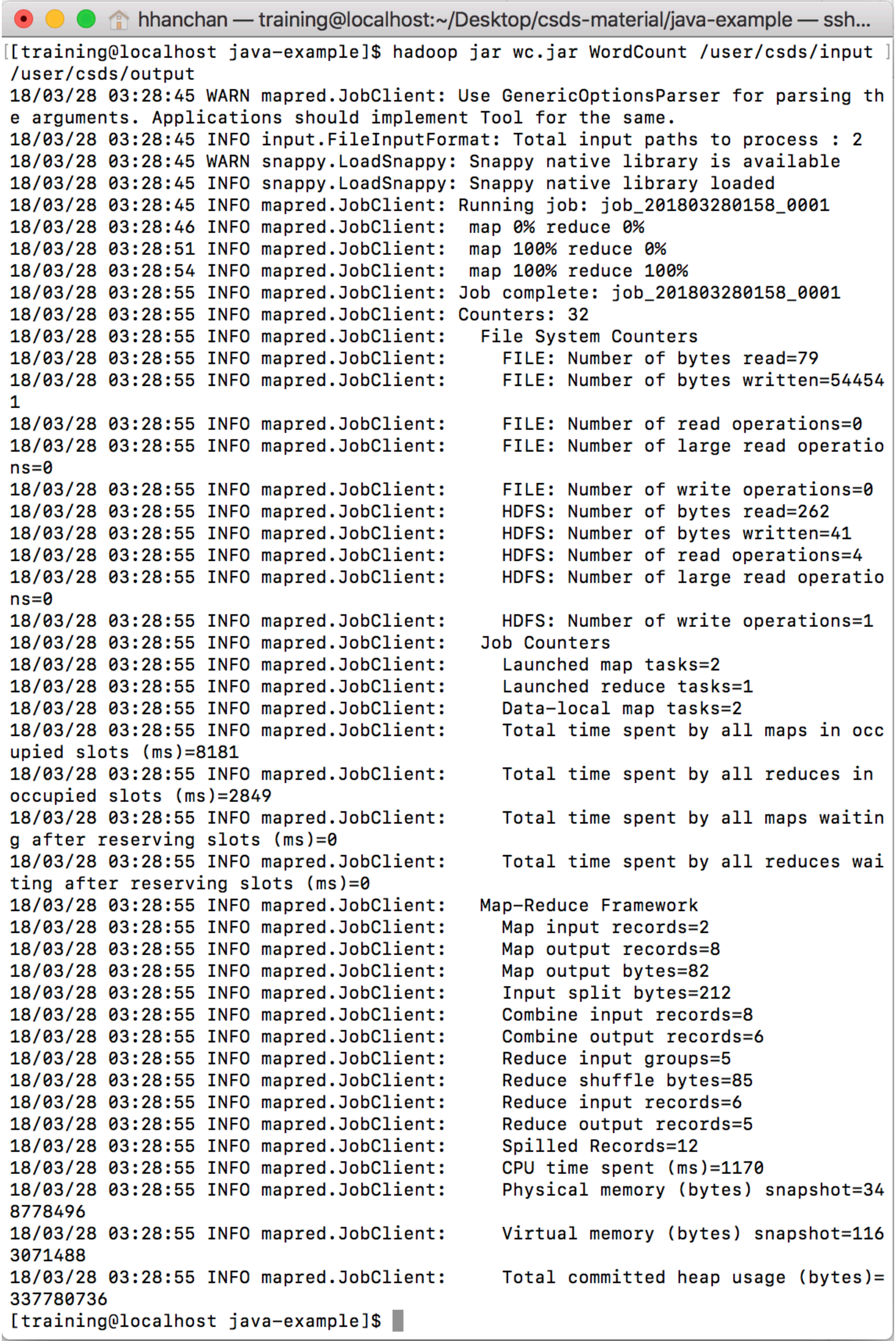
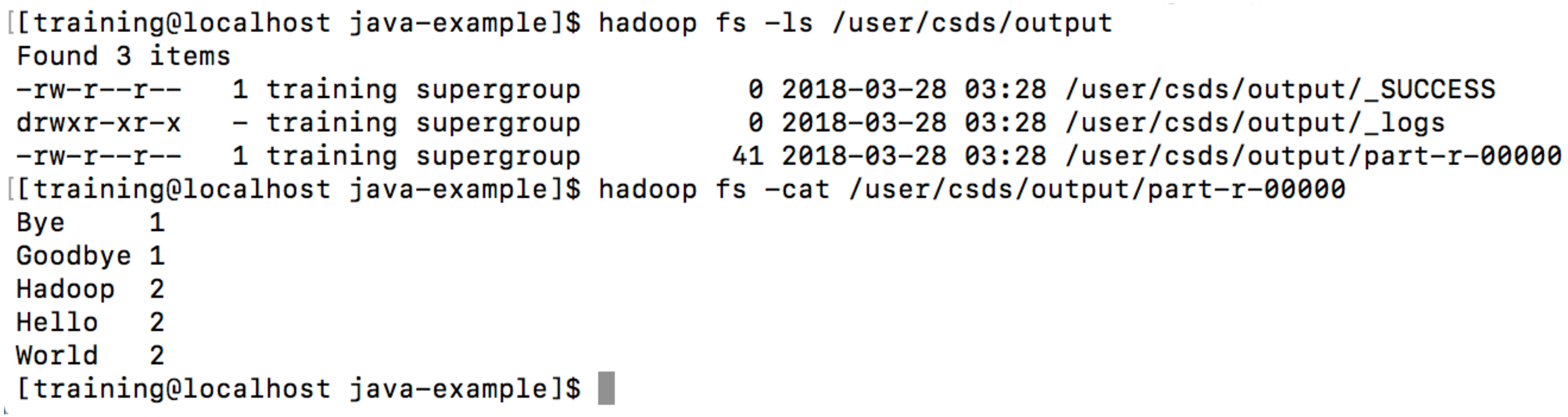
HW4: hadoop

NAME: Jinhan Cheng
UNI: jc4834


hhanchan — training@localhost:~/Desktop/csds-material/java-example — ssh...

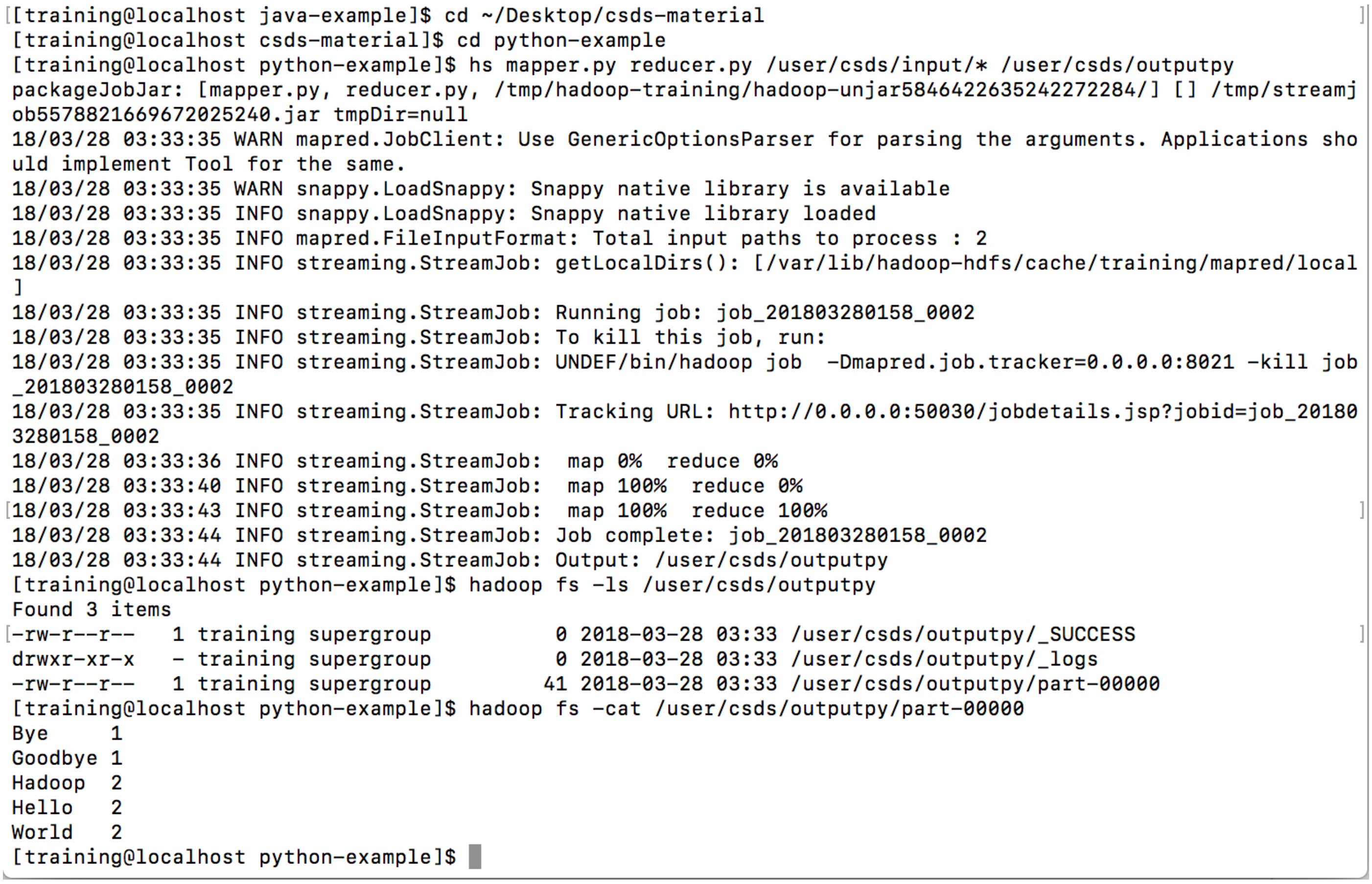
```
[training@localhost java-example]$ hadoop jar wc.jar WordCount /user/csds/input
/user/csds/output
18/03/28 03:28:45 WARN mapred.JobClient: Use GenericOptionsParser for parsing th
e arguments. Applications should implement Tool for the same.
18/03/28 03:28:45 INFO input.FileInputFormat: Total input paths to process : 2
18/03/28 03:28:45 WARN snappy.LoadSnappy: Snappy native library is available
18/03/28 03:28:45 INFO snappy.LoadSnappy: Snappy native library loaded
18/03/28 03:28:45 INFO mapred.JobClient: Running job: job_201803280158_0001
18/03/28 03:28:46 INFO mapred.JobClient:  map 0% reduce 0%
18/03/28 03:28:51 INFO mapred.JobClient:  map 100% reduce 0%
18/03/28 03:28:54 INFO mapred.JobClient:  map 100% reduce 100%
18/03/28 03:28:55 INFO mapred.JobClient: Job complete: job_201803280158_0001
18/03/28 03:28:55 INFO mapred.JobClient: Counters: 32
18/03/28 03:28:55 INFO mapred.JobClient:   File System Counters
18/03/28 03:28:55 INFO mapred.JobClient:     FILE: Number of bytes read=79
18/03/28 03:28:55 INFO mapred.JobClient:     FILE: Number of bytes written=54454
1
18/03/28 03:28:55 INFO mapred.JobClient:     FILE: Number of read operations=0
18/03/28 03:28:55 INFO mapred.JobClient:     FILE: Number of large read operatio
ns=0
18/03/28 03:28:55 INFO mapred.JobClient:     FILE: Number of write operations=0
18/03/28 03:28:55 INFO mapred.JobClient:     HDFS: Number of bytes read=262
18/03/28 03:28:55 INFO mapred.JobClient:     HDFS: Number of bytes written=41
18/03/28 03:28:55 INFO mapred.JobClient:     HDFS: Number of read operations=4
18/03/28 03:28:55 INFO mapred.JobClient:     HDFS: Number of large read operatio
ns=0
18/03/28 03:28:55 INFO mapred.JobClient:     HDFS: Number of write operations=1
18/03/28 03:28:55 INFO mapred.JobClient:   Job Counters
18/03/28 03:28:55 INFO mapred.JobClient:     Launched map tasks=2
18/03/28 03:28:55 INFO mapred.JobClient:     Launched reduce tasks=1
18/03/28 03:28:55 INFO mapred.JobClient:     Data-local map tasks=2
18/03/28 03:28:55 INFO mapred.JobClient:     Total time spent by all maps in occ
upied slots (ms)=8181
18/03/28 03:28:55 INFO mapred.JobClient:     Total time spent by all reduces in
occupied slots (ms)=2849
18/03/28 03:28:55 INFO mapred.JobClient:     Total time spent by all maps waitin
g after reserving slots (ms)=0
18/03/28 03:28:55 INFO mapred.JobClient:     Total time spent by all reduces wai
ting after reserving slots (ms)=0
18/03/28 03:28:55 INFO mapred.JobClient:   Map-Reduce Framework
18/03/28 03:28:55 INFO mapred.JobClient:     Map input records=2
18/03/28 03:28:55 INFO mapred.JobClient:     Map output records=8
18/03/28 03:28:55 INFO mapred.JobClient:     Map output bytes=82
18/03/28 03:28:55 INFO mapred.JobClient:     Input split bytes=212
18/03/28 03:28:55 INFO mapred.JobClient:     Combine input records=8
18/03/28 03:28:55 INFO mapred.JobClient:     Combine output records=6
18/03/28 03:28:55 INFO mapred.JobClient:     Reduce input groups=5
18/03/28 03:28:55 INFO mapred.JobClient:     Reduce shuffle bytes=85
18/03/28 03:28:55 INFO mapred.JobClient:     Reduce input records=6
18/03/28 03:28:55 INFO mapred.JobClient:     Reduce output records=5
18/03/28 03:28:55 INFO mapred.JobClient:     Spilled Records=12
18/03/28 03:28:55 INFO mapred.JobClient:     CPU time spent (ms)=1170
18/03/28 03:28:55 INFO mapred.JobClient:     Physical memory (bytes) snapshot=34
8778496
18/03/28 03:28:55 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=116
3071488
18/03/28 03:28:55 INFO mapred.JobClient:     Total committed heap usage (bytes)=
337780736
[training@localhost java-example]$
```

Java



```
[[training@localhost java-example]$ hadoop fs -ls /user/csds/output
Found 3 items
-rw-r--r--   1 training supergroup          0 2018-03-28 03:28 /user/csds/output/_SUCCESS
drwxr-xr-x   - training supergroup          0 2018-03-28 03:28 /user/csds/output/_logs
-rw-r--r--   1 training supergroup         41 2018-03-28 03:28 /user/csds/output/part-r-00000
[[training@localhost java-example]$ hadoop fs -cat /user/csds/output/part-r-00000
Bye     1
Goodbye 1
Hadoop  2
Hello   2
World   2
[training@localhost java-example]$ █
```

Python



```
[training@localhost java-example]$ cd ~/Desktop/csds-material
[training@localhost csds-material]$ cd python-example
[training@localhost python-example]$ hs mapper.py reducer.py /user/csds/input/* /user/csds/outputpy
packageJobJar: [mapper.py, reducer.py, /tmp/hadoop-training/hadoop-unjar5846422635242272284/] [] /tmp/streamj
ob5578821669672025240.jar tmpDir=null
18/03/28 03:33:35 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications sho
uld implement Tool for the same.
18/03/28 03:33:35 WARN snappy.LoadSnappy: Snappy native library is available
18/03/28 03:33:35 INFO snappy.LoadSnappy: Snappy native library loaded
18/03/28 03:33:35 INFO mapred.FileInputFormat: Total input paths to process : 2
18/03/28 03:33:35 INFO streaming.StreamJob: getLocalDirs(): [/var/lib/hadoop-hdfs/cache/training/mapred/local
]
18/03/28 03:33:35 INFO streaming.StreamJob: Running job: job_201803280158_0002
18/03/28 03:33:35 INFO streaming.StreamJob: To kill this job, run:
18/03/28 03:33:35 INFO streaming.StreamJob: UNDEF/bin/hadoop job  -Dmapred.job.tracker=0.0.0.0:8021 -kill job
_201803280158_0002
18/03/28 03:33:35 INFO streaming.StreamJob: Tracking URL: http://0.0.0.0:50030/jobdetails.jsp?jobid=job_20180
3280158_0002
18/03/28 03:33:36 INFO streaming.StreamJob:  map 0%  reduce 0%
18/03/28 03:33:40 INFO streaming.StreamJob:  map 100%  reduce 0%
18/03/28 03:33:43 INFO streaming.StreamJob:  map 100%  reduce 100%
18/03/28 03:33:44 INFO streaming.StreamJob: Job complete: job_201803280158_0002
18/03/28 03:33:44 INFO streaming.StreamJob: Output: /user/csds/outputpy
[training@localhost python-example]$ hadoop fs -ls /user/csds/outputpy
Found 3 items
-rw-r--r--   1 training supergroup          0 2018-03-28 03:33 /user/csds/outputpy/_SUCCESS
drwxr-xr-x   - training supergroup          0 2018-03-28 03:33 /user/csds/outputpy/_logs
-rw-r--r--   1 training supergroup         41 2018-03-28 03:33 /user/csds/outputpy/part-00000
[training@localhost python-example]$ hadoop fs -cat /user/csds/outputpy/part-00000
Bye     1
Goodbye 1
Hadoop  2
Hello   2
World   2
[training@localhost python-example]$ 
```

0 Hadoop Map/Reduce Administration

Tracking Jobs

Quick Links

State: RUNNING
Started: Wed Mar 28 01:58:22 EDT 2018
Version: 2.0.0-mr1-cdh4.1.1, Unknown
Compiled: Tue Oct 16 11:50:49 PDT 2012 by jenkins from Unknown
Identifier: 201803280158

Cluster Summary (Heap Size is 15.56 MB/193.38 MB)

Running Map Tasks	Running Reduce Tasks	Total Submissions	Nodes	Occupied Map Slots	Occupied Reduce Slots	Reserved Map Slots	Reserved Reduce Slots	Map Task Capacity	Reduce Task Capacity	Avg. Tasks/Node	Blacklisted Nodes	Excluded Nodes
0	0	2	1	0	0	0	0	2	2	4.00	0	0

Scheduling Information

Queue Name	State	Scheduling Information
default	running	N/A

Filter (Jobid, Priority, User, Name)

Example: 'user:smith 3200' will filter by 'smith' only in the user field and '3200' in all fields

Running Jobs

none

Completed Jobs

Jobid	Priority	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce Total	Reduces Completed	Job Scheduling Information	Diagnostic Info
job_201803280158_0001	NORMAL	training	word count	100.00%	2	2	100.00%	1	1	NA	NA
job_201803280158_0002	NORMAL	training	streamjob5578821669672025240.jar	100.00%	2	2	100.00%	1	1	NA	NA

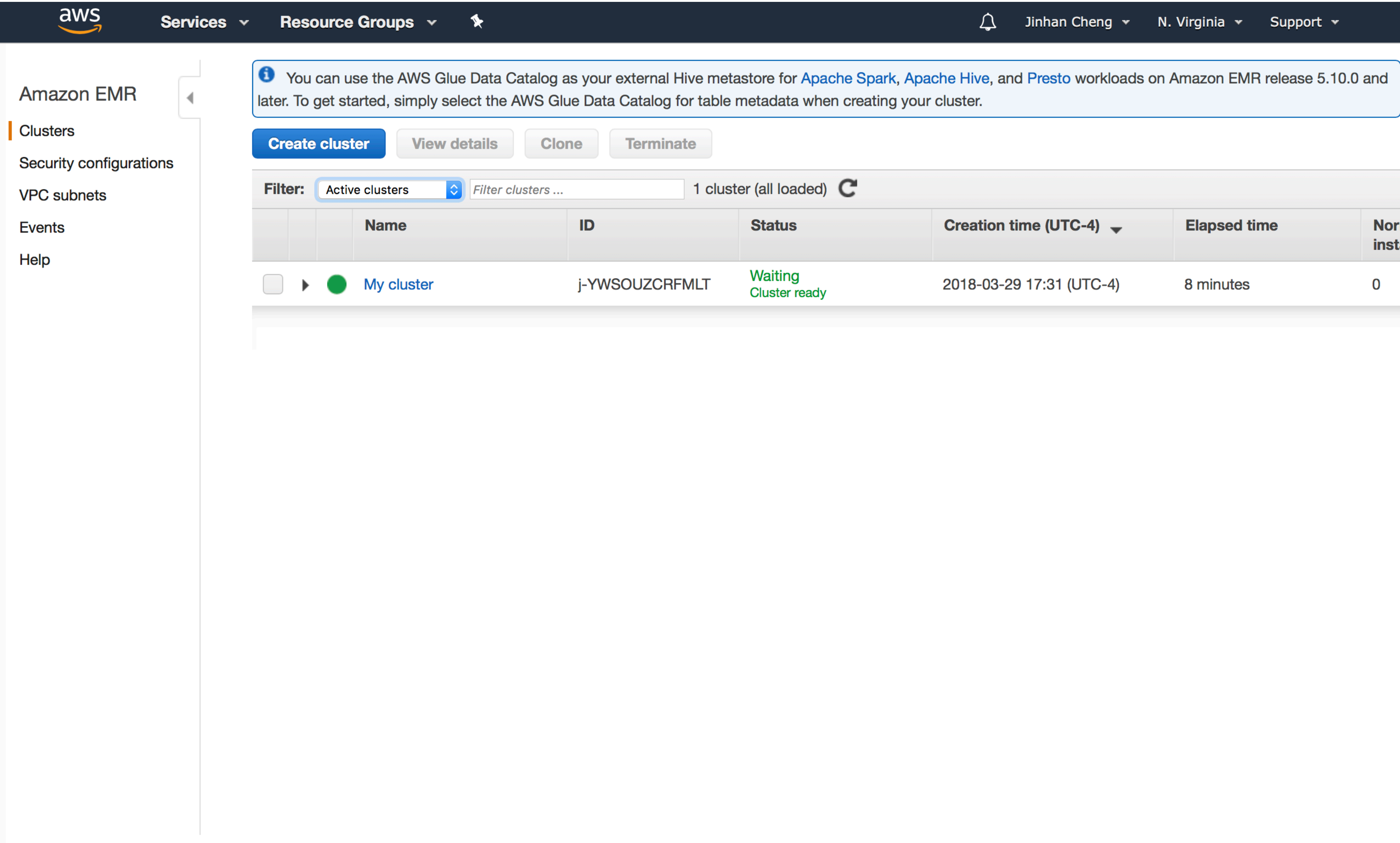
Retired Jobs

none

Local Logs

Log directory, Job Tracker History

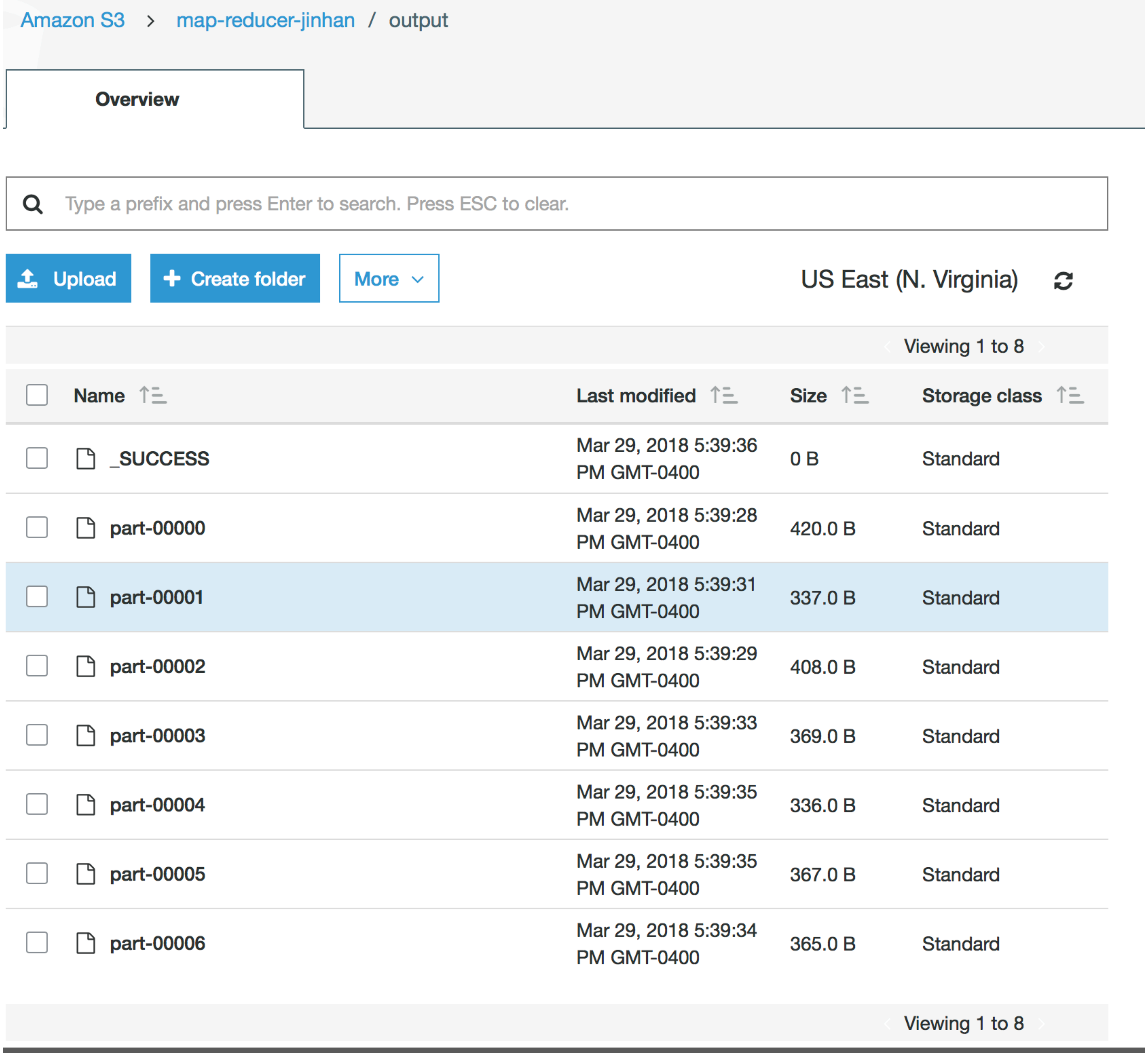
Hadoop, 2018.



You can use the AWS Glue Data Catalog as your external Hive metastore for Apache Spark, Apache Hive, and Presto workloads on Amazon EMR release 5.10.0 and later. To get started, simply select the AWS Glue Data Catalog for table metadata when creating your cluster.

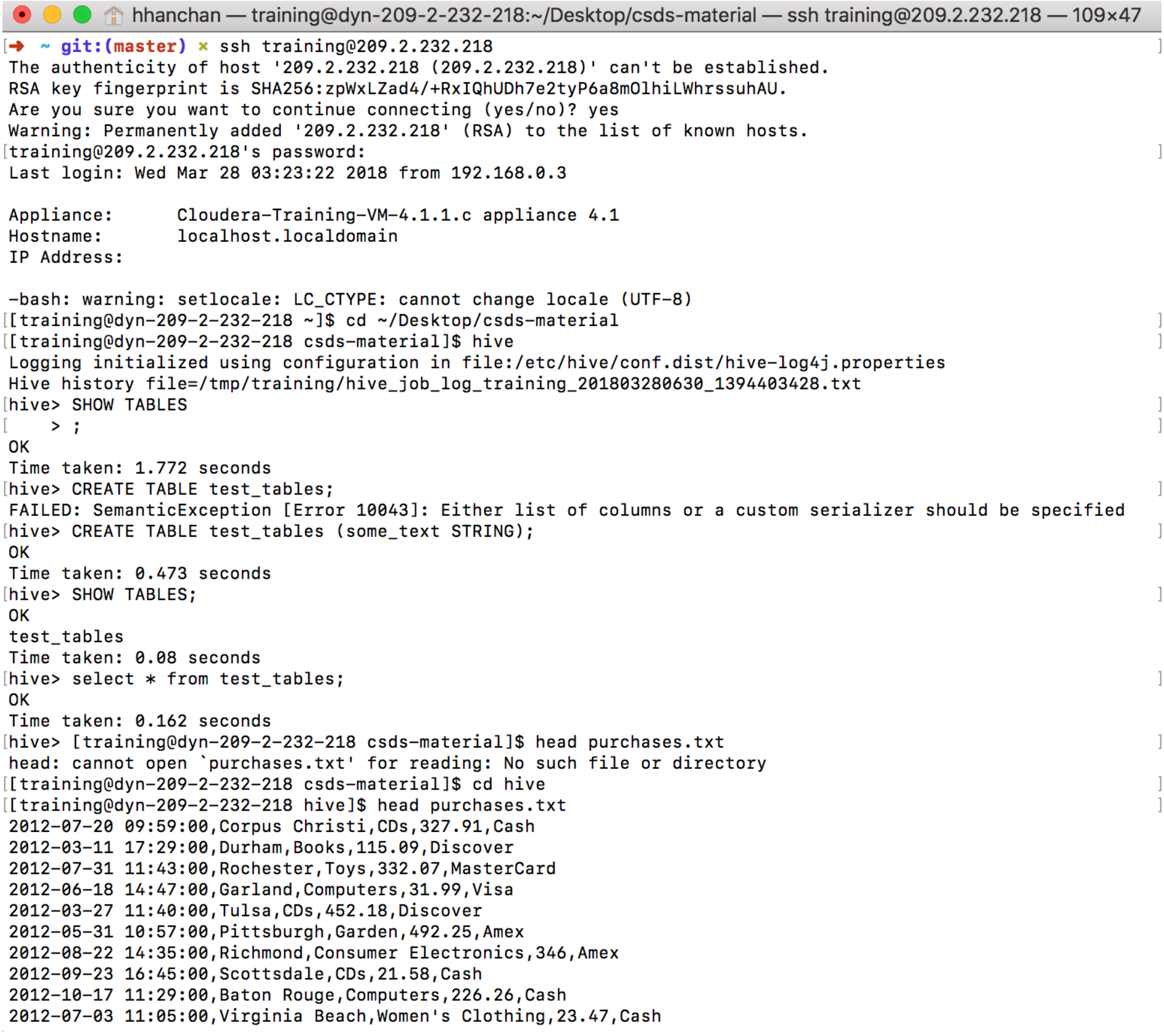
Filter: Active clusters | Filter clusters ... | 1 cluster (all loaded)

			Name	ID	Status	Creation time (UTC-4)	Elapsed time	Nor inst
□	▶	●	My cluster	j-YWSOUZCRFMLT	Waiting Cluster ready	2018-03-29 17:31 (UTC-4)	8 minutes	0


Amazon S3 > map-reducer-jinhan / output
Overview
Type a prefix and press Enter to search. Press ESC to clear.
Upload
+ Create folder
More
US East (N. Virginia)
Viewing 1 to 8

□	Name	Last modified	Size	Storage class
□	_SUCCESS	Mar 29, 2018 5:39:36 PM GMT-0400	0 B	Standard
□	part-00000	Mar 29, 2018 5:39:28 PM GMT-0400	420.0 B	Standard
□	part-00001	Mar 29, 2018 5:39:31 PM GMT-0400	337.0 B	Standard
□	part-00002	Mar 29, 2018 5:39:29 PM GMT-0400	408.0 B	Standard
□	part-00003	Mar 29, 2018 5:39:33 PM GMT-0400	369.0 B	Standard
□	part-00004	Mar 29, 2018 5:39:35 PM GMT-0400	336.0 B	Standard
□	part-00005	Mar 29, 2018 5:39:35 PM GMT-0400	367.0 B	Standard
□	part-00006	Mar 29, 2018 5:39:34 PM GMT-0400	365.0 B	Standard

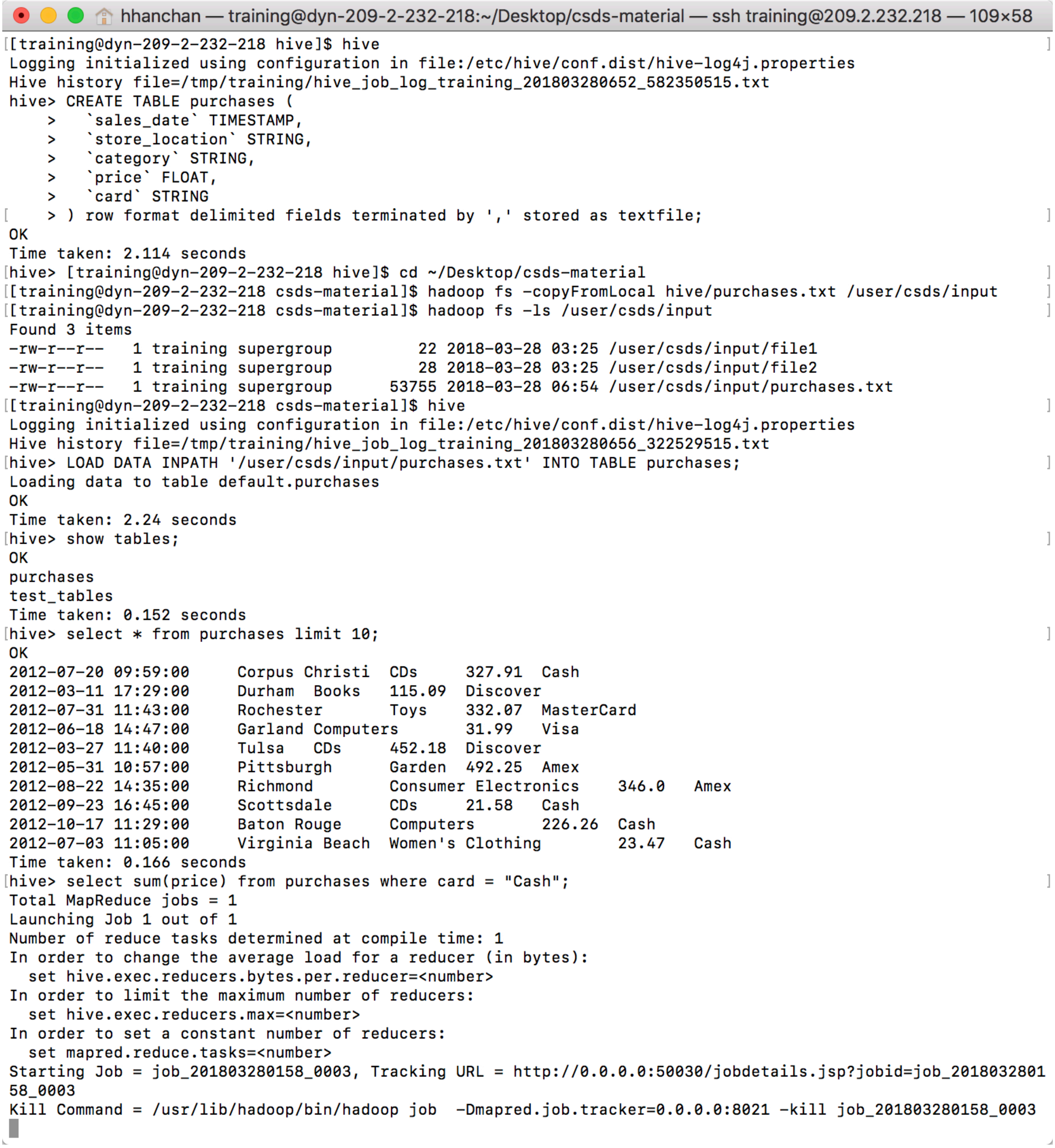
Hive


hhanchan — training@dyn-209-2-232-218:~/Desktop/csds-material — ssh training@209.2.232.218 — 109×47

```
[➜  ~ git:(master) ✗ ssh training@209.2.232.218
 The authenticity of host '209.2.232.218 (209.2.232.218)' can't be established.
 RSA key fingerprint is SHA256:zpWxLZad4/+RxIQhUDh7e2tyP6a8mOlhiLWhrssuhAU.
 Are you sure you want to continue connecting (yes/no)? yes
 Warning: Permanently added '209.2.232.218' (RSA) to the list of known hosts.
[training@209.2.232.218's password:
 Last login: Wed Mar 28 03:23:22 2018 from 192.168.0.3

 Appliance:      Cloudera-Training-VM-4.1.1.c appliance 4.1
 Hostname:       localhost.localdomain
 IP Address:

 -bash: warning: setlocale: LC_CTYPE: cannot change locale (UTF-8)
[[training@dyn-209-2-232-218 ~]$ cd ~/Desktop/csds-material
[[training@dyn-209-2-232-218 csds-material]$ hive
 Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
 Hive history file=/tmp/training/hive_job_log_training_201803280630_1394403428.txt
[hive> SHOW TABLES
[    > ;
 OK
 Time taken: 1.772 seconds
[hive> CREATE TABLE test_tables;
 FAILED: SemanticException [Error 10043]: Either list of columns or a custom serializer should be specified
[hive> CREATE TABLE test_tables (some_text STRING);
 OK
 Time taken: 0.473 seconds
[hive> SHOW TABLES;
 OK
 test_tables
 Time taken: 0.08 seconds
[hive> select * from test_tables;
 OK
 Time taken: 0.162 seconds
[hive> [training@dyn-209-2-232-218 csds-material]$ head purchases.txt
 head: cannot open `purchases.txt' for reading: No such file or directory
[[training@dyn-209-2-232-218 csds-material]$ cd hive
[[training@dyn-209-2-232-218 hive]$ head purchases.txt
 2012-07-20 09:59:00,Corpus Christi,CDs,327.91,Cash
 2012-03-11 17:29:00,Durham,Books,115.09,Discover
 2012-07-31 11:43:00,Rochester,Toys,332.07,MasterCard
 2012-06-18 14:47:00,Garland,Computers,31.99,Visa
 2012-03-27 11:40:00,Tulsa,CDs,452.18,Discover
 2012-05-31 10:57:00,Pittsburgh,Garden,492.25,Amex
 2012-08-22 14:35:00,Richmond,Consumer Electronics,346,Amex
 2012-09-23 16:45:00,Scottsdale,CDs,21.58,Cash
 2012-10-17 11:29:00,Baton Rouge,Computers,226.26,Cash
 2012-07-03 11:05:00,Virginia Beach,Women's Clothing,23.47,Cash
```


hhanchan — training@dyn-209-2-232-218:~/Desktop/csds-material — ssh training@209.2.232.218 — 109×58

```
[training@dyn-209-2-232-218 hive]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
Hive history file=/tmp/training/hive_job_log_training_201803280652_582350515.txt
hive> CREATE TABLE purchases (
    >  `sales_date` TIMESTAMP,
    >  `store_location` STRING,
    >  `category` STRING,
    >  `price` FLOAT,
    >  `card` STRING
    > ) row format delimited fields terminated by ',' stored as textfile;
OK
Time taken: 2.114 seconds
hive> [training@dyn-209-2-232-218 hive]$ cd ~/Desktop/csds-material
[training@dyn-209-2-232-218 csds-material]$ hadoop fs -copyFromLocal hive/purchases.txt /user/csds/input
[training@dyn-209-2-232-218 csds-material]$ hadoop fs -ls /user/csds/input
Found 3 items
-rw-r--r--   1 training supergroup         22 2018-03-28 03:25 /user/csds/input/file1
-rw-r--r--   1 training supergroup         28 2018-03-28 03:25 /user/csds/input/file2
-rw-r--r--   1 training supergroup      53755 2018-03-28 06:54 /user/csds/input/purchases.txt
[training@dyn-209-2-232-218 csds-material]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
Hive history file=/tmp/training/hive_job_log_training_201803280656_322529515.txt
hive> LOAD DATA INPATH '/user/csds/input/purchases.txt' INTO TABLE purchases;
Loading data to table default.purchases
OK
Time taken: 2.24 seconds
hive> show tables;
OK
purchases
test_tables
Time taken: 0.152 seconds
hive> select * from purchases limit 10;
OK
2012-07-20 09:59:00     Corpus Christi  CDs     327.91  Cash
2012-03-11 17:29:00     Durham  Books   115.09  Discover
2012-07-31 11:43:00     Rochester       Toys    332.07  MasterCard
2012-06-18 14:47:00     Garland Computers       31.99   Visa
2012-03-27 11:40:00     Tulsa   CDs     452.18  Discover
2012-05-31 10:57:00     Pittsburgh      Garden  492.25  Amex
2012-08-22 14:35:00     Richmond        Consumer Electronics    346.0   Amex
2012-09-23 16:45:00     Scottsdale      CDs     21.58   Cash
2012-10-17 11:29:00     Baton Rouge     Computers       226.26  Cash
2012-07-03 11:05:00     Virginia Beach  Women's Clothing        23.47   Cash
Time taken: 0.166 seconds
hive> select sum(price) from purchases where card = "Cash";
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_201803280158_0003, Tracking URL = http://0.0.0.0:50030/jobdetails.jsp?jobid=job_2018032801
58_0003
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=0.0.0.0:8021 -kill job_201803280158_0003
```