

# Hadoop Fundamentals

## *Unit 2: Hadoop Architecture*



---

# Contents

<b>LAB 2</b>	<b>HADOOP ARCHITECTURE .....</b>	<b>4</b>
	2.1 GETTING STARTED .....	4
	2.2 BASIC HDFS INTERACTIONS USING THE COMMAND LINE .....	7
	2.3 SUMMARY .....	11

---

## Lab 2 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 Ambari Web Console to start services
- Use Hadoop commands 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 **IBM BigInsights Quick Start image**. Throughout this lab you will be using the following account login information.


	Username	Password
Rvm login	root	password
Ambari login	Admin	Admin

### 2.1 Getting Started

- \_\_1. Open up VMware Workstation Player and play the BigInsights QuickStart image. It will take a couple of minutes to load.
- \_\_2. Once loaded, type "root" as the username and "password" as password.

**Note:** To leave the VM and return to your local machine at any point press Ctrl+Alt

- \_\_3. As there is no GUI mode in this config, to use Ambari you will need to determine the IP address of the VM. Find it by using the **ifconfig** command.



```

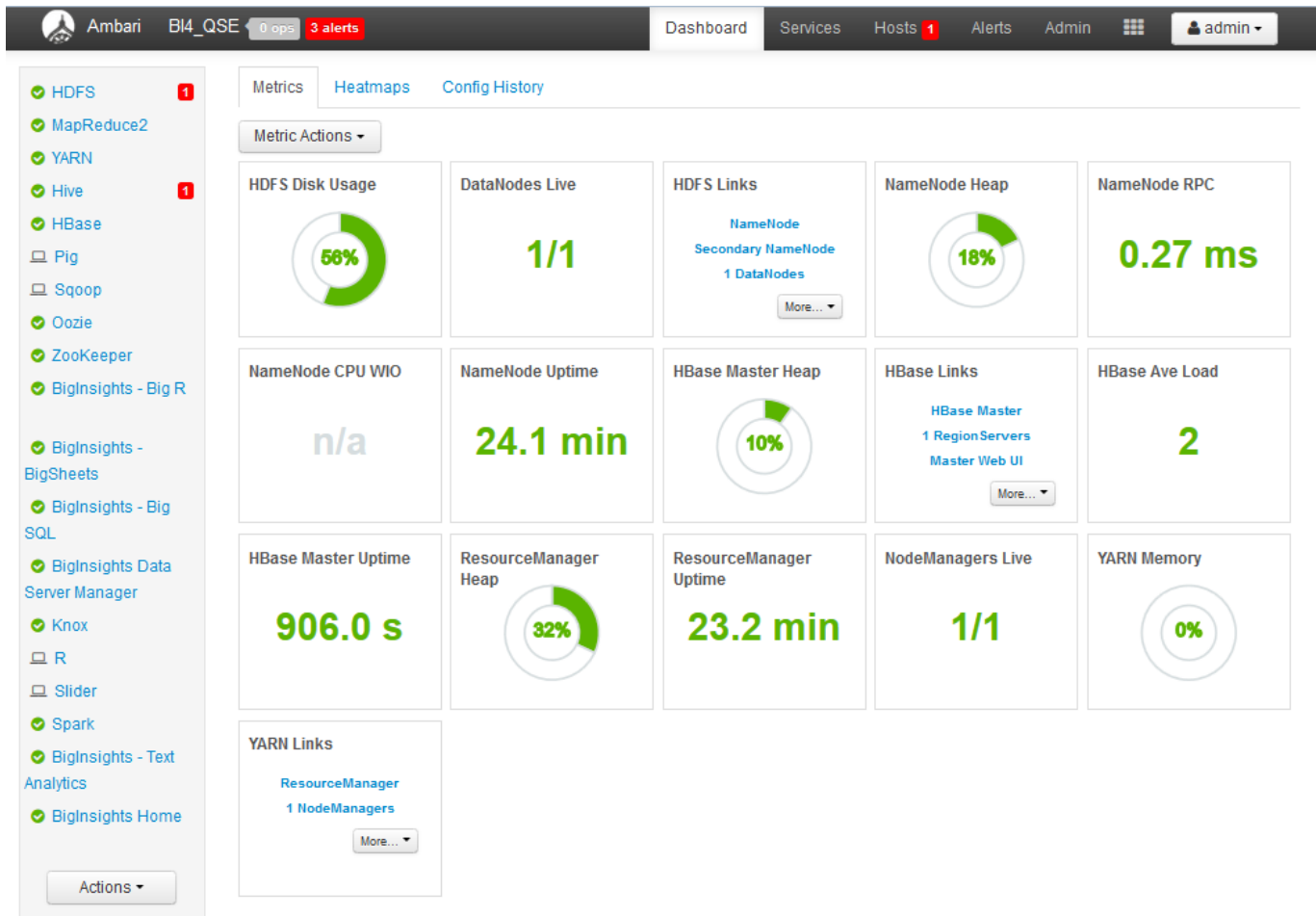
[root@rvm ~]# ifconfig
eth1      Link encap:Ethernet  HWaddr 00:0C:29:C8:05:35
          inet addr:192.168.245.130  Bcast:192.168.245.255  Mask:255.255.255.0
          inet6 addr: fe80::20c:29ff:fec8:535/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:107 errors:0 dropped:0 overruns:0 frame:0
          TX packets:90 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:10122 (9.8 KiB)  TX bytes:7718 (7.5 KiB)

lo        Link encap:Local Loopback
          inet addr:127.0.0.1  Mask:255.0.0.0
          inet6 addr: ::1/128 Scope:Host
          UP LOOPBACK RUNNING  MTU:65536  Metric:1
          RX packets:46587 errors:0 dropped:0 overruns:0 frame:0
          TX packets:46587 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:0
          RX bytes:29562153 (28.1 MiB)  TX bytes:29562153 (28.1 MiB)

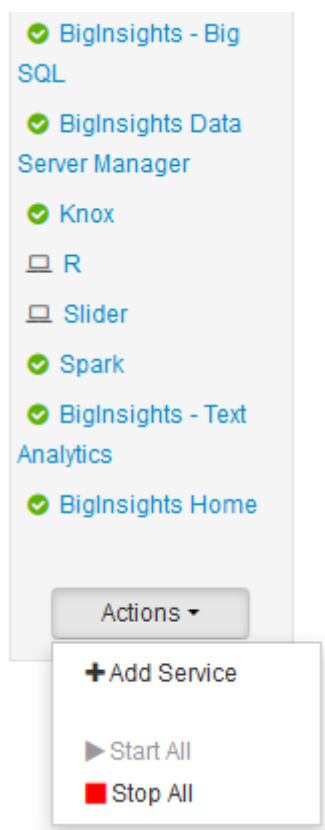
[root@rvm ~]#

```

- \_\_4. Return to your local machine using Ctrl+Alt and open up Notepad as an Administrator.
- \_\_5. Within notepad open up the /etc/hosts file, on Windows 7 it can be found in **C:\Windows\System32\drivers\etc**
- \_\_6. Add the following line to the hosts file, then save and close.  
`<IP_address_from_step_3> rvm.svl.ibm.com`
- \_\_7. From your local machine open up a web browser and navigate to: <http://rvm.svl.ibm.com:8080/>
- \_\_8. Log in to the Ambari console using **admin/admin**
- \_\_9. All the services should be started up automatically. Wait a few minutes if it hasn't started up yet to let everything start up.



If you need to start or stop your services, on the left side of the window, near the bottom, Click **Actions** and **Start All** or **Stop All**



It may take a while for all the services to start up. Once it does, you may proceed with the rest of the lab.

## 2.2 Basic HDFS interactions using the Command Line

The 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.

You can interact with HDFS at the command line:

### 1. `hdfs dfs command options`

Where command is the particular command (`ls`, `rm`, `mkdir`, ...) and options are variations on the particular command and may be followed by a list of files or a list of directories. The command is preceded by a single dash ("`-`") and the options may be preceded by a single dash.

Start with the `ls` command to list files and directories. In your VM, type the following command and hit **Enter**. Pause after each to review your results.

```
hdfs dfs -ls
hdfs dfs -ls .
```

```
hdfs dfs -ls /
```

```
[root@rpm ~]# hdfs dfs -ls
ls: '.': No such file or directory
[root@rpm ~]# hdfs dfs -ls .
ls: '.': No such file or directory
[root@rpm ~]# hdfs dfs -ls /
Found 9 items
drwxrwxrwx   - yarn      hadoop      0 2015-12-11 03:52 /app-logs
drwxr-xr-x   - hdfs      hdfs        0 2015-12-11 03:54 /apps
drwxrwxr-x   - hdfs      hdfs        0 2015-12-11 04:33 /biginsights
drwxr-xr-x   - hdfs      hdfs        0 2015-12-11 03:58 /ibmpacks
drwxr-xr-x   - hdfs      hdfs        0 2015-12-11 03:53 /iop
drwxr-xr-x   - mapred    hdfs        0 2015-12-11 03:52 /mapred
drwxr-xr-x   - hdfs      hdfs        0 2015-12-11 03:52 /mr-history
drwxrwxrwx   - hdfs      hdfs        0 2015-12-11 03:58 /tmp
drwxr-xr-x   - hdfs      hdfs        0 2015-12-11 05:21 /user
[root@rpm ~]# _
```

The first of these lists the files in the current directory — there are none. The second is a little more explicit since it asks for files in dot (“.”), a synonym for “here” (again the current directory). The third lists files at the root level within the HDFS (and there are eight directories).

Look at the directory, **/user** — this is where all “home” directories are kept for HDFS. The equivalent for Linux is **/home** — and note the spelling “**/user**” as this distinguishes this directory from the **/usr** directory in Linux that is used for executable binary programs.

```
hdfs dfs -ls /user
```

```
[root@rpm ~]# hdfs dfs -ls /user
Found 10 items
drwxrwx---   - ambari-qa hdfs        0 2015-12-11 03:52 /user/ambari-qa
drwxrwxrwx   - biadmin  hdfs        0 2015-12-11 05:21 /user/biadmin
drwxrwxrwx   - bigr     hdfs        0 2015-12-11 03:58 /user/bigr
drwxr-xr-x   - bigsql   hdfs        0 2015-12-11 04:33 /user/bigsql
drwxr-xr-x   - hbase    hdfs        0 2015-12-11 03:52 /user/hbase
drwxr-xr-x   - hcat     hdfs        0 2015-12-11 03:54 /user/hcat
drwx-----  - hive     hdfs        0 2015-12-11 03:54 /user/hive
drwxrwxr-x   - oozie    hdfs        0 2015-12-11 03:55 /user/oozie
drwxr-xr-x   - spark    hadoop      0 2015-12-11 04:05 /user/spark
drwxrwxr-x   - tauser   hadoop      0 2015-12-11 03:54 /user/tauser
```

Switch to the hdfs user using this command

```
su hdfs
```

Create a directory called *test* in the **/user/spark/** directory

```
hdfs dfs -mkdir /user/spark/test
```



Check the contents of your home directory before and after the command to see that is created.

```
hdfs dfs -ls /user/spark/
```

```
[root@rvm ~]# su hdfs
[hdfs@rvm root]$ hdfs dfs -mkdir /user/spark/test
[hdfs@rvm root]$ hdfs dfs -ls /user/spark/
Found 1 items
drwxr-xr-x  - hdfs hadoop          0 2016-01-07 13:55 /user/spark/test
[hdfs@rvm root]$ _
```

Exit out of the hdfs user to get back to root: Type **exit** in the terminal.

Create a file in your Linux home directory. Execute the following commands: (Ctrl-c means to press-and-hold the **Ctrl** key and then the **c** key.)

```
cd sampleData/spark
cat > myfile.txt
this is some data
in my file
Ctrl-c
```

```
[hdfs@rvm ~]# exit
exit
[root@rvm /]# cd sampleData/spark
[root@rvm spark]# ls
labfiles.zip  numbers
[root@rvm spark]# cat > myfile.txt
this is some data
in my file
^C
[root@rvm spark]# _
```

Next upload this newly created file to the *test* directory that you just created. Don't forget to switch to the hdfs user first or else you will get that error message below.

```
su hdfs
hdfs dfs -put *.txt /user/spark/test
```

```
[root@rvm spark]# hdfs dfs -put *.txt /user/spark/test/
put: Permission denied: user=root, access=WRITE, inode="/user/spark/test/myfile.txt._COPYING_":hdfs:hadoop:drwxr-xr-x
[root@rvm spark]# su hdfs
[hdfs@rvm spark]$ hdfs dfs -put *.txt /user/spark/test
[hdfs@rvm spark]$ _
```

Now list your *test* directory in HDFS. You can use either of the following commands:

```
hdfs dfs -ls -R /user/spark/test
```

```
[hdfs@rvm spark]$ hdfs dfs -ls -R /user/spark/test
-rw-r--r-- 1 hdfs hadoop      29 2016-01-07 14:04 /user/spark/test/myfile.t
xt
[hdfs@rvm spark]$ _
```

Note the number 1 that follows the permissions. This is the replication factor for that data file. Normally, in a cluster this is 3, but sometimes in a single-node cluster such as the one that you are running with, there might be only one copy of each block (“split”) of the this file.

The value 1 (or 3, or something else) is the result of a configuration setting of HDFS that sets the number of replicants by default.

To view the contents of the uploaded file, execute

```
hdfs dfs -cat /user/spark/test/myfile.txt
```

```
[hdfs@rvm spark]$ hdfs dfs -cat /user/spark/test/myfile.txt
this is some data
in my file
[hdfs@rvm spark]$ _
```

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

```
hdfs dfs -cat /user/spark/test/myfile.txt | grep my
```

```
[hdfs@rvm spark]$ hdfs dfs -cat /user/spark/test/myfile.txt | grep my
in my file
[hdfs@rvm spark]$ _
```

Or,

```
hdfs dfs -ls -R /user/spark/test/ | grep test
```

```
[hdfs@rvm spark]$ hdfs dfs -cat /user/spark/test/ | grep test
cat: '/user/spark/test': Is a directory
```

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

```
hdfs dfs -du /user/spark/test/myfile.txt
```

Or, to get the size of all files in a directory by using a directory name rather than a file name.

```
hdfs dfs -du /user/spark
```

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

```
hdfs dfs -du -s /user/spark
```

```
[hdfs@rvm spark]$ hdfs dfs -du /user/spark/test/myfile.txt
29 /user/spark/test/myfile.txt
[hdfs@rvm spark]$ hdfs dfs -du /user/spark
29 /user/spark/test
[hdfs@rvm spark]$ hdfs dfs -du -s /user/spark
29 /user/spark
[hdfs@rvm spark]$ _
```

Remember that you can always use the **-help** parameter to get more help:

```
hdfs dfs -help
hdfs dfs -help du
```

```
[hdfs@rvm spark]$ hdfs dfs -help du
-du [-s] [-h] <path> ... :
  Show the amount of space, in bytes, used by the files that match the specified
  file pattern. The following flags are optional:

  -s  Rather than showing the size of each individual file that matches the
       pattern, shows the total (summary) size.

  -h  Formats the sizes of files in a human-readable fashion rather than a number
       of bytes.

  Note that, even without the -s option, this only shows size summaries one level
  deep into a directory.

  The output is in the form
      size    name(full path)
[hdfs@rvm spark]$ _
```

You can close the command line window.

## 2.3 Summary

Congratulations! You are now familiar with the Hadoop Distributed File System (HDFS). You now know how to manipulate files within HDFS by using the command line.

You may move on to the next unit.

## NOTES

[illegible]

## NOTES

[illegible]



---

© Copyright IBM Corporation 2016.

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

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



Please Recycle

---