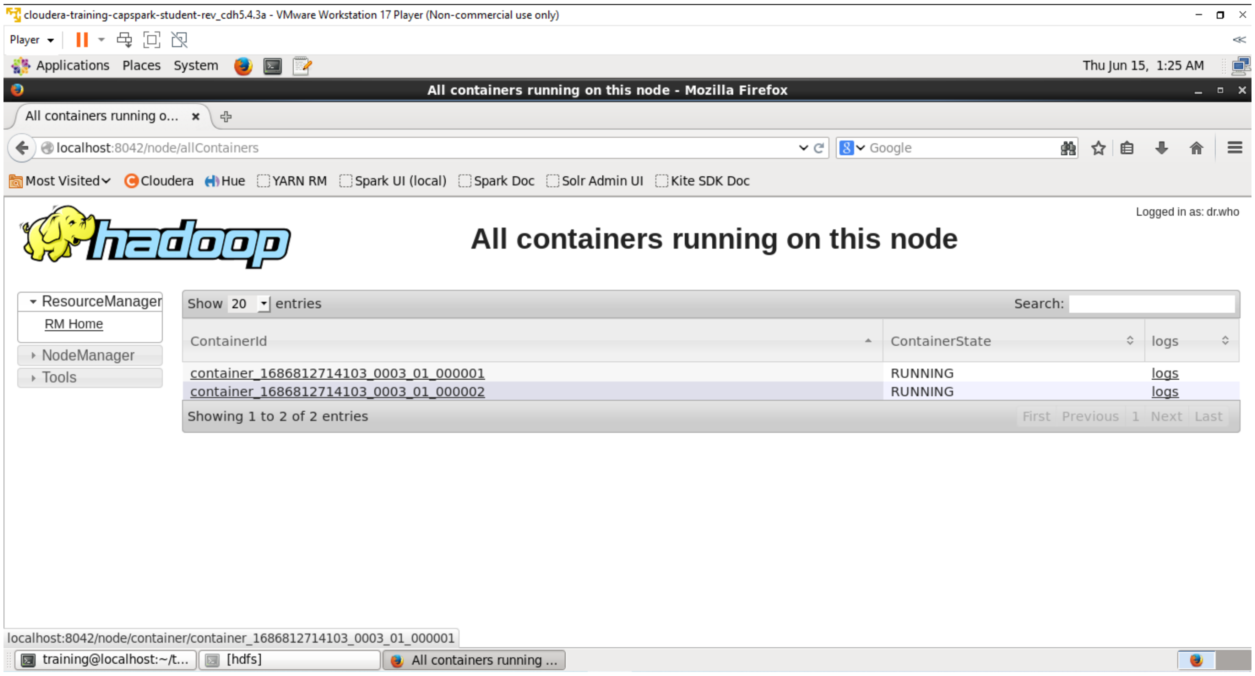
LAB 3 BIG DATA

Submitted by:
ABHISHEK ARYA
NET ID: axa220149

Answer: The desirable output of LAB 3 is as follows in the screenshot:


cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 17 Player (Non-commercial use only)
Player
Applications Places System
Thu Jun 15, 1:25 AM
All containers running on this node - Mozilla Firefox
All containers running o...
localhost:8042/node/allContainers
Google
Most Visited
Cloudera
Hue
YARN RM
Spark UI (local)
Spark Doc
Solr Admin UI
Kite SDK Doc
Logged in as: dr.who
hadoop
ResourceManager
RM Home
NodeManager
Tools
training@localhost:~/t...
[hdfs]
All containers running ...

All containers running on this node

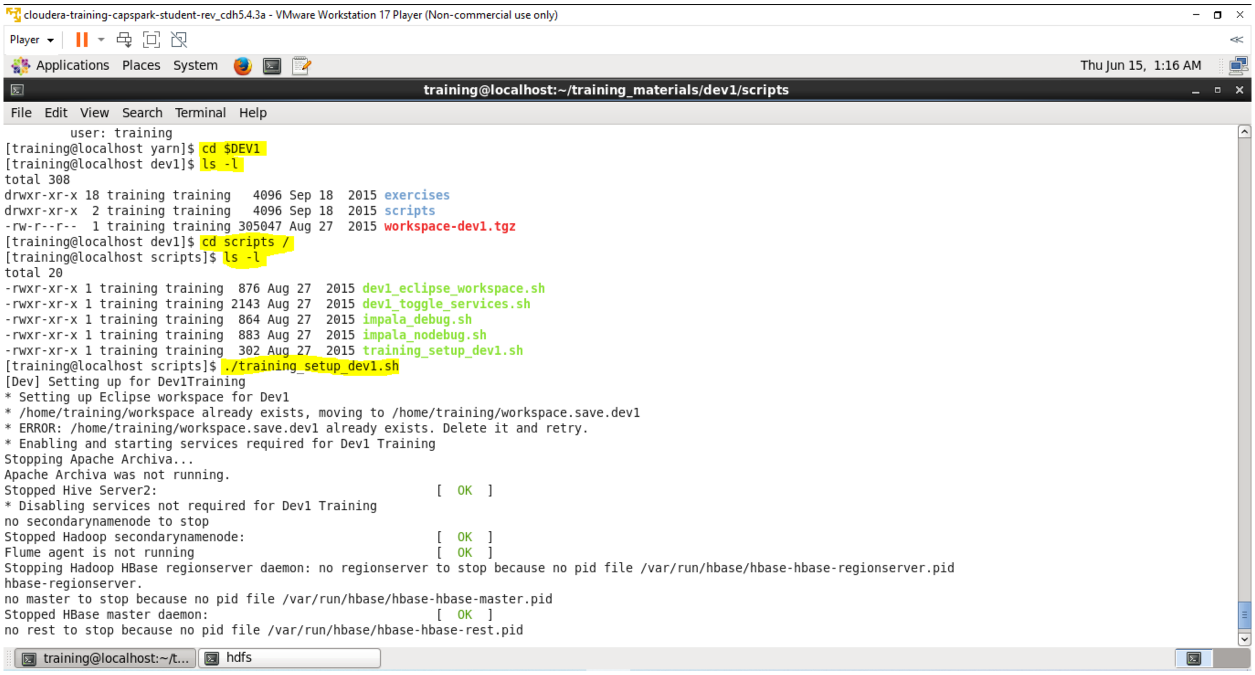
Show 20 entries Search:

ContainerId	ContainerState	logs
container_1686812714103_0003_01_000001	RUNNING	logs
container_1686812714103_0003_01_000002	RUNNING	logs

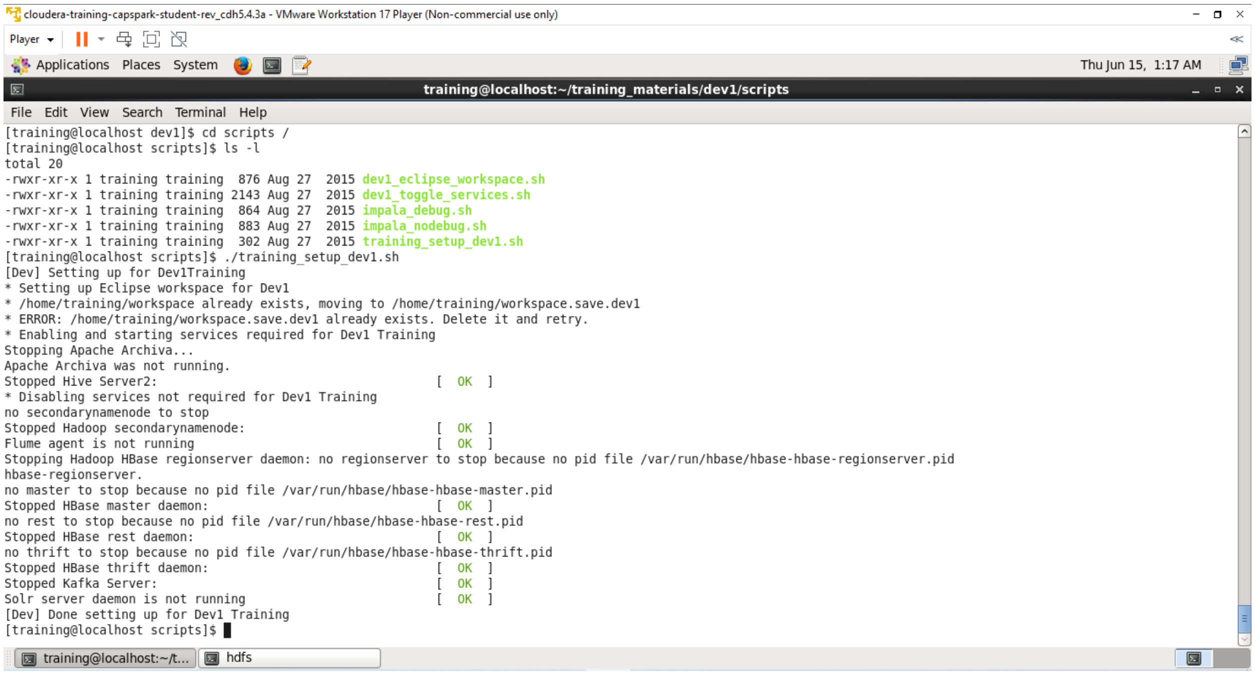
Showing 1 to 2 of 2 entries First Previous 1 Next Last

localhost:8042/node/container/container_1686812714103_0003_01_000001

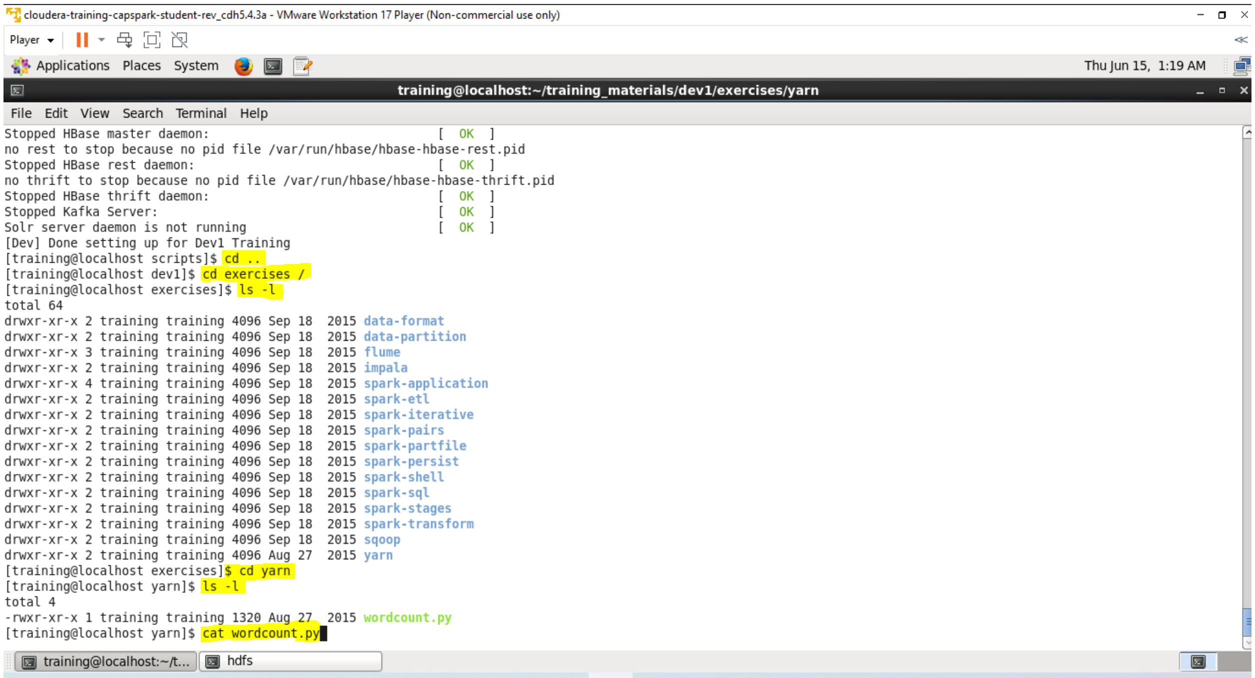
I am adding all the steps which led me to this output just for the safer side; please find the screenshots of all the steps outputs below:


cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 17 Player (Non-commercial use only)
Player
Applications Places System
Thu Jun 15, 1:16 AM
training@localhost:~/training_materials/dev1/scripts
File Edit View Search Terminal Help
training@localhost:~/t...
hdfs

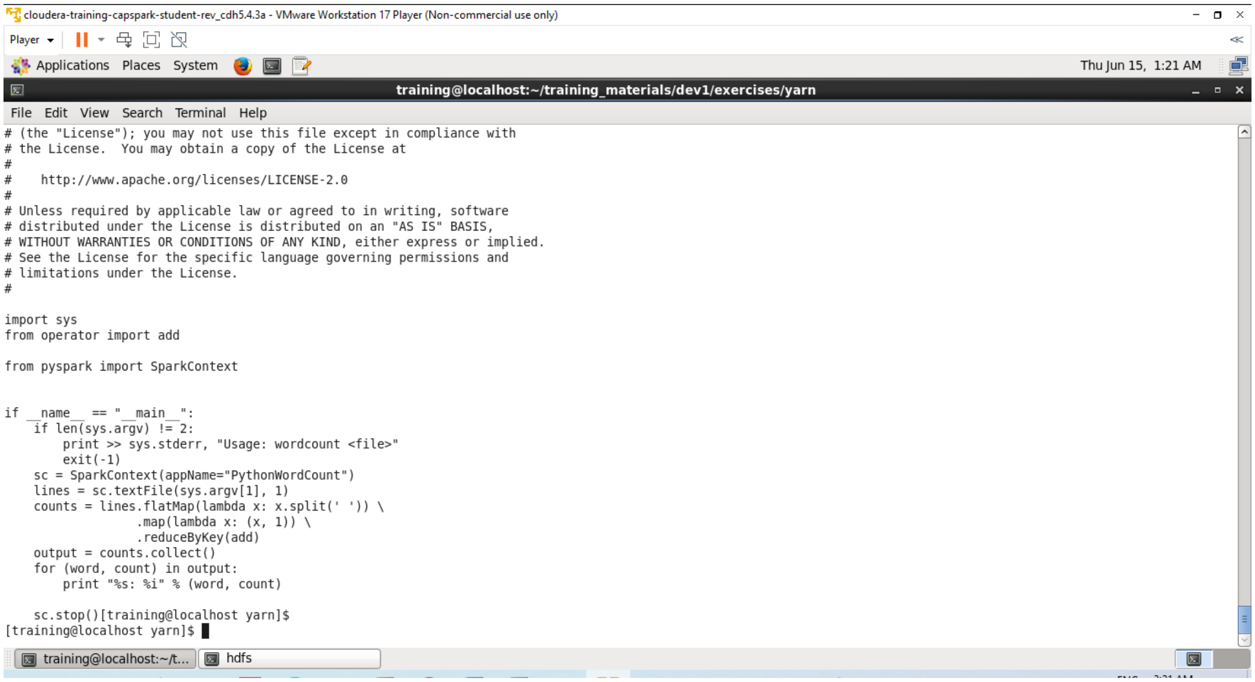
```
        user: training
[training@localhost yarn]$ cd $DEV1
[training@localhost dev1]$ ls -l
total 308
drwxr-xr-x 18 training training   4096 Sep 18  2015 exercises
drwxr-xr-x  2 training training   4096 Sep 18  2015 scripts
-rw-r--r--  1 training training 305047 Aug 27  2015 workspace-dev1.tgz
[training@localhost dev1]$ cd scripts /
[training@localhost scripts]$ ls -l
total 20
-rwxr-xr-x 1 training training  876 Aug 27  2015 dev1_eclipse_workspace.sh
-rwxr-xr-x 1 training training 2143 Aug 27  2015 dev1_toggle_services.sh
-rwxr-xr-x 1 training training  864 Aug 27  2015 impala_debug.sh
-rwxr-xr-x 1 training training  883 Aug 27  2015 impala_nodebug.sh
-rwxr-xr-x 1 training training  302 Aug 27  2015 training_setup_dev1.sh
[training@localhost scripts]$ ./training_setup_dev1.sh
[Dev] Setting up for Dev1Training
* Setting up Eclipse workspace for Dev1
* /home/training/workspace already exists, moving to /home/training/workspace.save.dev1
* ERROR: /home/training/workspace.save.dev1 already exists. Delete it and retry.
* Enabling and starting services required for Dev1 Training
Stopping Apache Archiva...
Apache Archiva was not running.
Stopped Hive Server2:                                      [  OK  ]
* Disabling services not required for Dev1 Training
no secondarynamenode to stop
Stopped Hadoop secondarynamenode:                          [  OK  ]
Flume agent is not running                                 [  OK  ]
Stopping Hadoop HBase regionserver daemon: no regionserver to stop because no pid file /var/run/hbase/hbase-hbase-regionserver.pid
hbase-regionserver.
no master to stop because no pid file /var/run/hbase/hbase-hbase-master.pid
Stopped HBase master daemon:                               [  OK  ]
no rest to stop because no pid file /var/run/hbase/hbase-hbase-rest.pid
```


cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 17 Player (Non-commercial use only)
Player
Applications Places System
Thu Jun 15, 1:17 AM
training@localhost:~/training_materials/dev1/scripts
File Edit View Search Terminal Help
training@localhost:~/t...
hdfs

```
[training@localhost dev1]$ cd scripts /
[training@localhost scripts]$ ls -l
total 20
-rwxr-xr-x 1 training training  876 Aug 27  2015 dev1_eclipse_workspace.sh
-rwxr-xr-x 1 training training 2143 Aug 27  2015 dev1_toggle_services.sh
-rwxr-xr-x 1 training training  864 Aug 27  2015 impala_debug.sh
-rwxr-xr-x 1 training training  883 Aug 27  2015 impala_nodebug.sh
-rwxr-xr-x 1 training training  302 Aug 27  2015 training_setup_dev1.sh
[training@localhost scripts]$ ./training_setup_dev1.sh
[Dev] Setting up for Dev1Training
* Setting up Eclipse workspace for Dev1
* /home/training/workspace already exists, moving to /home/training/workspace.save.dev1
* ERROR: /home/training/workspace.save.dev1 already exists. Delete it and retry.
* Enabling and starting services required for Dev1 Training
Stopping Apache Archiva...
Apache Archiva was not running.
Stopped Hive Server2:                                      [  OK  ]
* Disabling services not required for Dev1 Training
no secondarynamenode to stop
Stopped Hadoop secondarynamenode:                          [  OK  ]
Flume agent is not running                                 [  OK  ]
Stopping Hadoop HBase regionserver daemon: no regionserver to stop because no pid file /var/run/hbase/hbase-hbase-regionserver.pid
hbase-regionserver.
no master to stop because no pid file /var/run/hbase/hbase-hbase-master.pid
Stopped HBase master daemon:                               [  OK  ]
no rest to stop because no pid file /var/run/hbase/hbase-hbase-rest.pid
Stopped HBase rest daemon:                                 [  OK  ]
no thrift to stop because no pid file /var/run/hbase/hbase-hbase-thrift.pid
Stopped HBase thrift daemon:                               [  OK  ]
Stopped Kafka Server:                                      [  OK  ]
Solr server daemon is not running                          [  OK  ]
[Dev] Done setting up for Dev1 Training
[training@localhost scripts]$ 
```


cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 17 Player (Non-commercial use only)
Player
Applications
Places
System
Thu Jun 15, 1:19 AM
training@localhost:~/training_materials/dev1/exercises/yarn
File
Edit
View
Search
Terminal
Help
training@localhost:~/t...
hdfs

```
Stopped HBase master daemon:                               [  OK  ]
no rest to stop because no pid file /var/run/hbase/hbase-hbase-rest.pid
Stopped HBase rest daemon:                                 [  OK  ]
no thrift to stop because no pid file /var/run/hbase/hbase-hbase-thrift.pid
Stopped HBase thrift daemon:                               [  OK  ]
Stopped Kafka Server:                                      [  OK  ]
Solr server daemon is not running                          [  OK  ]
[Dev] Done setting up for Dev1 Training
[training@localhost scripts]$ cd ..
[training@localhost dev1]$ cd exercises /
[training@localhost exercises]$ ls -l
total 64
drwxr-xr-x 2 training training 4096 Sep 18  2015 data-format
drwxr-xr-x 2 training training 4096 Sep 18  2015 data-partition
drwxr-xr-x 3 training training 4096 Sep 18  2015 flume
drwxr-xr-x 2 training training 4096 Sep 18  2015 impala
drwxr-xr-x 4 training training 4096 Sep 18  2015 spark-application
drwxr-xr-x 2 training training 4096 Sep 18  2015 spark-etl
drwxr-xr-x 2 training training 4096 Sep 18  2015 spark-iterative
drwxr-xr-x 2 training training 4096 Sep 18  2015 spark-pairs
drwxr-xr-x 2 training training 4096 Sep 18  2015 spark-partfile
drwxr-xr-x 2 training training 4096 Sep 18  2015 spark-persist
drwxr-xr-x 2 training training 4096 Sep 18  2015 spark-shell
drwxr-xr-x 2 training training 4096 Sep 18  2015 spark-sql
drwxr-xr-x 2 training training 4096 Sep 18  2015 spark-stages
drwxr-xr-x 2 training training 4096 Sep 18  2015 spark-transform
drwxr-xr-x 2 training training 4096 Sep 18  2015 sqoop
drwxr-xr-x 2 training training 4096 Aug 27  2015 yarn
[training@localhost exercises]$ cd yarn
[training@localhost yarn]$ ls -l
total 4
-rwxr-xr-x 1 training training 1320 Aug 27  2015 wordcount.py
[training@localhost yarn]$ cat wordcount.py
```


cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 17 Player (Non-commercial use only)
Player
Applications
Places
System
Thu Jun 15, 1:21 AM
training@localhost:~/training_materials/dev1/exercises/yarn
File
Edit
View
Search
Terminal
Help
training@localhost:~/t...
hdfs

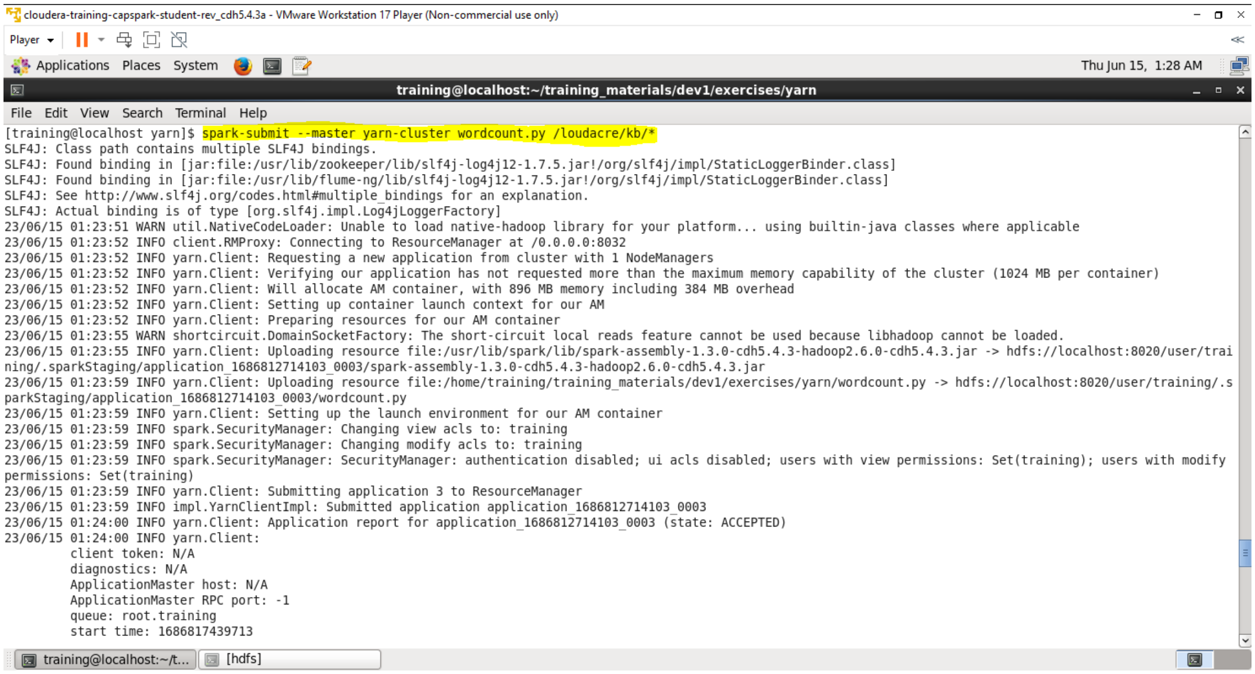
```
# (the "License"); you may not use this file except in compliance with
# the License.  You may obtain a copy of the License at
#
#    http://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.
#

import sys
from operator import add

from pyspark import SparkContext


if __name__ == "__main__":
    if len(sys.argv) != 2:
        print >> sys.stderr, "Usage: wordcount <file>"
        exit(-1)
    sc = SparkContext(appName="PythonWordCount")
    lines = sc.textFile(sys.argv[1], 1)
    counts = lines.flatMap(lambda x: x.split(' ')) \
                  .map(lambda x: (x, 1)) \
                  .reduceByKey(add)
    output = counts.collect()
    for (word, count) in output:
        print "%s: %i" % (word, count)

    sc.stop()[training@localhost yarn]$
[training@localhost yarn]$
```


cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 17 Player (Non-commercial use only)
Player
Applications
Places
System
Thu Jun 15, 1:28 AM
training@localhost:~/training_materials/dev1/exercises/yarn
File
Edit
View
Search
Terminal
Help
training@localhost:~/t...
[hdfs]

```
[training@localhost yarn]$ spark-submit --master yarn-cluster wordcount.py /loudacre/kb/*
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/flume-ng/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
23/06/15 01:23:51 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
23/06/15 01:23:52 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
23/06/15 01:23:52 INFO yarn.Client: Requesting a new application from cluster with 1 NodeManagers
23/06/15 01:23:52 INFO yarn.Client: Verifying our application has not requested more than the maximum memory capability of the cluster (1024 MB per container)
23/06/15 01:23:52 INFO yarn.Client: Will allocate AM container, with 896 MB memory including 384 MB overhead
23/06/15 01:23:52 INFO yarn.Client: Setting up container launch context for our AM
23/06/15 01:23:52 INFO yarn.Client: Preparing resources for our AM container
23/06/15 01:23:55 WARN shortcircuit.DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be loaded.
23/06/15 01:23:55 INFO yarn.Client: Uploading resource file:/usr/lib/spark/lib/spark-assembly-1.3.0-cdh5.4.3-hadoop2.6.0-cdh5.4.3.jar -> hdfs://localhost:8020/user/trai
ning/.sparkStaging/application_1686812714103_0003/spark-assembly-1.3.0-cdh5.4.3-hadoop2.6.0-cdh5.4.3.jar
23/06/15 01:23:59 INFO yarn.Client: Uploading resource file:/home/training/training_materials/dev1/exercises/yarn/wordcount.py -> hdfs://localhost:8020/user/training/.s
parkStaging/application_1686812714103_0003/wordcount.py
23/06/15 01:23:59 INFO yarn.Client: Setting up the launch environment for our AM container
23/06/15 01:23:59 INFO spark.SecurityManager: Changing view acls to: training
23/06/15 01:23:59 INFO spark.SecurityManager: Changing modify acls to: training
23/06/15 01:23:59 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(training); users with modify
permissions: Set(training)
23/06/15 01:23:59 INFO yarn.Client: Submitting application 3 to ResourceManager
23/06/15 01:23:59 INFO impl.YarnClientImpl: Submitted application application_1686812714103_0003
23/06/15 01:24:00 INFO yarn.Client: Application report for application_1686812714103_0003 (state: ACCEPTED)
23/06/15 01:24:00 INFO yarn.Client:
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: N/A
         ApplicationMaster RPC port: -1
         queue: root.training
         start time: 1686817439713
```