Table of Contents

1 Introduction

IBM InfoSphere BigInsights is a software platform consisting of various IBM and open-source tools designed to help companies discover and analyze information within data that is too large and complex to process using traditional tools.
Planning the layout of the installation and ensuring there are enough resources for all components are essential to proper functionality of the BigInsights software, as well as ease of future administration.

IBM's InfoSphere BigInsights Enterprise Edition enables firms to store, process, and analyze large volumes of various types of data. In this lab, you'll see how you can work with its Web console to administer your system, launch jobs (applications), monitor the status of jobs, and perform other functions. For further details on the Web console or BigInsights, visit the Info Center at s.

2 About this Lab

In this lab, we will explore various tools within the BigInsights Web Console, which help to administer and work with a BigInsights cluster.

3 Objective

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

1. Launch the BigInsights Web console (Web console).
2. Work with popular resources accessible through the Welcome page.
3. Administer BigInsights by inspecting the status of your cluster, starting and stopping components, and accessing administrative tools available for open source components provided with BigInsights.
4. Work with the distributed file system. In particular, you'll explore the HDFS directory structure, create subdirectories, and upload files to HDFS.
5. Launch applications (jobs) and inspect their status. You'll also learn how to view the output of one job in BigSheets, a spreadsheet-like tool.

4 Environment Setup Requirements

To complete this lab you will need the following:

1. IBM InfoSphere BigInsights Bootcamp VMware® image
2. VMware Player 2.x or VMware Workstation 5.x or later

5 Preparing for the lab

1. Start the VMware image by clicking the 

button in VMware Workstation if it is not already on.
2. Log in to the VMware virtual machine using the following information:
   - User: **biadmin**
   - Password: **password**
3. Open a terminal window by right-clicking on the Desktop area and choose the "Open in Terminal" item.

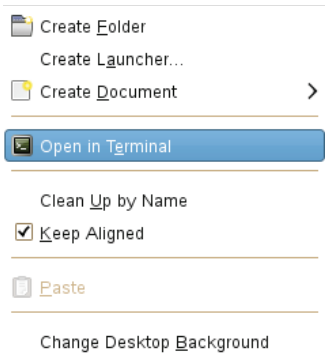


Figure 1

4. Start the BigInsights services by executing the following command:

```
start-all.sh
```

5. Verify all services have started successfully by checking the result of the above start command. There should not be any failed components:

```
[INFO] DeployManager - Start; SUCCEEDED components: [zookeeper, hadoop, derby, h
ive, hbase, bigsql, oozie, orchestrator, console, httpfs, monitoring]; Consumes
: 252804ms
```

6 BigInsights Installation

IBM InfoSphere BigInsights has been pre-installed on the virtual image you have been provided with. The installation has been deployed in a pseudo-distributed mode in a single node configuration.
The installation directory has been kept as default in **/opt/ibm/biginsights** and is easily identifiable via the **$BIGINSIGHTS_HOME** environment variable.

7 Launching the BigInsights Web Console

1. Launching the web console is done by entering a URL into a web browser. The format for the URL is:

   `http://<host>:<port>` **or** `http://<host>:<port>/data/html/index.html`

   The default is:

```
http://localhost:8080/data/html/index.html
```

   For convenience, there is a shortcut on the biadmin's user Desktop, which will launch the web console when double-clicked.
2. Security has been configured with security disabled. So you will not be prompted to enter a user ID and password.
3. Verify that the BigInsights Web Console looks like this:

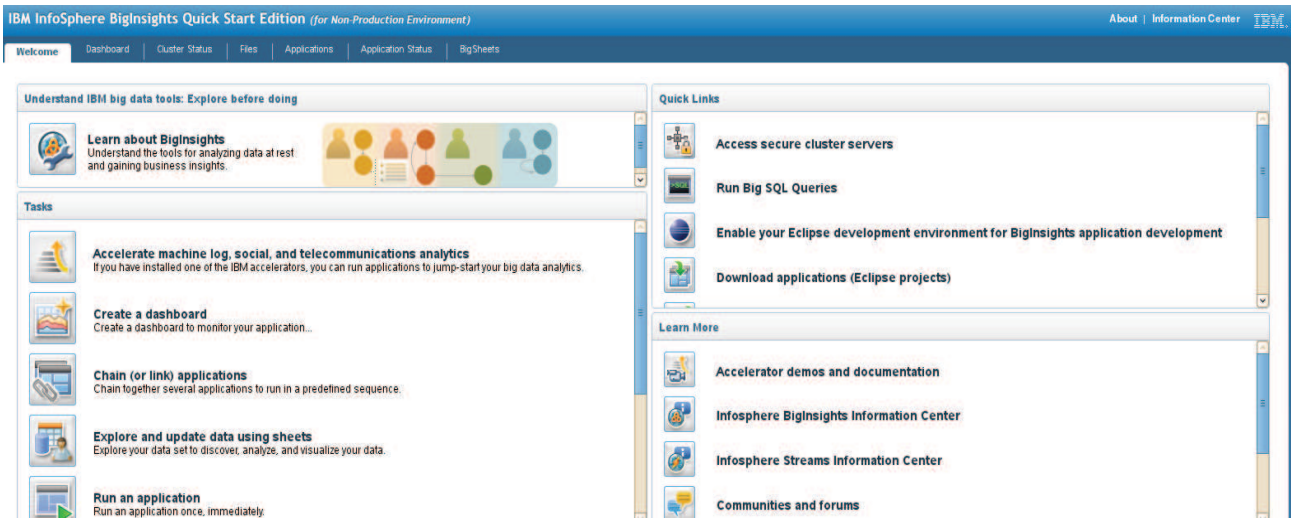


Figure 2 - InfoSphere BigInsights Web Console

8 Working with the Welcome page

This section introduces you to the Web console's main page accessible from the **Welcome** tab. The Welcome page features links to common tasks, many of which can also be launched from other areas of the console. In addition, the Welcome page includes links to popular external resources, such as the BigInsights Information Center (product documentation) and community forum. You'll have a chance to work with several important aspects of this page.

1. Verify that the **Welcome** tab at the upper left corner is active (or highlighted). If necessary, click it to change focus to it.

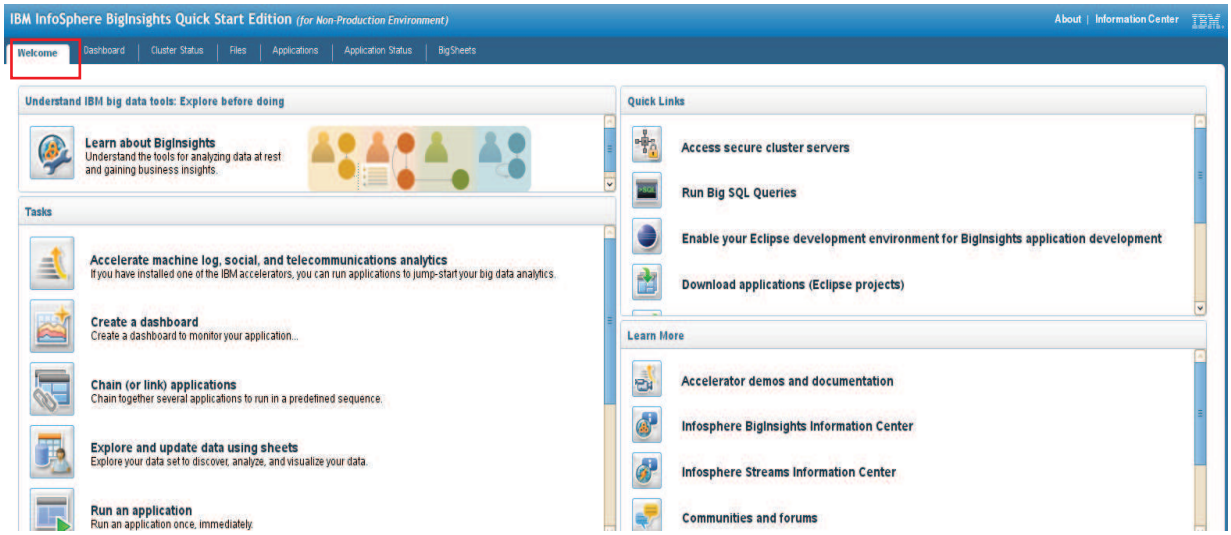


Figure 3

2. Inspect the **Tasks** pane at left and use its vertical scroll bar (if necessary) to become familiar with the various tasks you can launch from this area. Each task button provided here (1) executes a function accessible through one of the Web console's tabs at top and (2) displays context-sensitive help information. Later in this lab, you'll work directly with various Web console functions. For now, you'll only launch one.
3. Click the **View, start, or stop a service** button towards the end of the **Tasks** list.

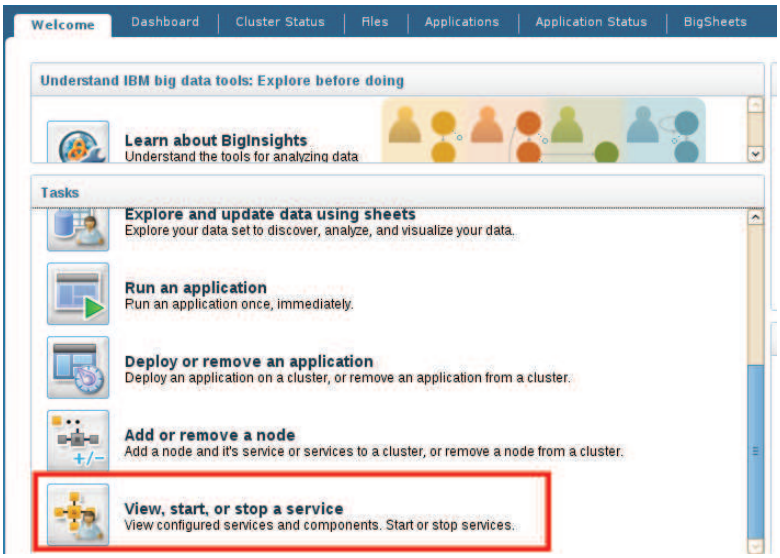


Figure 4

4. Verify that a list of services appears at left.

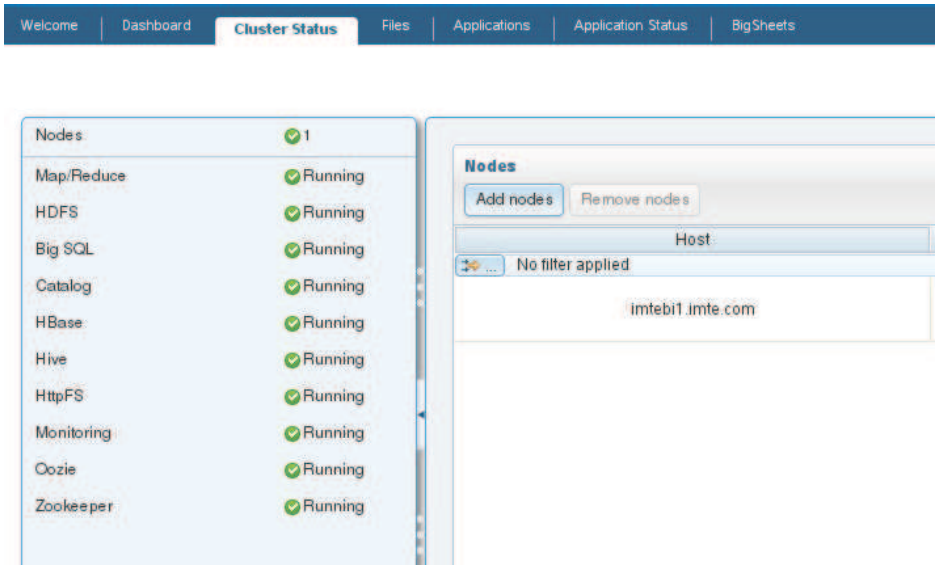


Figure 5

5. Click the **Welcome** tab to return back to the main page.
6. Inspect the **Quick Links** pane at top right and use its vertical scroll bar (if necessary) to become familiar with the various resources accessible through this pane. The first several links simply activate different tabs in the Web console, while subsequent links enable you to perform set-up functions, such as adding BigInsights plug-ins to your Eclipse development environment.

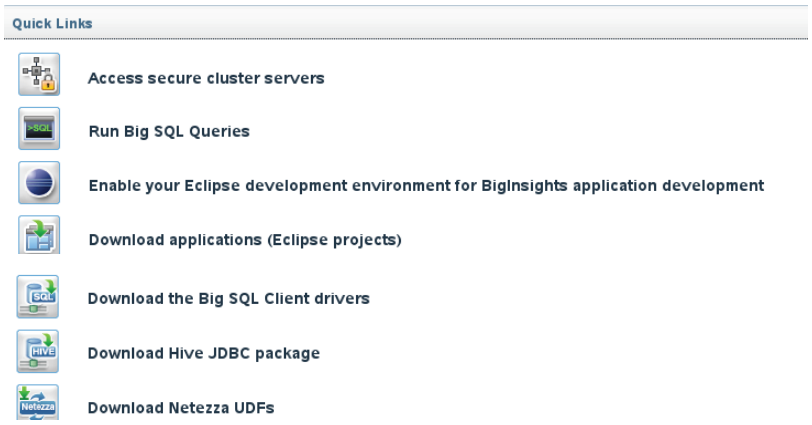


Figure 6

7. Verify that the **Files** tab of the Web console is active and that information about the distributed file system is shown as illustrated below.

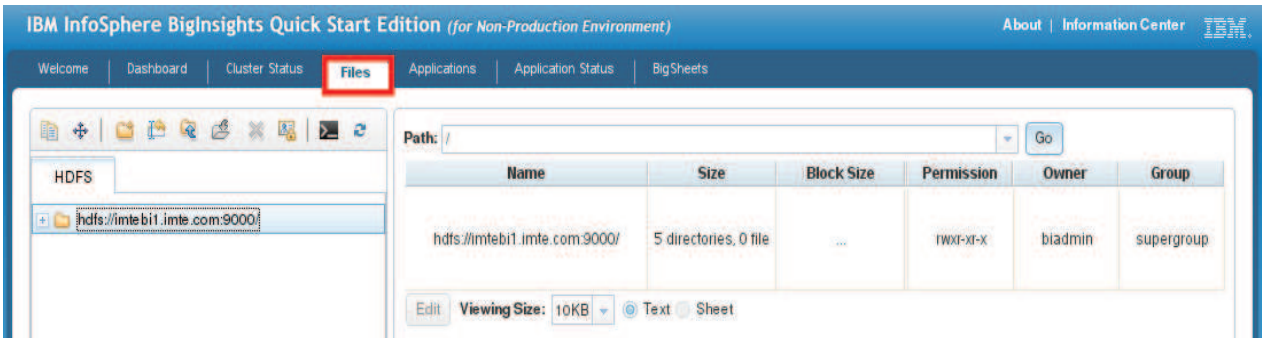


Figure 7

8. Click the **Welcome** tab to return back to the main page.
9. Inspect the **Learn more** pane at lower right. Links in this area access external Web resources that you may find useful, such as the BigInsights Information Center, a public discussion forum, IBM support, and IBM's BigInsights product site. If desired, click one or more of these links to see what's available to you.

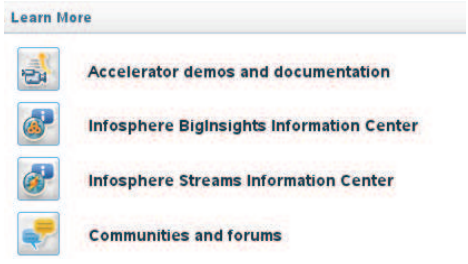


Figure 8

9 Administering InfoSphere BigInsights

The Web console allows administrators to inspect the overall health of the system as well as perform basic functions, such as starting and stopping specific servers (or components), adding nodes to your cluster, and so on. You'll explore a subset of these capabilities here.

9.1 Inspecting the status of your cluster

1. Click the **Cluster Status** tab at the top of the page.

Figure 9

2. Inspect the overall status of your cluster. The figure below was taken on a single-node cluster that had several services running.

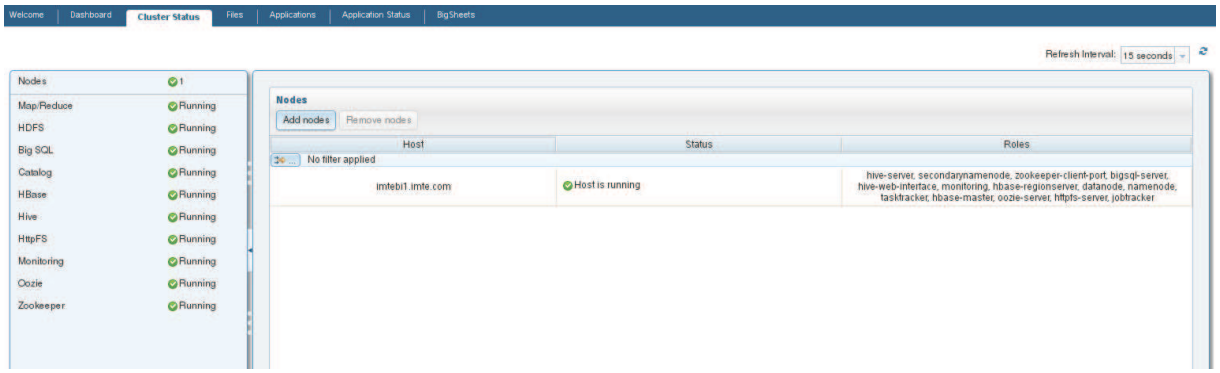


Figure 10

3. Click the Hive service and note the detailed information provided for this service in the pane at right. For example, you can see the URL for Hive's Web interface and its process ID.

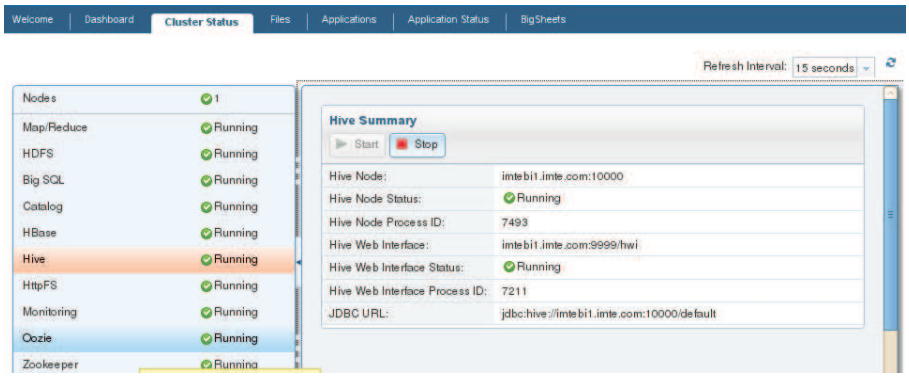


Figure 11

4. Optionally, cut-and-paste the Hive URL into a new tab of your browser. You'll see an open source tool provided with Hive for administration purposes, as shown below. Close this tab and return to the **Cluster Status** section of the BigInsights Web console.

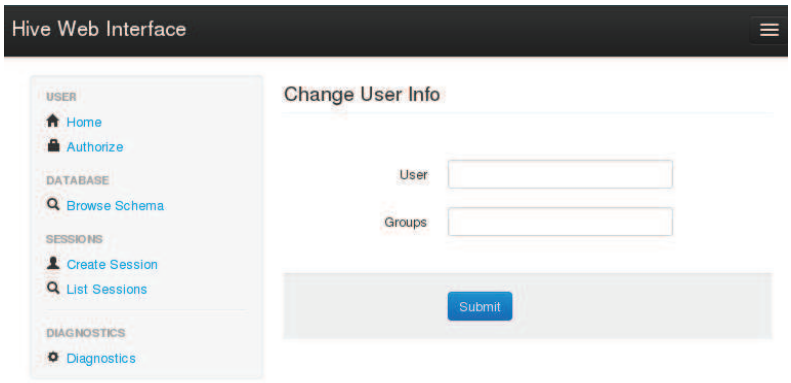


Figure 12

9.2 Starting and Stopping a Component

The web console provides means for toggling the status of a component directly through the browser.

Alternatively, the command line can be used as well by executing `start.sh or stop.sh` followed by the component name via the Administration user, i.e. biadmin. For example, to stop Hive via the command-line, execute:

```
stop.sh hive
```

```
biadmin@imtebi1:~> stop.sh hive
[INFO] Progress - Stop hive
[INFO] @imtebi1.imte.com - hive-web-interface stopped
[INFO] @imtebi1.imte.com - hive-server stopped
[INFO] Progress - 100%
[INFO] DeployManager - Stop; SUCCEEDED components: [hive]; FAILED components: []
biadmin@imtebi1:~>
```

To start it once again, execute:

```
start.sh hive
```

```
biadmin@imtebi1:~> start.sh hive
[INFO] Progress - Start hive
[INFO] @imtebi1.imte.com - derby already running, pid 2377
[INFO] Progress - 10%
[INFO] @imtebi1.imte.com - hive-web-interface started, pid 5380
[INFO] @imtebi1.imte.com - hive-server started, pid 5652
[INFO] Progress - 100%
[INFO] DeployManager - Start; SUCCEEDED components: [hive]; FAILED components: [
]
biadmin@imtebi1:~>
```

To perform the task via the web console, perform the following steps:

1. If necessary, click the Hive service to display its status. 

2. In the pane at right (which displays the Hive status), click the red Stop button to stop the service.

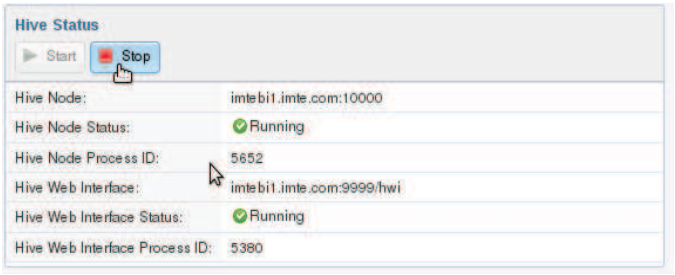


Figure 13

3. When prompted to confirm that you want to stop the Hive service, click **OK** and wait for the operation to complete. The right pane should appear similar to the following image. If nothing seems to happen after a few minutes, refresh your browser.

Figure 14

4. Restart the Hive service by clicking the green arrow just beneath the Hive Status heading. (See figure in Step #2.) When the operation completes, the Web console will indicate that Hive is running again, likely under a process ID that differs from the earlier Hive process ID shown at the beginning of this lab module.

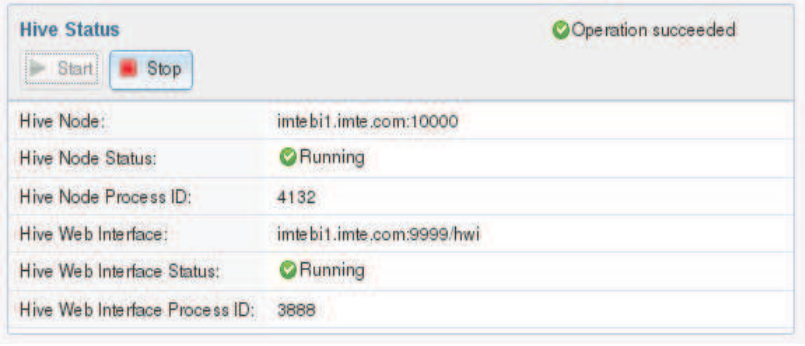


Figure 15

9.3 Accessing Service-specific Web Tools

1. Click on the **Welcome** tab.
2. Click the **Access secure cluster servers** button in the **Quick Links** section at right. A pop-up window will appear. Make sure your browser is not blocking pop-ups from this server.

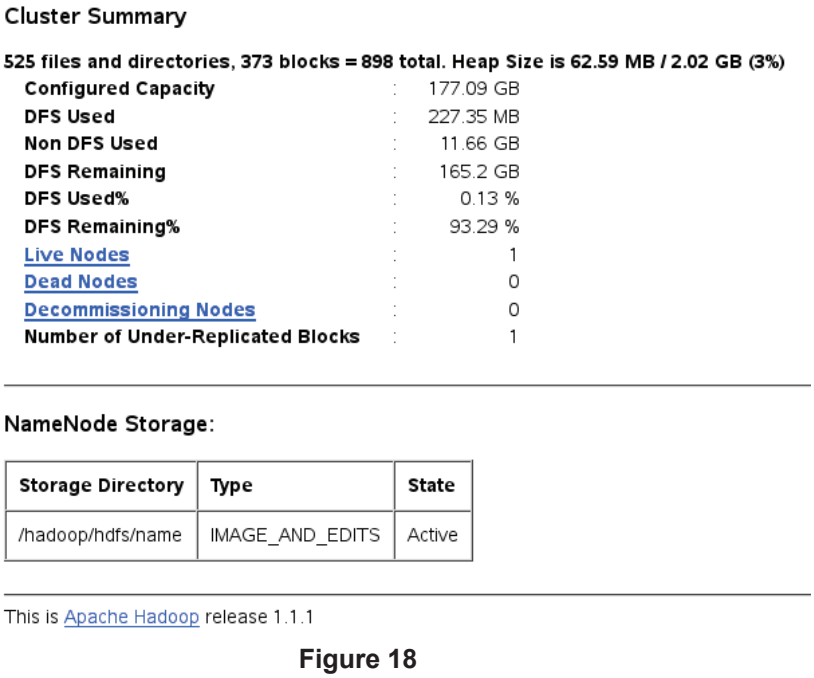
Figure 16

3. Inspect the list of server components for which there are additional Web-based tools. The BigInsights console displays the URLs you can use to access each of these Web sites directly.

URL	Alias
http://imtebi1.imte.com:9999/hwi	hive
http://imtebi1.imte.com:50030	jobtracker
http://imtebi1.imte.com:50060	tasktracker
http://imtebi1.imte.com:50070	namenode
http://imtebi1.imte.com:50075	datanode
http://imtebi1.imte.com:50090	secondarynamenode
http://imtebi1.imte.com:60010	hbase-master
http://imtebi1.imte.com:60030	hbase-regionserver

Figure 17

4. Click the **namenode** alias. Your browser will display the standard NameNode Web interface provided by Apache Hadoop.



Cluster Summary

525 files and directories, 373 blocks = 898 total. Heap Size is 62.59 MB / 2.02 GB (3%)

Configured Capacity	:	177.09 GB
DFS Used	:	227.35 MB
Non DFS Used	:	11.66 GB
DFS Remaining	:	165.2 GB
DFS Used%	:	0.13 %
DFS Remaining%	:	93.29 %
Live Nodes	:	1
Dead Nodes	:	0
Decommissioning Nodes	:	0
Number of Under-Replicated Blocks	:	1

NameNode Storage:

Storage Directory	Type	State
/hadoop/hdfs/name	IMAGE_AND_EDITS	Active

This is Apache Hadoop release 1.1.1

Figure 18

5. If desired, return to the list of server components (shown in Step 3) and explore additional Web-based interfaces provided for these open source software components.

9.4 Working with Files

The **Files** tab of the console enables you to explore the contents of your file system, create new subdirectories, upload small files for test purposes, and perform other file-related functions. In this module, you'll learn how to perform such tasks against the Hadoop Distributed File System (HDFS) of BigInsights.
Examples in this section use the **biadmin** user which has a **/user/biadmin** directory in its distributed file system. If you're accessing a BigInsights cluster using a different user ID, adapt the instructions in this exercise to work with your home directory in HDFS.

1. Click the **Files** tab of the console to begin exploring the distributed file system.

Figure 19

2. Expand the directory tree shown in the pane at left. If files have been uploaded to HDFS before, you'll be able to navigate through the directory to locate them. In our case, the figure below shows **/user/biadmin** containing default subdirectories.

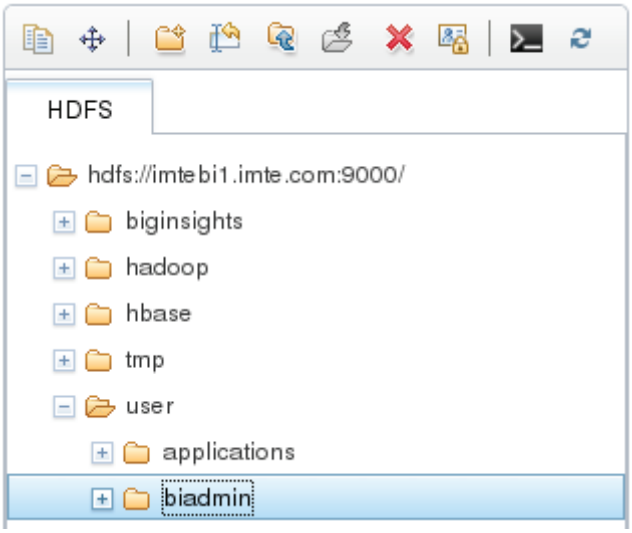


Figure 20

3. Become familiar with the functions provided through the icons at the top of this pane, as we'll refer to some of these in subsequent sections of this module. Simply hover your cursor over the icon to learn its function.

From left to right, the icons enable you to create a directory, upload a file to HDFS, download a file from HDFS to your local file system, delete a file from HDFS, open a command window to launch HDFS shell commands, and refresh the Web console page.

Figure 21

4. Position your cursor on the **user/biadmin** directory and click the **Create Directory** icon (at far left) to create a subdirectory for test purposes.

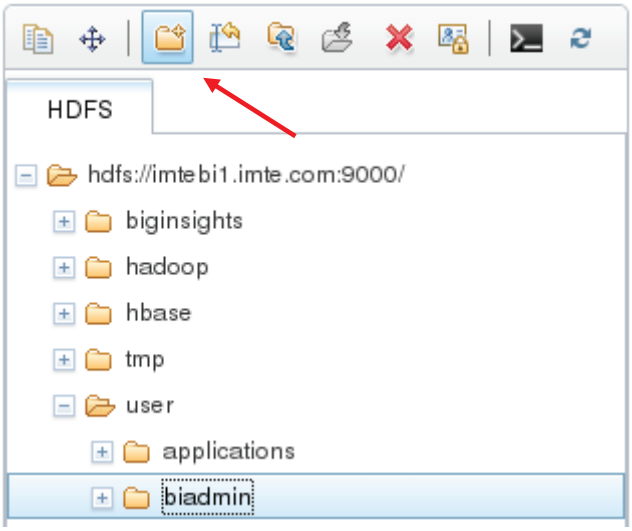


Figure 22

5. When a pop-up window appears prompting you for a directory name, enter **MyDirectory** and click **OK.**

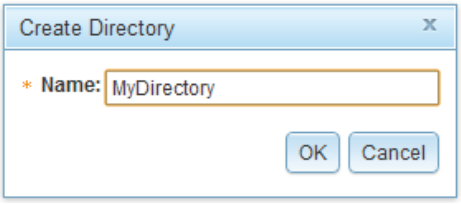


Figure 23

6. Expand the directory hierarchy to verify that your new subdirectory was created.

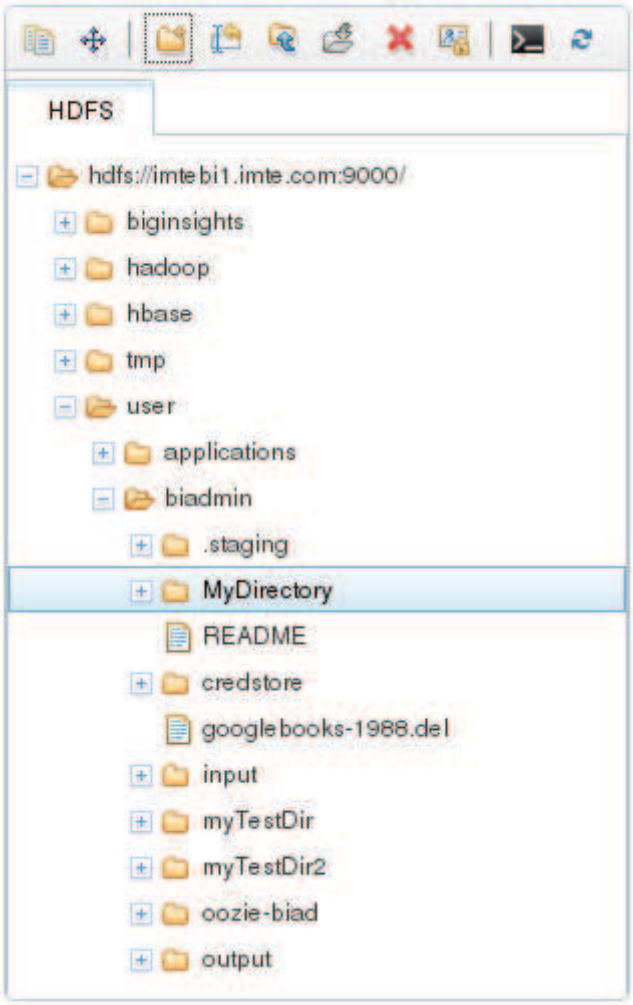


Figure 24

7. Let's copy a sample file from the local file system to HDFS using the command line. Open a terminal window and issue the command:

```
hadoop fs -put /home/biadmin/bootcamp/input/lab02_WebConsole/CHANGES.txt
MyDirectory/
```

8. Now go back to the Web Console and verify that CHANGES.txt file appears in the directory tree at left. Select the file. On the right, you should see a subset of the file's contents displayed in text format. You might have to refresh the view.

   Do so by pressing 

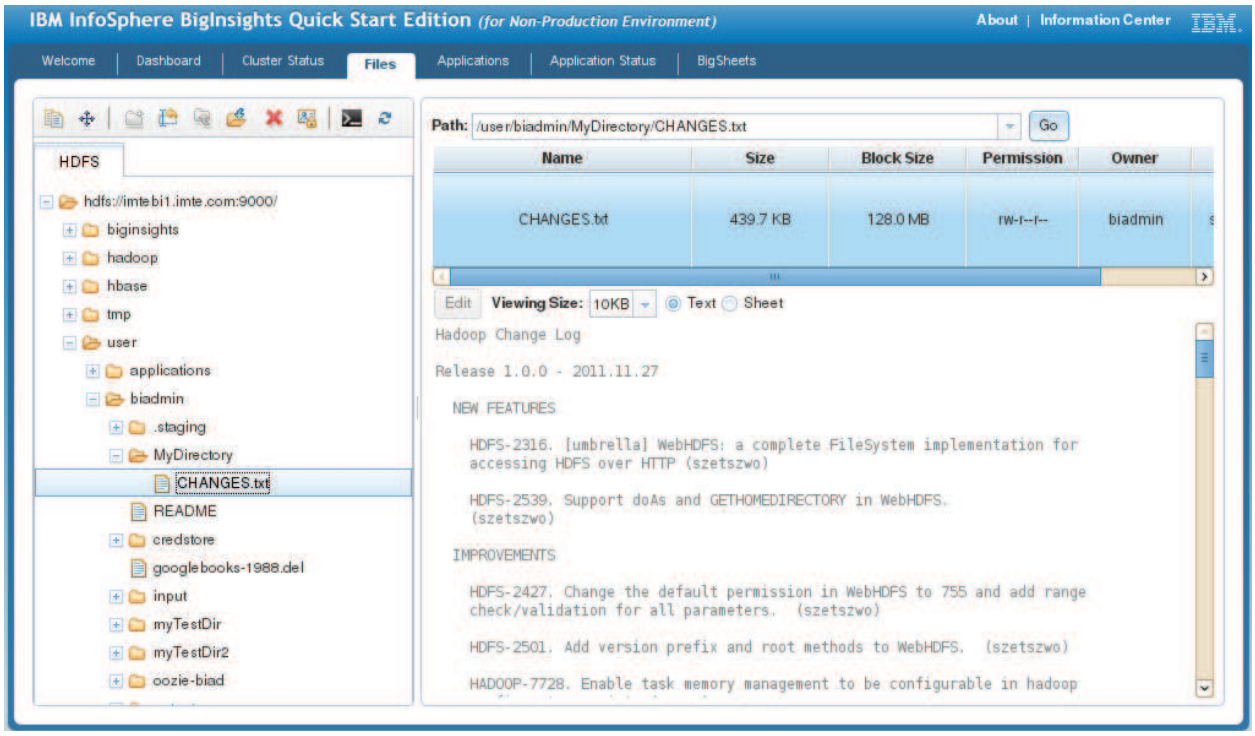


Figure 25

9. Let's now download the same file from HDFS to your local file system. Highlight the ***CHANGES.txt*** file in your ***MyDirectory*** directory and click the **Download** button. (This button is between the **Upload** and **Delete** buttons.)

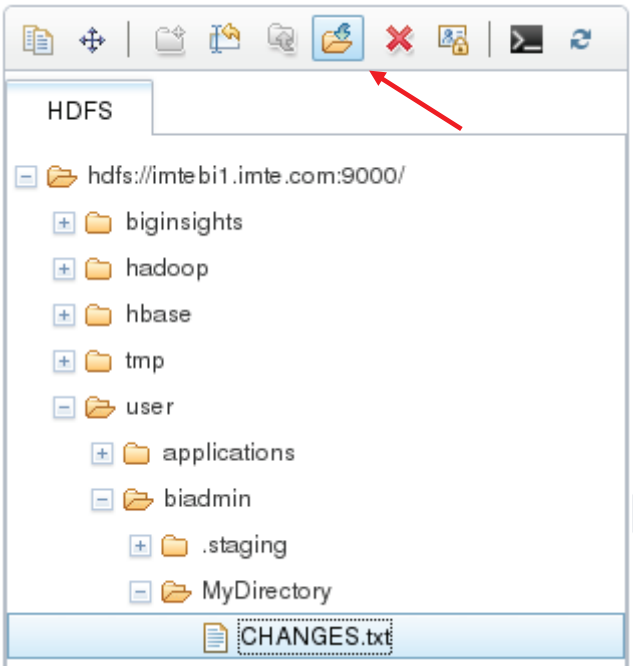


Figure 26

10. When prompted, click the **Save File** button. Then select Save.

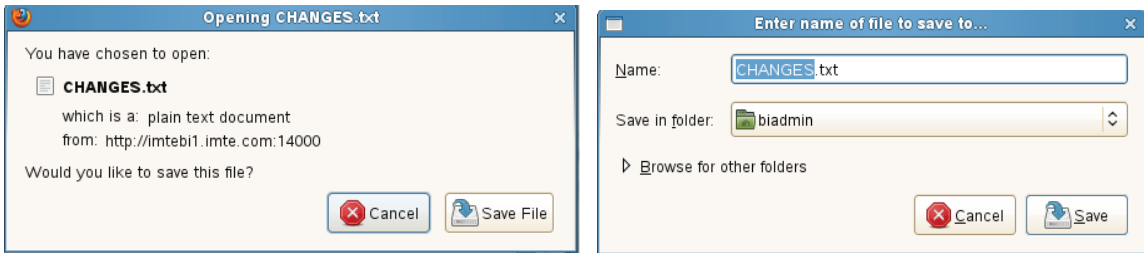


Figure 27

When prompted, identify the folder (directory) of your local file system where the file should be stored and click Save. Wait until the Web browser indicates that the file download is complete.

10 Working with Jobs or Applications

The **Applications** and **Applications Status** tabs of the console enable you to launch deployed applications (including sample applications provided by IBM), inspect the status of applications and workflows, and review execution details. In this module, you'll use the data you uploaded in the prior section as input to a sample job. The job will invoke the WordCount application, a simple open source MapReduce application for Hadoop-based environments. WordCount reads text-based documents from a user-specified input directory and counts how often words occur. So, if your input directory contained 100 documents that each included the word "IBM" 5 times and "BigInsights" 3 times, WordCount would return a count of 500 for IBM and 300 for BigInsights. (Further details about the sample WordCount application are available from the Hadoop wiki at http://wiki.apache.org/hadoop/WordCount.)
After executing the WordCount job, you'll have the option to visualize its output as a collection displayed through the BigSheets tab. (BigInsights provides a spreadsheet-like analysis tool called "BigSheets". While details about this tool are beyond the scope of this Web console lab, you'll have a chance to launch it here and create a simple collection)

10.1 Preparing to launch a sample application (job)

1. Click the **Applications** tab, then click the **Manage** link on the left side. An application catalog is displayed in the pane at left. In the image below, the catalog contains a collection of sample applications provided with BigInsights. (Administrators can upload and "publish" in-house or third-party applications as desired, as well as remove sample applications.)

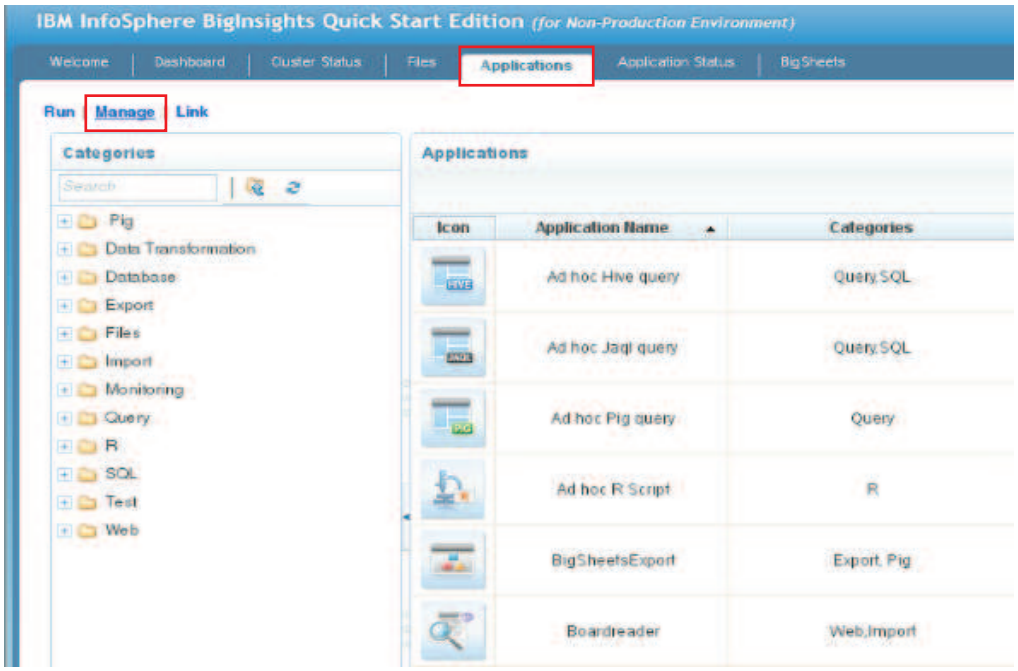


Figure 28

2. Inspect the status of each displayed application icon. An icon showing 'Not deployed' in the status column indicates that the application isn't ready for immediate use because it hasn't been deployed on the cluster. An icon with the status '**Deployed'** has been deployed and is ready for use. When you first install BigInsights, you'll need to deploy any sample applications you wish to use. Scroll down and click the **WordCount** application. Click the **Deploy** button, and **Deploy** again when the popup appears.

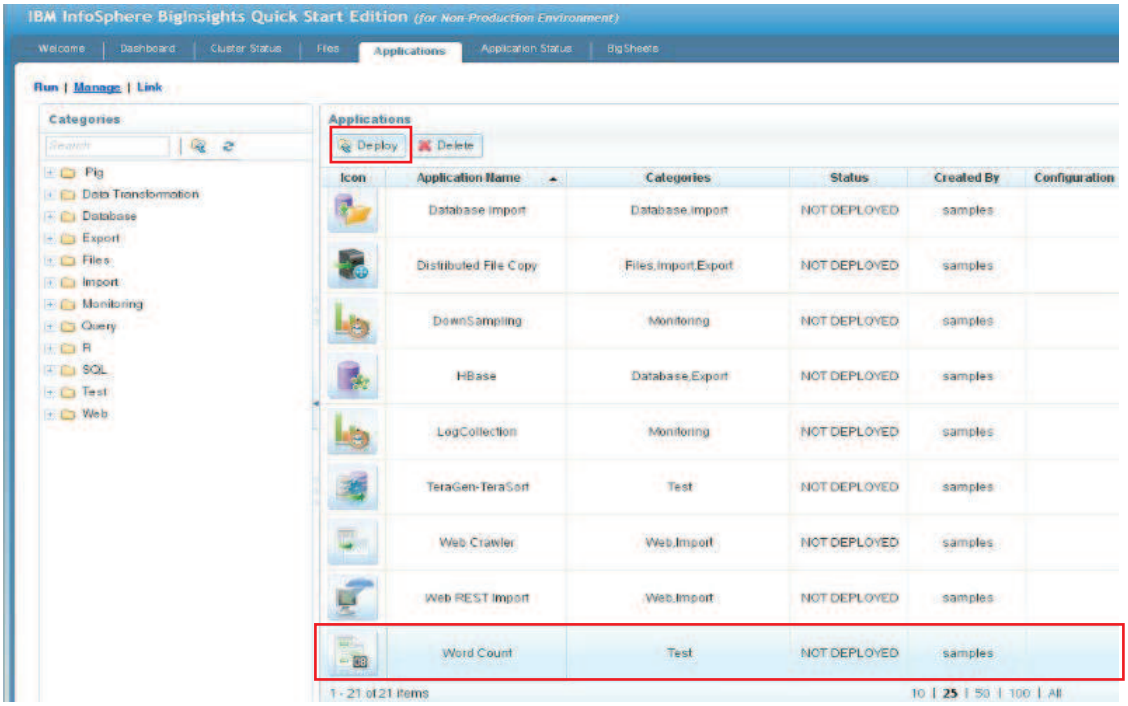


Figure 29

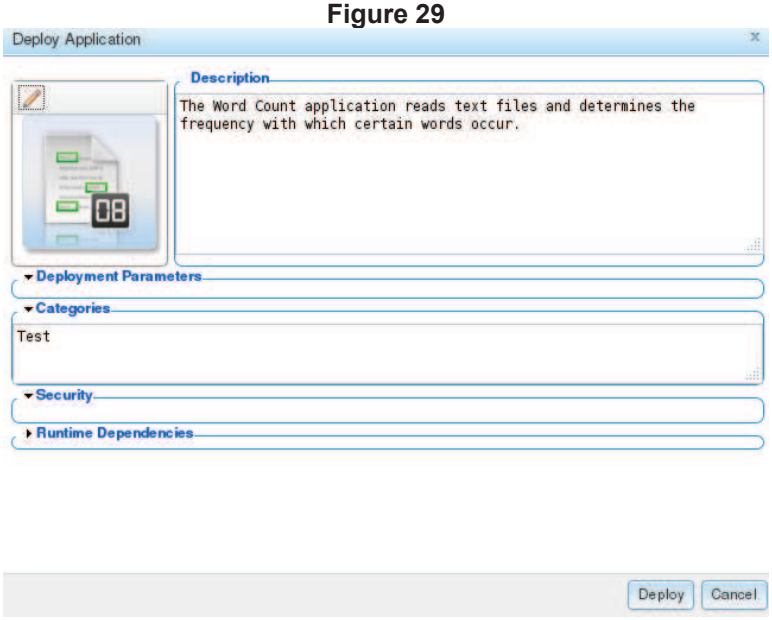


Figure 30

3. Click the **Configure** button  in the same row as the **WordCount** application.

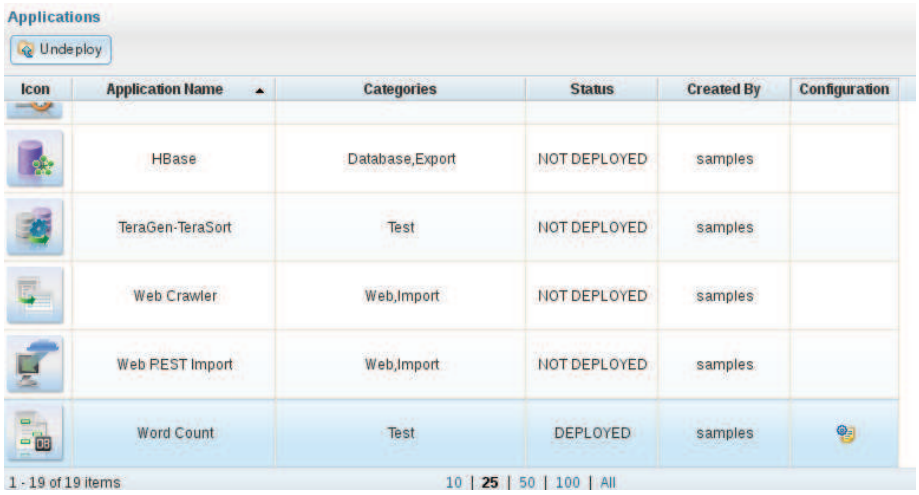


Figure 31

4. In the pop-up window, since security has been disabled in this BigInsights installation, the Security section cannot be expanded. Had security been enabled, this section would show check boxes for both the **supergroup** and **users** boxes. When checked, once the application has been deployed, this will cause the application's icon to be visible to members of either group when they visit the Applications page of the Web console. Furthermore, members of either group will be able to launch the application. Since you're working on a BigInsights cluster that was configured without security, anyone who accesses the Web console will be able to see and launch any deployed application. Close the window by pressing Cancel.

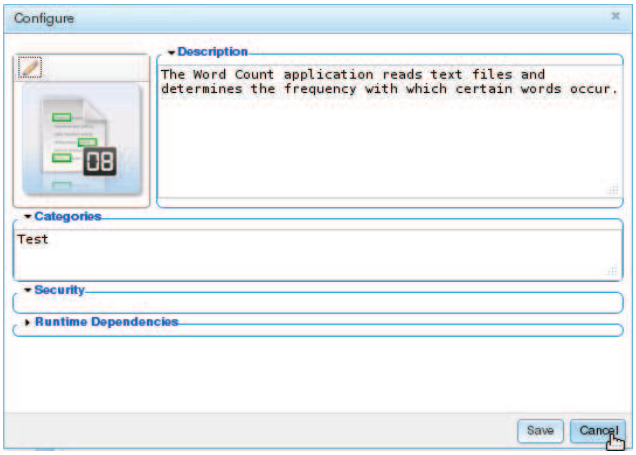


Figure 32

5. Click the **Run** link, then select the **WordCount** application. Notice that your **WordCount** application has been deployed, and any applications deployed will show up in this area.

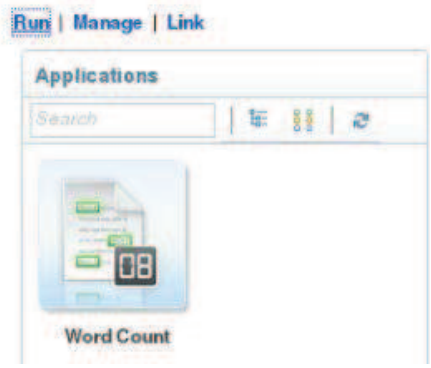


Figure 33

6. The Web console displays information about this application in the pane at right.

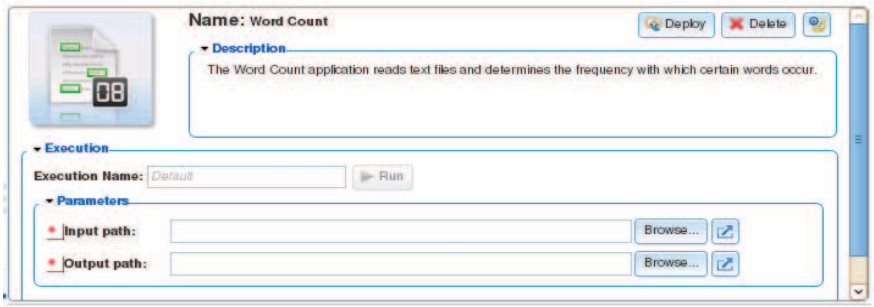


Figure 34

10.2 Launching the WordCount sample application

1. In the **Execution Name** box, enter **ConsoleLabTest**. **Do NOT** click the **Run** button yet.

Figure 35

2. Click on **Browse** button at the end of the input directory field. When a pop-up window appears, expand the HDFS directory tree to locate the **/user/biadmin/MyDirectory** directory. (As you'll recall, you created this directory in a previous section of this lab and uploaded a sample text file to it) Highlight the directory and click **OK**.

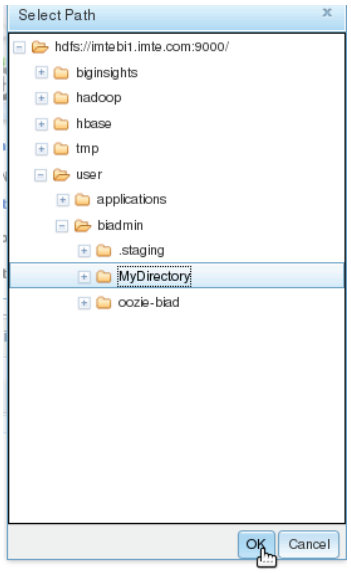


Figure 36

3. Verify that the **Input path** parameter is set to your target directory **(/user/biadmin/MyDirectory** in the figure below).

Figure 37

4. For the **Output path**, specify **/user/biadmin/MyDirectory_WC**
5. Verify that your WordCount settings are consistent with those shown below, and press the green **Run** button. This will cause a simple Oozie-based workflow to be generated for your application, and the application will begin to run.

Figure 38

6. While the application is running, monitor its progress in the bottom right pane (titled **Application History**). Note that you can stop the application by click the red Stop button in this pane. However, for this exercise, allow the job to run to completion.

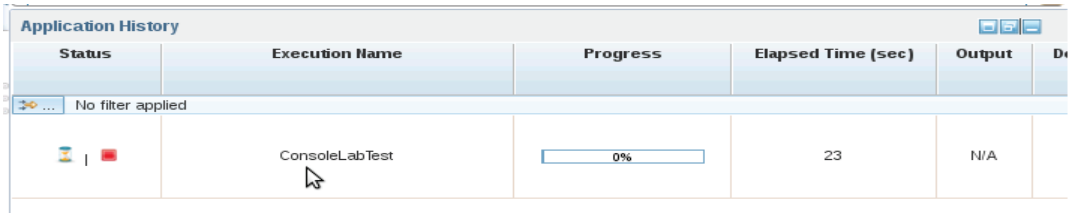


Figure 39

7. Depending on your machine resources, this may take a few minutes. While it is running, you can also view the Application Status by selecting the tab from the top of the screen, if desired.

10.3 Inspecting the Output of your Application

Many applications, including WordCount, generate output when successfully executed.

1. In the **Application History** pane (where you monitored the execution status of your **ConsoleLabTest** application), click the file icon  in the **Output** column of your job. Your console will switch to the **Files** view and display the directory you specified as the output path for your application (**/user/biadmin/MyDirectory_WC** in our example).

2. Expand the contents of this directory until you reach the output file, which is named **part-r-00000**. Click this file, and the right pane will display a portion of its contents in text format.

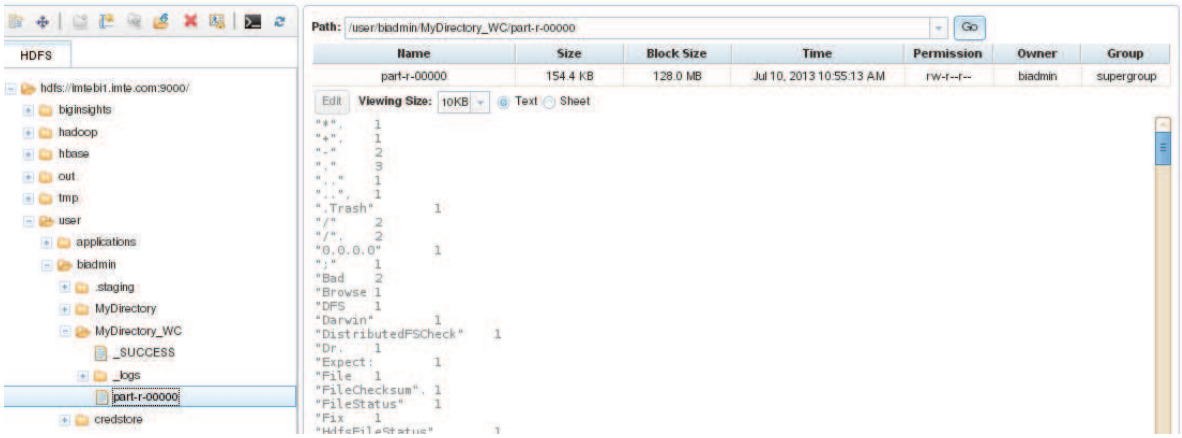


Figure 40

3. Because WordCount generates output in a tab-delimited format, you can easily view this data using BigInsight's spreadsheet-like tool.
4. In the right pane, change the reset the display style from **Text** to **Sheet** by clicking the **Sheet** button.

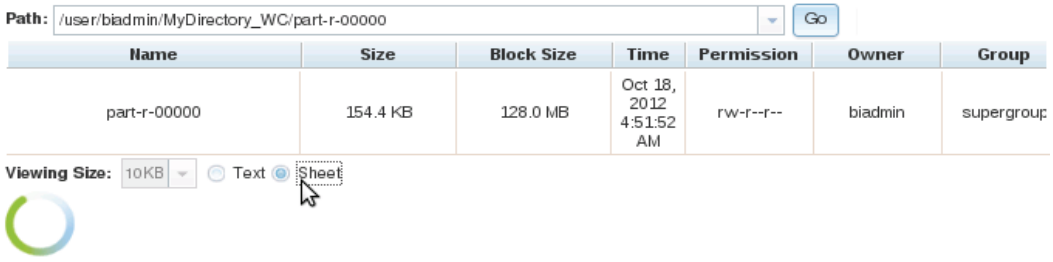


Figure 41

5. A “reader” translates your data format into the tabular format used by BigInsight's spreadsheet tool. By default, the Line Reader is used, but this can be changed by clicking on the ***Edit collection reader*** button.

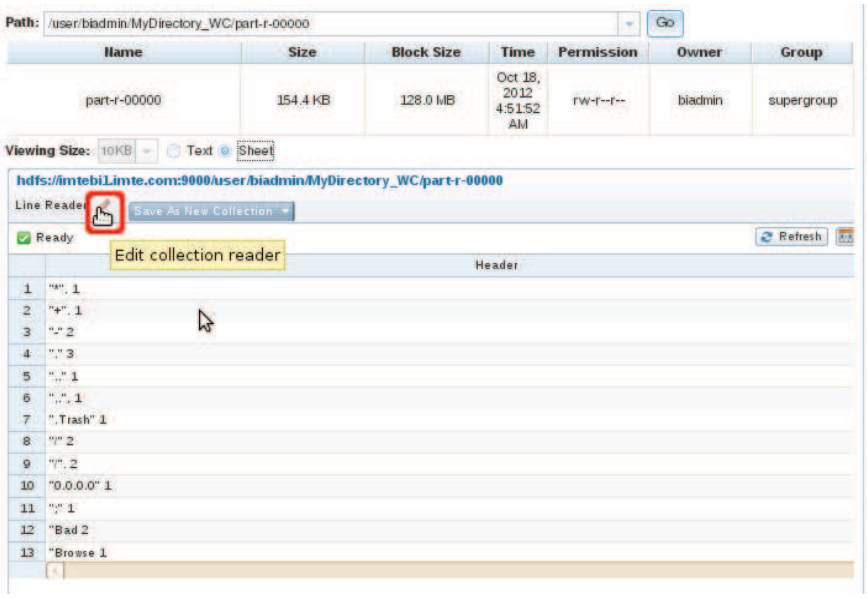


Figure 42

6. Use the drop-down menu to specify **Tab Separated Value (TSV) Data** as your reader type, deselect the ***Headers Included?*** check box, then click the green check mark at the bottom to continue.

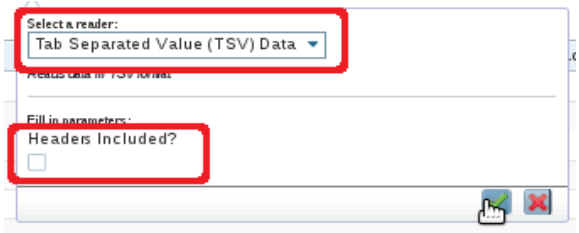


Figure 43

7. Verify that your output appears in a spreadsheet-style format.

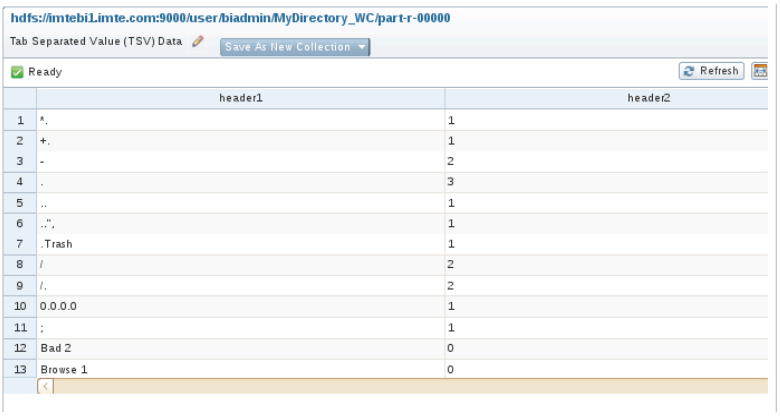


Figure 44

8. Optionally, save your sheet for future work. Click the **Save as Master Workbook** button just above the displayed data. When prompted, specify **ConsoleLab_Workbook** as the name for your new collection (or sheet), and provide a description of your choice. Click **Save**.

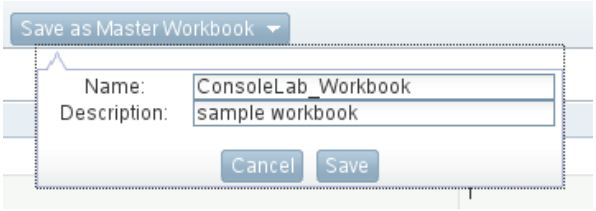


Figure 45

9. When you save the output of your WordCount run as a sheet, BigInsights switches to the **Bigsheets** tab, which enables you to analyze and manipulate this data in various ways. An exploration of BigSheets technology is beyond the scope of this lab, so we won't investigate further options available through this tool.

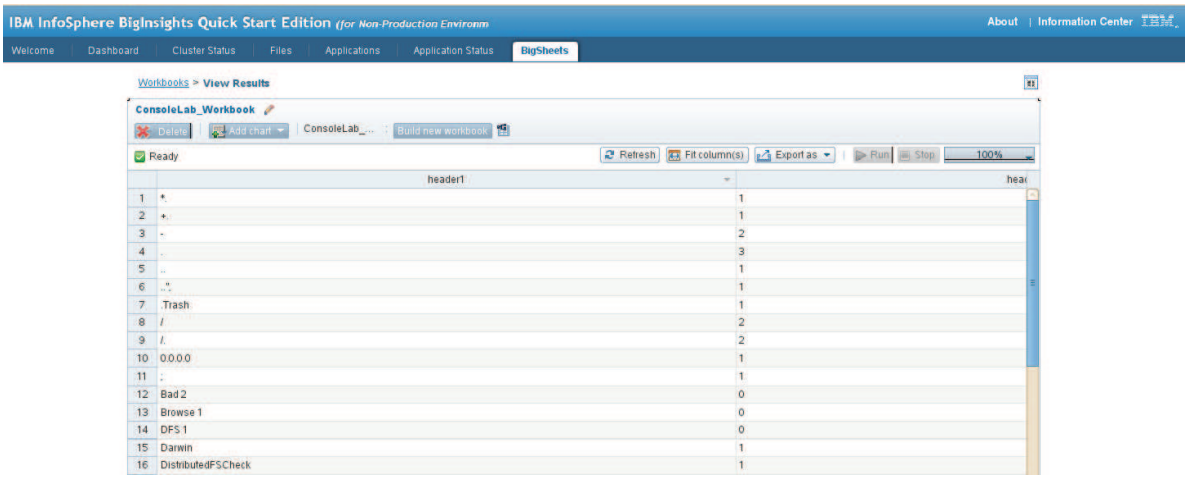


Figure 46

10.4 Exploring the Details of your Application's Execution

1. Return to the **Applications** tab.
2. If necessary, click the **WordCount** application icon in the left pane to refresh the contents of information displayed about this application at right.

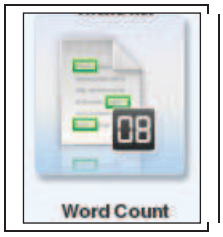


Figure 47

3. Locate the status of your test application (**ConsoleLabTest**) in the **Application History** section and click the arrow in its **Details** column.

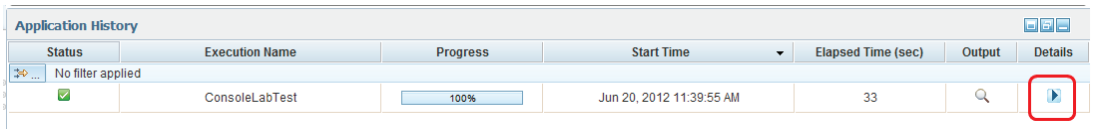


Figure 48

4. When the console switches to the **Application Status** tab, inspect the summary displayed about your application's workflow. (The Web console initiates an Oozie workflow for each application you launch). Note that the workflow completed successfully, and information about its start time, end time, ID, and other data is displayed.

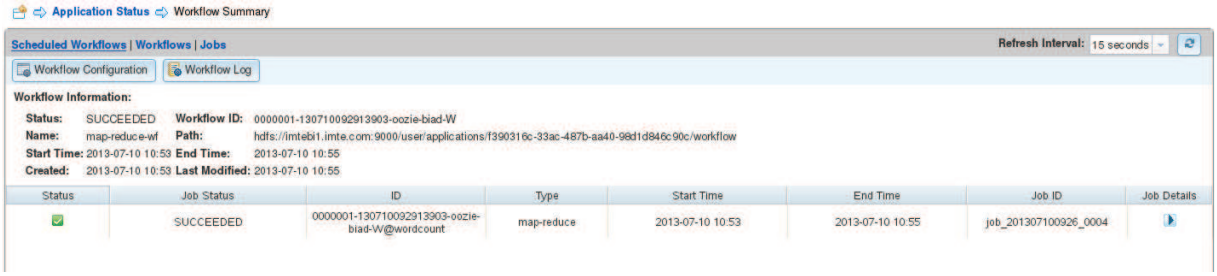


Figure 49

5. Click on the **Workflow Configuration** button to inspect its contents. Note that important properties about your job are defined here, including the input and output paths you specified as invocation parameters earlier.

Figure 50

6. Click on the **Workflow Log** button to inspect its contents. Had the job failed, you might locate useful information in this log. In this case, you can see entries indicating successful completion of an action and output generated by an action.

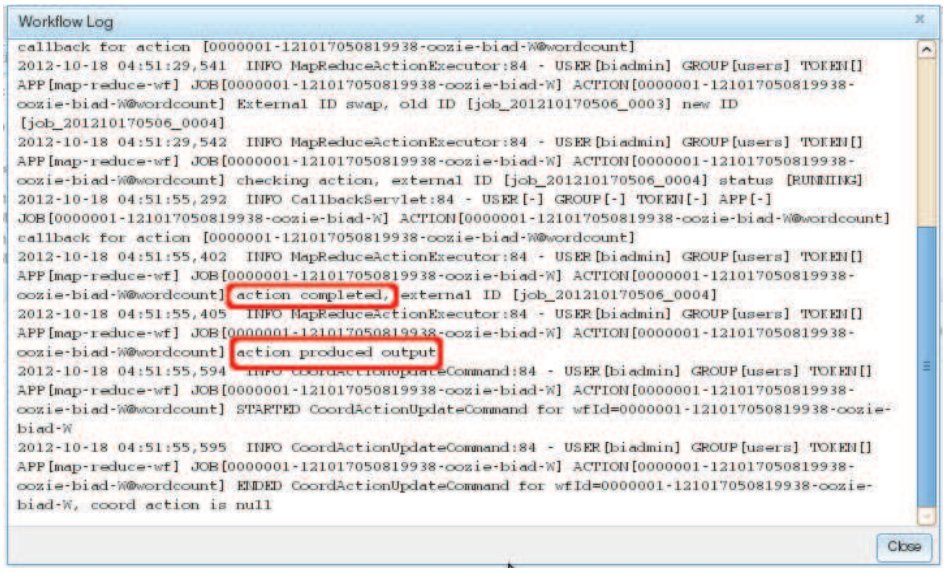


Figure 51

7. Scroll up to the top of the **Application Status** page and click the arrow in the **Details** column in the upper right corner.

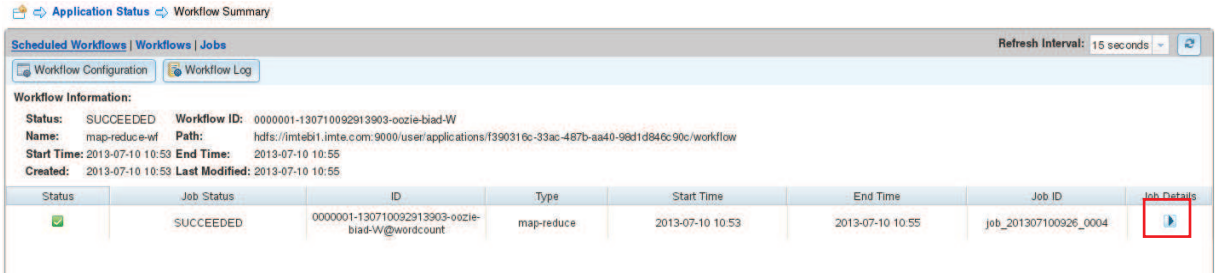


Figure 52

8. Inspect the information displayed for the job(s) associated with your workflow. In this case, the workflow that BigInsights generated on your behalf invoked two jobs. Select a given job, and more detail will be displayed at the bottom of the window. Each job requires set up, Map, Reduce, and clean up tasks, just as you would expect from any Hadoop-based job.

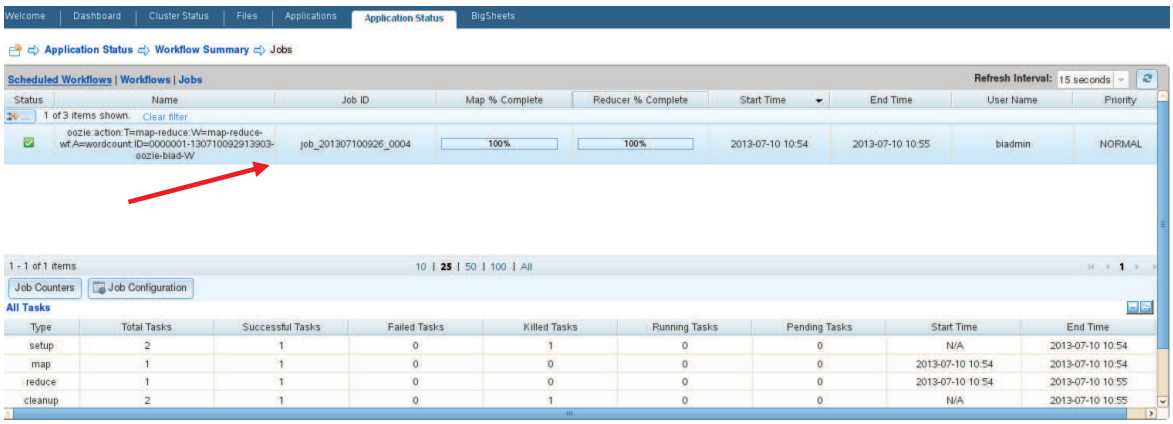


Figure 53

9. At the bottom of the window, just above the ***All Tasks*** link, there are two buttons, Job Counters and Job Configuration. These buttons can be used to learn more about the statistics and configuration associated with this job.

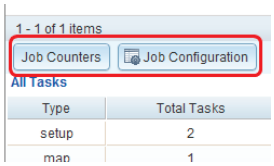


Figure 54

10. Click the **Job Counters** button to display details about the number of bytes read and written, the number of various types of input and output records produced by the MapReduce framework, and so on. Scroll through the pop-up window, if needed, to become familiar with the various statistical data collected. Click the **x** at the top of the Counters window when you're done.

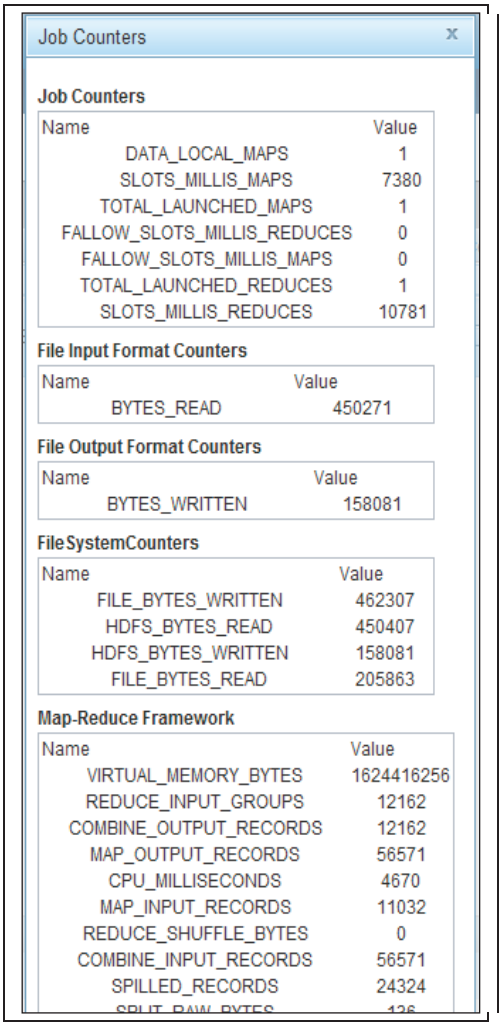


Figure 55

11. Next, click the **Job Configuration** button to display information about the configuration parameters associated with this job.

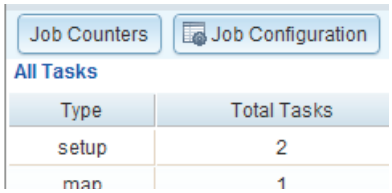


Figure 56

12. Inspect the window that appears, scrolling down as needed to review the information collected. You can also type on the *Search by Name* field at the top right corner a keyword to search, for example, type the word *input* and only the configuration information with this keyword will be displayed.

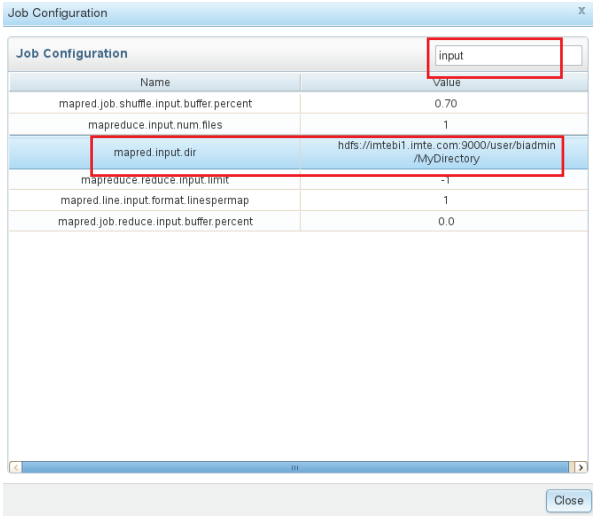


Figure 57

13. If desired, click one of the tasks under ***All Tasks*** at the bottom of the window to display further details. In many cases you can continue to drill down into further details by clicking an entry. For example, the next images below display information about the Reduce task for this job. Log information represents the final level of detail available for this task.

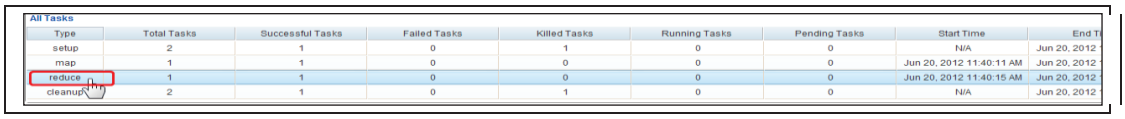


Figure 58

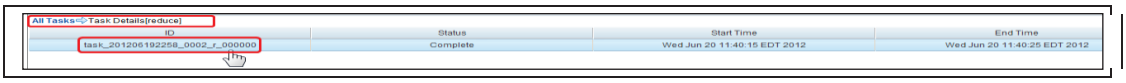


Figure 59

Figure 60

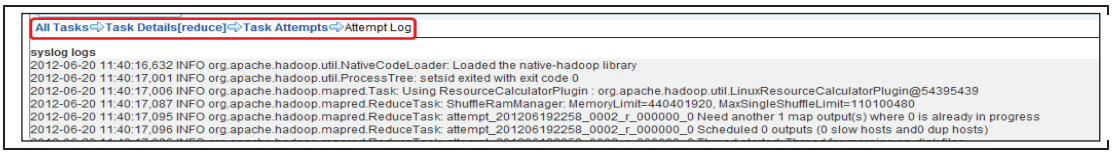
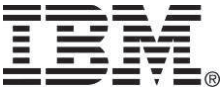


Figure 61

11 Summary

In this lab you have seen how you can work with BigInsights' Web console to administer your system, launch jobs (applications), monitor the status of jobs, and perform other functions. In next labs, you will learn how to work with unstructured text, perform data discovery, and more.



IBM Canada
8200 Warden Avenue
Markham, ON
L6G 1C7
Canada

IBM Software
Information Management

Lab 03: BigSheets

Hands-On Lab