

# Hadoop Basics with InfoSphere BigInsights

*Lesson 5: Hadoop administration*



# An IBM Proof of Technology

---

Catalog Number

---

## Contents

Lab 1Hadoop Administration.....	4
1.1Managing a Hadoop Cluster.....	5
1.1.1Adding/Removing a node from the cluster.....	5
1.1.2Setting up Master/Slave nodes.....	6
1.1.3Adding a node from Web Console.....	8
1.1.4Adding a node from the Terminal.....	11
1.1.5Removing a node.....	11
1.1.6Health of a Cluster.....	12
1.1.7Visual Health Check.....	12
1.1.8DFS Disk Check.....	14
1.2Hadoop Administration.....	14
1.2.1Administering Specific Services.....	15
1.2.2Configuring Hadoop Default Settings.....	16
1.2.3Increasing Storage Block Size.....	16
1.2.4Limit Data nodes disk usage.....	17
1.2.5Configuring the replication factor.....	18
1.3Importing Large Amounts of Data.....	18
1.3.1Moving Data to and from HDFS.....	18
1.3.2Hadoop commands through terminal.....	19
1.3.3Hadoop commands through WebConsole.....	20
1.4Summary.....	22

---

## Lab 1 Hadoop Administration

IBM's InfoSphere BigInsights 2.0 Enterprise Edition enables firms to store, process, and analyze large volumes of various types of data using a wide array of machines working together as a cluster. In this exercise, you'll learn some essential Hadoop administration tasks from expanding a cluster to ingesting large amounts of data into the Hadoop Distributed File System (HDFS).

After completing this hands-on lab, you'll be able to:

- Manage a cluster running BigInsights to add or remove nodes as necessary
- Cover essential hadoop administration tasks such as expanding disk space and how to start and stop services

Allow 60 minutes to 90 minutes to complete this lab.

This version of the lab was designed using the InfoSphere BigInsights 2.1 Quick Start Edition. Throughout this lab you will be using the following account login information:

	<b>Username</b>	<b>Password</b>
VM image setup screen	root	password
Linux	biadmin	biadmin

For this lab all Hadoop components should be up and running. If all components are running you may move on to Section 2 of this lab. Otherwise please refer to Hadoop Basics Unit 1: Exploring Hadoop Distributed File System Section 1.1 to get started. (All Hadoop components should be started)

## 1.1 Managing a Hadoop Cluster

In this section you will learn how to:

- Add and remove nodes through Web Console, and Terminal
- Check the health of the cluster and individual nodes within that cluster
- Perform checks on the disk and storage of the HDFS

Typical Hadoop clusters rely on being able to use multiple cheap computers/devices as nodes working together as a Hadoop cluster. Because of this, and the way in which hardware and hard disk drives operate from a mechanical point, the hardware is bound to fail over the years – which hadoop handles efficiently by replicating the data across the various nodes (3-node replication by default).

### 1.1.1 Adding/Removing a node from the cluster

One of the key parts of managing a Hadoop cluster is being able to scale the cluster with ease, adding and removing nodes as needed. Adding a node can be done through a range of methods, of which we will cover adding from a Web Console, and from a terminal. Each of these methods can achieve the same results.

Before proceeding with adding a node, you should first verify that you can access the node you are trying to add. This can be done by simply “sshing” the given node(s) as follows.

\_\_1. Open a terminal window by clicking BigInsights Shell icon, then on the Terminal icon.



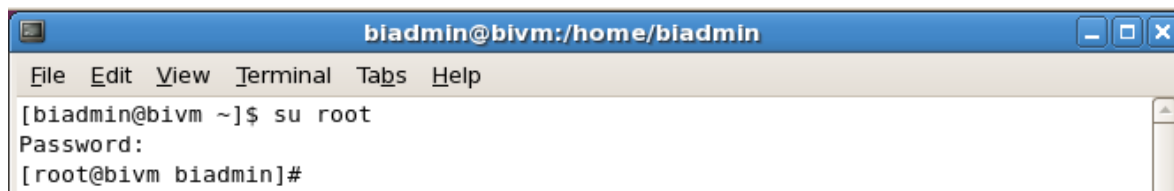
\_\_2. You will need to switch users to root. Type the following:

```
su root
```

When prompted for the password enter:

```
password
```

You should now be the root user



```
biadmin@bivm:/home/biadmin
File Edit View Terminal Tabs Help
[biadmin@bivm ~]$ su root
Password:
[root@bivm biadmin]#
```

- \_\_3. Type ssh followed by the IP address that you wish to use as your node. For example, the first node added in this case has an IP address of 192.168.44.15:

```
ssh 192.168.44.158
```

When doing ssh on a new IP you will get an authenticity message:

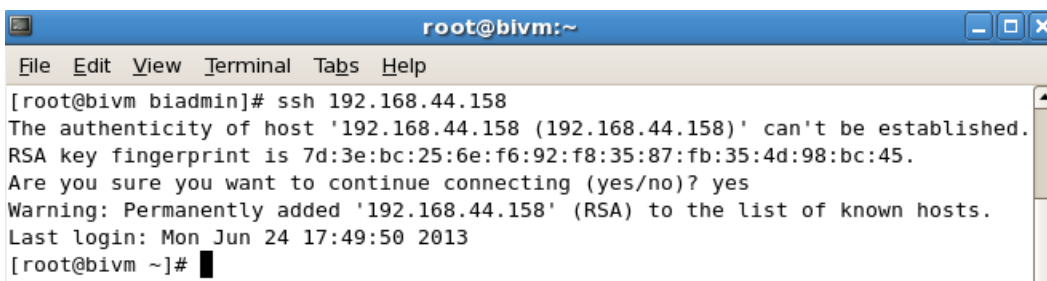
```
The authenticity of host '192.168.44.158 (192.168.44.158)' can't be established.
RSA key fingerprint is 29:2f:72:9f:f4:97:16:89:cf:d9:cc:09:d3:16:d9:bf.
Are you sure you want to continue connecting (yes/no)?
```

Go ahead and type yes, you will then get a warning:

```
Warning: Permanently added '192.168.44.158' (RSA) to the list of known hosts.
```

Enter the password for the VM/node that you are adding and you should have access.

If you are successful in the above steps then your terminal should look similar to the image below:



```
root@bivm:~
File Edit View Terminal Tabs Help
[root@bivm biadmin]# ssh 192.168.44.158
The authenticity of host '192.168.44.158 (192.168.44.158)' can't be established.
RSA key fingerprint is 7d:3e:bc:25:6e:f6:92:f8:35:87:fb:35:4d:98:bc:45.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added '192.168.44.158' (RSA) to the list of known hosts.
Last login: Mon Jun 24 17:49:50 2013
[root@bivm ~]#
```

- \_\_4. Exit the ssh connection then open a new terminal.

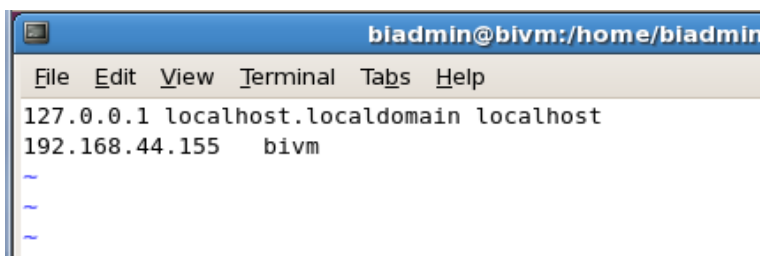
```
exit
```

### 1.1.2 Setting up Master/Slave nodes

Now that we know that the node we want to add is reachable via ssh we need to configure some files on the master and slave nodes. The master node is the node from which you will do the adding, the slave node is the node being added to the cluster. In this case multiple Quick Start Editions will be used for Master and Slave nodes. This method should also work for non-Quick Start Editions.

- \_\_1. On your Master node open a new terminal. Execute the following:

```
sudo vi /etc/hosts
```



```
biadmin@bivm:/home/biadmin
File Edit View Terminal Tabs Help
127.0.0.1 localhost.localdomain localhost
192.168.44.155 bivm
~
~
~
```

- \_\_2. Press i and move the cursor using the arrow keys to the end of line, press <Enter> and type the IP address of your node VM and give it a name of node1.

```

biadmin@bivm:/home/biadmin
File Edit View Terminal Tabs Help
127.0.0.1 localhost.localdomain localhost
192.168.44.155 bivm
192.168.44.158 node1
~

```

- \_\_3. Press the <Esc> key then press <Shift+;> then type "x!" and hit <Enter>. This will save the changes made to the /etc/hosts file. You should now be at the terminal.

```

biadmin@bivm:~
File Edit View Terminal Tabs Help
[biadmin@bivm ~]$ sudo vi /etc/hosts
[biadmin@bivm ~]$

```

- \_\_4. Verify that the changes were indeed saved, execute the following:

```
cat /etc/hosts
```

```

biadmin@bivm:~
File Edit View Terminal Tabs Help
[biadmin@bivm ~]$ cat /etc/hosts
127.0.0.1 localhost.localdomain localhost
192.168.44.155 bivm
192.168.44.158 node1
[biadmin@bivm ~]$

```



**NOTE:** The following step will UNINSTALL your BigInsights distribution on your Slave node. This step is done on a Quick Start Edition which you can download multiple times.

- \_\_5. On your slave node open a terminal and execute the following:

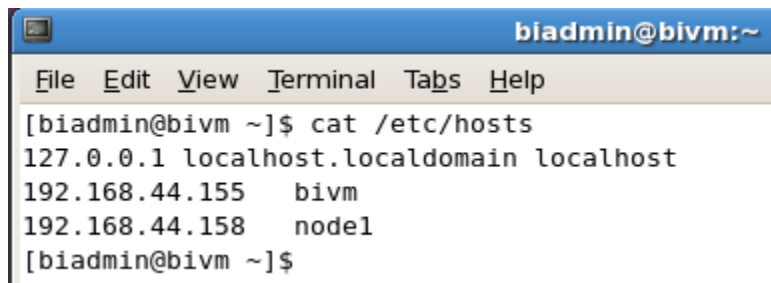
```
/opt/ibm/biginsights/bin/uninstall.sh
```

When prompted to stop all BigInsights processes type yes and hit <Enter>

When warned about erasing all data type yes and hit <Enter>

```
BigInsights uninstallation will end all InfoSphere BigInsights processes on all
cluster nodes, and DELETE all files in the $BIGINSIGHTS_HOME directory, $BIGINSI
GHTS_VAR directory, and all other Hadoop data directories. Do you want to contin
ue (yes/no)?
yes
Warning : All data will be deleted and cannot be recovered. Are you sure you wan
t to continue (yes/no)?
yes
```

- \_\_6. BigInsights will now uninstall from the VM. This may take a few minutes.
- \_\_7. Now in your Slave node repeat steps \_\_1-- \_\_4 but instead of adding the Slave node IP address you add the Master node IP address and change the name from bivm to node1. Your changes in the /etc/hosts file on the Slave node should look like the image below. Notice that it looks exactly like the Master node file.



```
biadmin@bivm:~
File Edit View Terminal Tabs Help
[biadmin@bivm ~]$ cat /etc/hosts
127.0.0.1 localhost.localdomain localhost
192.168.44.155 bivm
192.168.44.158 node1
[biadmin@bivm ~]$
```

- \_\_8. Exit the Terminal on your Slave node

**exit**

You must always keep the /etc/hosts file updated when adding and removing nodes on all nodes (Master and Slave nodes)

### 1.1.3 Adding a node from Web Console

One of the great features of IBM's InfoSphere BigInsights, is the web console. The web console provides an interface to not only the data in HDFS, but also a user-friendly way for performing the tasks associated with simple and advanced hadoop scripts as well as extensive visualizations.

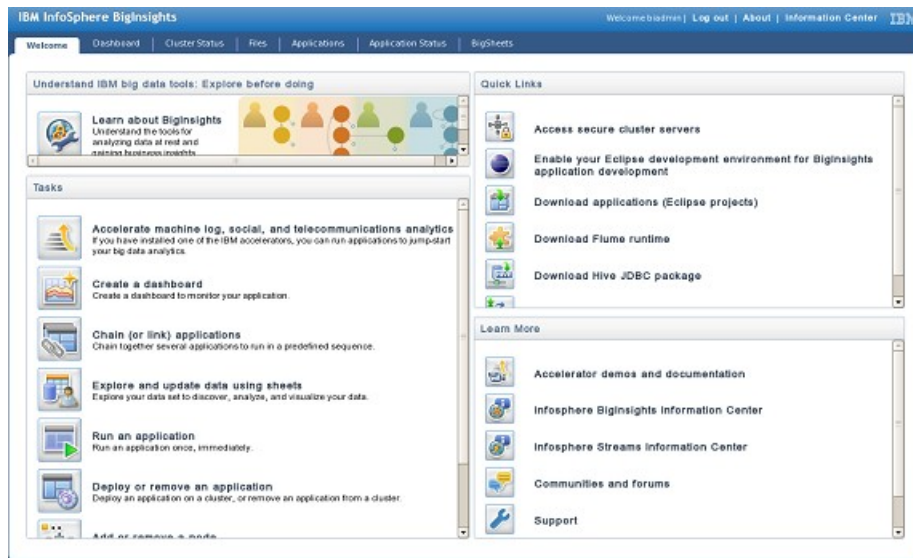
All of the following steps will be done on your Master node.

- \_\_1. Launch the Web Console by clicking on the BigInsights WebConsole icon





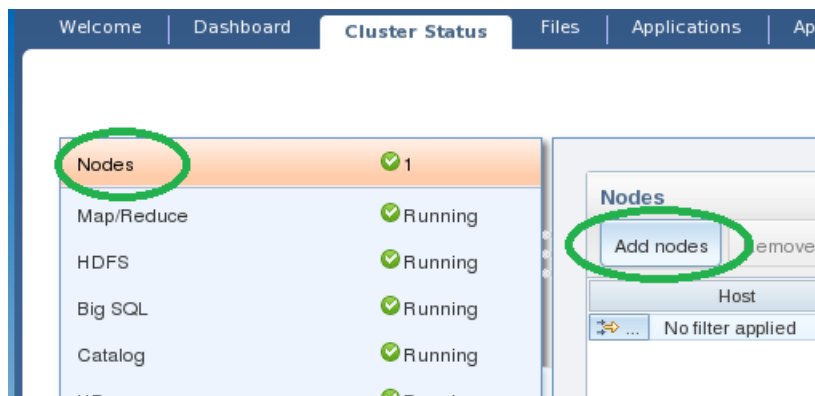
- \_\_2. For the Quick Start Edition, you will not require to log into the Web Console. You should now be at the Welcome Page



- \_\_3. Click on the Cluster Status tab.

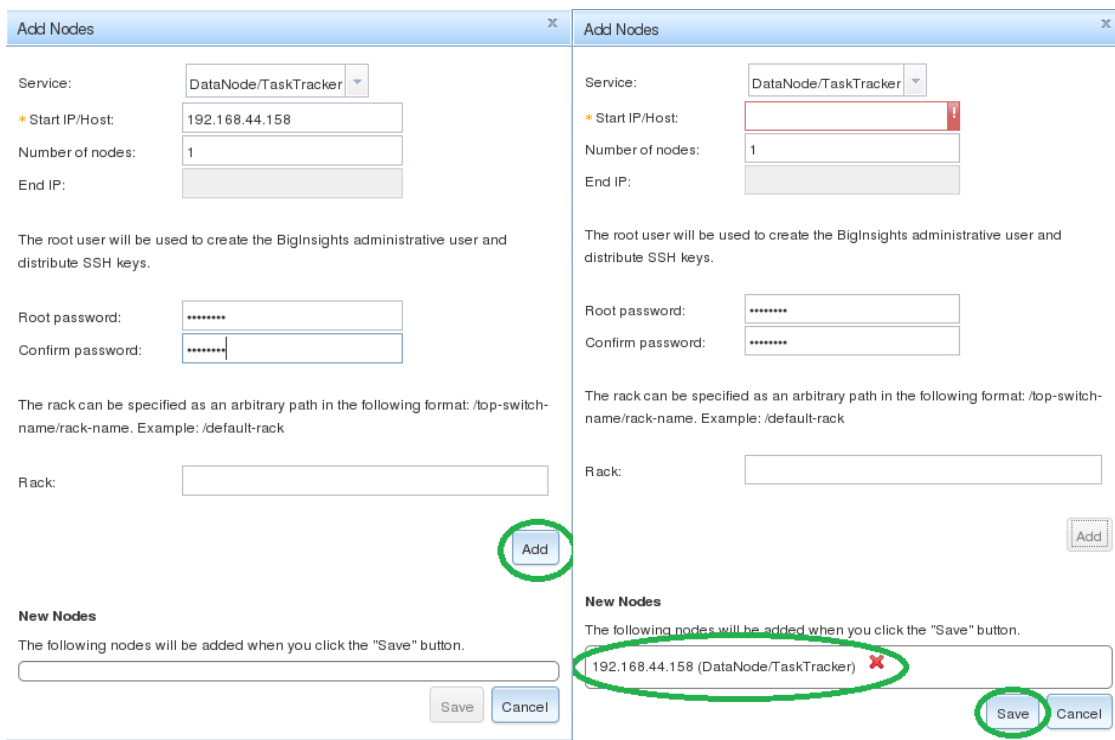


- \_\_4. Click on Nodes then click the Add nodes button

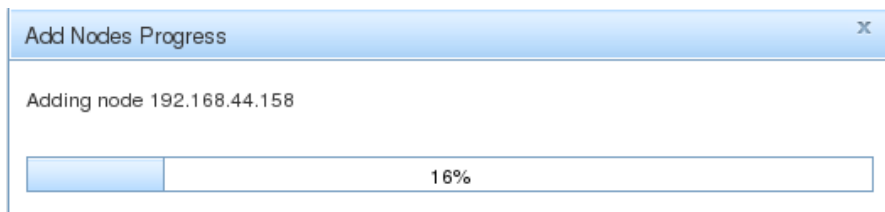


- \_\_5. Enter the IP address of the first node. This node must be online and reachable.

6. After you enter the IP address and password click on the *add* button (Left) you will now have a new IP address in the *New Nodes* section (Right). Now click save. The root password is 'password'



An add node progress bar will appear. Be patient as this may take some time.



You have now successfully added 1 Slave node to your cluster. The method which we just used is one of the simplest manners to expand your cluster, however we will cover another very useful method below. You can quickly see which nodes are running by navigating back to the *Cluster Status* tab in your BigInsights console.

Host	Status	Roles
No filter applied		
bivm	✔ Host is running	hive-server, secondarynamenode, zookeeper-client-port, bigsql-server, hive-web-interface, monitoring, hbase-regionserver, datanode, namenode, tasktracker, hbase-master, oozie-server, https-server, jobtracker
192.168.44.158	✔ Host is running	monitoring, datanode, tasktracker

### 1.1.4 Adding a node from the Terminal

You may also choose to add a node from the terminal. This can prove useful for a variety of different scenarios, such as real-time error logs if a node is not able to add successfully. Additionally, if you are not running the 'Console' service within BigInsights, or are using a remote connect program such as Putty to ssh into your cluster– this proves very useful. **REMEMBER** to update the /etc/host file for the Master node and new Slave node before trying to add a new node.

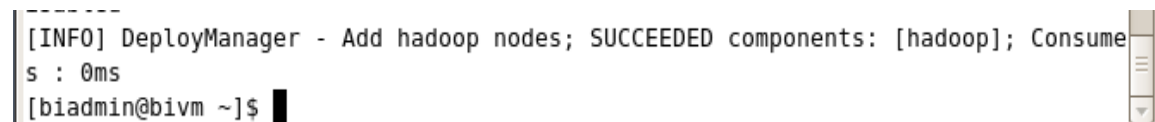
\_\_1. Open a new terminal.

\_\_2. Execute the following:

```
addnode.sh hadoop <IP_Addr OR Hostname>,password
```

IP\_Addr is the IP address of the new node you want to add, and Hostname is the name you have the node in your /etc/hosts file. If you are using another Quick Start Edition then the password will just be 'password'.

Adding a node through the terminal will take some time. After the node has been added you will get a message at the end:



```
-----
[INFO] DeployManager - Add hadoop nodes; SUCCEEDED components: [hadoop]; Consumed resources: 0ms
[biadmin@bivm ~]$
```

You have now successfully added a second node. You now have 2 Slave nodes.

If your machine is capable of running 4 VMs then go ahead and add a third Slave node through any method of your choosing.

### 1.1.5 Removing a node

Removing a node is as simple as adding a node, as the steps are very similar. We will show how to remove a node through the terminal in a quick manner. If a node has more than one service running, such as hadoop and flume, the specific service to be removed may be specified in the script – or if no service is specified the node is removed completely. **REMEMBER** to update the /etc/hosts file before removing.

\_\_1. Open a terminal.

\_\_2. You can remove a node by executing the following script:

```
removenode.sh --f <IP_Addr OR Hostname>
```

Where IP\_Addr is the IP address of the Slave node you want to remove and Hostname is the host name of the Slave node you wish to remove.

```

biadmin@bivm:~
File Edit View Terminal Tabs Help
[biadmin@bivm ~]$ removenode.sh node2
Warning: use command 'removenode.sh node2' will remove 'node2' from Biginsights
cluster ( any components configured to be running on these nodes will be decommi
ssioned, but application binaries/data will be retain ), are you sure to continu
e ? yes/y or no/n
yes
    
```

\_\_3. When prompted type 'yes' and hit <Enter>. This node will now be removed.

\_\_4. To verify that the node is now removed you can run the listnode.sh script.

**listnode.sh**

```

biadmin@bivm:~
File Edit View Terminal Tabs Help
[biadmin@bivm ~]$ listnode.sh
[INFO] DeployCmdline - [ IBM InfoSphere BigInsights QuickStart Edition ]
[INFO] DeployManager - All cluster nodes: [192.168.44.158, bivm]
[INFO] DeployManager - bigsql nodes: {bivm}
[INFO] DeployManager - console nodes: {bivm}
[INFO] DeployManager - derby nodes: {bivm}
[INFO] DeployManager - eclipsetooling nodes: {bivm}
[INFO] DeployManager - hadoop nodes: {192.168.44.158=[datanode, tasktracker], bi
    
```

1.1.6 Health of a Cluster

Servers, machines, and disk drives are all prone to a physical failure over time. When running a large cluster with dozens of nodes, it is crucial to over time maintain a constant health check of hardware and take appropriate actions when necessary. BigInsights 2.1 allows for a quick and simple way to perform these types of health checks on a cluster.

1.1.7 Visual Health Check

You can visually check the status of your cluster by following these simple steps:

\_\_1. Open a Web Console window by clicking the BigInsights WebConsole icon.



\_\_2. You should now be in the Welcome page. Click on the Cluster Status tab.



From here you can check the status of your nodes

The screenshot displays the Hadoop cluster management interface. On the left, a sidebar shows the status of various components. On the right, a 'Nodes' table provides details for each node in the cluster.

Component	Status
Nodes	2
Map/Reduce	Running
HDFS	Running
Big SQL	Running
Catalog	Running
HBase	Unavailable
Hive	Running
HttpFS	Running
Monitoring	Unavailable
Oozie	Running
Zookeeper	Running

Nodes		
Host	Status	Roles
No filter applied		
bvm	Host is running	hive-server, secondarynamenode, zookeeper-client-port, bigsql-server, hive-web-interface, monitoring, hbase-regionserver, datanode, namenode, tasktracker, hbase-master, oozie-server, https-server, jobtracker
192.168.44.158	Host is running	monitoring, datanode, tasktracker

You can also check the status of each component.

Nodes	2
Map/Reduce	Running
HDFS	Running
Big SQL	Running
Catalog	Running
HBase	Unavailable
Hive	Running
HttpFS	Running
Monitoring	Unavailable
Oozie	Running
Zookeeper	Running

### 1.1.8 DFS Disk Check

There are various ways to monitoring the DFS Disk, and this should be done occasionally to avoid space issues which can arise if there is low disk storage remaining. One such issue can occur if the “hadoop healthcheck” or heartbeat as it is also referred to sees that a node has gone offline. If a node is offline for a certain period of time, the data that the offline node was storing will be replicated to other nodes (since there is a 3node replication, the data is still available on the other 2 nodes). If there is limited disk space, this can quickly cause an issue.

- \_\_\_1. From a terminal window you can quickly access the dfs report by entering the following command:

```
hadoop dfsadmin -report
```

```

biadmin@bivm:~
File Edit View Terminal Tabs Help
[biadmin@bivm ~]$ hadoop dfsadmin -report
Configured Capacity: 76108426446 (70.88 GB)
Present Capacity: 49409740910 (46.02 GB)
DFS Remaining: 49233817600 (45.85 GB)
DFS Used: 175923310 (167.77 MB)
DFS Used%: 0.36%
Under replicated blocks: 0
Blocks with corrupt replicas: 0
Missing blocks: 0

-----
Datanodes available: 3 (4 total, 1 dead)

Name: 192.168.44.156:50010
Decommission Status : Normal
Configured Capacity: 25369475482 (23.63 GB)
DFS Used: 28782 (28.11 KB)
Non DFS Used: 8546314540 (7.96 GB)
DFS Remaining: 16823132160 (15.67 GB)
DFS Used%: 0%
DFS Remaining%: 66.31%
Last contact: Tue Jun 25 14:53:52 EDT 2013

```

## 1.2 Hadoop Administration

After completing this section, you’ll be able to:

- Start and stop individual services to best optimize the cluster performance
- Change default parameters within Hadoop such as the HDFS Block Size
- Manage service-specific slave nodes

### 1.2.1 Administering Specific Services

A single node can have a wide variety of services running at any given time, as seen in the screenshot below. Depending on your system and needs, it may not always be necessary to have all of the services running, as the more services running the more resources and computing power is being consumed by them.

```
hive-server, secondarynamenode,
zookeeper-client-port, hive-web-interface,
monitoring, flume-node, flume-master,
hbase-regionserver, datanode,
namenode, tasktracker, jaqlserver, hbase-
master, jobtracker
```

Stopping specific services can be done easily through the terminal, as well as through the web console. For the purpose of this lab, we will stop the 2 services, *hadoop* and *console* which should have been previously started.

- \_\_1. Open a terminal window.
- \_\_2. Stop the hadoop and console services by entering the following:

```
stop.sh hadoop console
```

```
bladmin@bigdata:/opt/ibm/biginsights/bin
File Edit View Terminal Tabs Help
[bladmin@crbigdata ~]$ cd $BIGINSIGHTS_HOME/bin
[bladmin@crbigdata bin]$ ./stop.sh hadoop console
[INFO] Progress - Stop console
[INFO] Deployer - stopping pigserver
[INFO] Deployer - no pigserver to stop
[INFO] Deployer - Stopping BigInsights Management Console...
[INFO] Deployer - Server waslp-server stopped.
[INFO] Progress - 50%
[INFO] Progress - Stop hadoop
[INFO] @bigdata - jobtracker stopped
[INFO] Progress - 60%
[INFO] @bigdata - tasktracker stopped
[INFO] Progress - 63%
[INFO] @192.168.44.129 - tasktracker already stopped
[INFO] Progress - 65%
[INFO] @bigdata - secondarynamenode stopped
[INFO] Progress - 70%
[INFO] Progress - 80%
[INFO] @bigdata - namenode stopped
[INFO] Progress - 85%
[INFO] @bigdata - datanode stopped
[INFO] Progress - 88%
[INFO] @192.168.44.129 - datanode already stopped
[INFO] Progress - 90%
[INFO] Progress - 100%
[INFO] DeployManager - Stop; SUCCEEDED components: [console, hadoop]; Consumes : 25094ms
[bladmin@crbigdata bin]$
```

The output should look similar to the image above.

## 1.2.2 Configuring Hadoop Default Settings

There are certain attributes from Apache Hadoop which are imported, and some have been changed to improve performance. One such attribute is the default block size used for storing large files.

Consider the following short example. You have a 1GB file, on a 3-node replication cluster. With a block-size of 128MB, this file will be split into 24 blocks (8 blocks, each replicated 3 times), and then stored on the hadoop cluster accordingly by the master node. Increasing and decreasing the block size can have very specific use-case implications, however for the sake of this lab we will not cover those Hadoop specific questions, but rather how to change these default values.

## 1.2.3 Increasing Storage Block Size

Hadoop uses a standard block storage system to store the data across its data nodes. Since block size is slightly more of an advanced topic, we will not cover the specifics as to what and why the data is stored as blocks throughout the cluster.

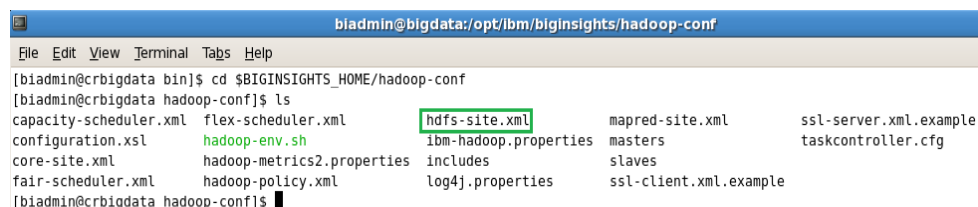
The default block size value for IBM BigInsights 2.1 is currently set at 128MB (as opposed to the Hadoop default of 64MB as you will see in the steps below). If your specific use-case requires you to change this, it can be easily modified through Hadoop configuration files.

\_\_1. When making any Hadoop core changes, it is good practice (and a requirement for most) to stop the services you are changing before making any necessary changes. For the block size, you must stop the “Hadoop” and “Console” services before proceeding if you have not done so in the previous steps, and re-start them after you have made the changes.

\_\_2. Move to the directory where Hadoop configuration files are stored

```
cd $BIGINSIGHTS_HOME/hadoop-conf
```

```
ls
```



```

biadmin@bigdata:/opt/ibm/biginsights/hadoop-conf
File Edit View Terminal Tabs Help
[biadmin@crbigdata bin]$ cd $BIGINSIGHTS_HOME/hadoop-conf
[biadmin@crbigdata hadoop-conf]$ ls
capacity-scheduler.xml  flex-scheduler.xml  hdfs-site.xml  mapred-site.xml  ssl-server.xml.example
configuration.xml       hadoop-env.sh       ibm-hadoop.properties  masters           taskcontroller.cfg
core-site.xml           hadoop-metrics2.properties  includes        slaves
fair-scheduler.xml     hadoop-policy.xml    log4j.properties  ssl-client.xml.example
[biadmin@crbigdata hadoop-conf]$

```

\_\_3. Within this directory, you will see a file named “hdfs-site.xml”, one of the site-specific configuration files, which is on every host in your cluster.

```
gedit hdfs-site.xml &
```



- \_\_4. Navigate to the property called *dfs.block.size*, and you will see the value is set to 128MB, the default block size for BigInsights. For the purpose of this lab, we will not change the value.

```

<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file. -->

<configuration>

  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>

  <property>
    <!-- The default block size for new files. Overrides default 64MB. -->
    <name>dfs.block.size</name>
    <value>134217728</value><!-- 128MB -->
  </property>

  <property>
    <!-- The number of server threads for the namenode. Overrides default 10. -->
    <name>dfs.namenode.handler.count</name>
    <value>64</value>
  </property>

```

#### 1.2.4 Limit Data nodes disk usage

- \_\_1. Navigate to the property named *dfs.datanode.du.reserved*. This value represents reserved space in bytes per volume. HDFS will always leave this much space free for non-dfs use.

```

<property>
  <name>dfs.secondary.http.address</name>
  <value>bigdata:50090</value>
</property>
<property>
  <name>dfs.datanode.du.reserved</name>
  <value>6012954214</value>
</property>
<property>
  <name>dfs.hosts</name>
  <value>/opt/ibm/biginsights/hadoop-conf/includes</value>
</property>
</configuration>

```



**NOTE:** This configuration file is site-specific which means it only is affective for a node this file belongs to. Read-only default configuration is stored at `$BIGINSIGHTS_HOME/IHC/src/hdfs/hdfs-default.xml`

### 1.2.5 Configuring the replication factor

- \_\_1. Navigate to the property named *dfs.replication*. If it is not specified in this file, then default replication value is used which is specified in *hdfs-default.xml*.
- \_\_2. *hdfs-default.xml* is stored at `$BIGINSIGHTS_HOME/IHC/src/hdfs`
- \_\_3. Current default replication factor value is 3. You can overwrite the default value by adding the following lines to this file (*hdfs-site.xml*). The value will be the number of your choice.

```
<property>  
  <name>dfs.replication</name>  
  <value>3</value>  
</property>
```

- \_\_4. For the purpose of this lab, we will not save this configuration change. This part of the lab is just to let you browse how to change some of the configuration values when you need it later on.

## 1.3 Importing Large Amounts of Data

After completing this section, you'll be able to:

- Easily copy entire directories from a live HDFS to a Linux file system
- Load entire directories from any file system onto the cluster (HDFS)

### 1.3.1 Moving Data to and from HDFS

If you have followed some of the other IMAZ BigInsights Labs, you will have come across loading a single file onto the Hadoop Distributed File System at various points using the console tool. One key point worth noting about that approach is that while it works excellently for a small number of files, when loading entire directories with dozens of files it can be rather tedious.

The Hadoop shell provides a great alternative to this previous approach for loading files and can be accessed in a wide number of ways such as through a terminal, through a jaql command, through java code, from the BigInsights console, and many others. For this section, we will cover loading data through the linux terminal and through the BigInsights console, and compare the differences and limitations of each approach.

To do this, we will follow a simple use case of copying an entire directory from the cluster to the local file system and back to the cluster via different methods at each step.

### 1.3.2 Hadoop commands through terminal

Now that we have implemented the program, we can run the program. The tooling provides the capabilities to run the Java map-reduce program locally, or it can be run remotely on the cluster. First we will run the program locally.

- \_\_1. Open a terminal window.
- \_\_2. First, we will copy an entire directory from the cluster onto our local file system. This can be done in one step by invoking the `hadoop fs` from the terminal as follows:

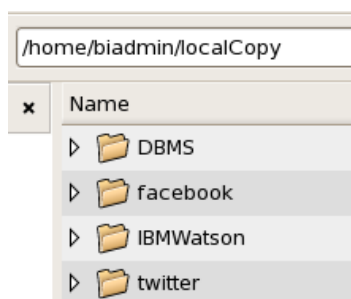
```
hadoop fs -copyToLocal /user/biadmin/sampleData/
/home/biadmin/localCopy/
```

This will copy all the files and directories recursively within `sampleData` to a local directory. Common linux regular expressions may also be used in the naming, such as `*`.



A terminal window titled `biadmin@bigdata:/opt/ibm/biginsights/bin` showing the command `hadoop fs -copyToLocal /user/biadmin/sampleData/ /home/biadmin/localCopy/` being executed. The prompt changes from `[biadmin@crbigdata bin]$` to `[biadmin@crbigdata bin]$` after the command is run.

- \_\_3. Navigate to the folder you just created to verify that the files were moved successfully. You may leave the terminal window open as it will be used again.



#### Can I enter a different URL to reach an external cluster?



The file system `'hadoop fs'` commands can be run against any cluster (not just one local to your machine), granted you have firewall/network access to that given hadoop file system. It is also possible to communicate directly from a hadoop filesystem to another if they are running the same compatible version of hadoop.

### 1.3.3 Hadoop commands through WebConsole

An alternate way to invoke hadoop commands is through the BigInsights WebConsole.

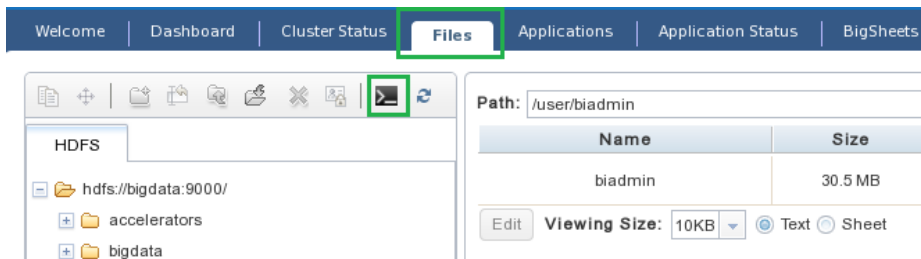


**NOTE:** Since you are invoking these commands from the cluster side, any commands which refer to “local” will cause an error as local does not apply since this can be run from a cluster of machines and the concept of ‘local’ no longer applies.

- \_\_4. Open a Web Console window by clicking on the BigInsights WebConsole icon.

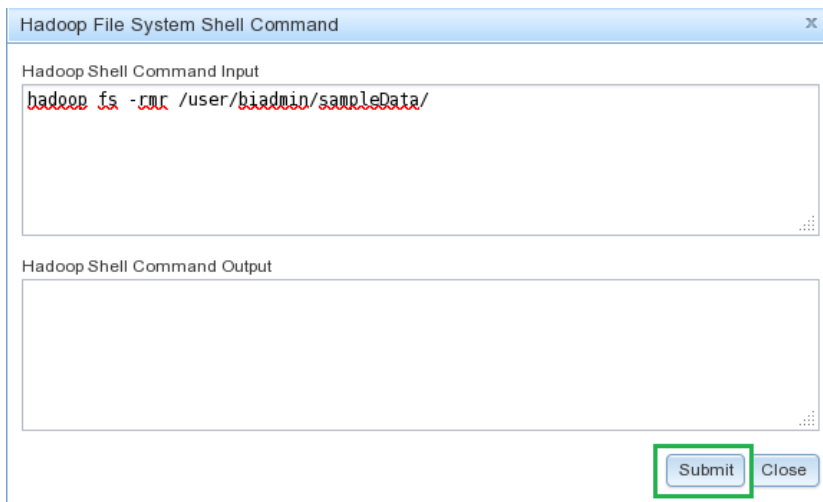


- \_\_5. Click on the Files tab, then select the Hadoop Shell icon.

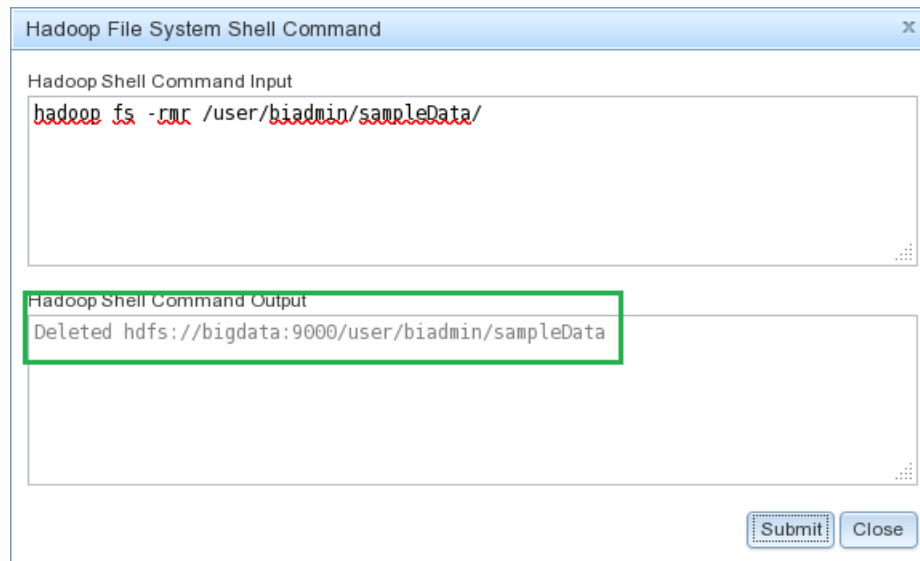


- \_\_6. Enter the following command to completely remove the directory from the cluster which you created a local backup in section 4.1.1. then click *Submit*

```
hadoop fs -rmr /user/biadmin/sampleData/
```



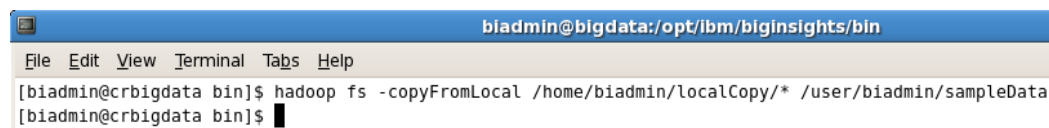
Your hadoop shell should now look like the image below.



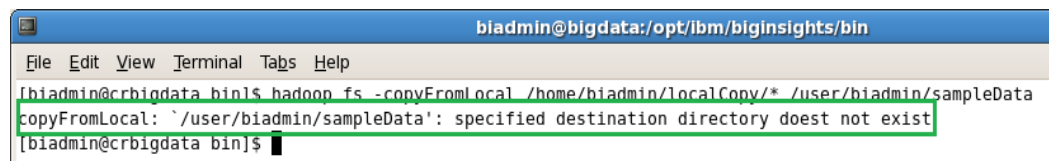
\_\_7. Verify by navigating to the `/user/biadmin/` directory from the files tab. You should no longer see the `sampleData` directory or any of its contents. Return to the terminal window.

\_\_8. You can now load the backup files onto the cluster with a simple command:

```
hadoop fs -copyFromLocal /home/biadmin/localCopy/*
/user/biadmin/sampleData/
```



You may come across an error that says specified destination does not exist



\_\_9. If this error occurs then you must create the directory `sampleData` under `/user/biadmin/`. Enter the following:

```
hadoop fs -mkdir /user/biadmin/sampleData
```

\_\_10. Now you may enter the command:

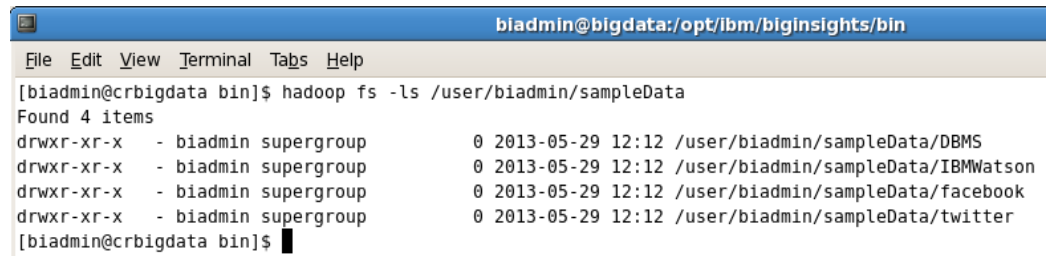
```
hadoop fs -copyFromLocal /home/biadmin/localCopy/*
/user/biadmin/sampleData/
```



\_\_11. Verify the files have been moved by issuing an ls command:

```
hadoop fs -ls /user/biadmin/sampleData
```

All the files should be there.



```
biadmin@bigdata:/opt/ibm/biginsights/bin
File Edit View Terminal Tabs Help
[biadmin@crbigdata bin]$ hadoop fs -ls /user/biadmin/sampleData
Found 4 items
drwxr-xr-x - biadmin supergroup      0 2013-05-29 12:12 /user/biadmin/sampleData/DBMS
drwxr-xr-x - biadmin supergroup      0 2013-05-29 12:12 /user/biadmin/sampleData/IBMWatson
drwxr-xr-x - biadmin supergroup      0 2013-05-29 12:12 /user/biadmin/sampleData/facebook
drwxr-xr-x - biadmin supergroup      0 2013-05-29 12:12 /user/biadmin/sampleData/twitter
[biadmin@crbigdata bin]$
```

## 1.4 Summary

Congratulations! You have now experience some common tasks of hadoop administrations.









---

© Copyright IBM Corporation 2013.

The information contained in these materials is provided for informational purposes only, and is provided AS IS without warranty of any kind, express or implied. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, these materials. Nothing contained in these materials is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers or licensors, or altering the terms and conditions of the applicable license agreement governing the use of IBM software. References in these materials to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates. This information is based on current IBM product plans and strategy, which are subject to change by IBM without notice. Product release dates and/or capabilities referenced in these materials may change at any time at IBM's sole discretion based on market opportunities or other factors, and are not intended to be a commitment to future product or feature availability in any way.

IBM, the IBM logo and [ibm.com](http://ibm.com) are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at [www.ibm.com/legal/copytrade.shtml](http://www.ibm.com/legal/copytrade.shtml).



Please Recycle

---