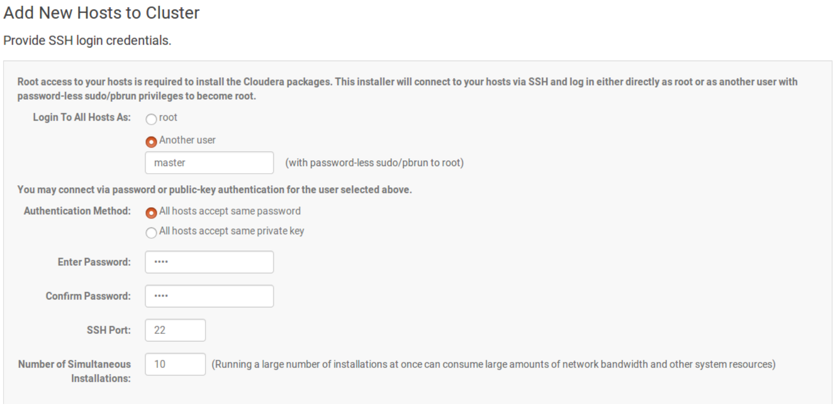
Hosts should be specified using the same hostname (FQDN) that they will identify themselves with.

Cloudera recommends including Cloudera Manager Server's host. This also enables health monitoring for that host.

Hint: Search for hostnames and IP addresses using patterns.

2 hosts scanned, 2 running SSH.

☑	Expanded Query	Hostname (FQDN)	IP Address	Currently Managed	Result
☑	192.168.0.27	slave1	192.168.0.27	No	✓ Host ready: 3 ms response time.
☑	192.168.0.28	master	192.168.0.28	No	✓ Host ready: 0 ms response time.



Add New Hosts to Cluster

Provide SSH login credentials.

Root access to your hosts is required to install the Cloudera packages. This installer will connect to your hosts via SSH and log in either directly as root or as another user with password-less sudo/pbrun privileges to become root.

Login To All Hosts As: ○ root

● Another user

master (with password-less sudo/pbrun to root)

You may connect via password or public-key authentication for the user selected above.

Authentication Method: ● All hosts accept same password

○ All hosts accept same private key

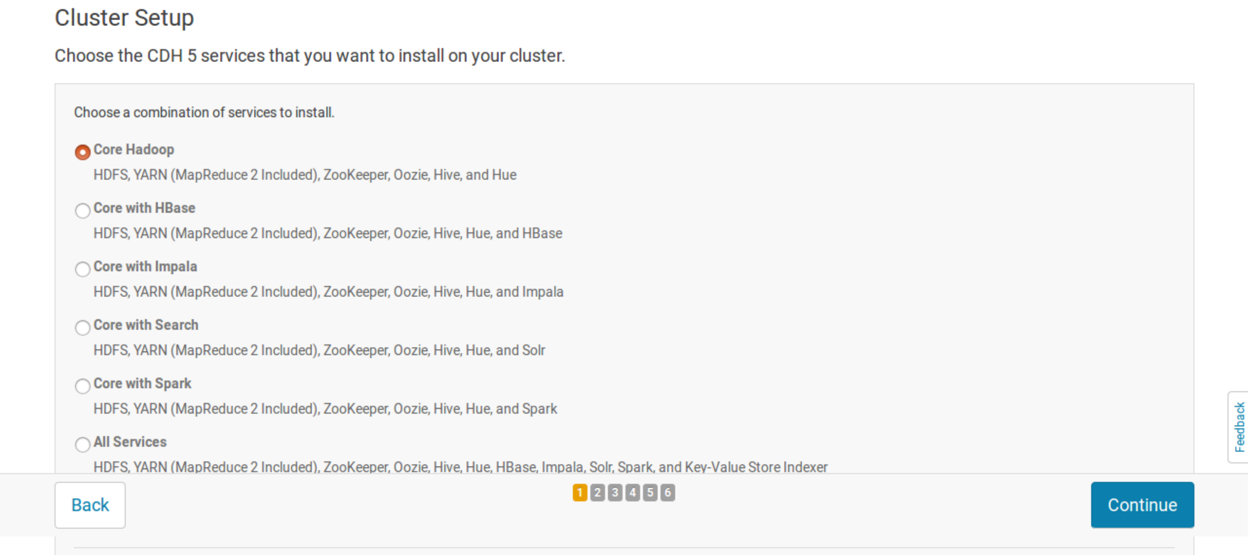
Enter Password: ····

Confirm Password: ····

SSH Port: 22

Number of Simultaneous Installations: 10 (Running a large number of installations at once can consume large amounts of network bandwidth and other system resources)

Each node containing same username and password would make the process easier.



Cluster Setup

Choose the CDH 5 services that you want to install on your cluster.

Choose a combination of services to install.

● Core Hadoop
HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, and Hue

○ Core with HBase
HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, and HBase

○ Core with Impala
HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, and Impala

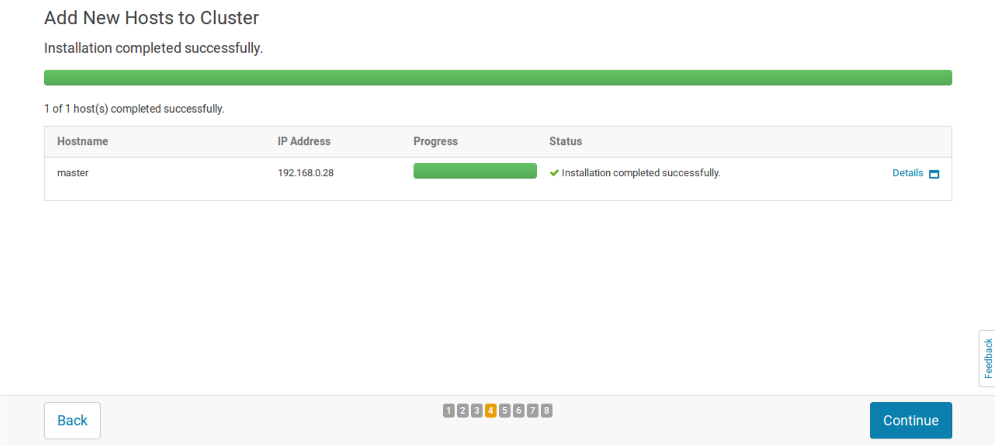
○ Core with Search
HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, and Solr

○ Core with Spark
HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, and Spark

○ All Services
HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, HBase, Impala, Solr, Spark, and Key-Value Store Indexer

Manually distribute role instances to nodes.

(https://www.cloudera.com/documentation/enterprise/5-8-x/topics/cm_ig_host_allocations.html)

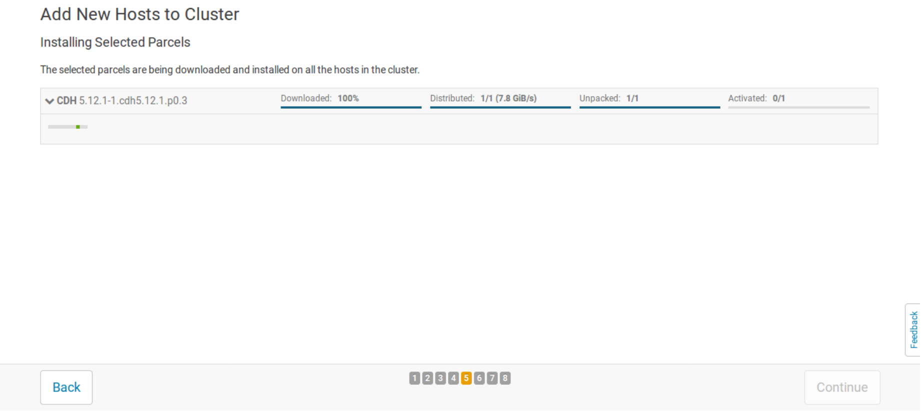


Add New Hosts to Cluster

Installation completed successfully.

1 of 1 host(s) completed successfully.

Hostname	IP Address	Progress	Status	
master	192.168.0.28		✓ Installation completed successfully.	Details

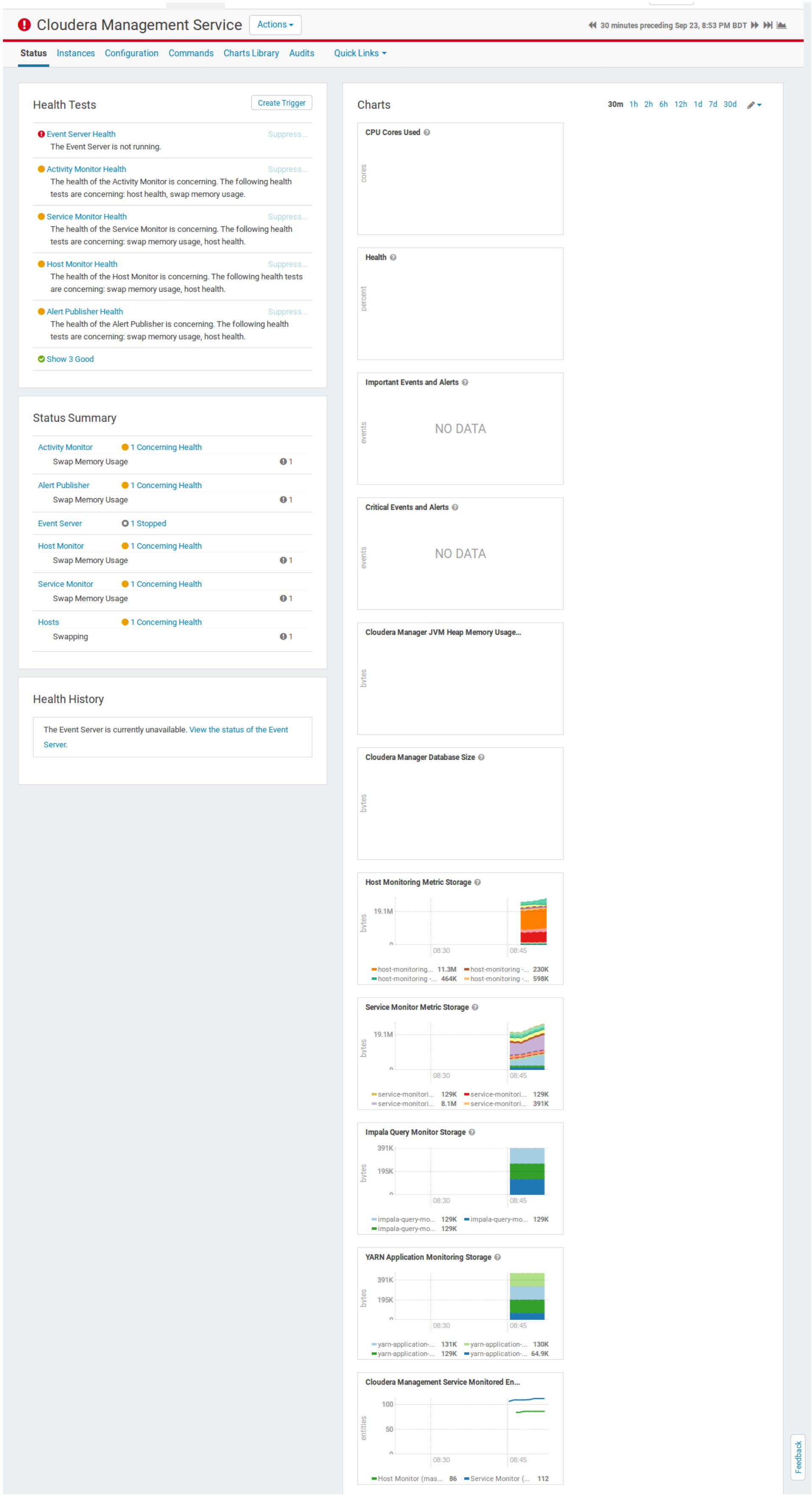


Add New Hosts to Cluster

Installing Selected Parcels

The selected parcels are being downloaded and installed on all the hosts in the cluster.

CDH 5.12.1-1.cdh5.12.1.p0.3 Downloaded: 100% Distributed: 1/1 (7.8 GiB/s) Unpacked: 1/1 Activated: 0/1



Health Tests

- Event Server Health — Suppress...
  The Event Server is not running.
- Activity Monitor Health — Suppress...
  The health of the Activity Monitor is concerning. The following health tests are concerning: host health, swap memory usage.
- Service Monitor Health — Suppress...
  The health of the Service Monitor is concerning. The following health tests are concerning: swap memory usage, host health.
- Host Monitor Health — Suppress...
  The health of the Host Monitor is concerning. The following health tests are concerning: swap memory usage, host health.
- Alert Publisher Health — Suppress...
  The health of the Alert Publisher is concerning. The following health tests are concerning: swap memory usage, host health.
- ✓ Show 3 Good

Status Summary

Activity Monitor	● 1 Concerning Health	
Swap Memory Usage		1
Alert Publisher	● 1 Concerning Health	
Swap Memory Usage		1
Event Server	○ 1 Stopped	
Host Monitor	● 1 Concerning Health	
Swap Memory Usage		1
Service Monitor	● 1 Concerning Health	
Swap Memory Usage		1
Hosts	● 1 Concerning Health	
Swapping		1

Health History

The Event Server is currently unavailable. View the status of the Event Server.

Charts

30m
1h
2h
6h
12h
1d
7d
30d
CPU Cores Used
cores
Health
percent
Important Events and Alerts
events
NO DATA
Critical Events and Alerts
events
NO DATA
Cloudera Manager JVM Heap Memory Usage...
bytes
Cloudera Manager Database Size
bytes
Host Monitoring Metric Storage
bytes
19.1M
0
08:30
08:45
host-monitoring... 11.3M
host-monitoring -... 230K
host-monitoring -... 464K
host-monitoring -... 598K
Service Monitor Metric Storage
bytes
19.1M
0
08:30
08:45
service-monitori... 129K
service-monitori... 129K
service-monitori... 8.1M
service-monitori... 391K
Impala Query Monitor Storage
bytes
391K
195K
0
08:30
08:45
impala-query-mo... 129K
impala-query-mo... 129K
impala-query-mo... 129K
YARN Application Monitoring Storage
bytes
391K
195K
0
08:30
08:45
yarn-application-... 131K
yarn-application-... 130K
yarn-application-... 129K
yarn-application-... 64.9K
Cloudera Management Service Monitored En...
entities
100
50
0
08:30
08:45
Host Monitor (mas... 86
Service Monitor (... 112
Feedback