Setting up an on premise Cloudera Hadoop Cluster

1. Install the Linux OS on all the Machines. We are using CentOS 6.5 for this guide.

2. Set the static IP's and make sure it is connected to internet for yum update and all.

   Why? - In a cluster every node is connected with other node and all services are accessed by hostname and IP address. Even you are maintaining DNS, it is recommended to have static IP at least for Journal nodes

   a. **Configure eth0 and IP** - */etc/sysconfig/network-scripts/ifcfg-eth0*

   What? - It is a most Ethernet interface configuration file which controls the first Ethernet *network interface card* or NIC system

   #Sample (From our cluster)

   ```
   #vi /etc/sysconfig/network-scripts/ifcfg-eth0
   TYPE=Ethernet
   ONBOOT=yes
   BOOTPROTO=STATIC
   IPV6INIT=no
   USERCTL=no
   NM_CONTROLLED=yes
   PEERDNS=yes
   IPADDR=10.132.3.127
   NETMASK=255.255.252.0
   GATEWAY=10.132.2.254
   DNS1=10.10.1.1
   DNS2=8.8.8.8


   ~
   ```

b. **Configure the Default Gateway** - */etc/sysconfig/network*

What? - Used to specify the desired network configuration about network

Sample: One from our cluster

```
# vi /etc/sysconfig/network
NETWORKING=yes
NETWORKING_IPV6=no
HOSTNAME=twi.datalab.node3
GATEWAY=10.132.2.254

~
~
~
```

c. Set the hostname in */etc/hosts*

What? - Hosts is an operating system file that maps the hostname to ip address

Sample: One from our cluster

```
127.0.0.1          localhost.localdomain localhost
::1      localhost6.localdomain6 localhost6
10.132.3.125     twi.datalab.node1      node1
10.132.3.126     twi.datalab.node2      node2
10.132.3.127     twi.datalab.node3      node3
10.132.3.128     twi.datalab.node4      node4


~
```

d. **Restart Network Interface** - `service network restart`

e. **Start ssh daemon** - `service sshd start`
By default, it is off.
Start ssh daemon to give access to other nodes to able to ssh in to the system.
Also do `chkconfig sshd on` - It will start ssh service on every reboot.

f. Make sure`hostname` and `hostname -f` should give same output.

g. **Disable SELinux** - */etc/selinux/config*

What? - SELinux is a Linux kernel security module that provides a mechanism for access control security policies.

By default, it is enforced, make it disabled.
It needs a reboot to make the selinux changes.

```
# This file controls the state of SELinux on the system.
# SELINUX= can take one of these three values:
#        enforcing - SELinux security policy is enforced.
#        permissive - SELinux prints warnings instead of enforcing.
#        disabled - SELinux is fully disabled.
SELINUX=disabled
# SELINUXTYPE= type of policy in use. Possible values are:
#        targeted - Only targeted network daemons are protected.
#        strict - Full SELinux protection.
SELINUXTYPE=targeted
```

h. **Set ntp configurations** - /etc/ntp.conf
Synchronize the system clock with a remote server over the Network Time Protocol (NTP). By default it will be like 0.centos.pool.ntp.org iburst...

```
# Hosts on local network are less restricted.
#restrict 192.168.1.0 mask 255.255.255.0 nomodify notrap

# Use public servers from the pool.ntp.org project.
# Please consider joining the pool (http://www.pool.ntp.org/join.html).
server 0.centos.pool.ntp.org iburst
server 1.centos.pool.ntp.org iburst
server 2.centos.pool.ntp.org iburst
server 3.centos.pool.ntp.org iburst

#broadcast 192.168.1.255 autokey        # broadcast server
#broadcastclient                        # broadcast client
#broadcast 224.0.1.1 autokey            # multicast server
#multicastclient 224.0.1.1              # multicast client
#manycastserver 239.255.254.254         # manycast server
#manycastclient 239.255.254.254 autokey # manycast client

# Enable public key cryptography.
#crypto

includefile /etc/ntp/crypto/pw

# Key file containing the keys and key identifiers used when operating
# with symmetric key cryptography.
```

Change it to the following

```
# Hosts on local network are less restricted.
#restrict 192.168.1.0 mask 255.255.255.0 nomodify notrap

# Use public servers from the pool.ntp.org project.
# Please consider joining the pool (http://www.pool.ntp.org/join.html).
server 0.pool.ntp.org
server 1.pool.ntp.org
server 2.pool.ntp.org
server 3.pool.ntp.org

#broadcast 192.168.1.255 autokey        # broadcast server
#broadcastclient                        # broadcast client
#broadcast 224.0.1.1 autokey            # multicast server
#multicastclient 224.0.1.1              # multicast client
#manycastserver 239.255.254.254         # manycast server
#manycastclient 239.255.254.254 autokey # manycast client

# Enable public key cryptography.
#crypto

includefile /etc/ntp/crypto/pw

# Key file containing the keys and key identifiers used when operating
# with symmetric key cryptography
```

    i.   Start the ntp daemon - `service ntpd start` & `chkconfig ntpd on`

    j.   start nscd daemon - `service nscd start` & `chkconfig nscd on`
    What? - nscd is a daemon that provides cache for the most common name service requests

    k.   Stop/Configure IP Tables:
    What? - iptables is a flexible firewall utility for Linux
    Why? - Allows ports to communicate with outside world

    l.   Change the swappiness settings - /etc/sysctl.conf
    What? - It's a Linux kernel parameter that controls the relative weight given to swapping out runtime memory, as opposed to dropping pages from the system page cache.
    Add the following line to the conf file. (Cloudera recommended)

```
vm.swappiness=10
```

3. Restart the machine after all the changes. Basic machine settings are done.

4. On the host machine install Cloudera manager by any one of the method from following

a. To download latest version:
http://www.cloudera.com/downloads/manager/5-5-1.html
b. wget
https://archive.cloudera.com/cm5/installer/latest/cloudera-manager-installer.bi
n

5. Install the cloudera manager by executing it.(have to make it executable
first)
a. chmod u+x cloudera-manager-installer.bin
b. ./cloudera-manager-installer.bin

After the installation of the cloudera manger server s service will start on the port 7180

6. *Go to the cloudera manager UI at http://twi-hadoop-cluster:7180*
a. Where the twi-hadoop-cluster is the hostname or IP of the host
machine where you have installed the cloudera manager.
b. Login with the credentials admin/admin
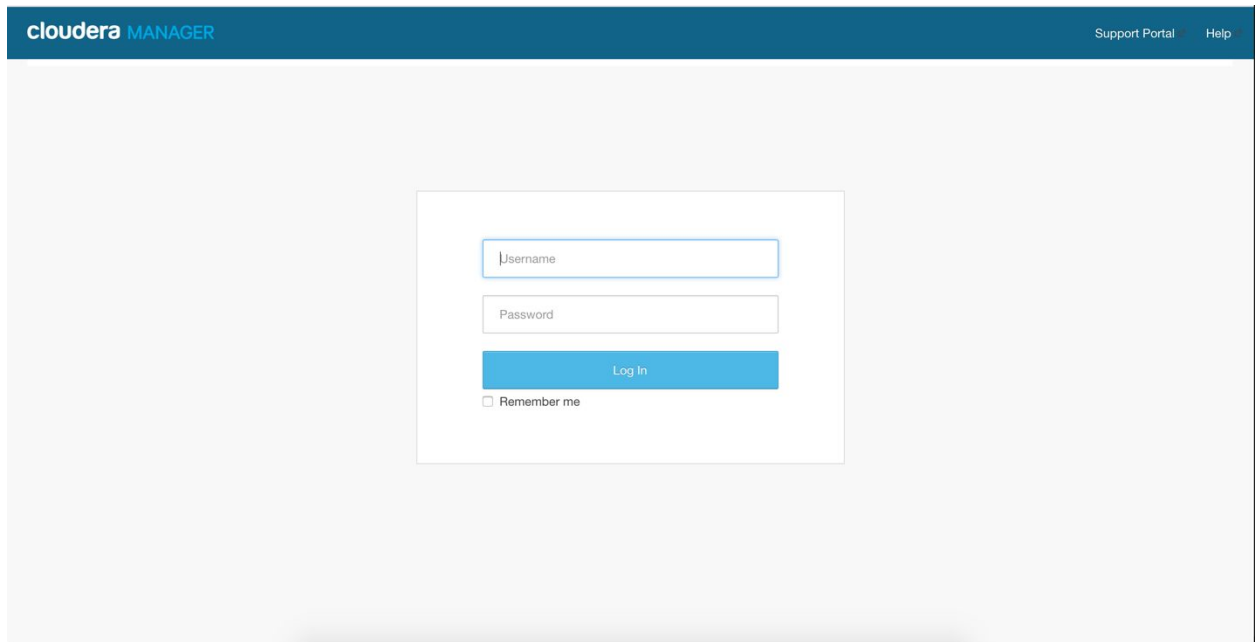
7. Accept the licenses on the cloudera Manager UI and move forward to
edition page. There are two flavors Cloudera Express and Cloudera
Enterprise. There is also enterprise trial edition. You can choose your
edition as Cloudera Enterprise Data Hub Edition trial which will move to
Cloudera Express in 60 days.

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			✔

◄ Back   **1 2**   ► Continue

---

cloudera MANAGER   Support ▾   admin ▾

Thank you for choosing Cloudera Manager and CDH.

This installer will install **Cloudera Enterprise Data Hub Edition Trial 5.10.0** and enable you to later choose packages for the services below (there may be some license implications).

- Apache Hadoop (Common, HDFS, MapReduce, YARN)
- Apache HBase
- Apache ZooKeeper
- Apache Oozie
- Apache Hive
- Hue (Apache licensed)
- Apache Flume
- Cloudera Impala (Apache licensed)
- Apache Sentry
- Apache Sqoop
- Cloudera Search (Apache licensed)
- Apache Spark

You are using Cloudera Manager to install and configure your system. You can learn more about Cloudera Manager by clicking on the **Support** menu above.

Continue

---

8. Specify the other host names or the IP pattern and do a search host. eg: 10.132.3.[125-128] for the range of the ips between 125 to 128 in that IP class.

Specify hosts for your CDH cluster installation.

Hint: Search for hostnames and IP addresses using patterns ⬚.

10.132.3.[125-128]

SSH Port: 22   Search

Back   Continue

9. Select all the listed node from the list.

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 ⬚.

4 hosts scanned, 4 running SSH.   New Search

	Expanded Query	Hostname (FQDN)	IP Address	Currently Managed	Result
☑	10.132.3.125	twi.datalab.node1	10.132.3.125	No	✔ Host ready: 0 ms response time.
☑	10.132.3.126	twi.datalab.node2	10.132.3.126	No	✔ Host ready: 1 ms response time.
☑	10.132.3.127	twi.datalab.node3	10.132.3.127	No	✔ Host ready: 1 ms response time.
☑	10.132.3.128	twi.datalab.node4	10.132.3.128	No	✔ Host ready: 1 ms response time.

Back   Continue

node1:7180/cmf/

10. On next page choose the Parcels for installation also add the KAFKA as additional parcel and leave other setting as default.

cloudera MANAGER

Support ▾ admin ▾

Cluster Installation

Select Repository

Cloudera recommends the use of parcels for installation over packages, because parcels enable Cloudera Manager to easily manage the software on your cluster, automating the deployment and upgrade of service binaries. Electing not to use parcels will require you to manually upgrade packages on all hosts in your cluster when software updates are available, and will prevent you from using Cloudera Manager's rolling upgrade capabilities.

Choose Method   ◯ Use Packages ❓

    ● Use Parcels (Recommended) ❓   [More Options]   [Proxy Settings]

Select the version of CDH

    ● CDH-5.10.0-1.cdh5.10.0.p0.41
    ◯ CDH-4.7.1-1.cdh4.7.1.p0.47
    Versions of CDH that are too new for this version of Cloudera Manager (5.10.0) will not be shown.

Additional Parcels   ◯ ACCUMULO-1.7.2-5.5.0.ACCUMULO5.5.0.p0.8
    ◯ ACCUMULO-1.4.4-1.cdh4.5.0.p0.65
    ● None

    ● KAFKA-2.1.1-1.2.1.1.p0.18
    ◯ KAFKA-2.1.0-1.2.1.0.p0.115
    ◯ None

[Back]   **1** 2 3   [Continue]

11. In next page select Install Oracle JDK and the Java Encryption Policy Files. For quick prototype, we can make use of embedded database and Java jdk but if you are going production mode, it is strongly recommended not to use embedded database or JDK. It's good to setup them manually and configure with cloudera manager while installing the cluster.

Cluster Installation

JDK Installation Options

Agreement. Source code may not be redistributed unless expressly provided for in this Agreement.

J. THIRD PARTY CODE. Additional copyright notices and license terms applicable to portions of the Software are set forth in the THIRDPARTYLICENSEREADME file accessible at http://www.oracle.com/technetwork/java/javase/documentation/index.html. In addition to any terms and conditions of any third party opensource/freeware license identified in the THIRDPARTYLICENSEREADME file, the disclaimer of warranty and limitation of liability provisions in paragraphs 4 and 5 of the Binary Code License Agreement shall apply to all Software in this distribution.

K. TERMINATION FOR INFRINGEMENT. Either party may terminate this Agreement immediately should any Software become, or in either party's opinion be likely to become, the subject of a claim of infringement of any intellectual property right.

L. INSTALLATION AND AUTO-UPDATE. The Software's installation and auto-update processes transmit a limited amount of data to Oracle (or its service provider) about those specific processes to help Oracle understand and optimize them. Oracle does not associate the data with personally identifiable information. You can find more information about the data Oracle collects as a result of your Software download at http://www.oracle.com/technetwork/java/javase/documentation/index.html.

For inquiries please contact: Oracle America, Inc., 500 Oracle Parkway,

Redwood Shores, California 94065, USA.

Last updated 02 April 2013

☑ Install Oracle Java SE Development Kit (JDK)
Check this box to accept the Oracle Binary Code License Agreement and install the JDK. Leave it unchecked to use a currently installed JDK.

☑ Install Java Unlimited Strength Encryption Policy Files
Check this checkbox if local laws permit you to deploy unlimited strength encryption and you are running a secure cluster.

Back   1 2 3 4 5 6 7   Continue

12. Don't opt for single User Mode because we want full installation.

Cluster Installation

Enable Single User Mode

Only supported for CDH 5.2 and above.

By default, service processes run as distinct users on the system. For example, HDFS DataNodes run as user "hdfs" and HBase RegionServers run as user "hbase." Enabling "single user mode" configures Cloudera Manager to run service processes as a single user, by default "cloudera-scm", thereby prioritizing isolation between managed services and the rest of the system over isolation between the managed services.

The **major benefit** of this option is that the Agent does not run as root. However, this mode complicates installation, which is described fully in the documentation. Most notably, directories which in the regular mode are created automatically by the Agent, must be created manually on every host with appropriate permissions, and sudo (or equivalent) access must be set up for the configured user.

Switching back and forth between single user mode and regular mode is not supported.

Single User Mode   ☐   ❓

Back   1 2 3 4 5 6 7   Continue

13. Give a user name/password or ssh key for logging in other nodes.

14. Wait for the installation.

15. Wait for downloading CDH and distributing them on all the hosts.

16. Installing all packages Hold on it will take several minutes. Continue when all installed.

17. See the health inspector report which will show some concerning problems if any.

18. Select the software requirements (we have chosen all services).

19. Important set the default roles for each host. Keep in mind that how much memory(RAM) you have and how does each services will take.

Cluster Setup

Customize Role Assignments

You can customize the role assignments for your new cluster here, but if assignments are made incorrectly, such as assigning too many roles to a single host, this can impact the performance of your services. Cloudera does not recommend altering assignments unless you have specific requirements, such as having pre-selected a specific host for a specific role.

You can also view the role assignments by host.   View By Host

🗄 HDFS

NN NameNode × 1 New	SNN SecondaryNameNode × 1 New	B Balancer × 1 New	HFS HttpFS
twi.datalab.node1	twi.datalab.node1	twi.datalab.node1	Select hosts

NFSG NFS Gateway	DN DataNode × 3 New
Select hosts	twi.datalab.node[2-4] ▾

🐝 Hive

G Gateway × 4 New	HMS Hive Metastore Server × 1 New	WHCS WebHCat Server	HS2 HiveServer2 × 1 New
twi.datalab.node[1-4]	twi.datalab.node1	Select hosts	twi.datalab.node1

🐝 Hue

Back   1 2 3 4 5 6   Continue

node1:7180/cmf/clusters/1/express-add-services/index#

View By Host   ✕

Hosts	Count	Existing Roles		Added Roles
twi.datalab.node1	1	ES	SM	B G HS G NM S
twi.datalab.node2	1	AM	AP	DN NN G HS G NM S
twi.datalab.node3	1	HM		DN SNN G HS2 OS G NM
twi.datalab.node4	1	RM		DN G HMS G RM NM JHS S

This table is grouped by hosts having the same roles assigned to them.

Close

🐝 Hive

G Gateway × 4 New	HMS Hive Metastore Server × 1 New	WHCS WebHCat Server	HS2 HiveServer2 × 1 New
twi.datalab.node[1-4]	twi.datalab.node4 ▾	Select hosts	twi.datalab.node3 ▾

🐝 Hue

Back   1 2 3 4 5 6   Continue

20. Selection of Database
   a. Embedded (Non Production)
   b. Custom(Production)
   c. We are selecting Embedded mode for ease.

Cluster Setup

Database Setup

Configure and test database connections. If using custom databases, create the databases first according to the **Installing and Configuring an External Database** section of the Installation Guide ✪ .

○ Use Custom Databases
● Use Embedded Database

When using the embedded database, passwords are automatically generated. Please copy them down.

Hive

Database Host Name:	Database Type:	Database Name :	Username:	Password:
twi.datalab.node1:7432	PostgreSQL	hive1	hive1	VCNgOfvfgV

Hue

Database Host Name:	Database Type:	Database Name :	Username:	Password:
twi.datalab.node1:7432	PostgreSQL	hue2	hue2	d6JDeCfBuz

Oozie Server

Currently assigned to run on **twi.datalab.node3**.

Back   1 2 3 4 5 6   Continue

21. Review the configuration and Start all the the Services

DataNode Data Directory	DataNode Default Group		
dfs.data.dir, dfs.datanode.data.dir	...and 2 others		
Edit Individual Values	/dfs/dn	+ -	?

| NameNode Data Directories | NameNode Default Group | | |
| dfs.name.dir, dfs.namenode.name.dir | /dfs/nn | + - | ? |

| HDFS Checkpoint Directories | SecondaryNameNode Default Group | | |
| fs.checkpoint.dir, dfs.namenode.checkpoint.dir | /dfs/snn | + - | ? |

| Hive Warehouse Directory | Hive (Service-Wide) | |
| hive.metastore.warehouse.dir | /user/hive/warehouse | ? |

| Hive Metastore Server Port | Hive Metastore Server Default Group | |
| hive.metastore.port | 9083 | ? |

| ShareLib Root Directory | Oozie (Service-Wide) | |
| oozie.service.WorkflowAppService.system.lib path | /user/oozie | ? |

| Oozie Server Data Directory | Oozie Server Default Group | |
| | /var/lib/oozie/data | ? |

Back   1 2 3 4 5 6   Continue

Cluster Setup

✿ First Run Command

Status: **Running**   Start Time: Mar 30, 5:46:39 PM   [Abort]

Details   Completed 1 of 9 step(s).

● All ○ Failed Only ○ Running Only

Step	Context	Start Time	Duration	Actions
❯ ✔ Run 2 steps in parallel Successfully completed 2 steps.		Mar 30, 5:46:39 PM	51ms	
❯ ✿ **Deploying Client Configuration**	↗ Cluster 1 ↗	Mar 30, 5:46:39 PM		[Abort]
❯ Start ZooKeeper				
❯ Start HDFS				
❯ Start YARN (MR2 Included)				
❯ Start Spark				
❯ Start Hive				
❯ Start Oozie				
❯ Start Hue				

[Back]   1 2 3 4 5 6   [Continue]

Status: **Finished**   Start Time: Mar 30, 6:23:28 PM   Duration: 4.6m

Finished First Run of the following services successfully: ZooKeeper, HDFS, YARN (MR2 Included), Spark, Hive, Oozie, Hue.

Details   Completed 9 of 9 step(s).

● All ○ Failed Only ○ Running Only

Step	Context	Start Time	Duration	Actions
❯ ✔ Run 2 steps in parallel Successfully completed 2 steps.		Mar 30, 6:23:28 PM	30ms	
❯ ✔ Deploying Client Configuration Successfully deployed all client configurations.	↗ Cluster 1 ↗	Mar 30, 6:23:28 PM	539ms	
❯ ✔ Start ZooKeeper Successfully completed 1 steps.		Mar 30, 6:23:28 PM	0ms	
❯ ✔ Start HDFS Successfully completed 1 steps.		Mar 30, 6:23:28 PM	43.77s	
❯ ✔ Start YARN (MR2 Included) Successfully completed 1 steps.		Mar 30, 6:24:12 PM	26.14s	
❯ ✔ Start Spark Successfully completed 1 steps.		Mar 30, 6:24:38 PM	32.94s	
❯ ✔ Start Hive Successfully completed 1 steps.		Mar 30, 6:25:11 PM	49.83s	
❯ ✔ Start Oozie Successfully completed 1 steps.		Mar 30, 6:26:01 PM	94.68s	
❯ ✔ Start Hue Successfully completed 1 steps.		Mar 30, 6:27:36 PM	27.92s	

[Back]   1 2 3 4 5   [Continue]

22. See the health of the instances and services.
    It will take couple of minutes to sync all the services. Don't panic by looking at red flags :)

cloudera MANAGER

Cluster Setup

Congratulations!

The services are installed, configured, and running on your cluster.

| Back | 1 2 3 4 5 | Finish |

---

cloudera MANAGER

Clusters ▾   Hosts ▾   Diagnostics ▾   Audits   Charts ▾   Backup ▾   Administration ▾   🔍 Search   Support ▾   admin ▾

● **twi-caspian** (CDH 5.10.1 , Parcels)   Actions ▾   Today, 3:18 PM IST

Status   Configuration ▾

Status

● ⊞ Hosts		
● ▣ HDFS	✕ 1	▾
● 🐝 Hive		▾
● 📣 Hue		▾
● ▣ Oozie		▾
● ☆ Spark		▾
● ▥ YARN (MR2 Inc...		▾
● 🐾 ZooKeeper		▾

Charts

30m 1h 2h 6h 12h 1d 7d 30d 📝▾

Cluster CPU

percent: 100%, 50%, 0% — 03 PM — 03:15

— twi-caspian, Host CPU Usage Across Hosts **5%**

Cluster Disk IO

bytes / second: 9.5M/s, 4.8M/s, 0 — 03 PM — 03:15

— Total Disk Byte... **4.5K/s**   — Total Disk Bytes ... **1M/s**

Cluster Network IO

bytes / second: 2.9M/s, 1.9M/s, 977K/s, 0 — 03 PM — 03:15

— Total Bytes Re... **57.7K/s**   — Total Bytes Tr... **54.8K/s**

HDFS IO

bytes / second: 391K/s, 195K/s, 0 — 03 PM — 03:15

— Total Bytes Read ... **1b/s**   — Total Bytes Writ... **8.1b/s**
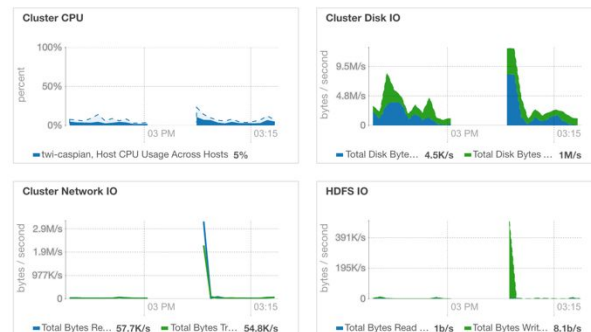
From
Balaji Sankar,
Abhishek Gupta.