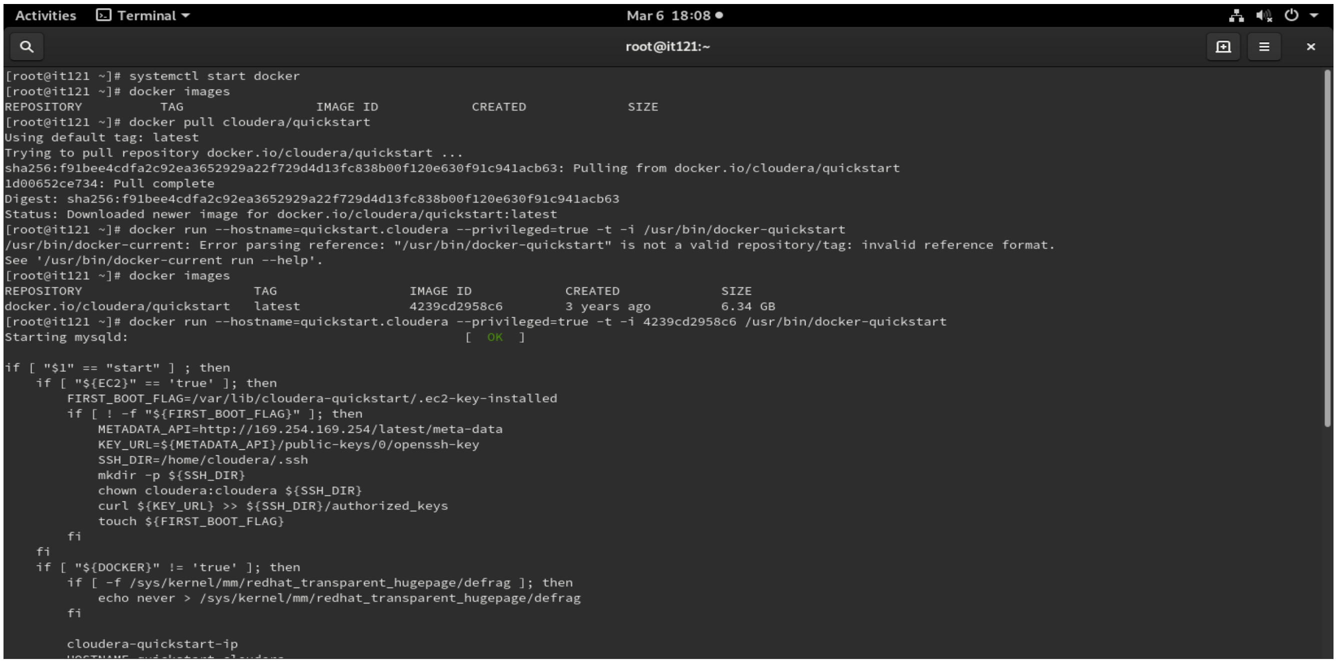
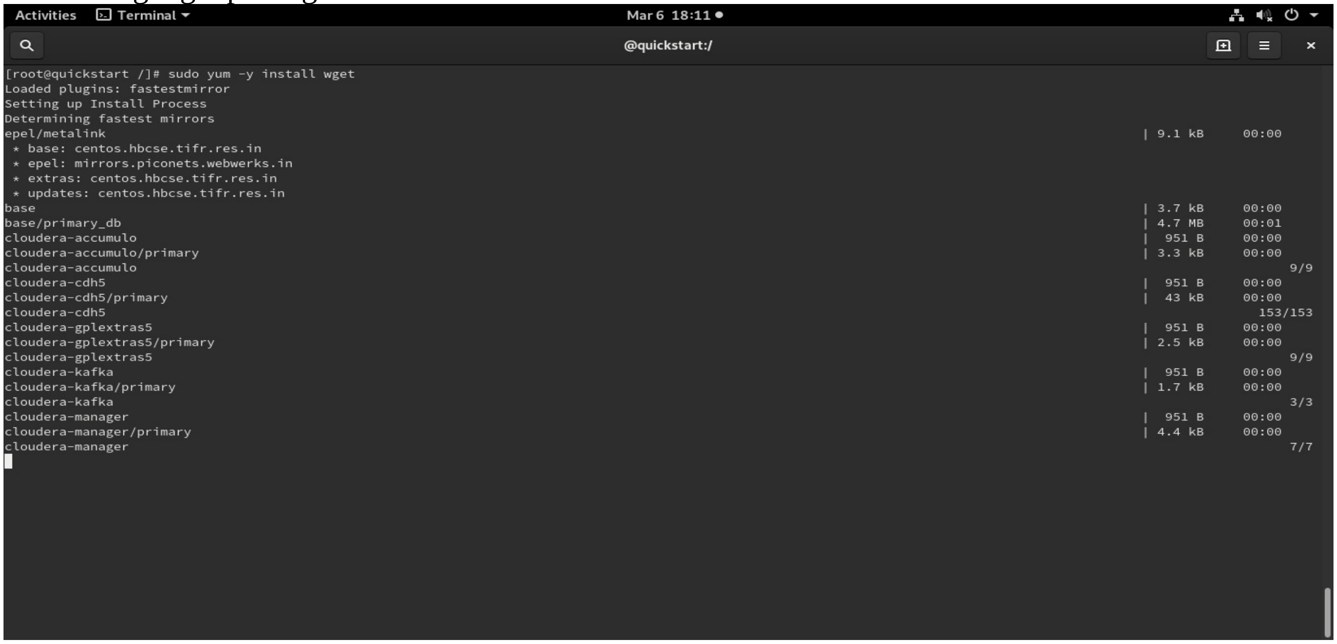
1. Starting Docker and pulling Cloudera quickstart vm



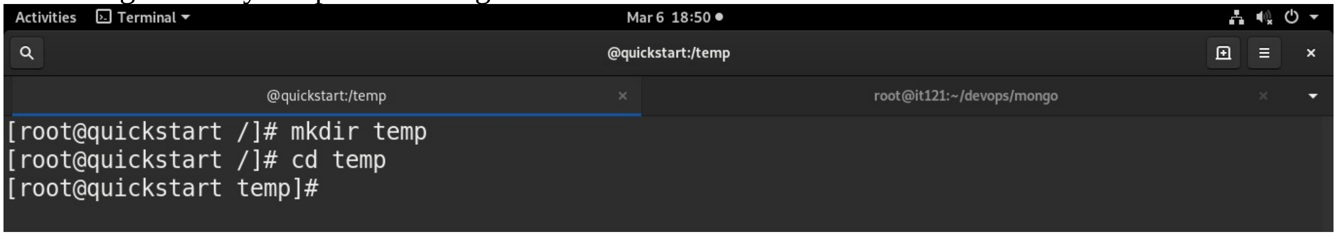
```
[root@it121 ~]# systemctl start docker
[root@it121 ~]# docker images
REPOSITORY          TAG                 IMAGE ID            CREATED             SIZE
[root@it121 ~]# docker pull cloudera/quickstart
Using default tag: latest
Trying to pull repository docker.io/cloudera/quickstart ...
sha256:f91bee4cdfa2c92ea3652929a22f729d4d13fc838b00f120e630f91c941acb63: Pulling from docker.io/cloudera/quickstart
1d00652ce734: Pull complete
Digest: sha256:f91bee4cdfa2c92ea3652929a22f729d4d13fc838b00f120e630f91c941acb63
Status: Downloaded newer image for docker.io/cloudera/quickstart:latest
[root@it121 ~]# docker run --hostname=quickstart.cloudera --privileged=true -t -i /usr/bin/docker-quickstart
/usr/bin/docker-current: Error parsing reference: "/usr/bin/docker-quickstart" is not a valid repository/tag: invalid reference format.
See '/usr/bin/docker-current run --help'.
[root@it121 ~]# docker images
REPOSITORY                      TAG                 IMAGE ID            CREATED             SIZE
docker.io/cloudera/quickstart   latest              4239cd2958c6        3 years ago         6.34 GB
[root@it121 ~]# docker run --hostname=quickstart.cloudera --privileged=true -t -i 4239cd2958c6 /usr/bin/docker-quickstart
Starting mysqld:                                           [  OK  ]

if [ "$1" == "start" ] ; then
    if [ "${EC2}" == 'true' ]; then
        FIRST_BOOT_FLAG=/var/lib/cloudera-quickstart/.ec2-key-installed
        if [ ! -f "${FIRST_BOOT_FLAG}" ]; then
            METADATA_API=http://169.254.169.254/latest/meta-data
            KEY_URL=${METADATA_API}/public-keys/0/openssh-key
            SSH_DIR=/home/cloudera/.ssh
            mkdir -p ${SSH_DIR}
            chown cloudera:cloudera ${SSH_DIR}
            curl ${KEY_URL} >> ${SSH_DIR}/authorized_keys
            touch ${FIRST_BOOT_FLAG}
        fi
    fi
    if [ "${DOCKER}" != 'true' ]; then
        if [ -f /sys/kernel/mm/redhat_transparent_hugepage/defrag ]; then
            echo never > /sys/kernel/mm/redhat_transparent_hugepage/defrag
        fi

        cloudera-quickstart-ip
        HOSTNAME=quickstart.cloudera
```

2.Installing wget package



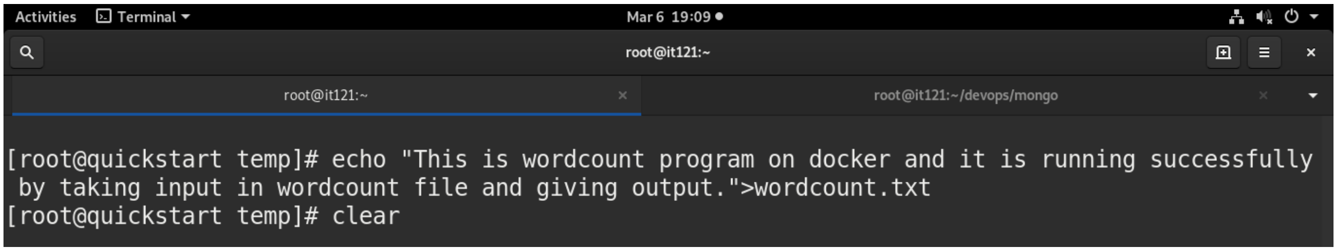
```
[root@quickstart /]# sudo yum -y install wget
Loaded plugins: fastestmirror
Setting up Install Process
Determining fastest mirrors
epel/metalink                                                                    | 9.1 kB     00:00
 * base: centos.hbcse.tifr.res.in
 * epel: mirrors.piconets.webwerks.in
 * extras: centos.hbcse.tifr.res.in
 * updates: centos.hbcse.tifr.res.in
base                                                                             | 3.7 kB     00:00
base/primary_db                                                                  | 4.7 MB     00:01
cloudera-accumulo                                                                |  951 B     00:00
cloudera-accumulo/primary                                                        | 3.3 kB     00:00
cloudera-accumulo                                                                                 9/9
cloudera-cdh5                                                                    |  951 B     00:00
cloudera-cdh5/primary                                                            |  43 kB     00:00
cloudera-cdh5                                                                                 153/153
cloudera-gplextras5                                                              |  951 B     00:00
cloudera-gplextras5/primary                                                      | 2.5 kB     00:00
cloudera-gplextras5                                                                               9/9
cloudera-kafka                                                                   |  951 B     00:00
cloudera-kafka/primary                                                           | 1.7 kB     00:00
cloudera-kafka                                                                                    3/3
cloudera-manager                                                                 |  951 B     00:00
cloudera-manager/primary                                                         | 4.4 kB     00:00
cloudera-manager                                                                                  7/7
```

3. Making directory temp and moving into it



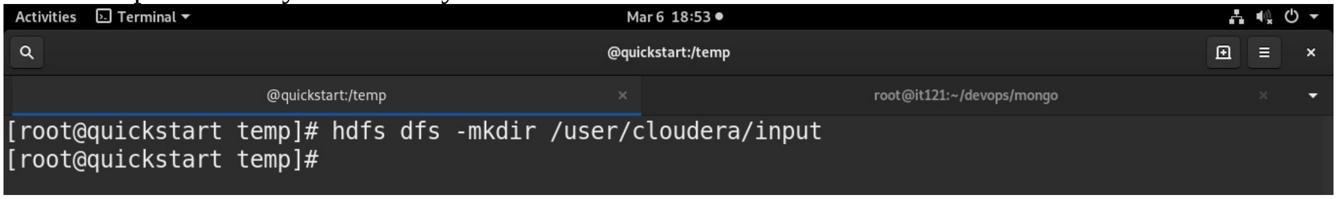
```
[root@quickstart /]# mkdir temp
[root@quickstart /]# cd temp
[root@quickstart temp]#
```

4.Create file with some content



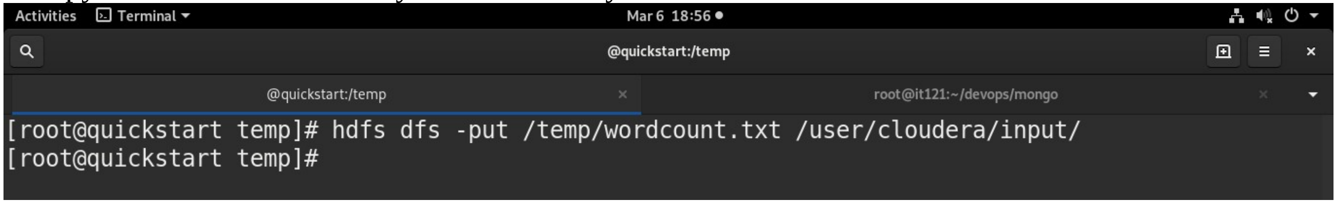
```
[root@quickstart temp]# echo "This is wordcount program on docker and it is running successfully
 by taking input in wordcount file and giving output.">wordcount.txt
[root@quickstart temp]# clear
```

5.Make input directory in HDFS system



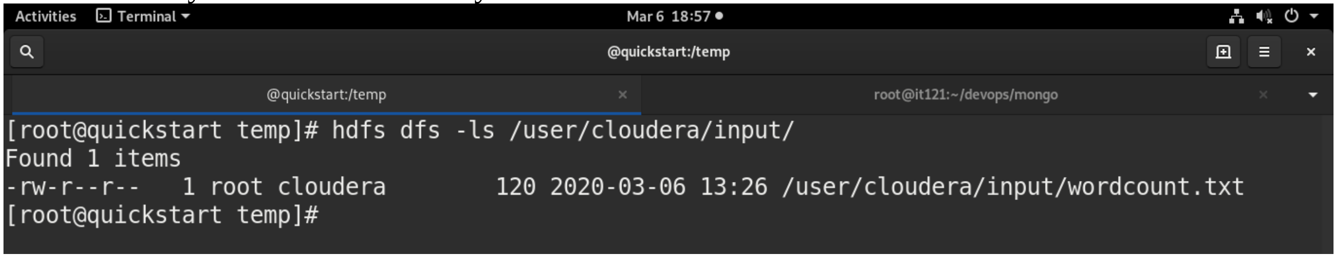
```
[root@quickstart temp]# hdfs dfs -mkdir /user/cloudera/input
[root@quickstart temp]#
```

6.Copy file from local directory to HDFS file system



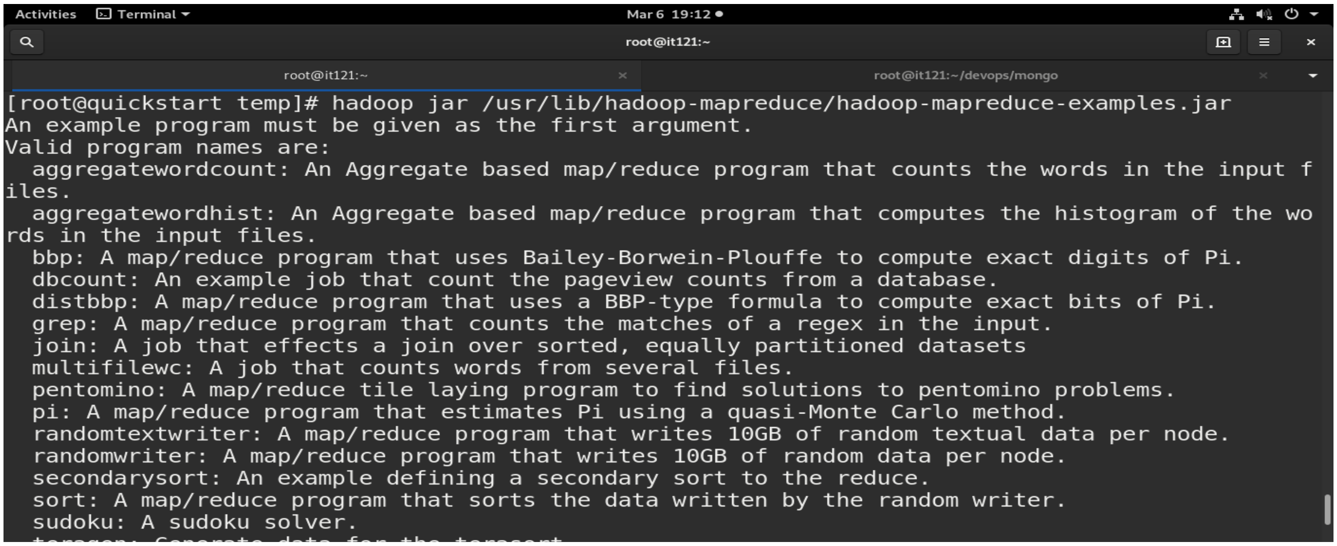
```
[root@quickstart temp]# hdfs dfs -put /temp/wordcount.txt /user/cloudera/input/
[root@quickstart temp]#
```

7.To check if your file is successfully or not



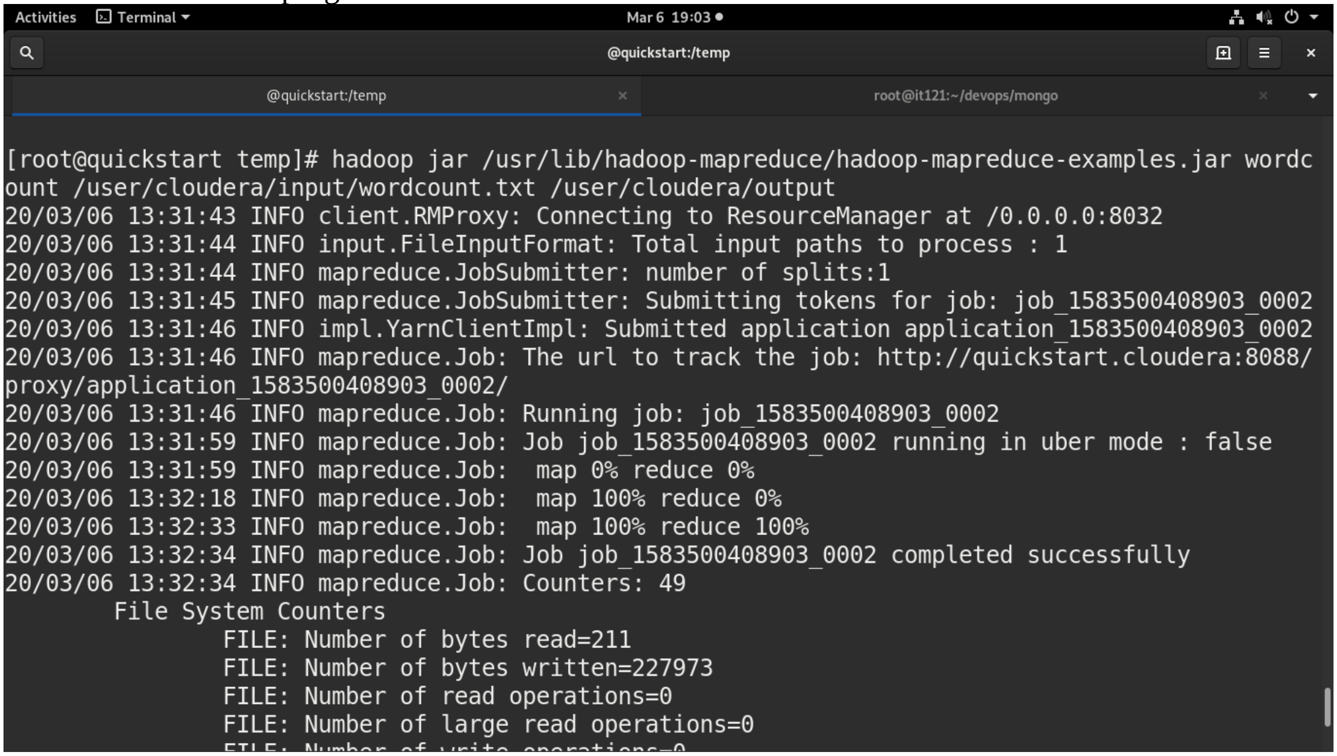
```
[root@quickstart temp]# hdfs dfs -ls /user/cloudera/input/
Found 1 items
-rw-r--r--   1 root cloudera        120 2020-03-06 13:26 /user/cloudera/input/wordcount.txt
[root@quickstart temp]#
```

8.To check Hadoop mapredue examples



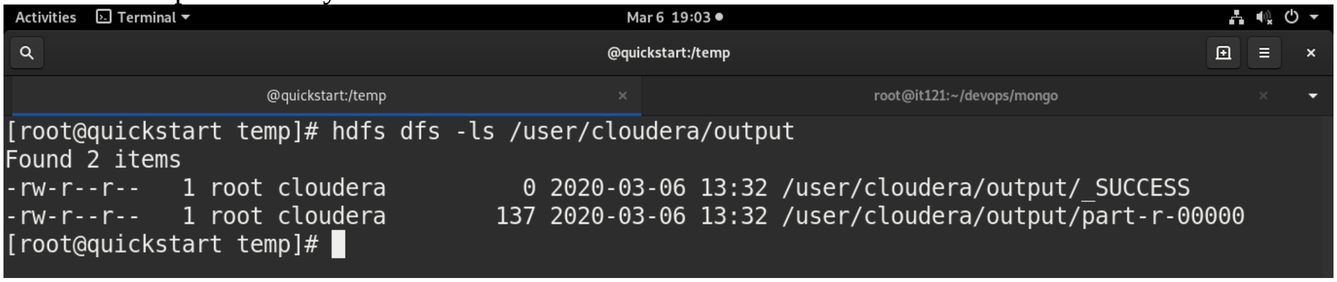
```
[root@quickstart temp]# hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar
An example program must be given as the first argument.
Valid program names are:
  aggregatewordcount: An Aggregate based map/reduce program that counts the words in the input f
iles.
  aggregatewordhist: An Aggregate based map/reduce program that computes the histogram of the wo
rds in the input files.
  bbp: A map/reduce program that uses Bailey-Borwein-Plouffe to compute exact digits of Pi.
  dbcount: An example job that count the pageview counts from a database.
  distbbp: A map/reduce program that uses a BBP-type formula to compute exact bits of Pi.
  grep: A map/reduce program that counts the matches of a regex in the input.
  join: A job that effects a join over sorted, equally partitioned datasets
  multifilewc: A job that counts words from several files.
  pentomino: A map/reduce tile laying program to find solutions to pentomino problems.
  pi: A map/reduce program that estimates Pi using a quasi-Monte Carlo method.
  randomtextwriter: A map/reduce program that writes 10GB of random textual data per node.
  randomwriter: A map/reduce program that writes 10GB of random data per node.
  secondarysort: An example defining a secondary sort to the reduce.
  sort: A map/reduce program that sorts the data written by the random writer.
  sudoku: A sudoku solver.
  teragen: Generate data for the terasort
```

9.Run the wordcount program



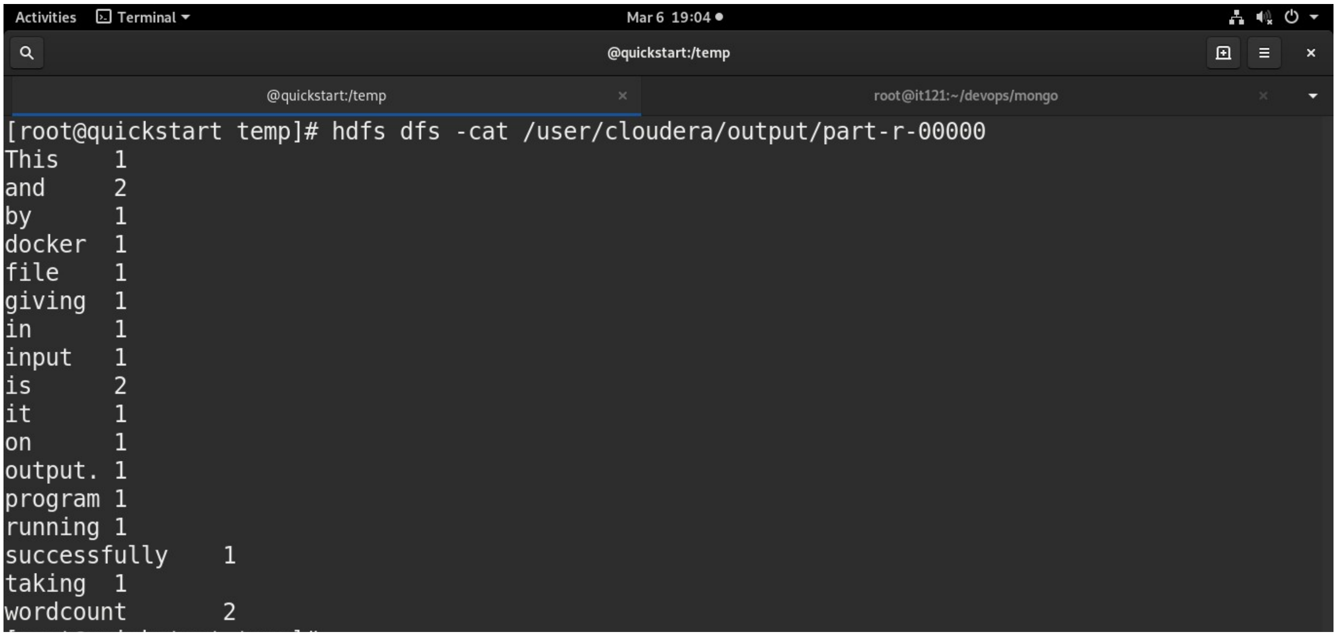
```
[root@quickstart temp]# hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar wordc
ount /user/cloudera/input/wordcount.txt /user/cloudera/output
20/03/06 13:31:43 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
20/03/06 13:31:44 INFO input.FileInputFormat: Total input paths to process : 1
20/03/06 13:31:44 INFO mapreduce.JobSubmitter: number of splits:1
20/03/06 13:31:45 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1583500408903_0002
20/03/06 13:31:46 INFO impl.YarnClientImpl: Submitted application application_1583500408903_0002
20/03/06 13:31:46 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/
proxy/application_1583500408903_0002/
20/03/06 13:31:46 INFO mapreduce.Job: Running job: job_1583500408903_0002
20/03/06 13:31:59 INFO mapreduce.Job: Job job_1583500408903_0002 running in uber mode : false
20/03/06 13:31:59 INFO mapreduce.Job:  map 0% reduce 0%
20/03/06 13:32:18 INFO mapreduce.Job:  map 100% reduce 0%
20/03/06 13:32:33 INFO mapreduce.Job:  map 100% reduce 100%
20/03/06 13:32:34 INFO mapreduce.Job: Job job_1583500408903_0002 completed successfully
20/03/06 13:32:34 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=211
                FILE: Number of bytes written=227973
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
```

10.Check output directory



```
[root@quickstart temp]# hdfs dfs -ls /user/cloudera/output
Found 2 items
-rw-r--r--   1 root cloudera          0 2020-03-06 13:32 /user/cloudera/output/_SUCCESS
-rw-r--r--   1 root cloudera        137 2020-03-06 13:32 /user/cloudera/output/part-r-00000
[root@quickstart temp]# 
```

11.Open part file to check output



```
[root@quickstart temp]# hdfs dfs -cat /user/cloudera/output/part-r-00000
This    1
and     2
by      1
docker  1
file    1
giving  1
in      1
input   1
is      2
it      1
on      1
output. 1
program 1
running 1
successfully    1
taking  1
wordcount       2
```