Hadoop HDFS & MapReduce

docker run command

```
172-16-40-107:~ rajdaiya$ docker stop hadoop
hadoop
172-16-40-107:~ rajdaiya$ docker rm hadoop
hadoop
172-16-40-107:~ rajdaiya$  docker run --hostname=quickstart.cloudera --privileged=true -t -i -d -p 8888:8888 -p 8000:80 -P --name hadoop cloudera/quickstart:latest /usr/bin/docker-quickstart
4c3e083736a3a748bbaaf2abe106bea7b8cb87d843daaeec7dcb2903e2b71aaa
172-16-40-107:~ rajdaiya$
```

docker

```
Last login: Thu Feb 22 09:55:11 on ttys000
Rajs-MacBook-Pro:~ rajdaiya$ docker

Usage:  docker COMMAND

A self-sufficient runtime for containers

Options:
      --config string      Location of client config files (default
                           "/Users/rajdaiya/.docker")
  -D, --debug              Enable debug mode
  -H, --host list          Daemon socket(s) to connect to
  -l, --log-level string   Set the logging level
                           ("debug"|"info"|"warn"|"error"|"fatal")
                           (default "info")
      --tls                Use TLS; implied by --tlsverify
      --tlscacert string   Trust certs signed only by this CA (default
                           "/Users/rajdaiya/.docker/ca.pem")
      --tlscert string     Path to TLS certificate file (default
                           "/Users/rajdaiya/.docker/cert.pem")
      --tlskey string      Path to TLS key file (default
                           "/Users/rajdaiya/.docker/key.pem")
      --tlsverify          Use TLS and verify the remote
  -v, --version            Print version information and quit

Management Commands:
  checkpoint  Manage checkpoints
  config      Manage Docker configs
  container   Manage containers
  image       Manage images
  network     Manage networks
  node        Manage Swarm nodes
  plugin      Manage plugins
  secret      Manage Docker secrets
  service     Manage services
  stack       Manage Docker stacks
  swarm       Manage Swarm
  system      Manage Docker
  trust       Manage trust on Docker images (experimental)
  volume      Manage volumes

Commands:
  attach      Attach local standard input, output, and error streams to a running container
  build       Build an image from a Dockerfile
  commit      Create a new image from a container's changes
  cp          Copy files/folders between a container and the local filesystem
  create      Create a new container
  deploy      Deploy a new stack or update an existing stack
  diff        Inspect changes to files or directories on a container's filesystem
  events      Get real time events from the server
  exec        Run a command in a running container
  export      Export a container's filesystem as a tar archive
  history     Show the history of an image
  images      List images
  import      Import the contents from a tarball to create a filesystem image
  info        Display system-wide information
  inspect     Return low-level information on Docker objects
  kill        Kill one or more running containers
  load        Load an image from a tar archive or STDIN
  login       Log in to a Docker registry
  logout      Log out from a Docker registry
  logs        Fetch the logs of a container
  pause       Pause all processes within one or more containers
```

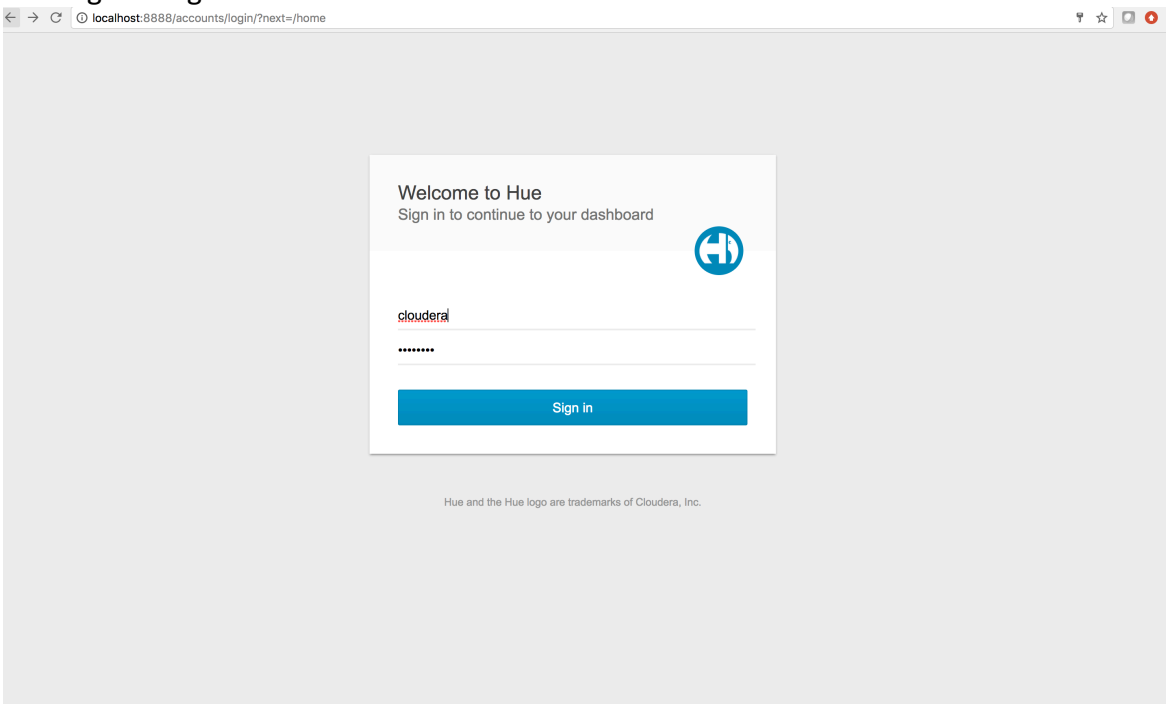
docker version command

```
  build       Build an image from a Dockerfile
  commit      Create a new image from a container's changes
  cp          Copy files/folders between a container and the local filesystem
  create      Create a new container
  deploy      Deploy a new stack or update an existing stack
  diff        Inspect changes to files or directories on a container's filesystem
  events      Get real time events from the server
  exec        Run a command in a running container
  export      Export a container's filesystem as a tar archive
  history     Show the history of an image
  images      List images
  import      Import the contents from a tarball to create a filesystem image
  info        Display system-wide information
  inspect     Return low-level information on Docker objects
  kill        Kill one or more running containers
  load        Load an image from a tar archive or STDIN
  login       Log in to a Docker registry
  logout      Log out from a Docker registry
  logs        Fetch the logs of a container
  pause       Pause all processes within one or more containers
  port        List port mappings or a specific mapping for the container
  ps          List containers
  pull        Pull an image or a repository from a registry
  push        Push an image or a repository to a registry
  rename      Rename a container
  restart     Restart one or more containers
  rm          Remove one or more containers
  rmi         Remove one or more images
  run         Run a command in a new container
  save        Save one or more images to a tar archive (streamed to STDOUT by default)
  search      Search the Docker Hub for images
  start       Start one or more stopped containers
  stats       Display a live stream of container(s) resource usage statistics
  stop        Stop one or more running containers
  tag         Create a tag TARGET_IMAGE that refers to SOURCE_IMAGE
  top         Display the running processes of a container
  unpause     Unpause all processes within one or more containers
  update      Update configuration of one or more containers
  version     Show the Docker version information
  wait        Block until one or more containers stop, then print their exit codes

Run 'docker COMMAND --help' for more information on a command.
Rajs-MacBook-Pro:~ rajdaiya$ docker version
Client:
 Version:      17.12.0-ce
 API version:  1.35
 Go version:   go1.9.2
 Git commit:   c97c6d6
 Built: Wed Dec 27 20:03:51 2017
 OS/Arch:      darwin/amd64

Server:
 Engine:
  Version:     17.12.0-ce
  API version: 1.35 (minimum version 1.12)
  Go version:  go1.9.2
  Git commit:  c97c6d6
  Built:       Wed Dec 27 20:12:29 2017
  OS/Arch:     linux/amd64
  Experimental: true
Rajs-MacBook-Pro:~ rajdaiya$ docker ps
CONTAINER ID        IMAGE                          COMMAND                  CREATED             STATUS              PORTS                                                NAMES
771dfa085eaf        cloudera/quickstart:latest   "/usr/bin/docker-qui…"   45 hours ago        Up 42 minutes       0.0.0.0:8888->8888/tcp, 0.0.0.0:8000->80/tcp   hadoop
```

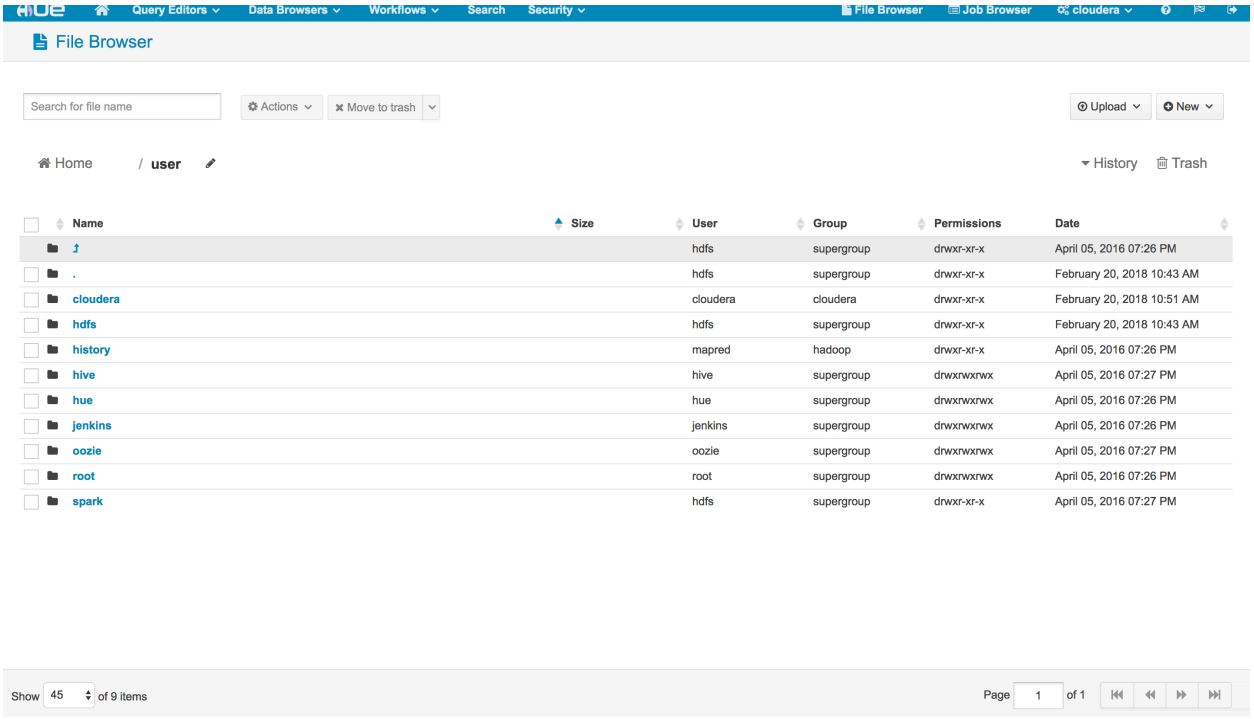
Running a version of Hadoop (docker or otherwise), create a directory on HDFS with your first name (e.g. mine will be ‘juan’). Submit a screen grab of the output of a Hadoop fs listing showing your home directory **and** your new directory in it

```
  rmi         Remove one or more images
  run         Run a command in a new container
  save        Save one or more images to a tar archive (streamed to STDOUT by default)
  search      Search the Docker Hub for images
  start       Start one or more stopped containers
  stats       Display a live stream of container(s) resource usage statistics
  stop        Stop one or more running containers
  tag         Create a tag TARGET_IMAGE that refers to SOURCE_IMAGE
  top         Display the running processes of a container
  unpause     Unpause all processes within one or more containers
  update      Update configuration of one or more containers
  version     Show the Docker version information
  wait        Block until one or more containers stop, then print their exit codes

Run 'docker COMMAND --help' for more information on a command.
Rajs-MacBook-Pro:~ rajdaiya$ docker version
Client:
 Version:      17.12.0-ce
 API version:  1.35
 Go version:   go1.9.2
 Git commit:   c97c6d6
 Built: Wed Dec 27 20:03:51 2017
 OS/Arch:      darwin/amd64

Server:
 Engine:
  Version:     17.12.0-ce
  API version: 1.35 (minimum version 1.12)
  Go version:  go1.9.2
  Git commit:  c97c6d6
  Built:       Wed Dec 27 20:12:29 2017
  OS/Arch:     linux/amd64
  Experimental: true
Rajs-MacBook-Pro:~ rajdaiya$ docker ps
CONTAINER ID        IMAGE                          COMMAND                  CREATED             STATUS              PORTS                                                NAMES
771dfa085eaf        cloudera/quickstart:latest   "/usr/bin/docker-qui…"   45 hours ago        Up 42 minutes       0.0.0.0:8888->8888/tcp, 0.0.0.0:8000->80/tcp   hadoop
Rajs-MacBook-Pro:~ rajdaiya$ docker ps -a
CONTAINER ID        IMAGE                          COMMAND                  CREATED             STATUS              PORTS                                                NAMES
771dfa085eaf        cloudera/quickstart:latest   "/usr/bin/docker-qui…"   45 hours ago        Up 42 minutes       0.0.0.0:8888->8888/tcp, 0.0.0.0:8000->80/tcp   hadoop
Rajs-MacBook-Pro:~ rajdaiya$ docker exec -i -t hadoop bash
[root@quickstart /]# hadoop fs -ls
[root@quickstart /]# hadoop version
Hadoop 2.6.0-cdh5.7.0
Subversion http://github.com/cloudera/hadoop -r c00978c67b0d3fe9f3b896b5030741bd40bf541a
Compiled by jenkins on 2016-03-23T18:36Z
Compiled with protoc 2.5.0
From source with checksum b2eabfa328e763c88cb14168f9b372
This command was run using /usr/jars/hadoop-common-2.6.0-cdh5.7.0.jar
[root@quickstart /]# hadoop fsd mkdir rajdaiya
Error: Could not find or load main class fsd
[root@quickstart /]# hadoop fs mkdir rajdaiya
mkdir: Unknown command
Did you mean -mkdir?  This command begins with a dash.
[root@quickstart /]# hadoop fs -mkdir rajdaiya
[root@quickstart /]# hadoop fs -ls
Found 1 items
drwxr-xr-x   - root supergroup          0 2018-02-22 15:02 rajdaiya
[root@quickstart /]# hadoop fs -mkdir raj
[root@quickstart /]# hadoop fs -ls
Found 2 items
drwxr-xr-x   - root supergroup          0 2018-02-22 15:11 raj
drwxr-xr-x   - root supergroup          0 2018-02-22 15:02 rajdaiya
[root@quickstart /]#
```

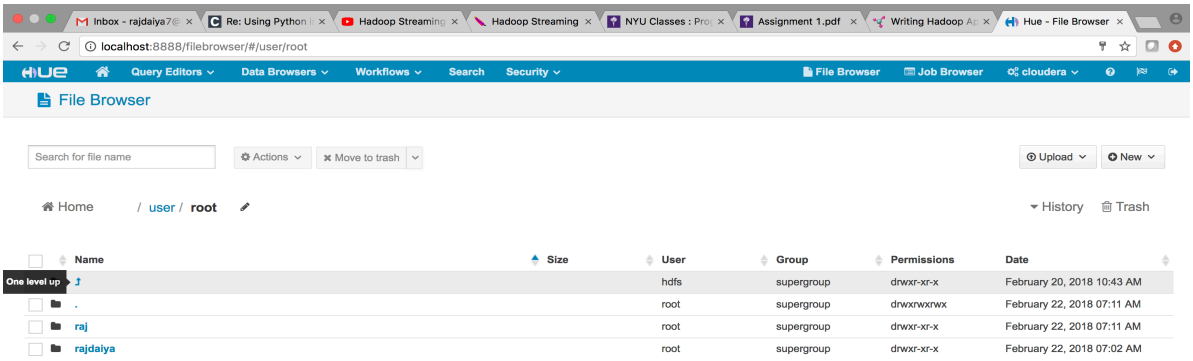
Hue login using cloudera



Users on Hue



The mkdir command directories on Hue

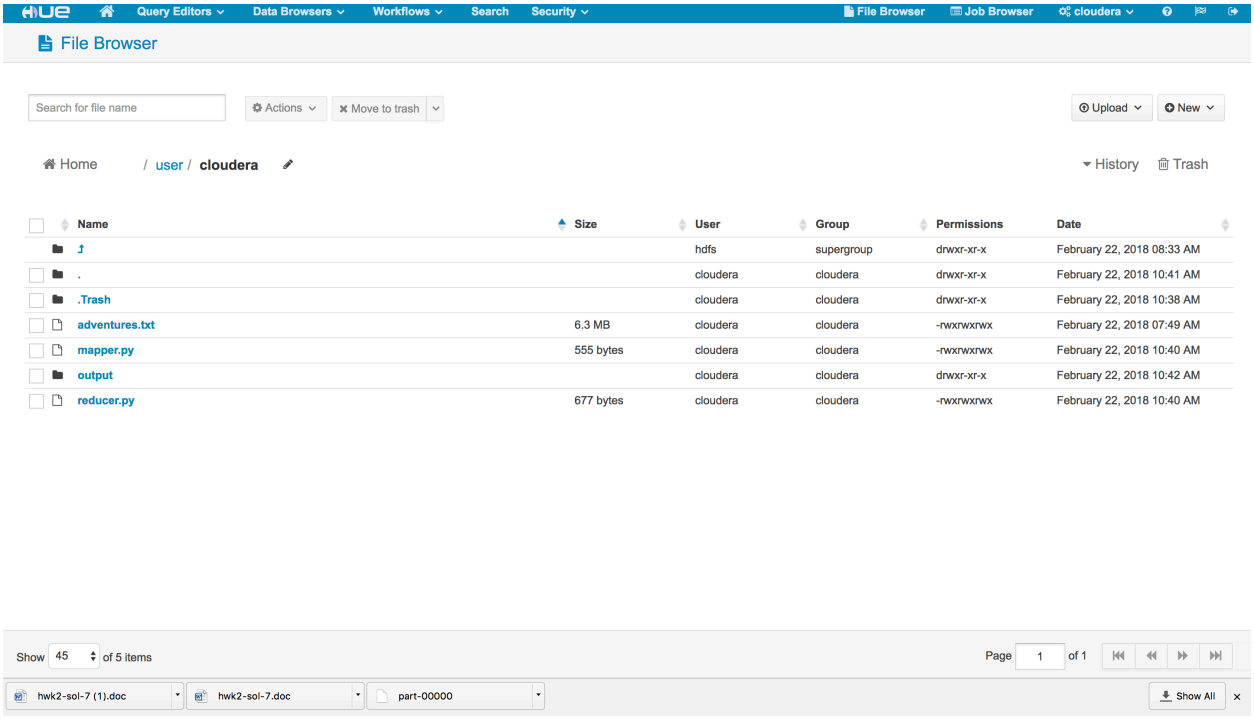


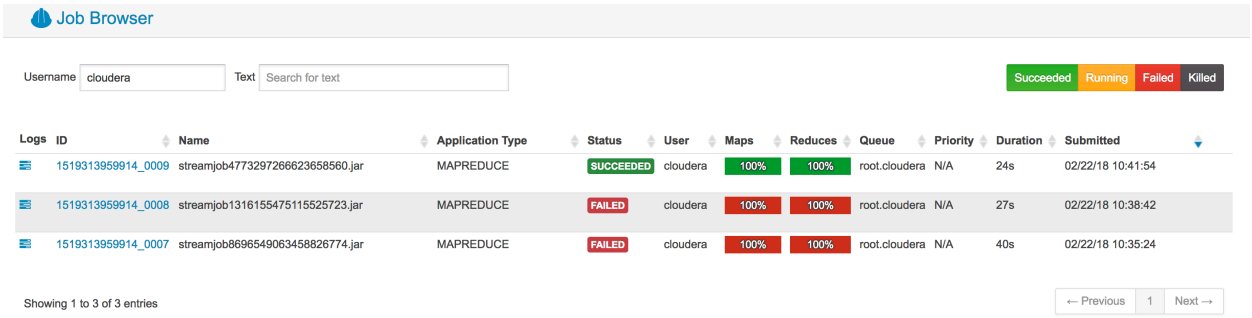
Submit a screen grab of your program running or completed in Hadoop. Hue has a jobs status page, use that one. Or use the command: 'hadoop jar':

- Running Hadoop using Hadoop-streaming:

```
18/02/22 20:39:01 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [mapper.py, reducer.py] [/usr/jars/hadoop-streaming-2.6.0-cdh5.7.0.jar] /tmp/streamjob2815961207967489724.jar tmpDir=null
18/02/22 20:39:02 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
18/02/22 20:39:03 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
18/02/22 20:39:03 INFO mapred.FileInputFormat: Total input paths to process : 1
18/02/22 20:39:03 INFO mapreduce.JobSubmitter: number of splits:2
18/02/22 20:39:03 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1519328626440_0003
18/02/22 20:39:04 INFO impl.YarnClientImpl: Submitted application application_1519328626440_0003
18/02/22 20:39:04 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1519328626440_0003/
18/02/22 20:39:04 INFO mapreduce.Job: Running job: job_1519328626440_0003
18/02/22 20:39:10 INFO mapreduce.Job: Job job_1519328626440_0003 running in uber mode : false
18/02/22 20:39:10 INFO mapreduce.Job:  map 0% reduce 0%
18/02/22 20:39:19 INFO mapreduce.Job:  map 50% reduce 0%
18/02/22 20:39:20 INFO mapreduce.Job:  map 100% reduce 0%
18/02/22 20:39:29 INFO mapreduce.Job:  map 100% reduce 100%
18/02/22 20:39:29 INFO mapreduce.Job: Job job_1519328626440_0003 completed successfully
18/02/22 20:39:29 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=15089778
                FILE: Number of bytes written=30530606
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=6621441
                HDFS: Number of bytes written=5418107
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=13878
                Total time spent by all reduces in occupied slots (ms)=7217
                Total time spent by all map tasks (ms)=13878
                Total time spent by all reduce tasks (ms)=7217
                Total vcore-seconds taken by all map tasks=13878
                Total vcore-seconds taken by all reduce tasks=7217
                Total megabyte-seconds taken by all map tasks=14211072
                Total megabyte-seconds taken by all reduce tasks=7390208
        Map-Reduce Framework
                Map input records=128457
                Map output records=991447
                Map output bytes=13106878
                Map output materialized bytes=15089784
                Input split bytes=224
                Combine input records=0
                Combine output records=0
                Reduce input groups=352966
                Reduce shuffle bytes=15089784
                Reduce input records=991447
                Reduce output records=352966
                Spilled Records=1982894
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=124
                CPU time spent (ms)=10350
```

```
18/02/22 18:39:11 INFO mapreduce.Job:  map 100% reduce 100%
18/02/22 18:39:11 INFO mapreduce.Job: Job job_1519313959914_0008 failed with state FAILED due to: Task failed task_1519313959914_0008_m_000000
Job failed as tasks failed. failedMaps:1 failedReduces:0

18/02/22 18:39:11 INFO mapreduce.Job: Counters: 13
        Job Counters
                Failed map tasks=7
                Killed map tasks=1
                Launched map tasks=8
                Other local map tasks=6
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=27754
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=27754
                Total vcore-seconds taken by all map tasks=27754
                Total megabyte-seconds taken by all map tasks=28420096
        Map-Reduce Framework
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
ntures.txt -output /user/hadoop fs -chmod 777 /user/cloudera/reducer.py
ntures.txt -output /user/cloudera/outputle reducer.py -input /user/cloudera/adven
18/02/22 18:41:52 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [mapper.py, reducer.py] [/usr/jars/hadoop-streaming-2.6.0-cdh5.7.0.jar] /tmp/streamjob4773297266623658560.jar tmpDir=null
18/02/22 18:41:53 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
18/02/22 18:41:53 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
18/02/22 18:41:54 INFO mapred.FileInputFormat: Total input paths to process : 1
18/02/22 18:41:54 INFO mapreduce.JobSubmitter: number of splits:2
18/02/22 18:41:54 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1519313959914_0009
18/02/22 18:41:54 INFO impl.YarnClientImpl: Submitted application application_1519313959914_0009
18/02/22 18:41:54 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1519313959914_0009/
18/02/22 18:41:54 INFO mapreduce.Job: Running job: job_1519313959914_0009
18/02/22 18:42:01 INFO mapreduce.Job: Job job_1519313959914_0009 running in uber mode : false
18/02/22 18:42:01 INFO mapreduce.Job:  map 0% reduce 0%
18/02/22 18:42:10 INFO mapreduce.Job:  map 50% reduce 0%
18/02/22 18:42:11 INFO mapreduce.Job:  map 100% reduce 0%
18/02/22 18:42:20 INFO mapreduce.Job:  map 100% reduce 100%
18/02/22 18:42:21 INFO mapreduce.Job: Job job_1519313959914_0009 completed successfully
18/02/22 18:42:21 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=15089778
                FILE: Number of bytes written=30530609
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=6621441
                HDFS: Number of bytes written=5418107
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=14408
                Total time spent by all reduces in occupied slots (ms)=7460
                Total time spent by all map tasks (ms)=14408
                Total time spent by all reduce tasks (ms)=7460
                Total vcore-seconds taken by all map tasks=14408
                Total vcore-seconds taken by all reduce tasks=7460
                Total megabyte-seconds taken by all map tasks=14753792
                Total megabyte-seconds taken by all reduce tasks=7639040
        Map-Reduce Framework
```




Job Browser
Username cloudera
Text Search for text
Succeeded Running Failed Killed
Logs ID Name Application Type Status User Maps Reduces Queue Priority Duration Submitted
1519313959914_0009 streamjob4773297266623658560.jar MAPREDUCE SUCCEEDED cloudera 100% 100% root.cloudera N/A 24s 02/22/18 10:41:54
1519313959914_0008 streamjob1316155475115525723.jar MAPREDUCE FAILED cloudera 100% 100% root.cloudera N/A 27s 02/22/18 10:38:42
1519313959914_0007 streamjob8696549063458826774.jar MAPREDUCE FAILED cloudera 100% 100% root.cloudera N/A 40s 02/22/18 10:35:24
Showing 1 to 3 of 3 entries
← Previous 1 Next →