Basic Hadoop Commands

```
$ hdfs dfs
```

Provides all the commands that you can run in HDFS

```
$ hdfs dfs -ls
```

Will list all of the files and directories in the HDFS

```
$ hdfs dfs -mkdir rawdata
```

Will make a new directory named ‘rawdata’ in HDFS

Copy the access_log file to the desktop in your VM (which adds it to the local filesystem on the VM). Now load the access_log file to the HDFS rawdata directory using the following command:

```
$ hdfs dfs -put Desktop/access_log rawdata
```

Now verify if the file is loaded

```
$ hdfs dfs -ls rawdata
```

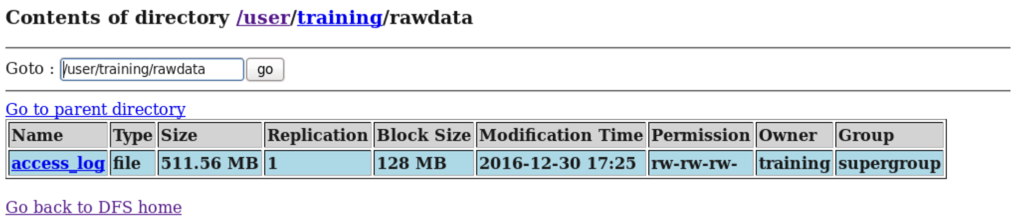
Now open the access_log file using the following command:

```
$ hdfs dfs -cat rawdata/access_log
```

To break the reading of the file, press Crtl+C

Now, go to the HDFS portal to see the details of how the files are stored. Use the URL

http://localhost:50070/dfshealth.jsp



Contents of directory /user/training/rawdata

Goto : /user/training/rawdata go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
access_log	file	511.56 MB	1	128 MB	2016-12-30 17:25	rw-rw-rw-	training	supergroup

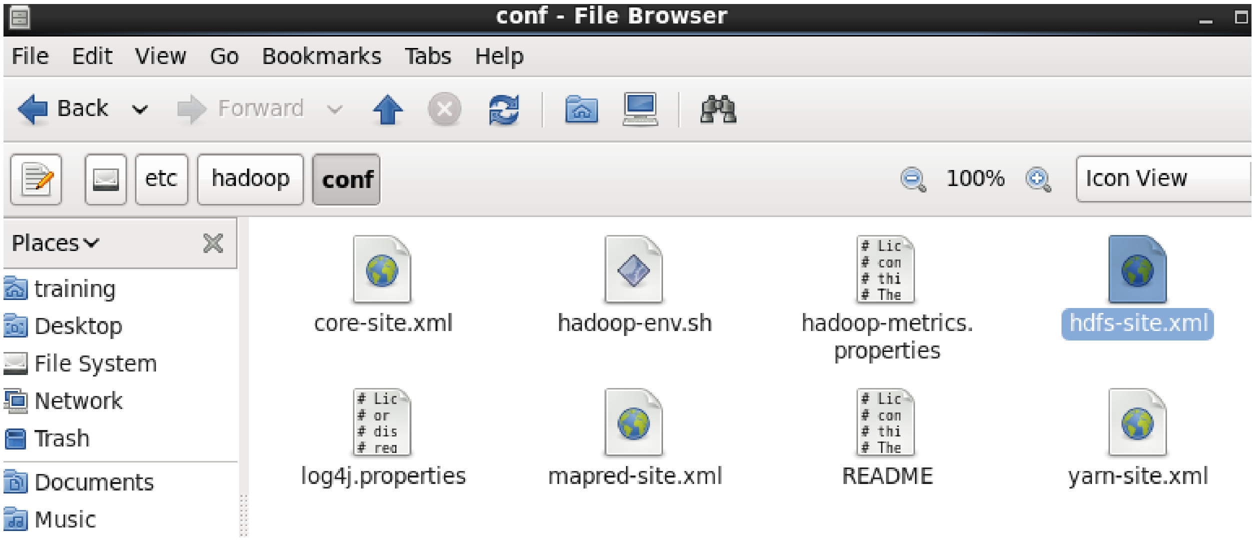
Go back to DFS home

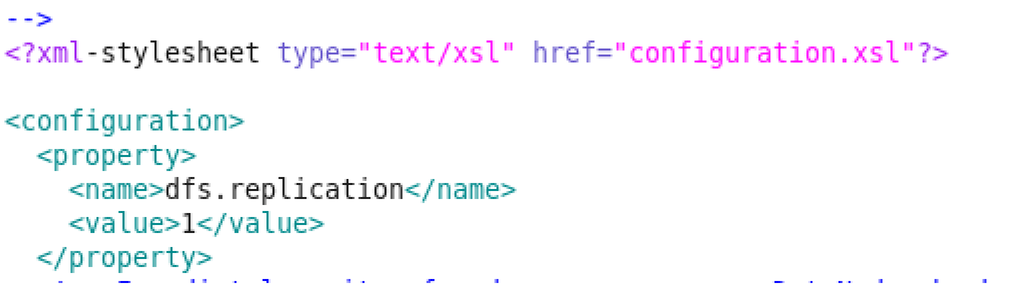
Notice that that replication for the Cloudera VM is set to 1 to save space and the block sizes are set to 128MB.

How can we change the default configurations on the VM to change the Replication and the Block size?

To change the default Hadoop settings, you will have to go to the local file system to the following directory:

/etc/hadoop/conf and modify the hdfs.xml file





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

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

dfs.replication is the name of the property that controls the replication factor in the Hadoop deployment.

Similarly, there is a property named `dfs.blocksize` which captures the size of the blocks in bytes for the HDFS. The default for the block size is 128MB

Modifying the file using the VI editor

Enter the sudo command to edit the file using the vi editor

```
$ sudo vi /etc/hadoop/conf/hdfs-site.xml
```

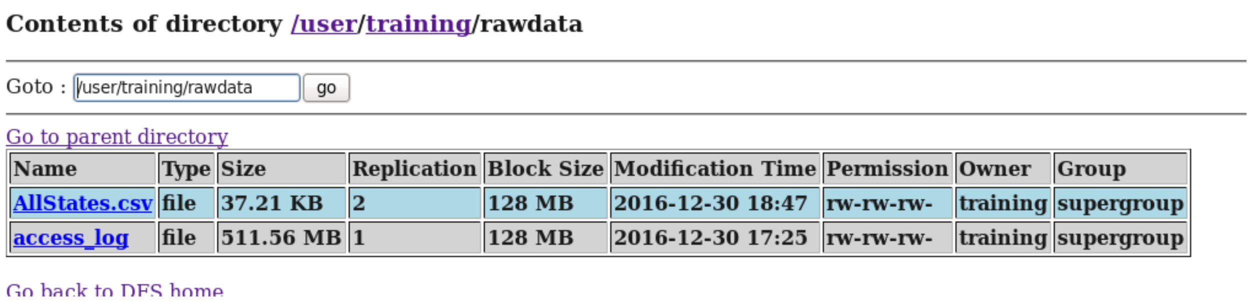
Press i to insert and make the change

Then press Esc and then write :wq! to save

Now, copy the AllStates.csv file to the VM and then use the put command to store it in HDFS rawdata folder.

```
$ hdfs dfs -put Desktop/AllStates.csv  rawdata
```

Now look at the details of how the file is stored using the HDFS portal.



Contents of directory /user/training/rawdata

Goto : /user/training/rawdata

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
AllStates.csv	file	37.21 KB	2	128 MB	2016-12-30 18:47	rw-rw-rw-	training	supergroup
access_log	file	511.56 MB	1	128 MB	2016-12-30 17:25	rw-rw-rw-	training	supergroup

Go back to DFS home

Notice how the replication of this file is changed to 2.

Since we are using only one machine in this cluster, there is no point in storing each file twice. Change the replication back to 1.