Overview Lab 2: Issuing basic Hadoc

saracco

Published on August 22, 2014 / *Updated on September 29, 2014*

0

Lab 2: Issuing basic Hadoop comma

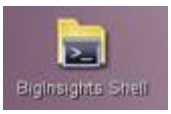
In this exercise, you'll work directly with Apache Hadoop to perform some basic tasks invc (HDFS) and launching a sample application. All the work you'll perform here involves comr from http://hadoop.apache.org. As mentioned earlier, Hadoop is part of IBM's InfoSphere I

Allow 15 minutes to complete this lab module. Prior to this lab, you should have set up a w Hadoop and BigInsights for details.

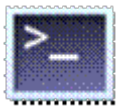
Please post questions or comments about this lab to the forum on Hadoop Dev at https://c /answers?community=hadoop.

2.1. Creating a directory in your distributed file sy

__1. Click the **BigInsights Shell** icon.



__2. Select the **Terminal** icon to open a terminal window.



__3. Execute the following Hadoop file system command to create a directory in HDFS for

hadoop fs -mkdir /user/biadmin/test

Note that HDFS is distinct from your Unix/Linux local file system directory, and working wit

2.2. Copying data into HDFS

__1. Using standard Unix/Linux file system commands, list the contents of the /home/biad

ls /home/biadmin/licenses

```
biadmin@bivm:~> ls /home/biadmin/licenses/
BIlicense_cs.txt  BIlicense_in.txt  BIlicense_pt.txt     LA.html
BIlicense_de.txt  BIlicense_it.txt  BIlicense_ru.txt     notices.html
BIlicense_el.txt  BIlicense_ja.txt  BIlicense_sl.txt     notices.txt
BIlicense_en.txt  BIlicense_ko.txt  BIlicense_tr.txt
BIlicense_es.txt  BIlicense_lt.txt  BIlicense_zh_TW.txt
BIlicense_fr.txt  BIlicense_pl.txt  BIlicense_zh.txt
```

Note the BIlicense_en.txt file. It contains license information in English, and it will serve as

__2. Copy the BIlicense_en.txt file into the /user/biadmin/test directory you just created

hadoop fs -put /home/biadmin/licenses/BIlicense_en.txt /user/biadmin/test

__3. List the contents of your target HDFS directory to verify that the file was successfully

hadoop fs -ls /user/biadmin/test

```
biadmin@bivm:~> hadoop fs -ls /user/biadmin/test
Found 1 items
-rw-r--r--   1 biadmin biadmin      62945 2014-07-02 16:23 /user/biadmin/test/BIlicense_en.txt
```

2.3. Running a sample MapReduce application

WordCount is one of several sample MapReduce applications provided for Apache Hadoop
document(s) and, for each word, returns the total number of occurrences found. You can r
(http://wiki.apache.org/hadoop/WordCount).

Since launching MapReduce applications (or jobs) is a common practice in Hadoop, you'll

__1. Execute the following command to launch the sample WordCount application provide

hadoop jar /opt/ibm/biginsights/IHC/hadoop-example.jar wordcount /user/biadr

This command specifies that the wordcount application contained in the specified .jar file application is in the /user/biadmin/test directory of HDFS. The output of this job will be s subdirectory of the user executing this command (biadmin). Thus, the output directory wil This directory will be created automatically as a result of executing this application.

NOTE: If the output folder already exists or if you try to rerun a successf MapReduce job with the same parameters, you will receive an error message. This is the default behavior of the sample WordCount application.

```
biadmin@bivm:~> hadoop jar /opt/ibm/biginsights/IHC/hadoop-example.jar wordcount /user/biadmin/t
est WordCount_output
14/07/02 18:42:29 INFO input.FileInputFormat: Total input paths to process : 1
14/07/02 18:42:30 INFO mapred.JobClient: Running job: job_201407021102_0010
14/07/02 18:42:31 INFO mapred.JobClient:  map 0% reduce 0%
14/07/02 18:42:39 INFO mapred.JobClient:  map 100% reduce 0%
14/07/02 18:42:48 INFO mapred.JobClient:  map 100% reduce 100%
14/07/02 18:42:49 INFO mapred.JobClient: Job complete: job_201407021102_0010
14/07/02 18:42:49 INFO mapred.JobClient: Counters: 29
14/07/02 18:42:49 INFO mapred.JobClient:   File System Counters
14/07/02 18:42:49 INFO mapred.JobClient:     FILE: BYTES_READ=33215
14/07/02 18:42:49 INFO mapred.JobClient:     FILE: BYTES_WRITTEN=418608
14/07/02 18:42:49 INFO mapred.JobClient:     HDFS: BYTES_READ=63069
14/07/02 18:42:49 INFO mapred.JobClient:     HDFS: BYTES_WRITTEN=24069
14/07/02 18:42:49 INFO mapred.JobClient:   org.apache.hadoop.mapreduce.JobCounter
14/07/02 18:42:49 INFO mapred.JobClient:     TOTAL_LAUNCHED_MAPS=1
14/07/02 18:42:49 INFO mapred.JobClient:     TOTAL_LAUNCHED_REDUCES=1
14/07/02 18:42:49 INFO mapred.JobClient:     DATA_LOCAL_MAPS=1
14/07/02 18:42:49 INFO mapred.JobClient:     SLOTS_MILLIS_MAPS=7009
14/07/02 18:42:49 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCES=8705
14/07/02 18:42:49 INFO mapred.JobClient:     FALLOW_SLOTS_MILLIS_MAPS=0
14/07/02 18:42:49 INFO mapred.JobClient:     FALLOW_SLOTS_MILLIS_REDUCES=0
14/07/02 18:42:49 INFO mapred.JobClient:   org.apache.hadoop.mapreduce.TaskCounter
14/07/02 18:42:49 INFO mapred.JobClient:     MAP_INPUT_RECORDS=755
14/07/02 18:42:49 INFO mapred.JobClient:     MAP_OUTPUT_RECORDS=9865
14/07/02 18:42:49 INFO mapred.JobClient:     MAP_OUTPUT_BYTES=102032
14/07/02 18:42:49 INFO mapred.JobClient:     MAP_OUTPUT_MATERIALIZED_BYTES=33215
14/07/02 18:42:49 INFO mapred.JobClient:     SPLIT_RAW_BYTES=124
14/07/02 18:42:49 INFO mapred.JobClient:     COMBINE_INPUT_RECORDS=9865
14/07/02 18:42:49 INFO mapred.JobClient:     COMBINE_OUTPUT_RECORDS=2322
14/07/02 18:42:49 INFO mapred.JobClient:     REDUCE_INPUT_GROUPS=2322
14/07/02 18:42:49 INFO mapred.JobClient:     REDUCE_SHUFFLE_BYTES=33215
14/07/02 18:42:49 INFO mapred.JobClient:     REDUCE_INPUT_RECORDS=2322
14/07/02 18:42:49 INFO mapred.JobClient:     REDUCE_OUTPUT_RECORDS=2322
14/07/02 18:42:49 INFO mapred.JobClient:     SPILLED_RECORDS=4644
14/07/02 18:42:49 INFO mapred.JobClient:     CPU_MILLISECONDS=2070
14/07/02 18:42:49 INFO mapred.JobClient:     PHYSICAL_MEMORY_BYTES=610246656
14/07/02 18:42:49 INFO mapred.JobClient:     VIRTUAL_MEMORY_BYTES=2491170816
14/07/02 18:42:49 INFO mapred.JobClient:     COMMITTED_HEAP_BYTES=2097152000
14/07/02 18:42:49 INFO mapred.JobClient:   File Input Format Counters
14/07/02 18:42:49 INFO mapred.JobClient:     Bytes Read=62945
14/07/02 18:42:49 INFO mapred.JobClient:   org.apache.hadoop.mapreduce.lib.output.FileOutputForm
at$Counter
14/07/02 18:42:49 INFO mapred.JobClient:     BYTES_WRITTEN=24069
```

__2. Inspect the output of your job.

hadoop fs -ls WordCount_output

```
biadmin@bivm:~> hadoop fs -ls WordCount_output
Found 3 items
-rw-r--r--   1 biadmin biadmin          0 2014-07-02 18:42 WordCount_output/_SUCCESS
drwx--x--x   - biadmin biadmin          0 2014-07-02 18:42 WordCount_output/_logs
-rw-r--r--   1 biadmin biadmin      24069 2014-07-02 18:42 WordCount_output/part-r-00000
```

In this case, the output was small and contained written to a single file. If you had run Wor output would have been split into multiple files (e.g., part-r-00001, part-r-00002, and so c

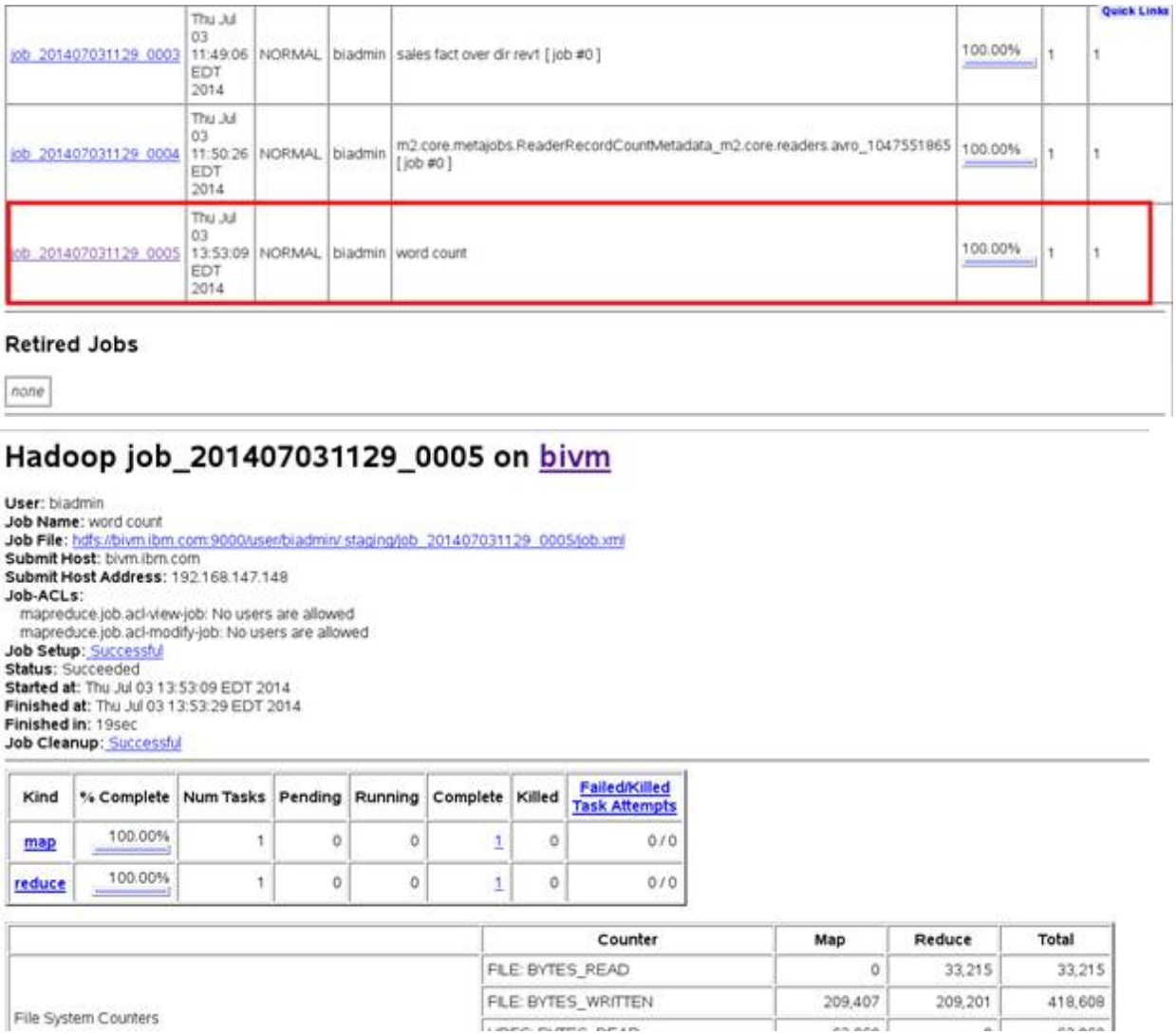
__3. To view the contents of *part-r-0000* file, issue this command:

hadoop fs -cat WordCount_output/*00

Partial output is shown here:

```
warrant 1
warranted       1
warranties      2
warranties,     2
warrants        1
warranty        7
was     8
wasted  1
way     1
web     1
website 1
well    1
were    2
what    1
when    4
where,  3
wherever        1
whether 2
which   37
who     5
whole   1
whom    10
will    72
willful 2
wishes  2
with    54
with,   1
within  8
without 20
word    2
words   1
would   1
writing 4
written 5
www.ibm.com/software/sla.       1
years   4
years,  2
```

__4. Optionally, inspect details about your job. Open a Web browser, or click on the web cc tab. Access the URL for Hadoop's Job Tracker (http://bivm.ibm.com:50030/jobtracker locate the Job ID associated with the Word Count application. Click on the Job ID link to re Reduce tasks launched for your application, the number of bytes read and written, etc. Par follows.



job_201407031129_0003	Thu Jul 03 11:49:06 EDT 2014	NORMAL	biadmin	sales fact over dir rev1 [job #0]	100.00%	1	1
job_201407031129_0004	Thu Jul 03 11:50:26 EDT 2014	NORMAL	biadmin	m2.core.metajobs.ReaderRecordCountMetadata_m2.core.readers.avro_1047551865 [job #0]	100.00%	1	1
job_201407031129_0005	Thu Jul 03 13:53:09 EDT 2014	NORMAL	biadmin	word count	100.00%	1	1

Quick Links

Retired Jobs

none

Hadoop job_201407031129_0005 on bivm

User: biadmin
Job Name: word count
Job File: hdfs://bivm.ibm.com:9000/user/biadmin/.staging/job_201407031129_0005/job.xml
Submit Host: bivm.ibm.com
Submit Host Address: 192.168.147.148
Job-ACLs:
mapreduce.job.acl-view-job: No users are allowed
mapreduce.job.acl-modify-job: No users are allowed
Job Setup: Successful
Status: Succeeded
Started at: Thu Jul 03 13:53:09 EDT 2014
Finished at: Thu Jul 03 13:53:29 EDT 2014
Finished in: 19sec
Job Cleanup: Successful

Kind	% Complete	Num Tasks	Pending	Running	Complete	Killed	Failed/Killed Task Attempts
map	100.00%	1	0	0	1	0	0 / 0
reduce	100.00%	1	0	0	1	0	0 / 0

	Counter	Map	Reduce	Total
File System Counters	FILE: BYTES_READ	0	33,215	33,215
	FILE: BYTES_WRITTEN	209,407	209,201	418,608

To find the other tutorials in this series, go to Overview tutorial.

TAGS BIGINSIGHTS VALUE ADD, MAPREDUCE, HADOOP, TUTORIAL, SARACCO, BIG DATA, HDFS, GETTING STARTED,

saracco

Join The Discussion

Your email address will not be published. Required fields are marked *

Enter your comments...

Name *

Email *

Website

Save my name, email, and website in this browser for the next time I comment.