

Hadoop Fundamentals

Unit 2: Hadoop Architecture

Contents

LAB 1	HADOOP ADMINISTRATION	4
1.1	MANAGING A HADOOP CLUSTER.....	ERROR! BOOKMARK NOT DEFINED.
1.1.1	ADDING/REMOVING A NODE FROM THE CLUSTER.....	ERROR! BOOKMARK NOT DEFINED.
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	ERROR! BOOKMARK NOT DEFINED.
1.1.5	REMOVING A NODE	ERROR! BOOKMARK NOT DEFINED.
1.1.6	HEALTH OF A CLUSTER	ERROR! BOOKMARK NOT DEFINED.
1.1.7	VISUAL HEALTH CHECK.....	ERROR! BOOKMARK NOT DEFINED.
1.1.8	DFS DISK CHECK	ERROR! BOOKMARK NOT DEFINED.
1.2	HADOOP ADMINISTRATION	ERROR! BOOKMARK NOT DEFINED.
1.2.1	ADMINISTERING SPECIFIC SERVICES	ERROR! BOOKMARK NOT DEFINED.
1.2.2	CONFIGURING HADOOP DEFAULT SETTINGS	ERROR! BOOKMARK NOT DEFINED.
1.2.3	INCREASING STORAGE BLOCK SIZE	ERROR! BOOKMARK NOT DEFINED.
1.2.4	LIMIT DATA NODES DISK USAGE.....	ERROR! BOOKMARK NOT DEFINED.
1.2.5	CONFIGURING THE REPLICATION FACTOR	ERROR! BOOKMARK NOT DEFINED.
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	11

Lab 1 Hadoop Architecture

The overwhelming trend towards digital services, combined with cheap storage, has generated massive amounts of data that enterprises need to effectively gather, process, and analyze. Data analysis techniques from the data warehouse and high-performance computing communities are invaluable for many enterprises, however often times their cost or complexity of scale-up discourages the accumulation of data without an immediate need. As valuable knowledge may nevertheless be buried in this data, related scaled-up technologies have been developed. Examples include Google's MapReduce, and the open-source implementation, Apache Hadoop.

Hadoop is an open-source project administered by the Apache Software Foundation. Hadoop's contributors work for some of the world's biggest technology companies. That diverse, motivated community has produced a collaborative platform for consolidating, combining and understanding data. After completing this hands-on lab, you'll be able to:

- Use Hadoop commands to explore HDFS on the Hadoop system
- Use the BigInsights Console to explore HDFS on the Hadoop system

Allow 60 minutes to 90 minutes to complete this lab.

This version of the lab was designed using the InfoSphere BigInsights 2.1.2 Quick Start Edition. Throughout this lab you will be using the following account login information. The assumptions are that the passwords for **root** and **biadmin** are as follows. If, when you setup the Quick Start Edition image, you specified different passwords, then you will have to make the appropriate mental translations throughout the exercise.

	Username	Password
VM image setup screen	root	password
Linux	biadmin	biadmin

1.1 Getting Started

First start your Quick Start image. You should allocate at least 5 gig of memory to your image.

IBM InfoSphere BigInsights v2.1.2 Quick Star... **Version:** BigInsights v2.1.2 Quick Start Edition
Author: IBM

Power on this virtual machine
Edit virtual machine settings
Upgrade this virtual machine

▼ **Devices**

Memory	6 GB
Processors	2
Hard Disk (SCSI)	40 GB
Network Adapter	NAT
USB Controller	Present
Display	Auto detect

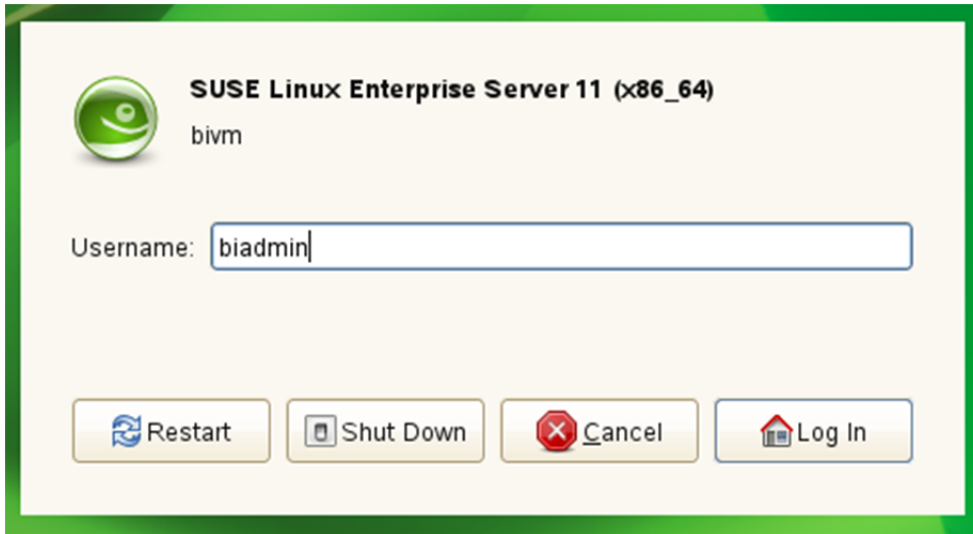
▼ **Description**

The IBM InfoSphere BigInsights v2.1.2 Quick Start Edition is designed to let you experiment with the features of InfoSphere BigInsights, while being able to use real data and run real applications. See the Readme file that accompanies this download for login and startup information.

▼ **Virtual Machine Details**

State: Powered off
Configuration file: E:\VMWareMa...\iibi2120_QuickStart_Single_VMware.vmx
Hardware compatibility: Workstation 6.5-7.x virtual machine

Log in with a username of **biadmin**.



Your desktop should look like the following:



- __4. You need to start the BigInsights components. With this image, you have two options. One option is to use the icon that was placed on the desktop. But that icon is unique to the Quick Start image. So start the components in the more traditional way.

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

__5. Type the following to start all BigInsights components:

\$BIGINSIGHTS_HOME/bin/start-all.sh

__6. Once all components have started successfully, you may continue.

```
[INFO] Progress - 100%
[INFO] DeployManager - Start; SUCCEEDED components: [hdm, zookeeper, hadoop, catalog, hbase, hive, bigsql, oozie, console, httpfs, alert]; Consumes : 192498ms
biadmin@bivm:~/Desktop>
```

1.2

Hadoop Distributed File System (HDFS) allows user data to be organized in the form of files and directories. It provides a command line interface called FS shell that lets a user interact with the data in HDFS accessible to Hadoop MapReduce programs.

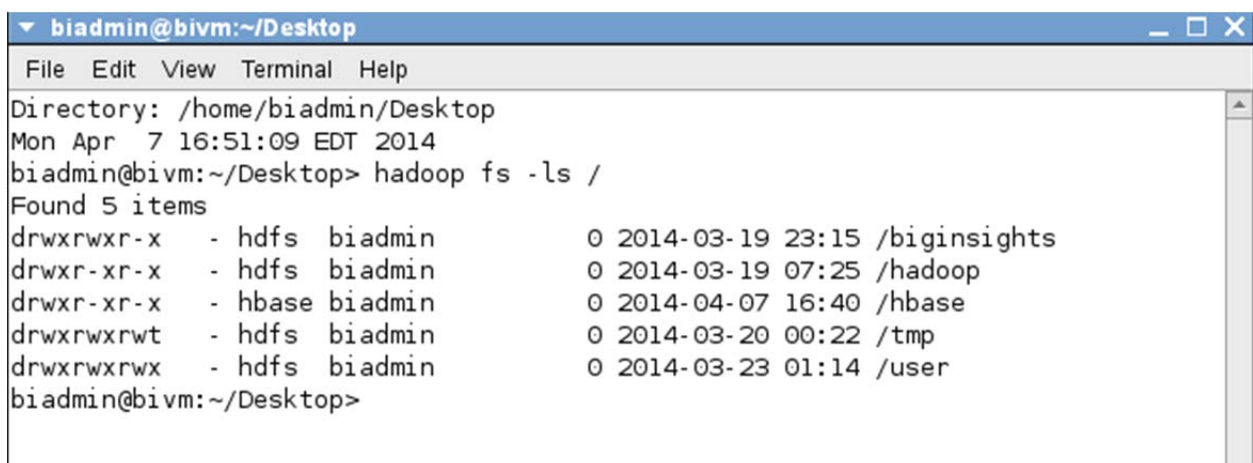
There are two ways that you can interact with HDFS:

1. You can use the command-line approach and invoke the File System (fs) shell.
2. You can use the BigInsights Console.

In this exercise you will work with both.

__1. Start with the command to list files and directories. In your command window, type the following:

hadoop fs -ls /



```
biadmin@bivm:~/Desktop
File Edit View Terminal Help
Directory: /home/biadmin/Desktop
Mon Apr 7 16:51:09 EDT 2014
biadmin@bivm:~/Desktop> hadoop fs -ls /
Found 5 items
drwxrwxr-x - hdfs biadmin 0 2014-03-19 23:15 /biginsights
drwxr-xr-x - hdfs biadmin 0 2014-03-19 07:25 /hadoop
drwxr-xr-x - hbase biadmin 0 2014-04-07 16:40 /hbase
drwxrwxrwt - hdfs biadmin 0 2014-03-20 00:22 /tmp
drwxrwxrwx - hdfs biadmin 0 2014-03-23 01:14 /user
biadmin@bivm:~/Desktop>
```

__2. Now list the files that are in biadmin's HDFS home directory .

hadoop fs -ls /user/biadmin

```

biadmin@bivm:~/Desktop
Directory: /home/biadmin/Desktop
Mon Apr 7 17:05:15 EDT 2014
biadmin@bivm:~/Desktop> hadoop fs -ls /user/biadmin
Found 4 items
drwx----- - biadmin biadmin      0 2014-03-20 00:51 /user/biadmin/.staging
drwxr-xr-x  - biadmin biadmin      0 2014-03-23 01:14 /user/biadmin/WordCount
drwxr-xr-x  - biadmin biadmin      0 2014-03-25 04:19 /user/biadmin/credstore
drwx--x--x  - biadmin biadmin      0 2014-03-20 00:51 /user/biadmin/oozie-oozi
biadmin@bivm:~/Desktop>

```

__3. Create a directory called *test*.

hadoop fs -mkdir /user/biadmin/test

__4. Then list biadmin's home directory again.

Hadoop fs -ls /user/biadmin

```

biadmin@bivm:~/Desktop
Directory: /home/biadmin/Desktop
Mon Apr 7 17:07:13 EDT 2014
biadmin@bivm:~/Desktop> hadoop fs -ls /user/biadmin
Found 5 items
drwx----- - biadmin biadmin      0 2014-03-20 00:51 /user/biadmin/.staging
drwxr-xr-x  - biadmin biadmin      0 2014-03-23 01:14 /user/biadmin/WordCount
drwxr-xr-x  - biadmin biadmin      0 2014-03-25 04:19 /user/biadmin/credstore
drwx--x--x  - biadmin biadmin      0 2014-03-20 00:51 /user/biadmin/oozie-oozi
drwxr-xr-x  - biadmin biadmin      0 2014-04-07 17:07 /user/biadmin/test
biadmin@bivm:~/Desktop>

```

__5. Create a file in the Linux home directory for biadmin. From a command line, execute the following commands. (ctrl-c means to press the Ctrl key and then the c key.)

cd ~

cat > myfile.txt

this is some data

in my file

ctrl-c

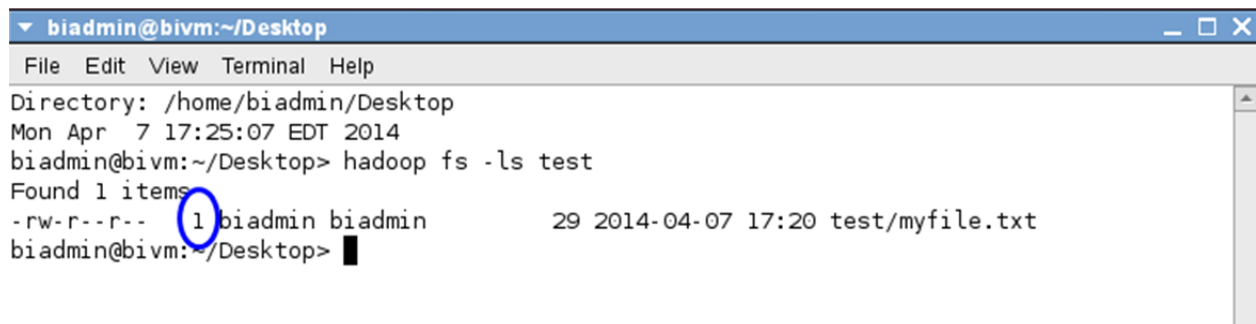
__6. Next upload this newly created file to the *test* directory that you just created.

```
hadoop fs -put ~/myfile.txt test/myfile.txt
```

__7. Now list the *test* directory.

```
hadoop fs -ls test
```

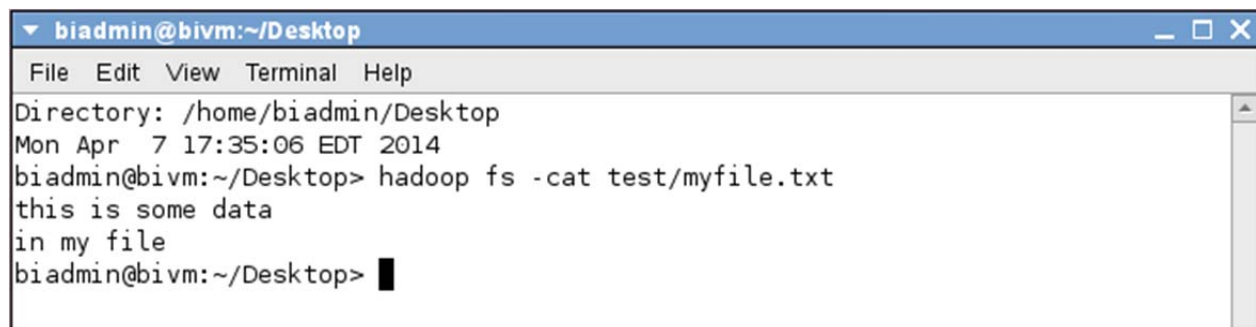
Note the number 1 that follows the permissions. This is the replication factor for that data file. Normally, that would be a 3 but since this is just a single node cluster, there is only going to be a single copy of the file.



```
biadmin@bivm:~/Desktop
File Edit View Terminal Help
Directory: /home/biadmin/Desktop
Mon Apr 7 17:25:07 EDT 2014
biadmin@bivm:~/Desktop> hadoop fs -ls test
Found 1 items
-rw-r--r-- 1 biadmin biadmin      29 2014-04-07 17:20 test/myfile.txt
biadmin@bivm:~/Desktop>
```

__8. To view the contents of the uploaded file, execute

```
hadoop fs -cat test/myfile.txt
```



```
biadmin@bivm:~/Desktop
File Edit View Terminal Help
Directory: /home/biadmin/Desktop
Mon Apr 7 17:35:06 EDT 2014
biadmin@bivm:~/Desktop> hadoop fs -cat test/myfile.txt
this is some data
in my file
biadmin@bivm:~/Desktop>
```

__9. You can pipe (using the | character) any HDFS command to be used with the Linux shell. For example, you can easily use *grep* with HDFS by doing the following.

```
hadoop fs -ls /user/biadmin ! grep test
```

```

biadmin@bivm:~/Desktop
File Edit View Terminal Help
Directory: /home/biadmin/Desktop
Mon Apr 7 17:31:28 EDT 2014
biadmin@bivm:~/Desktop> hadoop fs -ls /user/biadmin | grep test
drwxr-xr-x - biadmin biadmin          0 2014-04-07 17:20 /user/biadmin/test
biadmin@bivm:~/Desktop>

```

__10. To find the size of a particular file, like *myfile.txt*, execute the following:

`hadoop fs -du /user/biadmin/test/myfile.txt`

```

biadmin@bivm:~/Desktop
File Edit View Terminal Help
Directory: /home/biadmin/Desktop
Tue Apr 8 09:56:13 EDT 2014
biadmin@bivm:~/Desktop> hadoop fs -du /user/biadmin/test/myfile.txt
29 /user/biadmin/test/myfile.txt
biadmin@bivm:~/Desktop>

```

__11. Or to get the size of all files in a directory

`hadoop fs -du /user/biadmin`

```

biadmin@bivm:~/Desktop
File Edit View Terminal Help
Directory: /home/biadmin/Desktop
Tue Apr 8 10:05:27 EDT 2014
biadmin@bivm:~/Desktop> hadoop fs -du /user/biadmin
0 /user/biadmin/.staging
30 /user/biadmin/WordCount
137 /user/biadmin/credstore
0 /user/biadmin/oozie-oozi
29 /user/biadmin/test
biadmin@bivm:~/Desktop>

```

__12. Or to get a total file size value for all files in a directory:

`hadoop fs -du -s /user/biadmin`

```

biadmin@bivm:~/Desktop
File Edit View Terminal Help
Directory: /home/biadmin/Desktop
Tue Apr 8 10:06:17 EDT 2014
biadmin@bivm:~/Desktop> hadoop fs -du -s /user/biadmin
196 /user/biadmin
biadmin@bivm:~/Desktop>

```

__13. Remember that you can always use the `-help` parameter to get more help.

hadoop fs -help

hadoop fs -help du

```

biadmin@bivm:~/Desktop
File Edit View Terminal Help
Directory: /home/biadmin/Desktop
Tue Apr 8 10:10:38 EDT 2014
biadmin@bivm:~/Desktop> hadoop fs -help
Usage: hadoop fs [generic options]
    [-appendToFile <localsrc> ... <dst>]
    [-cat [-ignoreCrc] <src> ...]
    [-checksum <src> ...]
    [-chgrp [-R] GROUP PATH...]
    [-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]
    [-chown [-R] [OWNER][:[GROUP]] PATH...]
    [-copyFromLocal [-f] [-p] <localsrc> ... <dst>]
    [-copyToLocal [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
    [-count [-q] <path> ...]
    [-cp [-f] [-p] <src> ... <dst>]
    [-createSnapshot <snapshotDir> [<snapshotName>]]
    [-deleteSnapshot <snapshotDir> <snapshotName>]
    [-df [-h] [<path> ...]]
    [-du [-s] [-h] <path> ...]

```

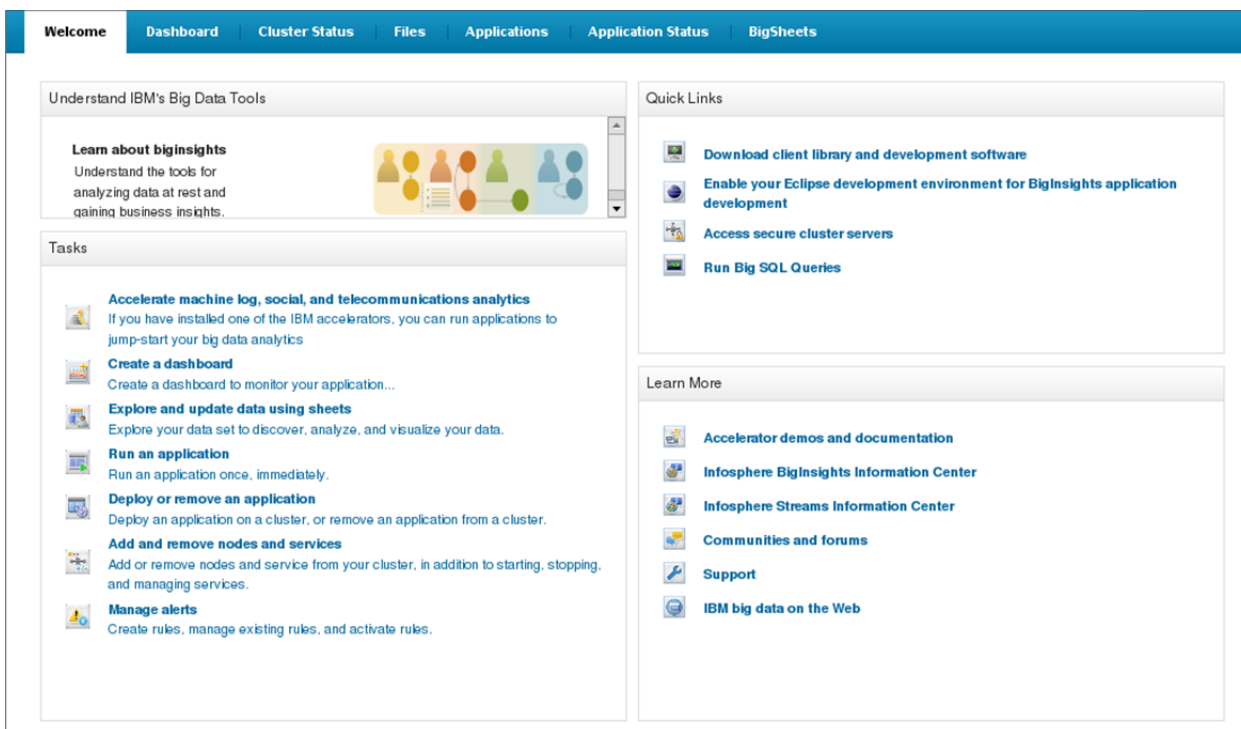
You can close the command line window.

1.3 BigInsights Console

__1. On the Quick Start Edition vmware image desktop, you can click an icon to start the BigInsights Console. Once again this is unique to this image. In real life, you would start Firefox and type in a URL of http://<host_name>:8080 where <host_name> is the host where the BigInsights Console is running.

For this exercise, double-click the **WebConsole** icon on the desktop.

2. Log in with a *User name* of **biadmin** and the password that you specified for *biadmin*. You should be taken to the *Welcome* tab which provides shortcuts to working with the console.



3. Click the **Cluster Status** tab. Here you can see the status of each component. On the left side, click **Nodes**. Here you can see the status of each node in the cluster, add and remove nodes and services. Selecting any of the components on the left, shows you detail information about that component.

Click **Map/Reduce**. On the right side, you can start and stop the JobTracker and any of the TaskTrackers. You also can see the host and port number for the JobTracker and each TaskTracker running on each node. In our case there is only a single node.

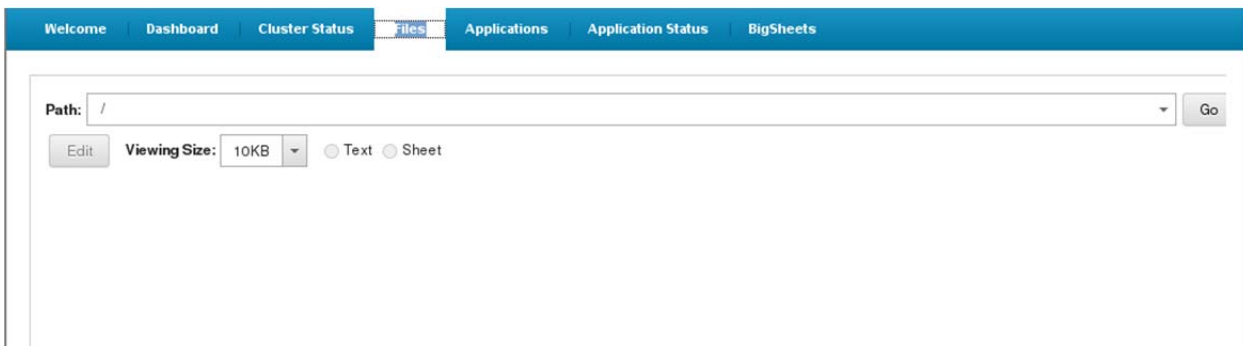
The screenshot shows the 'Cluster Status' page in the IBM Big Data Platform. The 'Map/Reduce' service is highlighted as 'Running'. The 'Map/Reduce Summary' section shows 'JobTracker: bivm.ibm.com:9001' and 'Status: Running'. Below this, the 'Tasktrackers' section shows a table with one entry: 'bivm.ibm.com:50060' with status 'Running' and 'Process ID' 25449. The 'Start' and 'Stop(1)' buttons are visible for the tasktrackers.

Host & Port	Status	Process ID
bivm.ibm.com:50060	Running	25449

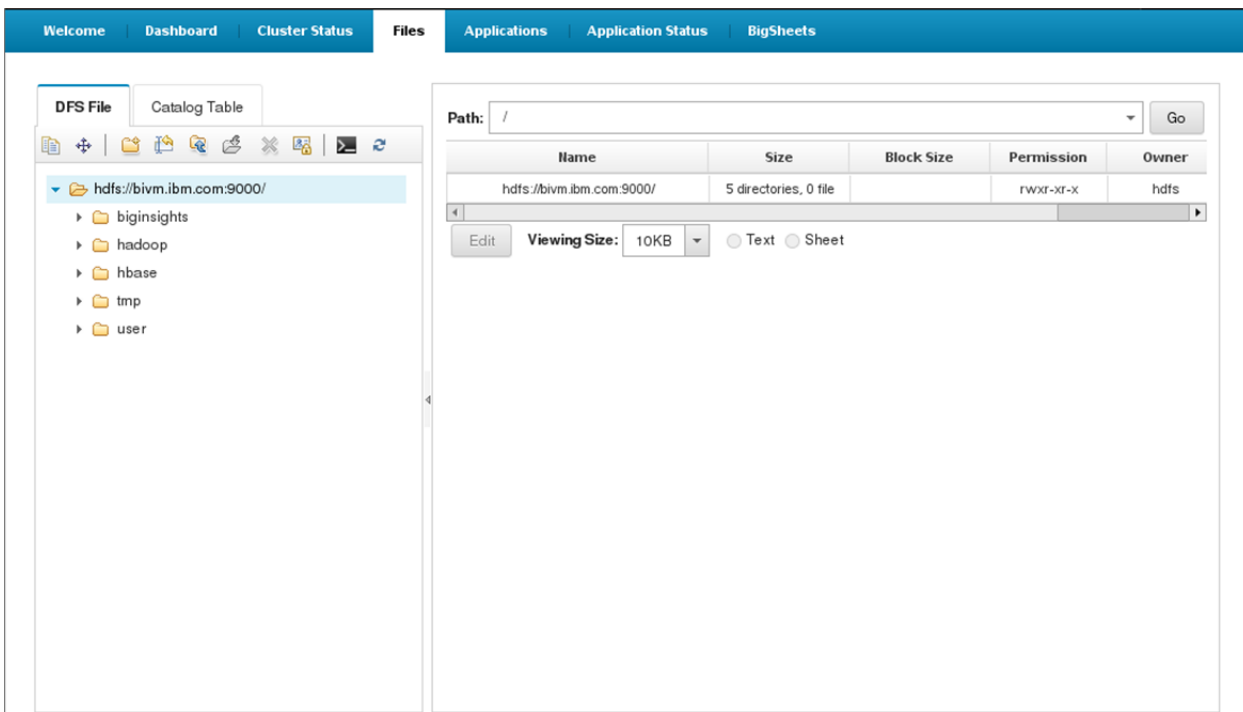
4. Open a new tab in Firefox by clicking on the plus sign tab. Type in a URL of **bivm.ibm.com:50060**. This is the web interface that comes with Hadoop.

The screenshot shows the Hadoop Task Tracker Status web interface. The URL is `bivm.ibm.com:50060/tasktracker.jsp`. The page title is `tracker_bivm.ibm.com:localhost.localdomain/127.0.0.1:50039 Task Tracker Status`. The Hadoop logo is displayed, along with the version `2.2.0, r81730d16ab61f7da9b11a94a87f8b8a08964ca8c` and the compile date `2014-03-04T00:10Z by jenkins`. The page is divided into sections for 'Running tasks', 'Non-Running Tasks', and 'Tasks from Running Jobs', each with buttons for 'Task Attempts', 'Status', 'Progress', and 'Errors'.

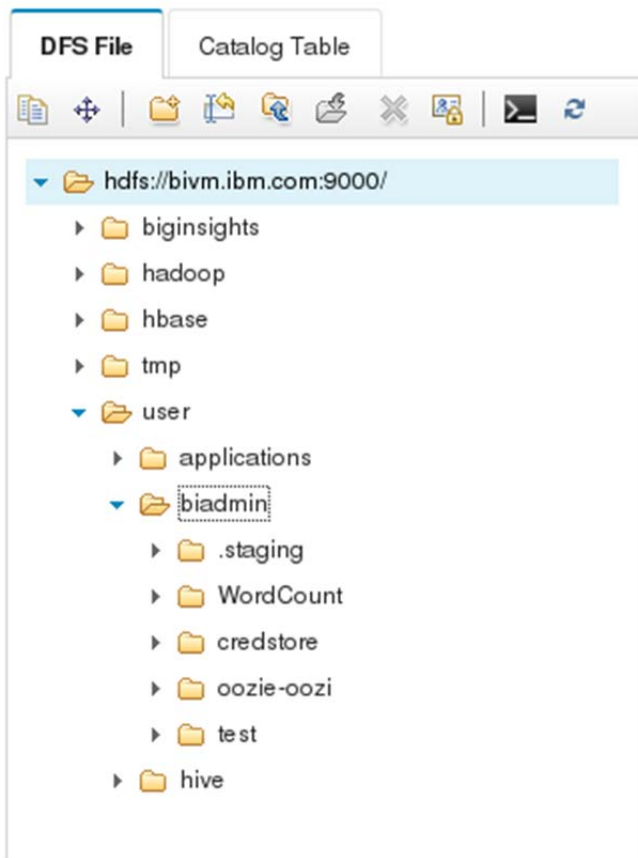
5. In Firefox, close the TaskTracker tab and return to **IBM InfoSphere BigInsights** tab
6. In the BigInsights Console, click the **Files** tab. (I have found on the Quick Start Edition image that there can be a problem having the information on the *Files tab* displayed properly. It appears that if I first display a tab other than the *Files tab* or the *BigSheets tab*, my display looks as follows. If you are experiencing the same problem, my suggestion is to either click the **BigSheets** tab and then the **Files** tab or log out of the console, log back in and select the **Files** tab as the first thing.)



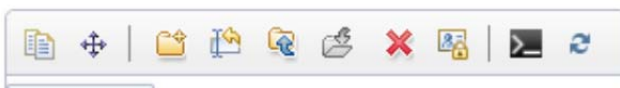
The contents of the *Files* tab should look as follows:



- __7. Expand the *user->biadmin* directories. You should be able to see the *test* directory that you created earlier.



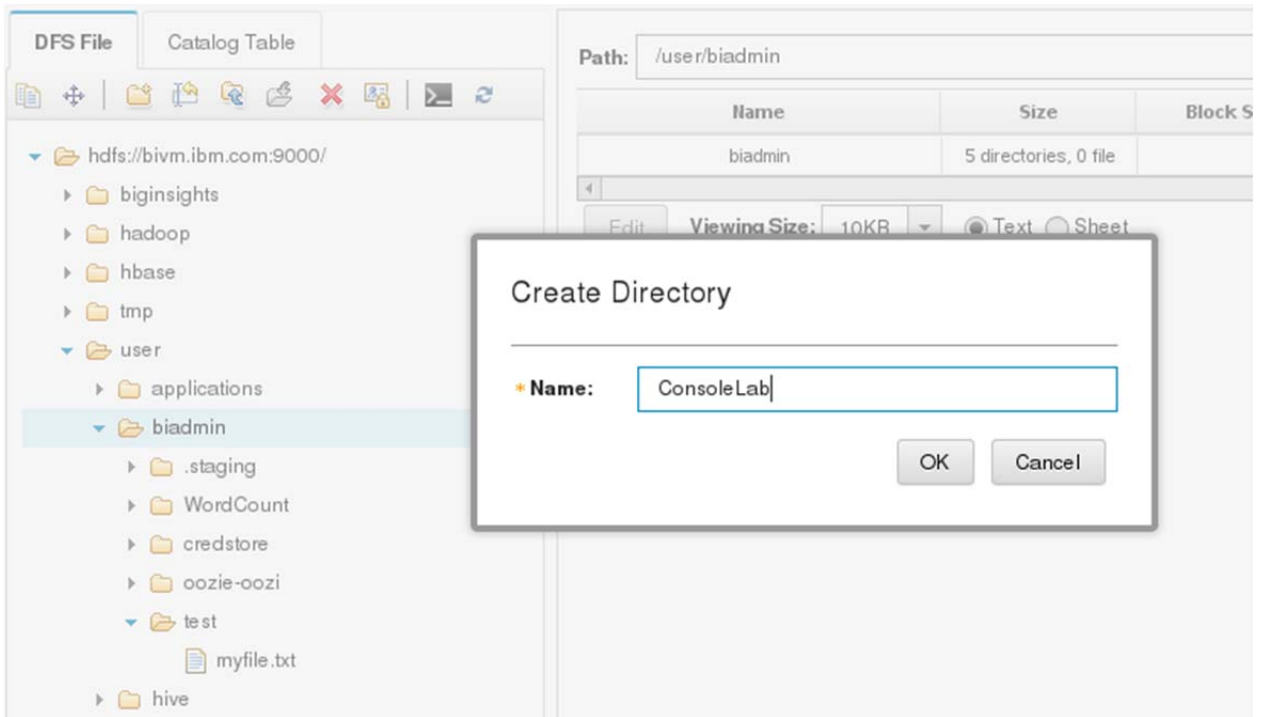
- __8. Expand the *test* directory and select **myfile.txt**. Note that the contents of the file are displayed on the right side. You can even edit it.
- __9. Become familiar with the functions provided by the icons at the top of the pane. Simply move your cursor over each icon and view the popup help.



- __10. Select the **biadmin** directory. Then click the **Create Directory** icon.

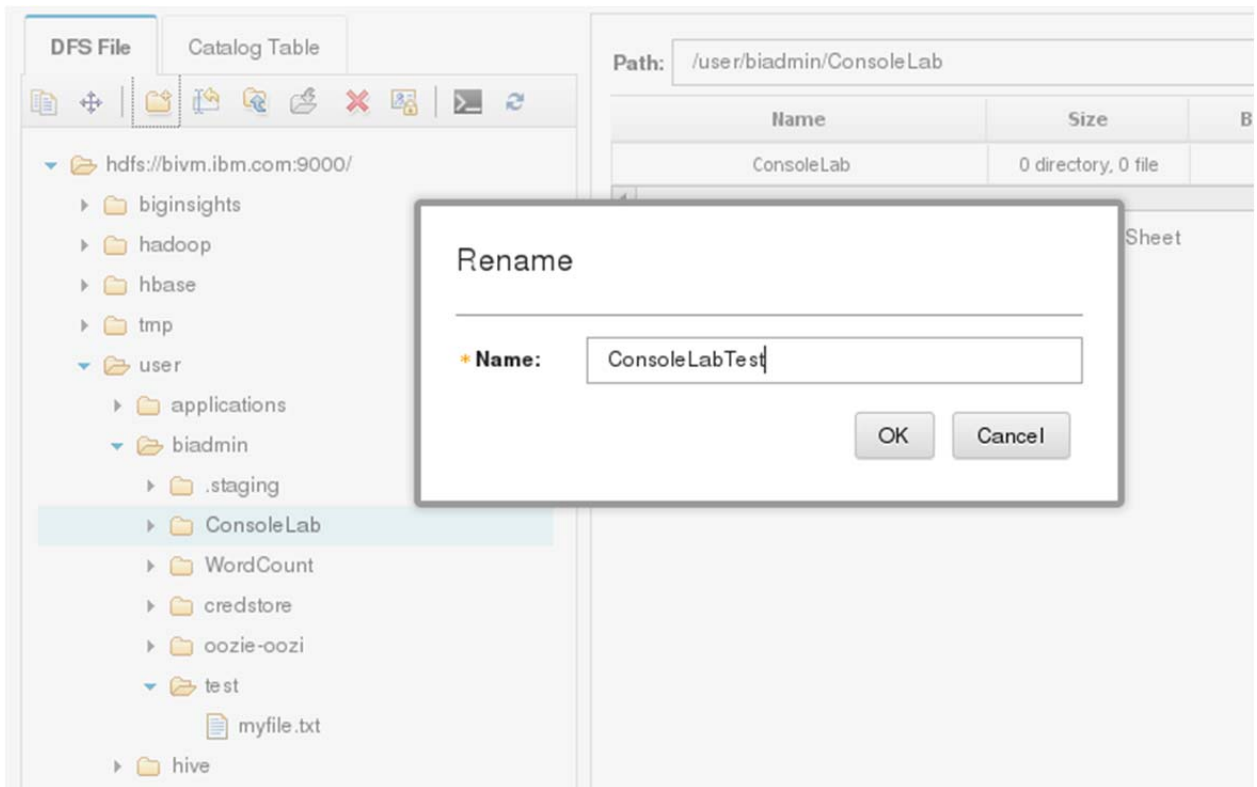


- __11. Type in a name of **ConsoleLab** and click **OK**.



- __12. Select the newly created *ConsoleLab* directory and click the **Rename** icon and change the directory name to **ConsoleLabTest**.

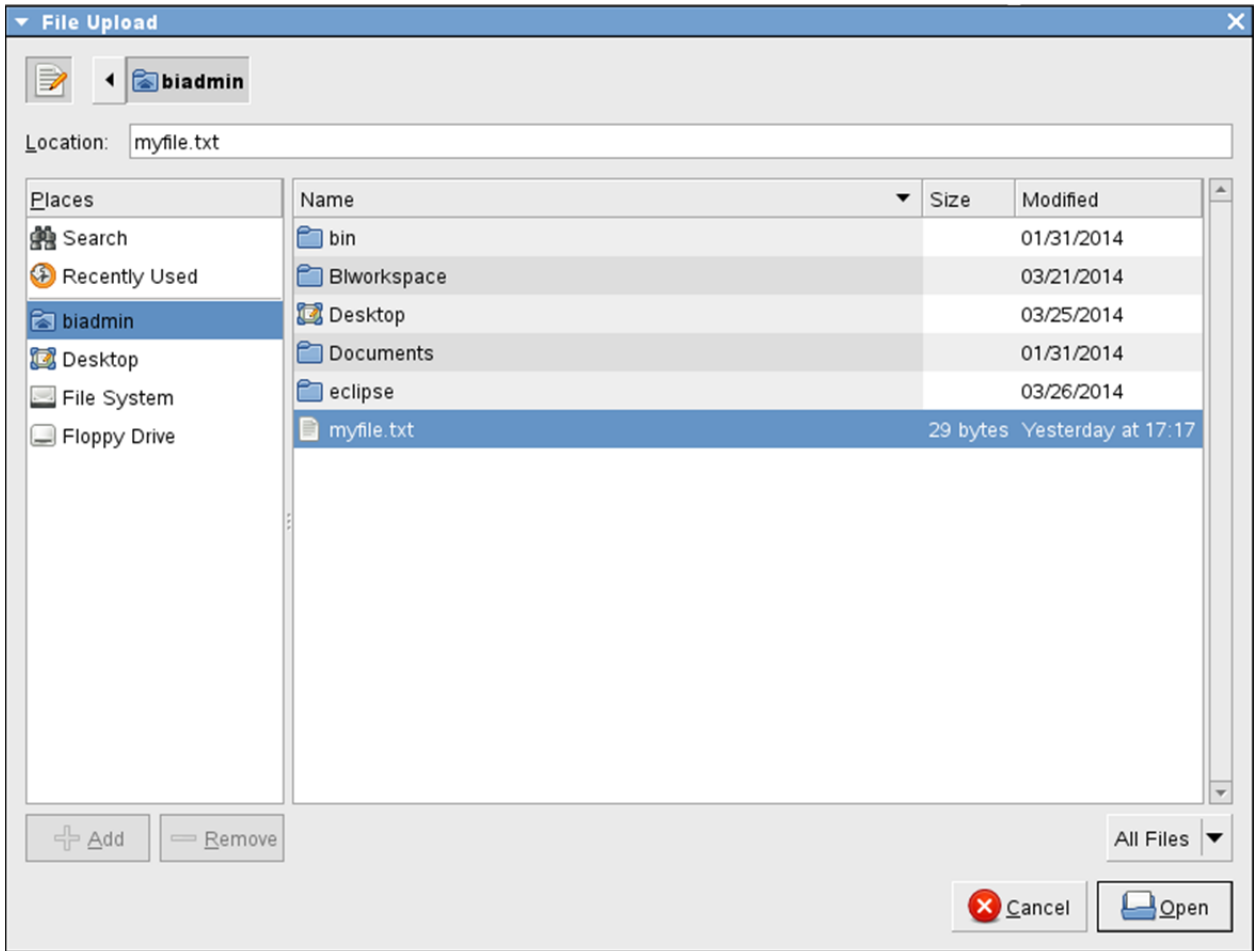




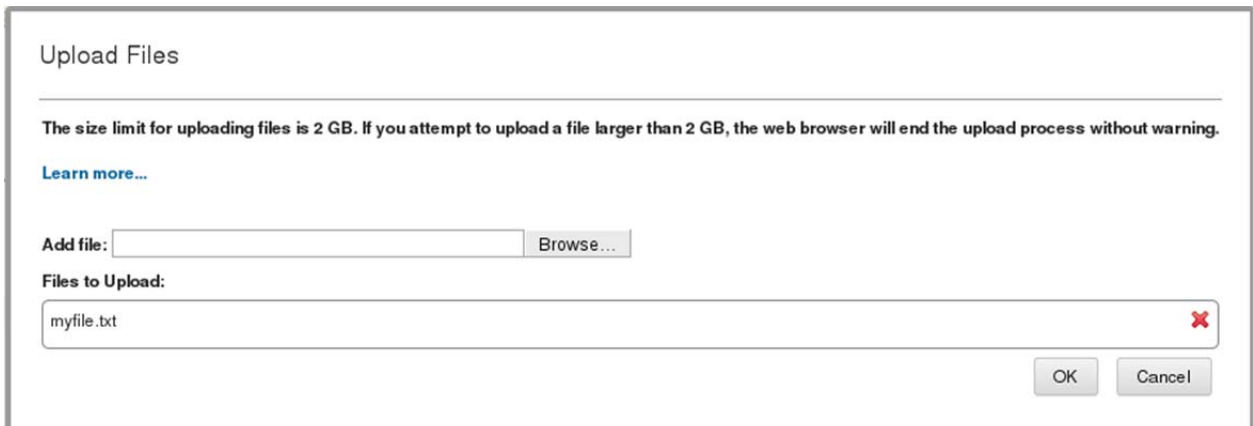
__13. With the *ConsoleLabTest* directory selected, click the **Upload** icon.



__14. Click the **Browse** pushbutton. Navigate to the home directory for *biadmin*. Select *myfile.txt* and click **Open**.



___15. *myfile.txt* has been added to the upload list. You could repeat this process and add additional files to the upload list. Click **OK**.



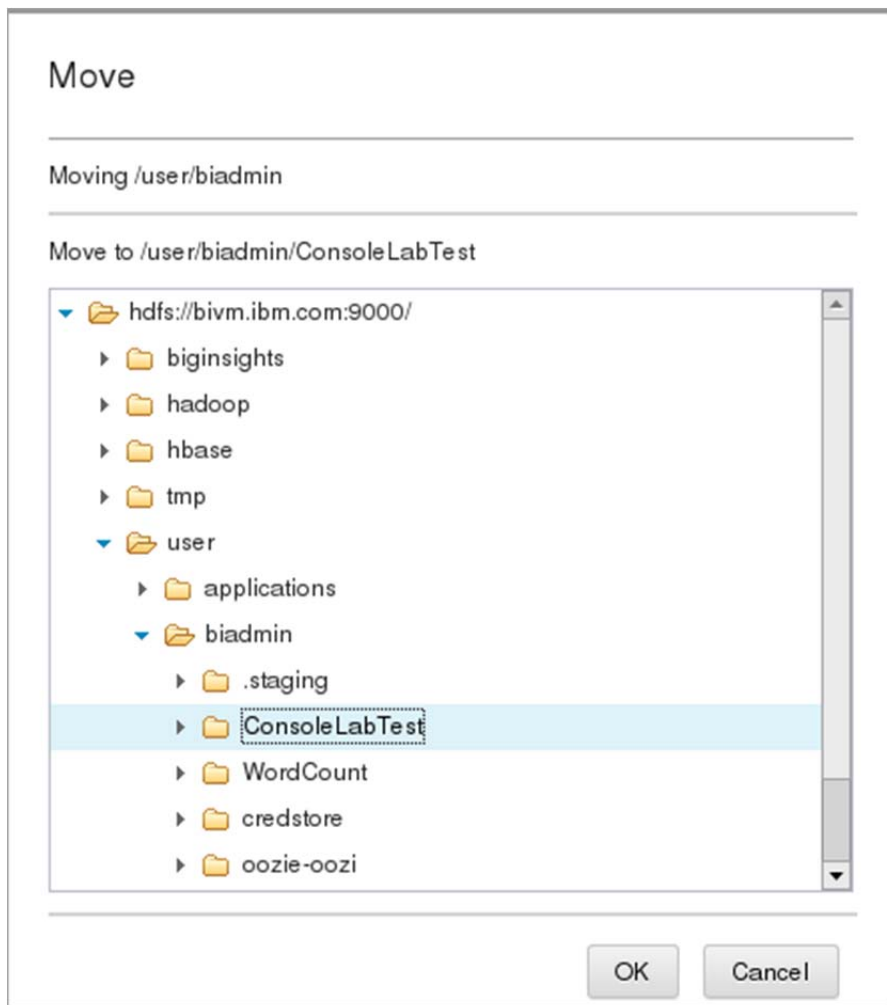
- __16. Select **myfile.txt** in the *ConsoleLabText* directory. Click the **Remove** icon. Click **Yes** on the confirmation dialog.



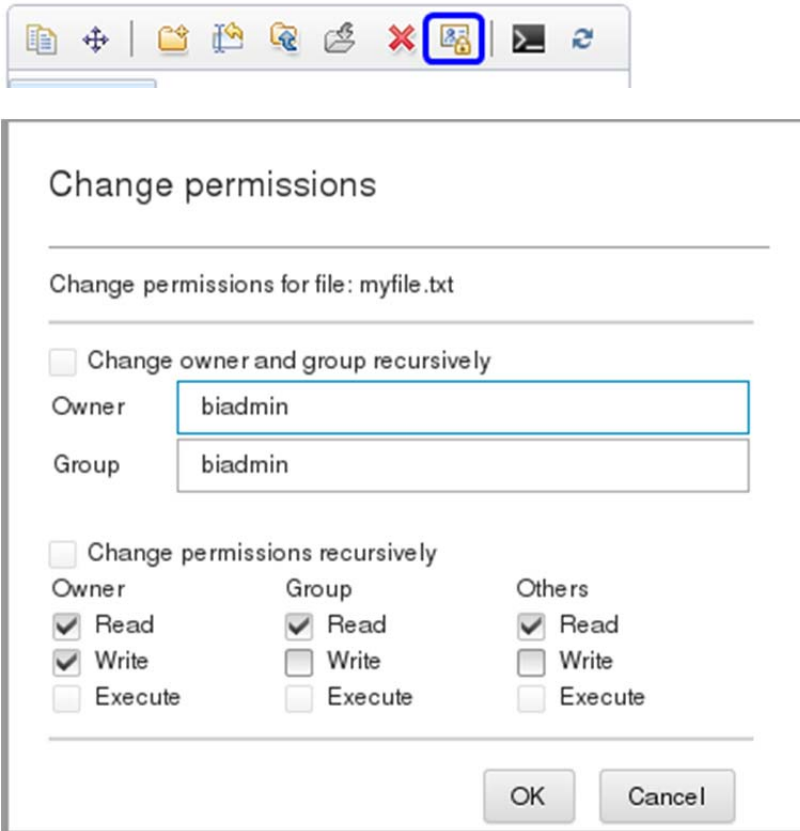
- __17. Select **myfile.txt** in the *test* directory. Then select the **Move** icon.



- __18. Drill down on *user->biadmin* and select **ConsoleLabText**. Click **OK**.



- __19. Select **myfile.txt** in the *ConsoleLabTest* directory and click the **Set permissions** icon. Here you can make changes to ownership and permissions. For this exercise, no changes are required.



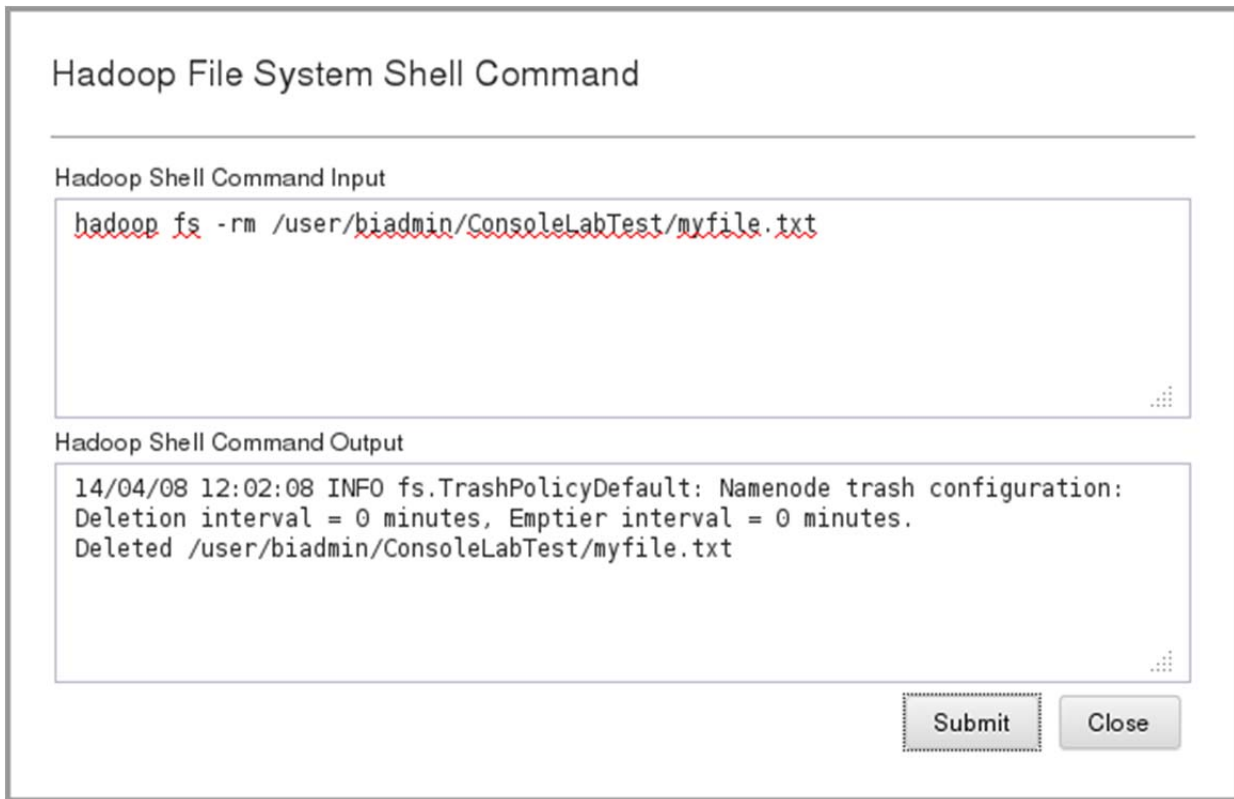
- __20. Click the **Cancel** pushbutton.
- __21. Finally, click the **Distributed File System Shell Commands** icon.



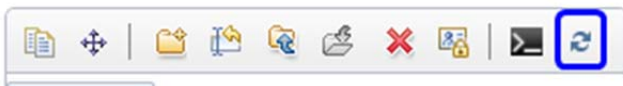
- __22. Type the following command to delete *myfile.txt*.

```
hadoop fs -rm /user/biadmin/ConsoleLabTest/myfile.txt
```

And click the **Submit** pushbutton.



- __23. Close the shell dialog. Then click the **Refresh** icon. You will probably then get a message that your file no longer exists. Close that dialog.



1.4 Summary

Congratulations! You're now familiar with the Hadoop Distributed File System. You now know how to manipulate files within by using the terminal and the BigInsights Console. You may move on to the next unit.



© 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
