

# Twitter

## Word Co-occurrence

```
DIC — @quickstart/src/data/input/tw — docker run --hostname=quickstart.cloudera --privileged=true -t -i -v ~/Documents/DockerMR/src --pub...
...wnloads/Academia- Local/University/SEM2/DIC — jupyter-notebook • python ... ---publish-all=true -p 8888 cloudera/quickstart /usr/bin/docker-quickstart +
[[root@quickstart tw]# hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.7.0.jar -file /src/coMapper.py -mapper /src/coM
apper.py -file /src/coReducer.py -reducer /src/coReducer.py -input /user/asishkak/MR/input/* -output /user/asishkak/MR/output/twc
19/04/21 03:38:35 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/src/coMapper.py, /src/coReducer.py] [/usr/jars/hadoop-streaming-2.6.0-cdh5.7.0.jar] /tmp/streamjob2005079676330314499
.jar tmpDir=null
19/04/21 03:38:36 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/04/21 03:38:37 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/04/21 03:38:37 INFO mapred.FileInputFormat: Total input paths to process : 4
19/04/21 03:38:38 INFO mapreduce.JobSubmitter: number of splits:4
19/04/21 03:38:38 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1555803922186_0003
19/04/21 03:38:38 INFO impl.YarnClientImpl: Submitted application application_1555803922186_0003
19/04/21 03:38:38 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1555803922186_0003/
19/04/21 03:38:38 INFO mapreduce.Job: Running job: job_1555803922186_0003
19/04/21 03:38:45 INFO mapreduce.Job: Job job_1555803922186_0003 running in uber mode : false
19/04/21 03:38:45 INFO mapreduce.Job: map 0% reduce 0%
19/04/21 03:38:56 INFO mapreduce.Job: map 25% reduce 0%
19/04/21 03:38:57 INFO mapreduce.Job: map 50% reduce 0%
19/04/21 03:38:58 INFO mapreduce.Job: map 100% reduce 0%
19/04/21 03:39:05 INFO mapreduce.Job: map 100% reduce 100%
19/04/21 03:39:05 INFO mapreduce.Job: Job job_1555803922186_0003 completed successfully
19/04/21 03:39:05 INFO mapreduce.Job: Counters: 49
File System Counters
  FILE: Number of bytes read=23721971
  FILE: Number of bytes written=48028870
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
```

```
DIC — @quickstart/src/data/input/tw — docker run --hostname=quickstart.cloudera --privileged=true -t -i -v ~/Documents/DockerMR/src --pub...
...wnloads/Academia- Local/University/SEM2/DIC — jupyter-notebook • python ... ---publish-all=true -p 8888 cloudera/quickstart /usr/bin/docker-quickstart +
Reduce input groups=14381
Reduce shuffle bytes=23721989
Reduce input records=1382966
Reduce output records=14381
Spilled Records=2765932
Shuffled Maps =4
Failed Shuffles=0
Merged Map outputs=4
GC time elapsed (ms)=179
CPU time spent (ms)=17170
Physical memory (bytes) snapshot=1275510784
Virtual memory (bytes) snapshot=6786428928
Total committed heap usage (bytes)=1277165568
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=2795577
File Output Format Counters
  Bytes Written=236962
19/04/21 03:39:05 INFO streaming.StreamJob: Output directory: /user/asishkak/MR/output/twc
[root@quickstart tw]#
```

## Word Co-occurrence output

```
DIC — @quickstart/src/data/input/tw — docker run --hostname=quickstart.cloudera --privileged=true -t -i -v ~/Documents/DockerMR/src --pub...
...wnloads/Academia- Local/University/SEM2/DIC — jupyter-notebook • python ... ---publish-all=true -p 8888 cloudera/quickstart /usr/bin/docker-quickstart +
Bytes Written=236962
19/04/21 03:39:05 INFO streaming.StreamJob: Output directory: /user/asishkak/MR/output/twc
[[root@quickstart tw]# hadoop fs -cat /user/asishkak/MR/output/twc/part* | sort -n -k2 -r | head -n1000
mueller|trump 15867
border|trump 12607
mueller|report 12235
border|wall 11580
report|trump 10102
trump|wall 10044
border|close 4418
presid|trump 4193
orang|trump 3707
mueller|orang 3511
build|wall 3456
close|trump 3383
donald|trump 3169
releas|report 2974
investig|mueller 2909
mueller|releas 2908
shut|trump 2683
border|build 2652
investig|trump 2599
releas|trump 2568
investig|orang 2417
border|shut 2167
close|wall 2112
```

## Word Count

```
DIC — @quickstart/src/data/input/tw — docker run --hostname=quickstart.cloudera --privileged=true -t -i -v ~/Documents/DockerMR/src --pub...
...wnloads/Academia- Local/University/SEM2/DIC — jupyter-notebook • python ... ---publish-all=true -p 8888 cloudera/quickstart /usr/bin/docker-quickstart +
[[root@quickstart tw]# hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.7.0.jar -file /src/mapper.py -mapper /src/mapper...
r.py -file /src/reducer.py -reducer /src/reducer.py -input /user/asishkak/MR/input/* -output /user/asishkak/MR/output/twcount
19/04/21 03:42:21 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/src/mapper.py, /src/reducer.py] [/usr/jars/hadoop-streaming-2.6.0-cdh5.7.0.jar] /tmp/streamjob6084708394353196027.jar
tmpDir=null
19/04/21 03:42:22 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/04/21 03:42:22 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/04/21 03:42:23 INFO mapred.FileInputFormat: Total input paths to process : 4
19/04/21 03:42:23 INFO mapreduce.JobSubmitter: number of splits:4
19/04/21 03:42:23 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1555803922186_0004
19/04/21 03:42:23 INFO impl.YarnClientImpl: Submitted application application_1555803922186_0004
19/04/21 03:42:23 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8080/proxy/application_1555803922186_0004/
19/04/21 03:42:23 INFO mapreduce.Job: Running job: job_1555803922186_0004
19/04/21 03:42:31 INFO mapreduce.Job: Job job_1555803922186_0004 running in uber mode : false
19/04/21 03:42:31 INFO mapreduce.Job: map 0% reduce 0%
19/04/21 03:42:39 INFO mapreduce.Job: map 25% reduce 0%
19/04/21 03:42:40 INFO mapreduce.Job: map 50% reduce 0%
19/04/21 03:42:41 INFO mapreduce.Job: map 100% reduce 0%
19/04/21 03:42:46 INFO mapreduce.Job: map 100% reduce 100%
19/04/21 03:42:47 INFO mapreduce.Job: Job job_1555803922186_0004 completed successfully
19/04/21 03:42:47 INFO mapreduce.Job: Counters: 49
File System Counters
FILE: Number of bytes read=4293136
FILE: Number of bytes written=9171125
FILE: Number of read operations=0
FILE: Number of large read operations=0
```

```
DIC — @quickstart/src/data/input/tw — docker run --hostname=quickstart.cloudera --privileged=true -t -i -v ~/Documents/DockerMR/src --pub...
...wnloads/Academia- Local/University/SEM2/DIC — jupyter-notebook • python ... ...--publish-all=true -p 8888 cloudera/quickstart /usr/bin/docker-quickstart +
Reduce input groups=1973
Reduce shuffle bytes=4293154
Reduce input records=406056
Reduce output records=1973
Spilled Records=812112
Shuffled Maps =4
Failed Shuffles=0
Merged Map outputs=4
GC time elapsed (ms)=313
CPU time spent (ms)=9350
Physical memory (bytes) snapshot=1454731264
Virtual memory (bytes) snapshot=6806159360
Total committed heap usage (bytes)=1402470400

Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0

File Input Format Counters
  Bytes Read=2795577

File Output Format Counters
  Bytes Written=20258
19/04/21 03:42:47 INFO streaming.StreamJob: Output directory: /user/asishkak/MR/output/twcount
[root@quickstart tw]#
```

## Word Count Output

```
DIC — @quickstart/src/data/input/tw — docker run --hostname=quickstart.cloudera --privileged=true -t -i -v ~/Documents/DockerMR/src --pub...
...wnloads/Academia- Local/University/SEM2/DIC — jupyter-notebook • python ... ...--publish-all=true -p 8888 cloudera/quickstart /usr/bin/docker-quickstart +
Bytes Written=20258
19/04/21 03:42:47 INFO streaming.StreamJob: Output directory: /user/asishkak/MR/output/twcount
[[root@quickstart tw]# hadoop fs -cat /user/asishkak/MR/output/twcount/part* | sort -n -k2 -r | head -n1000
trump 39236
mueller 19560
border 17479
wall 14508
report 12782
presid 5140
close 4931
orang 3978
build 3588
releas 3335
investig 3086
donald 2954
shut 2815
vote 2556
subpoena 2184
hous 2165
congress 2123
democrat 2061
mexico 1968
amp 1808
republican 1759
immigr 1642
plan 1603
```