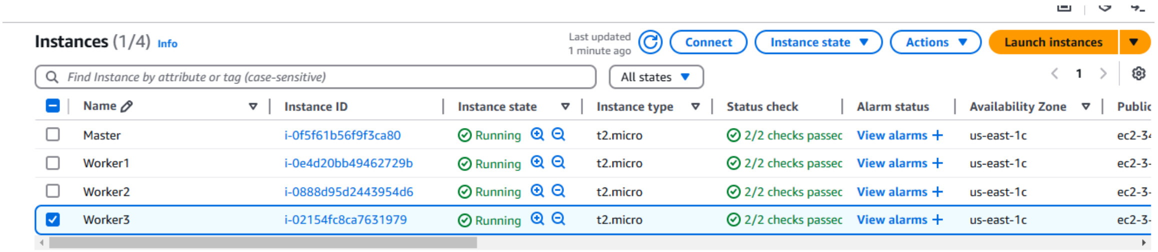
My Apache Spark Cluster

Master, 3 workers

On AWS EC2 Ubuntu Instance



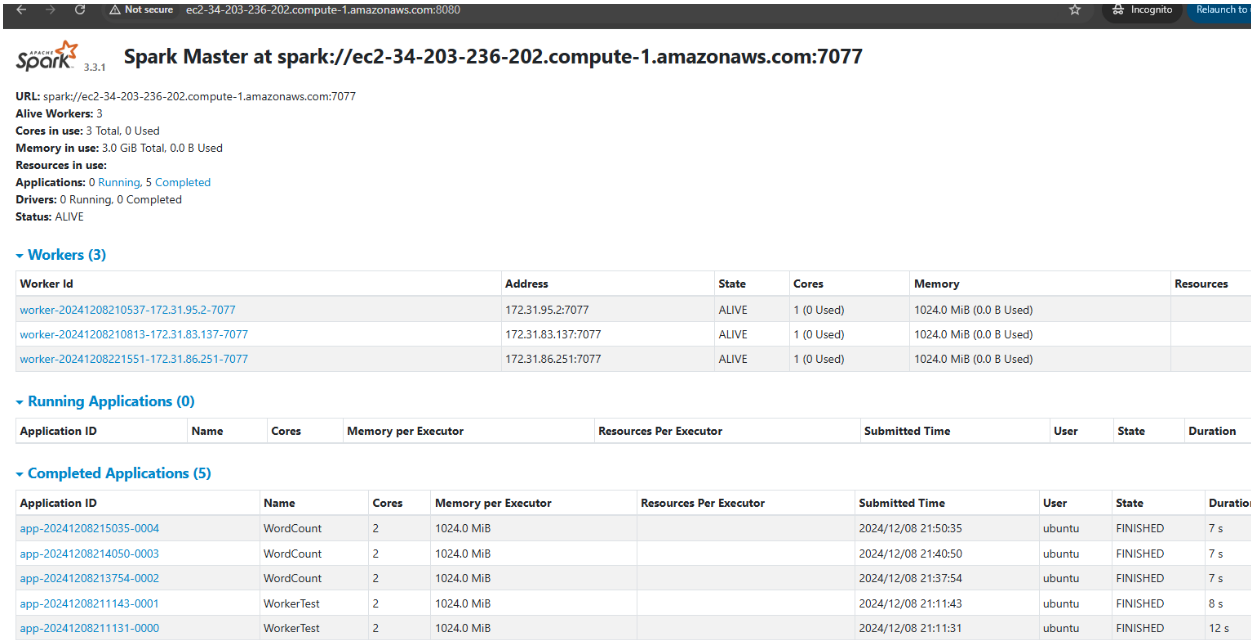
	Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone	Public
□	Master	i-0f5f61b56f9f3ca80	Running	t2.micro	2/2 checks passec	View alarms +	us-east-1c	ec2-3
□	Worker1	i-0e4d20bb49462729b	Running	t2.micro	2/2 checks passec	View alarms +	us-east-1c	ec2-3-
□	Worker2	i-0888d95d2443954d6	Running	t2.micro	2/2 checks passec	View alarms +	us-east-1c	ec2-3-
☑	Worker3	i-02154fc8ca7631979	Running	t2.micro	2/2 checks passec	View alarms +	us-east-1c	ec2-3-

C

I created EC2 instances, Installed java, downloaded apache spark, set environment variables for SPARK_HOME, PATH

I configured SPARK Environment variables and associated shell scripts to start and stop cluster.

My cluster is up and running



Spark Master at spark://ec2-34-203-236-202.compute-1.amazonaws.com:7077

URL: spark://ec2-34-203-236-202.compute-1.amazonaws.com:7077
Alive Workers: 3
Cores in use: 3 Total, 0 Used
Memory in use: 3.0 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 5 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (3)

Worker Id	Address	State	Cores	Memory	Resources
worker-20241208210537-172.31.95.2-7077	172.31.95.2:7077	ALIVE	1 (0 Used)	1024.0 MiB (0.0 B Used)	
worker-20241208210813-172.31.83.137-7077	172.31.83.137:7077	ALIVE	1 (0 Used)	1024.0 MiB (0.0 B Used)	
worker-20241208221551-172.31.86.251-7077	172.31.86.251:7077	ALIVE	1 (0 Used)	1024.0 MiB (0.0 B Used)	

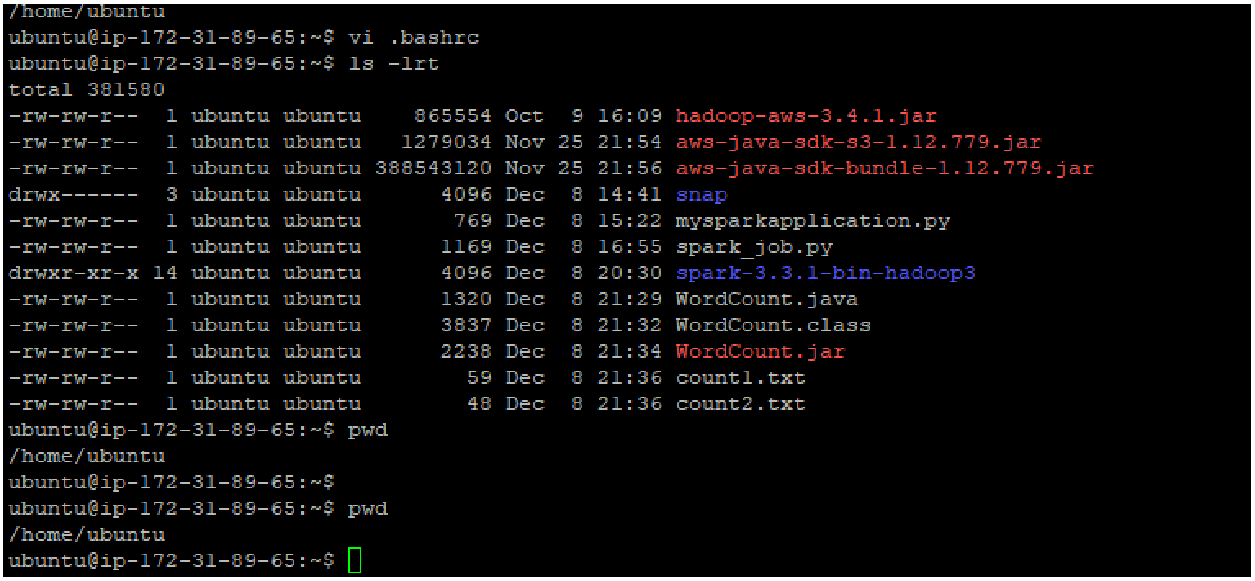
Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

Completed Applications (5)

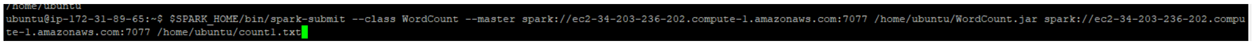
Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duratio
app-20241208215035-0004	WordCount	2	1024.0 MiB		2024/12/08 21:50:35	ubuntu	FINISHED	7 s
app-20241208214050-0003	WordCount	2	1024.0 MiB		2024/12/08 21:40:50	ubuntu	FINISHED	7 s
app-20241208213754-0002	WordCount	2	1024.0 MiB		2024/12/08 21:37:54	ubuntu	FINISHED	7 s
app-20241208211143-0001	WorkerTest	2	1024.0 MiB		2024/12/08 21:11:43	ubuntu	FINISHED	8 s
app-20241208211131-0000	WorkerTest	2	1024.0 MiB		2024/12/08 21:11:31	ubuntu	FINISHED	12 s

My java and Pyspark sample files



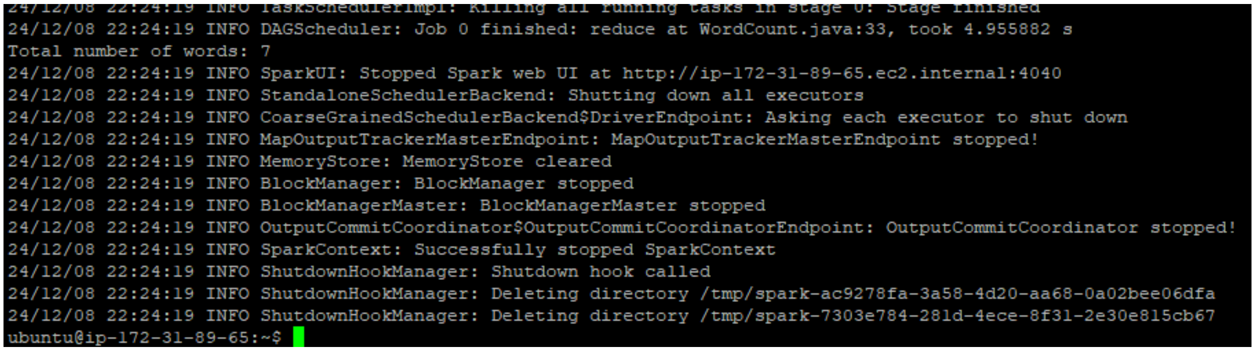
```
/home/ubuntu
ubuntu@ip-172-31-89-65:~$ vi .bashrc
ubuntu@ip-172-31-89-65:~$ ls -lrt
total 381580
-rw-rw-r--  1 ubuntu ubuntu    865554 Oct  9 16:09 hadoop-aws-3.4.1.jar
-rw-rw-r--  1 ubuntu ubuntu   1279034 Nov 25 21:54 aws-java-sdk-s3-1.12.779.jar
-rw-rw-r--  1 ubuntu ubuntu 388543120 Nov 25 21:56 aws-java-sdk-bundle-1.12.779.jar
drwx------  3 ubuntu ubuntu      4096 Dec  8 14:41 snap
-rw-rw-r--  1 ubuntu ubuntu       769 Dec  8 15:22 mysparkapplication.py
-rw-rw-r--  1 ubuntu ubuntu      1169 Dec  8 16:55 spark_job.py
drwxr-xr-x 14 ubuntu ubuntu      4096 Dec  8 20:30 spark-3.3.1-bin-hadoop3
-rw-rw-r--  1 ubuntu ubuntu      1320 Dec  8 21:29 WordCount.java
-rw-rw-r--  1 ubuntu ubuntu      3837 Dec  8 21:32 WordCount.class
-rw-rw-r--  1 ubuntu ubuntu      2238 Dec  8 21:34 WordCount.jar
-rw-rw-r--  1 ubuntu ubuntu        59 Dec  8 21:36 count1.txt
-rw-rw-r--  1 ubuntu ubuntu        48 Dec  8 21:36 count2.txt
ubuntu@ip-172-31-89-65:~$ pwd
/home/ubuntu
ubuntu@ip-172-31-89-65:~$
ubuntu@ip-172-31-89-65:~$ pwd
/home/ubuntu
ubuntu@ip-172-31-89-65:~$
```

Submitting sparkjob

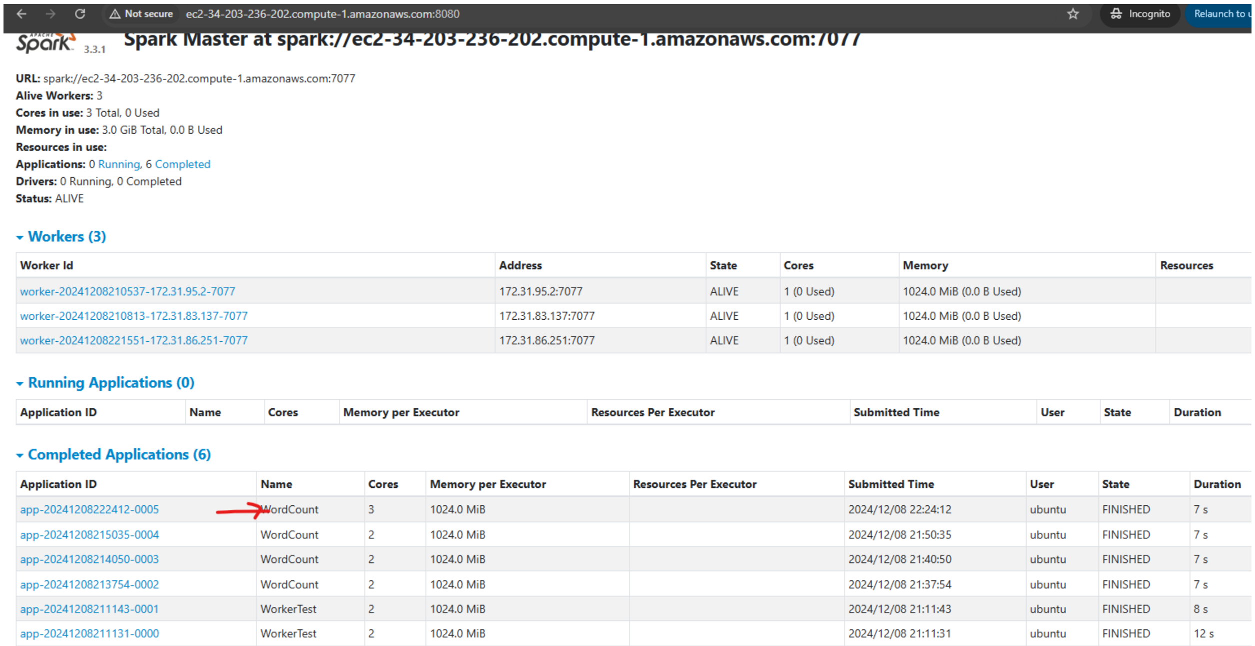


```
/home/ubuntu
ubuntu@ip-172-31-89-65:~$ $SPARK_HOME/bin/spark-submit --class WordCount --master spark://ec2-34-203-236-202.compute-1.amazonaws.com:7077 /home/ubuntu/WordCount.jar spark://ec2-34-203-236-202.compu
te-1.amazonaws.com:7077 /home/ubuntu/count1.txt
```

WordCount class run



```
24/12/08 22:24:19 INFO TaskSchedulerImpl: Killing all running tasks in stage 0: Stage finished
24/12/08 22:24:19 INFO DAGScheduler: Job 0 finished: reduce at WordCount.java:33, took 4.955882 s
Total number of words: 7
24/12/08 22:24:19 INFO SparkUI: Stopped Spark web UI at http://ip-172-31-89-65.ec2.internal:4040
24/12/08 22:24:19 INFO StandaloneSchedulerBackend: Shutting down all executors
24/12/08 22:24:19 INFO CoarseGrainedSchedulerBackend$DriverEndpoint: Asking each executor to shut down
24/12/08 22:24:19 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
24/12/08 22:24:19 INFO MemoryStore: MemoryStore cleared
24/12/08 22:24:19 INFO BlockManager: BlockManager stopped
24/12/08 22:24:19 INFO BlockManagerMaster: BlockManagerMaster stopped
24/12/08 22:24:19 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
24/12/08 22:24:19 INFO SparkContext: Successfully stopped SparkContext
24/12/08 22:24:19 INFO ShutdownHookManager: Shutdown hook called
24/12/08 22:24:19 INFO ShutdownHookManager: Deleting directory /tmp/spark-ac9278fa-3a58-4d20-aa68-0a02bee06dfa
24/12/08 22:24:19 INFO ShutdownHookManager: Deleting directory /tmp/spark-7303e784-281d-4ece-8f31-2e30e815cb67
ubuntu@ip-172-31-89-65:~$
```



Spark Master at spark://ec2-34-203-236-202.compute-1.amazonaws.com:7077

URL: spark://ec2-34-203-236-202.compute-1.amazonaws.com:7077
Alive Workers: 3
Cores in use: 3 Total, 0 Used
Memory in use: 3.0 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 6 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (3)

Worker Id	Address	State	Cores	Memory	Resources
worker-20241208210537-172.31.95.2-7077	172.31.95.2:7077	ALIVE	1 (0 Used)	1024.0 MiB (0.0 B Used)	
worker-20241208210813-172.31.83.137-7077	172.31.83.137:7077	ALIVE	1 (0 Used)	1024.0 MiB (0.0 B Used)	
worker-20241208221551-172.31.86.251-7077	172.31.86.251:7077	ALIVE	1 (0 Used)	1024.0 MiB (0.0 B Used)	

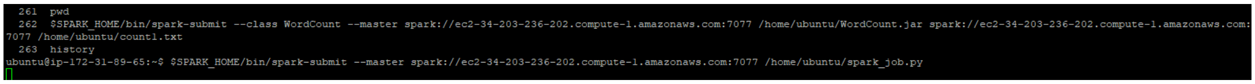
Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

Completed Applications (6)

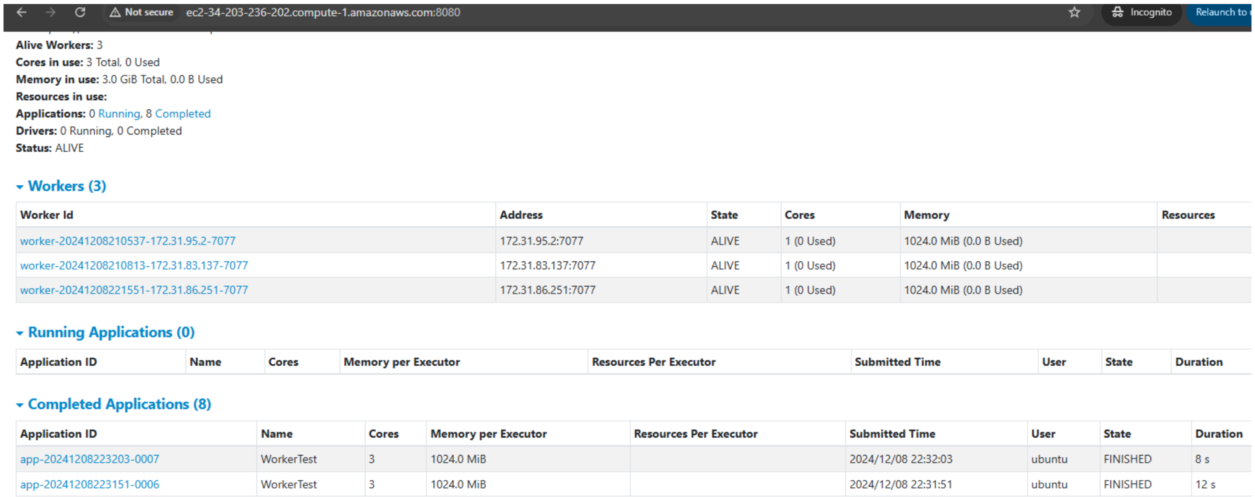
Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20241208222412-0005	WordCount	3	1024.0 MiB		2024/12/08 22:24:12	ubuntu	FINISHED	7 s
app-20241208215035-0004	WordCount	2	1024.0 MiB		2024/12/08 21:50:35	ubuntu	FINISHED	7 s
app-20241208214050-0003	WordCount	2	1024.0 MiB		2024/12/08 21:40:50	ubuntu	FINISHED	7 s
app-20241208213754-0002	WordCount	2	1024.0 MiB		2024/12/08 21:37:54	ubuntu	FINISHED	7 s
app-20241208211143-0001	WorkerTest	2	1024.0 MiB		2024/12/08 21:11:43	ubuntu	FINISHED	8 s
app-20241208211131-0000	WorkerTest	2	1024.0 MiB		2024/12/08 21:11:31	ubuntu	FINISHED	12 s

Pyspark job



```
  261  pwd
  262  $SPARK_HOME/bin/spark-submit --class WordCount --master spark://ec2-34-203-236-202.compute-1.amazonaws.com:7077 /home/ubuntu/WordCount.jar spark://ec2-34-203-236-202.compute-1.amazonaws.com:
7077 /home/ubuntu/count1.txt
  263  history
ubuntu@ip-172-31-89-65:~$ $SPARK_HOME/bin/spark-submit --master spark://ec2-34-203-236-202.compute-1.amazonaws.com:7077 /home/ubuntu/spark_job.py
```

Ran on 2 nodes for million rows



Alive Workers: 3
Cores in use: 3 Total, 0 Used
Memory in use: 3.0 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 8 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (3)

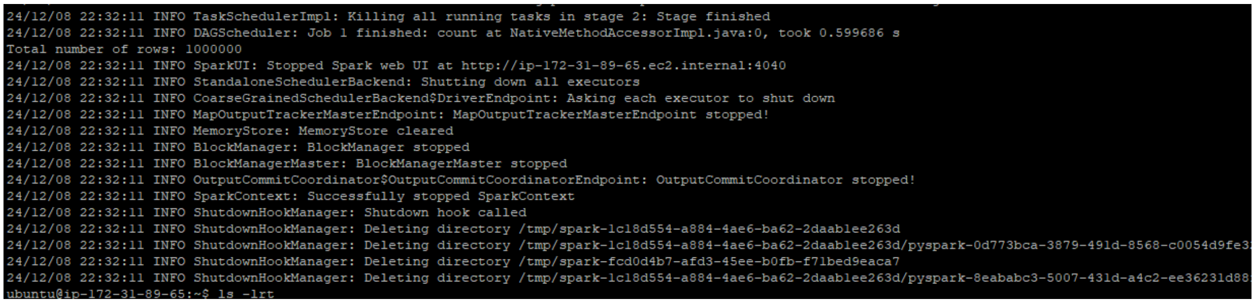
Worker Id	Address	State	Cores	Memory	Resources
worker-20241208210537-172.31.95.2-7077	172.31.95.2:7077	ALIVE	1 (0 Used)	1024.0 MiB (0.0 B Used)	
worker-20241208210813-172.31.83.137-7077	172.31.83.137:7077	ALIVE	1 (0 Used)	1024.0 MiB (0.0 B Used)	
worker-20241208221551-172.31.86.251-7077	172.31.86.251:7077	ALIVE	1 (0 Used)	1024.0 MiB (0.0 B Used)	

Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

Completed Applications (8)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20241208223203-0007	WorkerTest	3	1024.0 MiB		2024/12/08 22:32:03	ubuntu	FINISHED	8 s
app-20241208223151-0006	WorkerTest	3	1024.0 MiB		2024/12/08 22:31:51	ubuntu	FINISHED	12 s



```
24/12/08 22:32:11 INFO TaskSchedulerImpl: Killing all running tasks in stage 2: Stage finished
24/12/08 22:32:11 INFO DAGScheduler: Job 1 finished: count at NativeMethodAccessorImpl.java:0, took 0.599686 s
Total number of rows: 1000000
24/12/08 22:32:11 INFO SparkUI: Stopped Spark web UI at http://ip-172-31-89-65.ec2.internal:4040
24/12/08 22:32:11 INFO StandaloneSchedulerBackend: Shutting down all executors
24/12/08 22:32:11 INFO CoarseGrainedSchedulerBackend$DriverEndpoint: Asking each executor to shut down
24/12/08 22:32:11 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
24/12/08 22:32:11 INFO MemoryStore: MemoryStore cleared
24/12/08 22:32:11 INFO BlockManager: BlockManager stopped
24/12/08 22:32:11 INFO BlockManagerMaster: BlockManagerMaster stopped
24/12/08 22:32:11 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
24/12/08 22:32:11 INFO SparkContext: Successfully stopped SparkContext
24/12/08 22:32:11 INFO ShutdownHookManager: Shutdown hook called
24/12/08 22:32:11 INFO ShutdownHookManager: Deleting directory /tmp/spark-1c18d554-a884-4ae6-ba62-2daablee263d
24/12/08 22:32:11 INFO ShutdownHookManager: Deleting directory /tmp/spark-1c18d554-a884-4ae6-ba62-2daablee263d/pyspark-0d773bca-3879-491d-8568-c0054d9fe3
24/12/08 22:32:11 INFO ShutdownHookManager: Deleting directory /tmp/spark-fcd0d4b7-afd3-45ee-b0fb-f71bed9eaca7
24/12/08 22:32:11 INFO ShutdownHookManager: Deleting directory /tmp/spark-1c18d554-a884-4ae6-ba62-2daablee263d/pyspark-8eababc3-5007-431d-a4c2-ee36231d88
ubuntu@ip-172-31-89-65:~$ ls -lrt
```