

Hadoop Basics with InfoSphere BigInsights

Unit 4: Hadoop Administration



An IBM Proof of Technology

Catalog Number

Contents

LAB 1	HADOOP ADMINISTRATION	4
1.1	MANAGING A HADOOP CLUSTER.....	5
1.1.1	ADDING/REMOVING A NODE FROM THE CLUSTER.....	5
1.1.2	SETTING UP MASTER/SLAVE NODES	ERROR! BOOKMARK NOT DEFINED.
1.1.3	ADDING A NODE FROM WEB CONSOLE	ERROR! BOOKMARK NOT DEFINED.
1.1.4	ADDING A NODE FROM THE TERMINAL	14
1.1.5	REMOVING A NODE	15
1.1.6	HEALTH OF A CLUSTER	16
1.1.7	VISUAL HEALTH CHECK.....	16
1.1.8	DFS DISK CHECK	18
1.2	HADOOP ADMINISTRATION	18
1.2.1	ADMINISTERING SPECIFIC SERVICES	19
1.2.2	CONFIGURING HADOOP DEFAULT SETTINGS	20
1.2.3	INCREASING STORAGE BLOCK SIZE	20
1.2.4	LIMIT DATA NODES DISK USAGE.....	21
1.2.5	CONFIGURING THE REPLICATION FACTOR	21
1.3	IMPORTING LARGE AMOUNTS OF DATA	ERROR! BOOKMARK NOT DEFINED.
1.3.1	MOVING DATA TO AND FROM HDFS	ERROR! BOOKMARK NOT DEFINED.
1.3.2	HADOOP COMMANDS THROUGH TERMINAL	ERROR! BOOKMARK NOT DEFINED.
1.3.3	HADOOP COMMANDS THROUGH WEBCONSOLE.....	ERROR! BOOKMARK NOT DEFINED.
1.4	SUMMARY	23

Lab 1 Hadoop Administration

IBM's InfoSphere BigInsights 2.1.2 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 Cluster Capable Quick Start Edition and tested on BigInsights 2.1.2. Throughout this lab you will be using the following account login information:

NOTE: Make sure to get the Cluster Capable Quick Start Edition

	Username	Password
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:

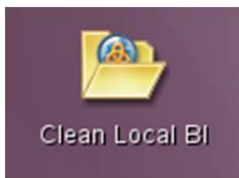
- Add and remove nodes through Web Console, and Terminal
- Check the health of the cluster and individual nodes within that cluster
- Perform checks on the disk and storage of the HDFS

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 Prepare your environment for a multi-node cluster

So far you have been working with just a single node cluster. To add a second node to the cluster, you need to have a second VMWare image. For clarification purposes, the existing image will be referred to

- __1. Unzip the image that you downloaded to a different directory. Boot it and go through the same process of accepting the licenses that you did for the Master image. Specify the same password for *root* and *biadmin* for the child image as you did for the Master image.
- __2. Once your child image boots up, log in with a username of **biadmin**.
- __3. For a node to be added to a BigInsights cluster, BigInsights cannot be installed. You need to uninstall BigInsights on the child image. Double click the **Clean Local BI** icon on the desktop to uninstall BigInsights.



- __4. Switch to the child image.

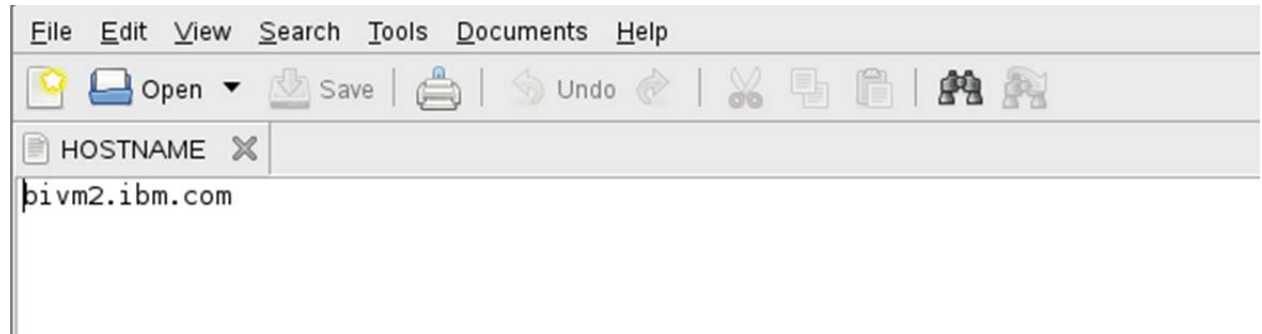
Child image

- __5. You need to update the hostname on this image. (It currently has the same hostname as the master.) You also need to update the `/etc/hosts` file so the child image can communicate with the master image. Right click the desktop and select **Open in Terminal**.
 - __a. Switch to root


```
su -
```
 - __b. Edit `/etc/HOSTNAME`

```
gedit /etc/HOSTNAME
```

- __c. Update the hostname from *bivm.ibm.com* to **bivm2.ibm.com**. Save your work and close the editor.



- __d. From the command line execute

```
hostname bivm2.ibm.com
```

- __e. Execute *hostname* without any parameters to verify that the hostname was changed.

- __f. Get the ipaddress from the master image. On the master image, right-click the desktop, select **Open in Terminal**. Then from the command line execute:

```
su -
```

```
ifconfig
```

- __g. On the child image, edit the */etc/hosts* file. (**gedit /etc/hosts**) Add the ipaddress and hostname from the master. Save your work and close the editor.

The following is an example. In this case, the ipaddress for the master was **192.168.70.202** and the ipaddress for the child was **192.168.70.201**.



Master image

- __6. On the master image, switch user to root. Then edit */etc/hosts* and add the hostname and ipaddress for the child image. Save your work and close the editor.

1.1.2 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 BigInsights 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.

- __7. On the *master* image, open a terminal window by right-clicking the desktop and select **Open in Terminal**.
- __8. Type the following ssh command to make sure that you have connectivity between the master and the child images:

```
ssh root@bivm2.ibm.com
```

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

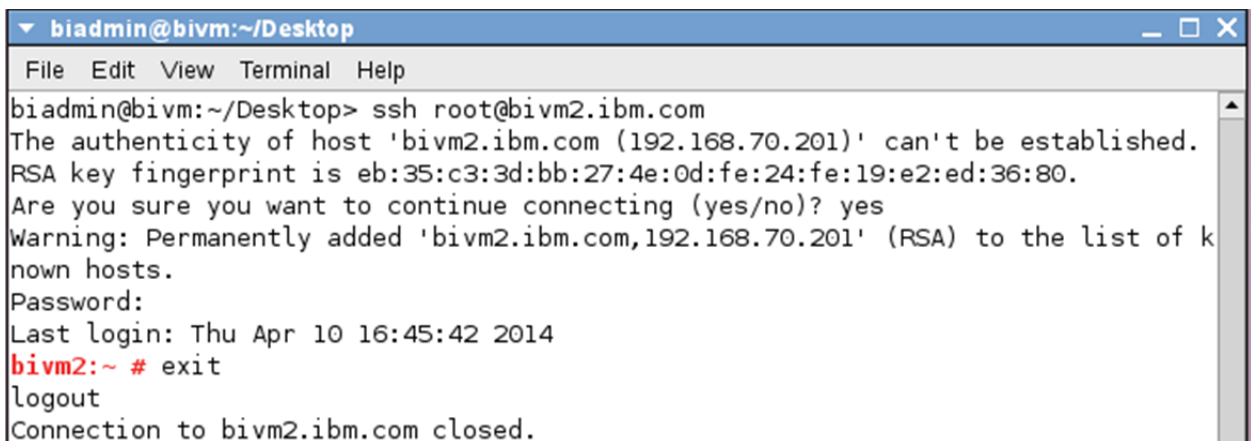
```
The authenticity of host 'bivm2.ibm.com (192.168.70.201)' can't be established.
RSA key fingerprint is 29:2f:72:9f:f4:97:16:89:cf:d9:cc:09:d3:16:d9:bf.
Are you sure you want to continue connecting (yes/no)?
```

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

```
Warning: Permanently added 'bivm2.ibm.com,192.168.70.201' (RSA) to the list of known
hosts.
```

Enter the password for root on the child image.

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



```
biadmin@bivm:~/Desktop
File Edit View Terminal Help
biadmin@bivm:~/Desktop> ssh root@bivm2.ibm.com
The authenticity of host 'bivm2.ibm.com (192.168.70.201)' can't be established.
RSA key fingerprint is eb:35:c3:3d:bb:27:4e:0d:fe:24:fe:19:e2:ed:36:80.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'bivm2.ibm.com,192.168.70.201' (RSA) to the list of k
nown hosts.
Password:
Last login: Thu Apr 10 16:45:42 2014
bivm2:~ # exit
logout
Connection to bivm2.ibm.com closed.
```

- __9. Exit the ssh connection then open a new terminal.

```
exit
```

1.1.3

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**.

BigInsights services must be started.

- __1. First start the BigInsights components. Right click the desktop of the master image and select **Open in Terminal**.
- __2. Start the BigInsight Components. You could use the *Start BigInsights* icon on the desktop. But this icon is only available with the Quick Start Edition image. When you install BigInsights, that icon will not be there. So let's use the technique that you will use in real-life. At the command line type:


```
$BIGINSIGHTS_HOME/bin/start-all.sh
```
- __3. Launch the Web Console by clicking on the BigInsights WebConsole icon. (Once again, this icon is only available with the Quick Start Edition image. In real-life you would open Firefox and specify a URL of <http://<hostname>:8080> (where <hostname> is the host where the console runs.)



- __4. For the Quick Start Edition, you will need to use the credentials to log into the BigInsights Console. Log in with a username of **biadmin** and specify the password that you assigned to *biadmin* when you configured the image. You should now be at the Welcome Page

IBM InfoSphere BigInsights Quick Start Edition (for Non-Production Environment) Welcome biadmin | Log out | About | Help

Welcome Dashboard Cluster Status Files Applications Application Status BigSheets

Understand IBM's Big Data Tools

Learn about biginsights Understand the tools for analyzing data at

Tasks

- Accelerate machine log, social, and telecommunications analytics**
If you have installed one of the IBM accelerators, you can run applications to jump-start your big data analytics.
- Create a dashboard**
Create a dashboard to monitor your application...
- Explore and update data using sheets**
Explore your data set to discover, analyze, and visualize your data.
- Run an application**
Run an application once, immediately.
- Deploy or remove an application**
Deploy an application on a cluster, or remove an application from a cluster.
- Add and remove nodes and services**
Add or remove nodes and service from your cluster, in addition to starting, stopping, and managing services.
- Manage alerts**
Create rules, manage existing rules, and activate rules.

Quick Links

- Download client library and development software
- Enable your Eclipse development environment for BigInsights application development
- Access secure cluster servers
- Run Big SQL Queries

Learn More

- Accelerator demos and documentation
- Infosphere BigInsights Information Center
- Infosphere Streams Information Center
- Communities and forums
- Support
- IBM big data on the Web

__5. Click on the Cluster Status tab.

IBM InfoSphere BigInsights Quick Start Edition (for Non-Production Environment)

Welcome Dashboard Cluster Status Files Applications Application

__6. Click on Nodes then click the Add nodes button

Welcome Dashboard Cluster Status Files Application

Monitor Services Manage Alerts Log Settings Backup and

Monitor Services	Status
Nodes	✓ 1
Map/Reduce	✓ Running
HDFS	✓ Running
Alert	✓ Running
Big SQL	✓ Running

Nodes

Add nodes

No filter a

__7. Enter the hostname of the first node. This node must be online and reachable.

__8. After you enter the IP address and password click on the *ok* button and then *accept* on the subsequent popup. Type the root password that you specified when you configured the child image.

Add Nodes

Hosts:

Hosts are public

The root user will be used to create the BigInsights administrative user and distribute SSH keys.

Root password:

The rack name may be used as a prefix to the host. BigInsights Console does not validate the rack information.
The rack can be specified as an arbitrary path in the following format: /top-switch-name/rack-name. Example: /default-rack

Rack:

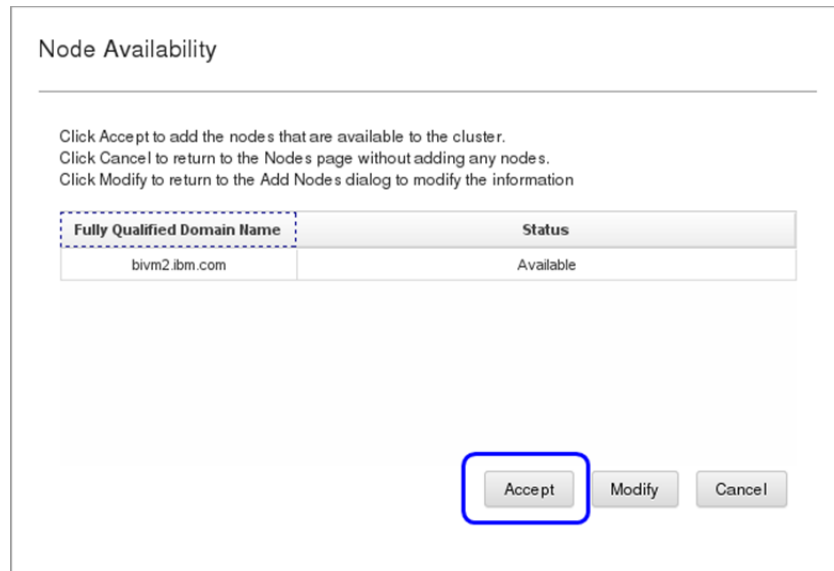
OK Cancel

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

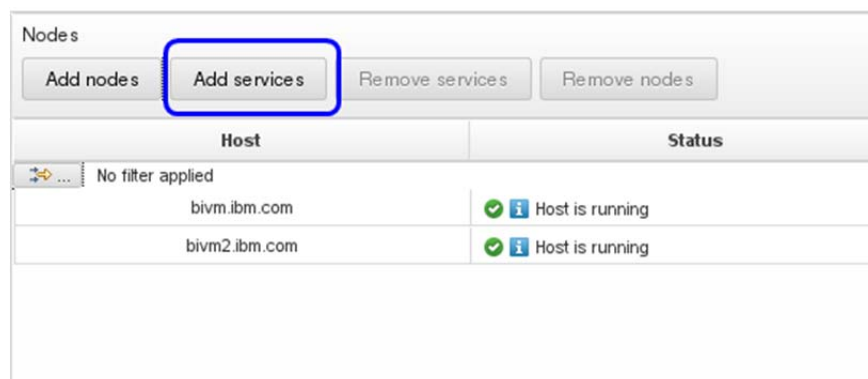
Resolve Hosts

70%

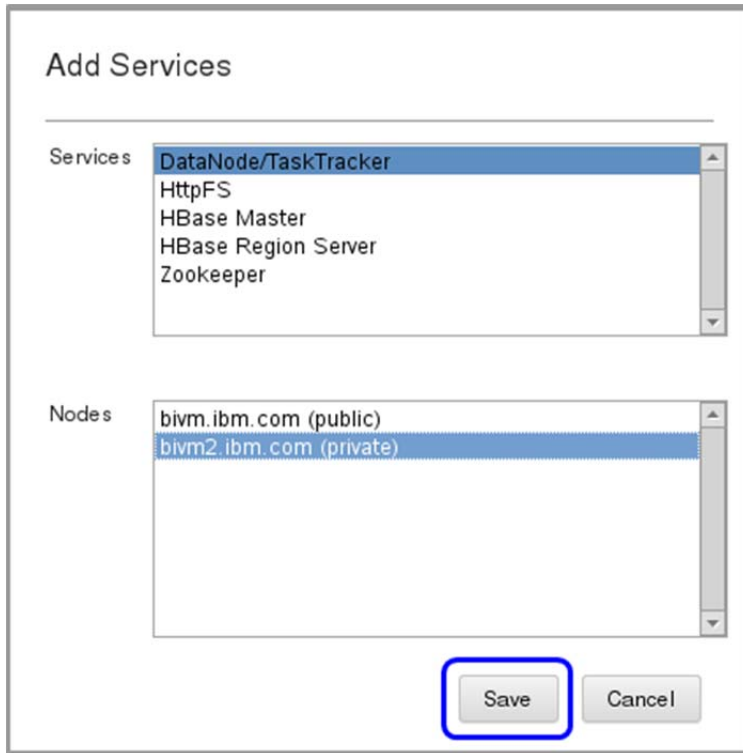
__9. A Node Availability window will pop up the nodes entered. Click **Accept** to proceed.



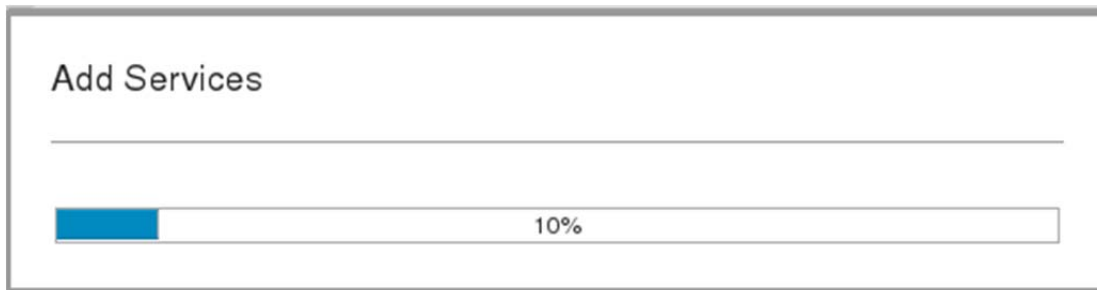
- ___10. It will take a few seconds for the nodes to appear up in the Node list. In BigInsights 2.1.2 there has been a slight change in how adding nodes works. First you must add the node, then you must give the node services. Click on **Add services**.



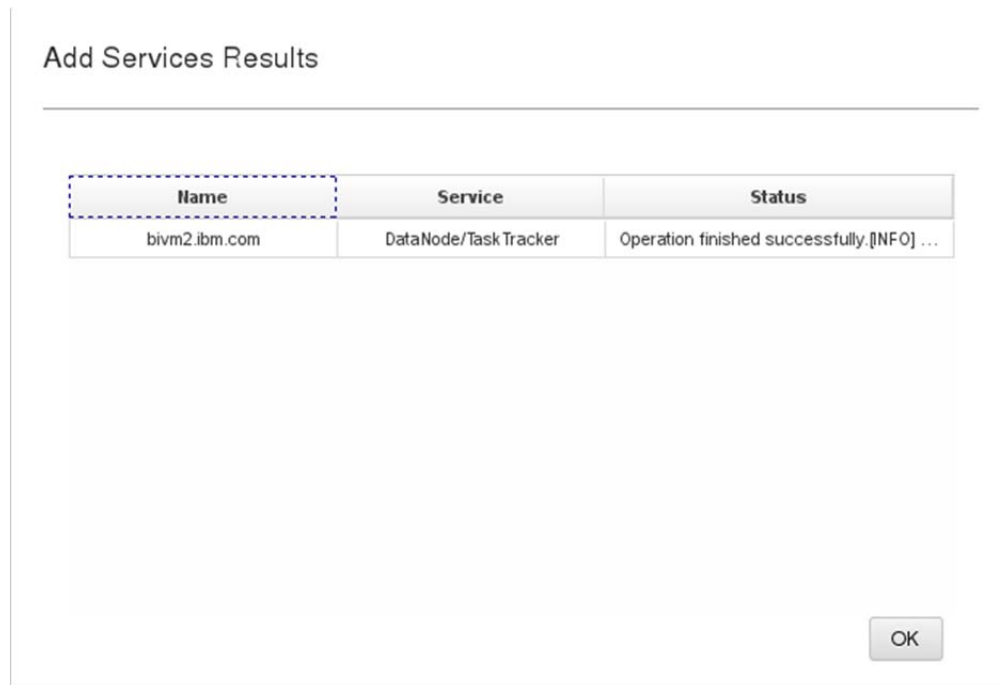
- ___11. For *Services* select **DataNode/TaskTracker**, and for *Nodes* select **bivm2.ibm.com** then click **Save**.



__12. You will now have a progress bar. This may take some time so be patient.



__13. After some time, you will get this window. Click **OK**.



You have now successfully added 1 child 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.

Nodes

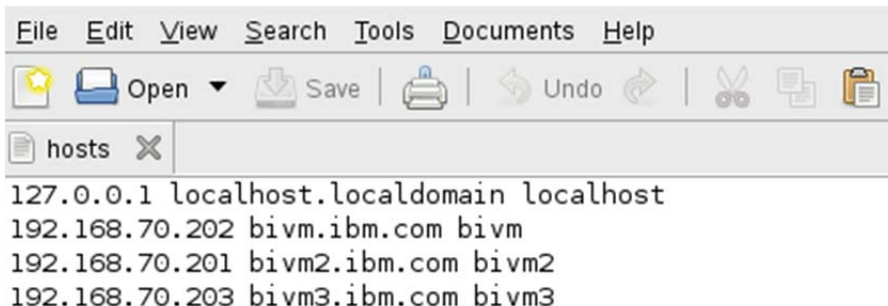
Add nodes Add services Remove services Remove nodes Refresh Interval: 15 seconds

Host	Status	Roles
No filter applied		
bivm.ibm.com	✓ Host is running	secondarynamenode, zookeeper-client-port, bigsql-server, hive-\
bivm2.ibm.com	✓ Host is running	monitoring, datanode, tasktracker

1.1.4 Adding a node from the Terminal Command Line

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 child node.

You may not have the resources on your PC to run three images. If you do not, then you can skip this section.

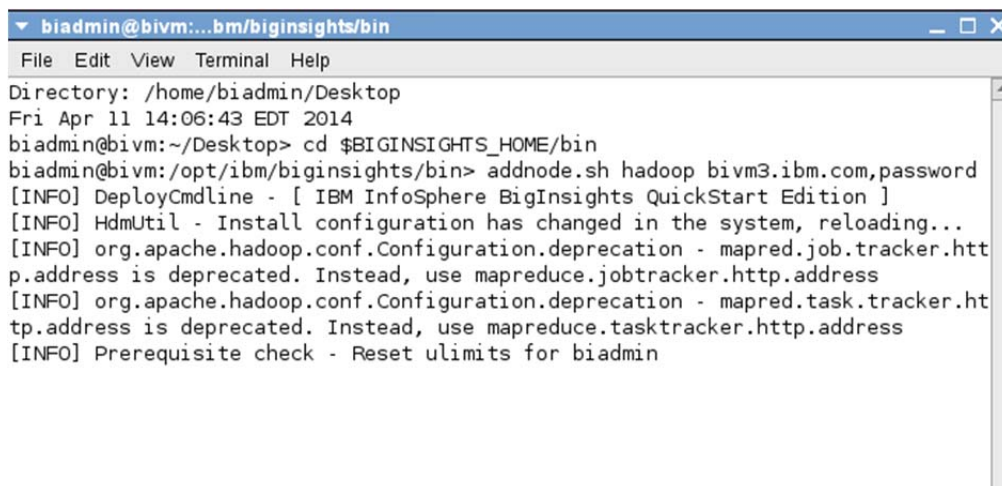


```
File Edit View Search Tools Documents Help
Open Save Undo
hosts
127.0.0.1 localhost.localdomain localhost
192.168.70.202 bivm.ibm.com bivm
192.168.70.201 bivm2.ibm.com bivm2
192.168.70.203 bivm3.ibm.com bivm3
```

__1. Right click the desktop and select **Open in Terminal**.

__2. Change directories to **\$BIGINSIGHTS_HOME/bin** and execute the following:

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



```
biadmin@bivm:...bm/biginsights/bin
File Edit View Terminal Help
Directory: /home/biadmin/Desktop
Fri Apr 11 14:06:43 EDT 2014
biadmin@bivm:~/Desktop> cd $BIGINSIGHTS_HOME/bin
biadmin@bivm:/opt/ibm/biginsights/bin> addnode.sh hadoop bivm3.ibm.com,password
[INFO] DeployCmdline - [ IBM InfoSphere BigInsights QuickStart Edition ]
[INFO] HdmUtil - Install configuration has changed in the system, reloading...
[INFO] org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.htt
p.address is deprecated. Instead, use mapreduce.jobtracker.http.address
[INFO] org.apache.hadoop.conf.Configuration.deprecation - mapred.task.tracker.ht
tp.address is deprecated. Instead, use mapreduce.tasktracker.http.address
[INFO] Prerequisite check - Reset ulimits for biadmin
```

<component> in our case is *hadoop*. *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. The password will be root's password on the child system.

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 - biginsights.properties unchanged, monitoring is already d
isabled
[INFO] DeployManager - Add hadoop nodes; SUCCEEDED components: [hadoop]; Consume
s : 421505ms
biadmin@bivm:/opt/ibm/biginsights/bin>

```

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

No filter applied		
bivm.ibm.com	✔ Host is running	secondarynamenode, zookeeper-client-port, bigsql-server, hive-...
bivm3.ibm.com	✔ Host is running	monitoring, datanode, tasktracker
bivm2.ibm.com	✔ Host is running	monitoring, datanode, tasktracker

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 or zookeeper, 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. The `--f` parameter says not to worry that some file chunks will not have a full set of replicas. Once again, you need to change to the `$BIGINSIGHTS_HOME` directory.

```
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:~/Desktop
Fri Apr 11 15:01:07 EDT 2014
biadmin@bivm:~/Desktop> cd $BIGINSIGHTS_HOME
biadmin@bivm:/opt/ibm/biginsights> removenode.sh --f bivm3.ibm.com
[INFO] DeployCmdline - [ IBM InfoSphere BigInsights QuickStart Edition ]
[INFO] HdmUtil - Install configuration has changed in the system, reloading...
[INFO] DeployManager - Stop monitoring agents on nodes:[bivm3.ibm.com]
[INFO] Deployer - Stop monitoring agents on [bivm3.ibm.com]
[INFO] Deployer - Monitoring agent is stopped on bivm3.ibm.com already.
[INFO] Deployer - monitoring service stopped
[INFO] Progress - RemoveNodes httpfs
[INFO] @bivm3.ibm.com - This node is not in httpfs cluster
[INFO] Progress - 16%
[INFO] Progress - RemoveNodes oozie

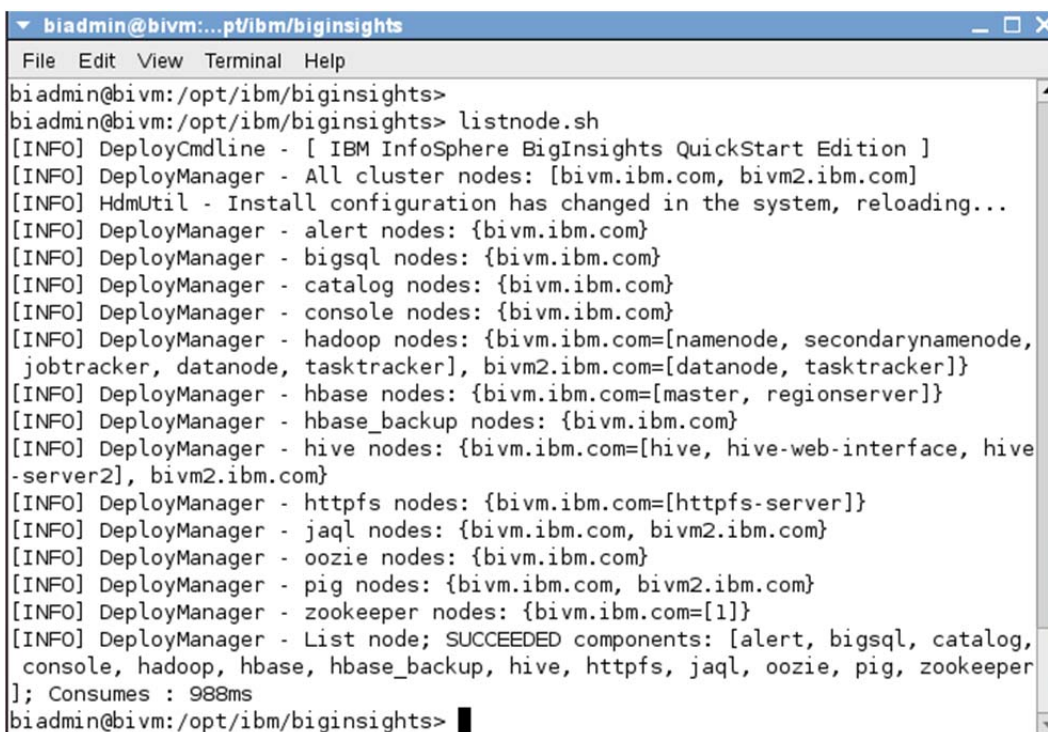
```

You should see this at the end:

```
[INFO] DeployManager - Cluster nodes removed [bivm3.ibm.com]
[INFO] DeployManager - RemoveNodes; SUCCEEDED components: [httpfs, oozie, hbase,
hadoop, zookeeper, monitoring]; Consumes : 64061ms
biadmin@bivm:/opt/ibm/biginsights>
```

__3. To verify that the node is now removed you can run the listnode.sh script.

listnode.sh



```
biadmin@bivm:/opt/ibm/biginsights>
biadmin@bivm:/opt/ibm/biginsights> listnode.sh
[INFO] DeployCmdline - [ IBM InfoSphere BigInsights QuickStart Edition ]
[INFO] DeployManager - All cluster nodes: [bivm.ibm.com, bivm2.ibm.com]
[INFO] HdmUtil - Install configuration has changed in the system, reloading...
[INFO] DeployManager - alert nodes: {bivm.ibm.com}
[INFO] DeployManager - bigsql nodes: {bivm.ibm.com}
[INFO] DeployManager - catalog nodes: {bivm.ibm.com}
[INFO] DeployManager - console nodes: {bivm.ibm.com}
[INFO] DeployManager - hadoop nodes: {bivm.ibm.com=[namenode, secondarynamenode,
jobtracker, datanode, tasktracker], bivm2.ibm.com=[datanode, tasktracker]}
[INFO] DeployManager - hbase nodes: {bivm.ibm.com=[master, regionserver]}
[INFO] DeployManager - hbase_backup nodes: {bivm.ibm.com}
[INFO] DeployManager - hive nodes: {bivm.ibm.com=[hive, hive-web-interface, hive
-server2], bivm2.ibm.com}
[INFO] DeployManager - httpfs nodes: {bivm.ibm.com=[httpfs-server]}
[INFO] DeployManager - jaql nodes: {bivm.ibm.com, bivm2.ibm.com}
[INFO] DeployManager - oozie nodes: {bivm.ibm.com}
[INFO] DeployManager - pig nodes: {bivm.ibm.com, bivm2.ibm.com}
[INFO] DeployManager - zookeeper nodes: {bivm.ibm.com=[1]}
[INFO] DeployManager - List node; SUCCEEDED components: [alert, bigsql, catalog,
console, hadoop, hbase, hbase_backup, hive, httpfs, jaql, oozie, pig, zookeeper
]; Consumes : 988ms
biadmin@bivm:/opt/ibm/biginsights>
```

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 2.1.2 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 BigInsights Console window by clicking the 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

Monitor Services | Manage Alerts | Log Settings | Backup and Restore HBase

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

Host	Status
No filter applied	
node1	✔ Host is running
bivm.ibm.com	✔ Host is running

You can also check the status of each component.

Nodes	Status
Nodes	✔ 2
Map/Reduce	✔ Running
HDFS	✔ Running
Alert	✔ Running
Big SQL	✔ Running
Catalog	✔ Running
HBase	✔ Running
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
```

```

Name: 192.168.70.201:50010 (bivm2.ibm.com)
Hostname: bivm2.ibm.com
Decommission Status : Normal
Configured Capacity: 37685021901 (35.10 GB)
DFS Used: 166364 (162.46 KB)
Non DFS Used: 8051958513 (7.50 GB)
DFS Remaining: 29632897024 (27.60 GB)
DFS Used%: 0.00%
DFS Remaining%: 78.63%
Last contact: Fri Apr 11 15:15:52 EDT 2014

Dead datanodes:
Name: 192.168.70.203:50010 (bivm3.ibm.com)
Hostname: bivm3.ibm.com
Decommission Status : Decommissioned
Configured Capacity: 0 (0 B)
DFS Used: 0 (0 B)
Non DFS Used: 0 (0 B)
DFS Remaining: 0 (0 B)
DFS Used%: 100.00%
DFS Remaining%: 0.00%
Last contact: Wed Dec 31 19:00:00 EST 1969

```

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

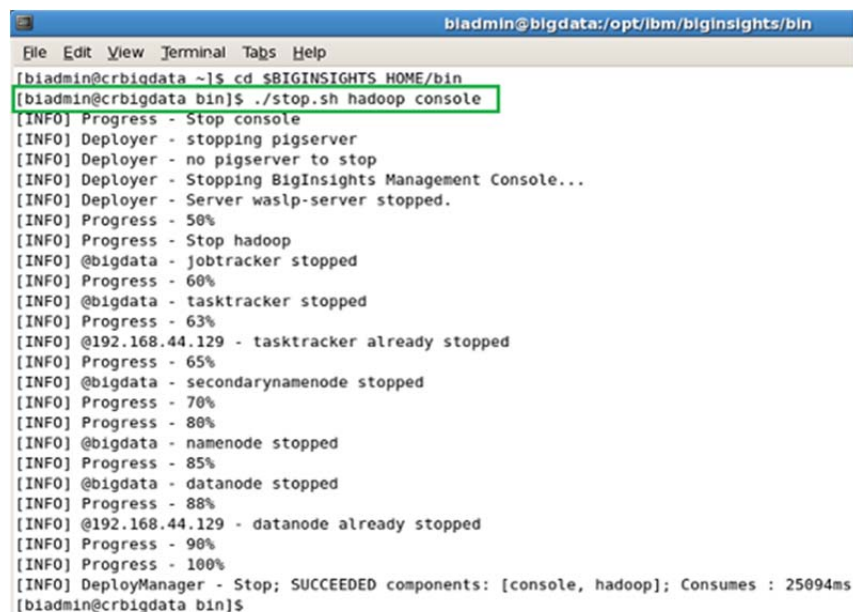
A single node can have a wide variety of services running at any given time, as seen in the screenshot below. Depending on your system and needs, it may not always be necessary to have all of the services running, as the more services running the more resources and computing power is being consumed by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

Configuration files for Hadoop are split into three files.

- `core-site.xml` – covers the Hadoop system
- `hdfs-site.xml` – covers HDFS specific configuration parameters
- `mapred-site.xml` – covers MapReduces specific configuration parameters

These configuration files reside in the `$(BIGINSIGHTS_HOME)/hadoop-conf` directory. Since there are multiple nodes in the cluster, when you change a configuration parameter, those changes need to be made on all nodes in the cluster, where it is appropriate. To automate this process, BigInsights includes a script, `syncconf.sh` that synchronizes the changes. For this to work, you do not modify the actual configuration files, but rather the *staging* configuration files. These files are located in the `$(BIGINSIGHTS_HOME)/hdm/ghadoop-conf-staging` directory.

1.2.3 Increasing Storage Block Size

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.

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 2.1.2 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 staging configuration files are stored

```
cd $(BIGINSIGHTS_HOME)/hdm/hadoop-conf-staging
```

```
ls
```

```

biadmin@bivm:...adoop-conf-staging
File Edit View Terminal Help
Directory: /opt/ibm/biginsights/hdm/hadoop-conf-staging
Fri Apr 11 15:45:06 EDT 2014
biadmin@bivm:/opt/ibm/biginsights/hdm/hadoop-conf-staging> ls
capacity-scheduler.xml  hadoop-metrics2.properties  masters
configuration.xml      hadoop-policy.xml           slaves
console-site.xml       hdfs-site.xml               ssl-client.xml.example
core-site.xml          ibm-hadoop.properties      ssl-server.xml.example
excludes               includes                     taskcontroller.cfg
fair-scheduler.xml    log4j.properties           zk-jaas.conf
flex-scheduler.xml    mapred-queue-acls.xml
hadoop-env.sh         mapred-site.xml
biadmin@bivm:/opt/ibm/biginsights/hdm/hadoop-conf-staging>

```

- __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 Configuring the replication factor

- __1. Navigate to the property named *dfs.replication*.

- __2. The current default replication factor will depend on the number of DataNodes that you have in your cluster. If you only have one, then the value is 1. If you have two DataNodes, then you will see a value of 2. For three or more DataNodes, the value will be 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>
```

1.2.5 Limit DataNodes 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 effective for a node this file belongs to. Read-only default configuration is stored at `$BIGINSIGHTS_HOME/IHC/src/hdfs/hdfs-default.xml`

- __2. 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. However, in real-life, once you have made changes to this file, you would then run the *synccconf.sh* script to synchronize this file across all appropriate nodes in the cluster.

1.3 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 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.



Please Recycle
