

## Common Crawl

## Word Count

```
cc -- @quickstart/src/data/output/cc/wc -- docker run --hostname=quickstart.cloudera --privileged=true -t -v ~/Documents/DockerMR/src -- ...  
-88 cloudera/quickstart /usr/bin/docker-quickstart ... jupyter-notebook - python ... -anta/DockerMR/data/output/cc/wc -- -bash  
Deleted /user/asishksh/MR/output/wc  
[root@quickstart cc]# hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.7.0.jar -file /src/mapper.py -mapper /src/mapper.py -file /src/reducer.py -reducer /src/reducer.py -input /user/asishksh/MR/input/*.txt -output /user/asishksh/MR/output/wc  
18/04/21 20:30:04 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/streamjob5288367258646699768.jar  
tmpDir=null  
18/04/21 20:30:06 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032  
18/04/21 20:30:06 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032  
18/04/21 20:30:07 INFO mapred.FileInputFormat: Total input paths to process : 1  
18/04/21 20:30:07 INFO mapreduce.JobSubmitter: number of splits:2  
18/04/21 20:30:07 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1555877373975_0002  
18/04/21 20:30:07 INFO impl.YarnClientImpl: Submitted application application_1555877373975_0002  
18/04/21 20:30:08 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8888/proxy/application_1555877373975_0002/  
18/04/21 20:30:08 INFO mapreduce.Job: Running job: job_1555877373975_0002  
18/04/21 20:30:15 INFO mapreduce.Job: Job job_1555877373975_0002 running in user mode : false  
18/04/21 20:30:15 INFO mapreduce.Job: map 0% reduce 0%  
18/04/21 20:30:22 INFO mapreduce.Job: map 50% reduce 0%  
18/04/21 20:30:23 INFO mapreduce.Job: map 100% reduce 0%  
18/04/21 20:30:29 INFO mapreduce.Job: map 100% reduce 100%  
18/04/21 20:30:29 INFO mapreduce.Job: Job job_1555877373975_0002 completed successfully  
18/04/21 20:30:29 INFO mapreduce.Job: Counters: 49  
File System Counters  
File: Number of bytes read=3848758  
File: Number of bytes written=6845358  
File: Number of read operations=0  
File: Number of large read operations=0
```

```
cc -- @quickstart/src/data/output/cc/wc -- docker run --hostname=quickstart.cloudera --privileged=true -t -v ~/Documents/DockerMR/src -- ...  
-88 cloudera/quickstart /usr/bin/docker-quickstart ... jupyter-notebook - python ... -anta/DockerMR/data/output/cc/wc -- -bash  
Combine output records=0  
Reduce input groups=10903  
Reduce shuffle bytes=2848764  
Reduce input records=263739  
Reduce output records=10903  
Spilled Records=527478  
Shuffled Maps=2  
Failed Shuffles=0  
Merged Map outputs=2  
GC time elapsed (ms)=118  
CPU time spent (ms)=5880  
Physical memory (bytes) snapshot=890778112  
Virtual memory (bytes) snapshot=4071820736  
Total committed heap usage (bytes)=658641408  
Shuffle Errors  
BAD_ID=0  
CONNECTION=0  
ID_ERROR=0  
WRONG_LENGTH=0  
WRONG_MAP=0  
WRONG_REDUCE=0  
File Input Format Counters  
Bytes Read=1811485  
File Output Format Counters  
Bytes Written=126631  
18/04/21 20:30:29 INFO streaming.StreamJob: Output directory: /user/asishksh/MR/output/wc
```

```
cc -- @quickstart/src/data/output/cc/wc -- docker run --hostname=quickstart.cloudera --privileged=true -t -v ~/Documents/DockerMR/src -- ...
-88 cloudera/quickstart /usr/bin/docker-quickstart ... jupyter-notebook • python ... ..anta/DockerMR/data/output/cc/wc -- -bash
[root@quickstart ~]# hadoop fs -cat /user/asishkak/MR/output/cc-topwc.txt
trump 3669
presid 1779
hour 1564
year 1171
us 1115
minut 1048
report 1026
compani 1014
deal 1010
day 983
the 938
state 792
peopl 861
nation 834
wallter 826
plan 615
week 612
news 580
work 578
review 555
star5 553
book 542
govern 532
today 516
feder 509
```

## Word Co-occurrence

```
cc -- @quickstart/src/data/output/cc/wcocc -- docker run --hostname=quickstart.cloudera --privileged=true -t -v ~/Documents/DockerMR/src -- ...
-88 cloudera/quickstart /usr/bin/docker-quickstart ... jupyter-notebook • python ... ..ta/DockerMR/data/output/cc/wcocc -- -bash
[root@quickstart ~]# 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/*txt -output /user/asishkak/MR/output/ccwc
occ
18/04/21 20:58:55 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/streamjob906326436289726466
.jar tmpDir=null
18/04/21 20:58:55 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
18/04/21 20:58:57 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
18/04/21 20:58:57 INFO mapred.FileInputFormat: Total input paths to process : 1
18/04/21 20:58:57 INFO mapreduce.JobSubmitter: number of splits:2
18/04/21 20:58:57 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1555877373975_0004
18/04/21 20:58:58 INFO impl.YarnClientImpl: Submitted application application_1555877373975_0004
18/04/21 20:58:58 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8888/proxy/application_1555877373975_0004/
18/04/21 20:58:58 INFO mapreduce.Job: Running job: job_1555877373975_0004
18/04/21 20:58:58 INFO mapreduce.Job: Job job_1555877373975_0004 running in user mode : false
18/04/21 20:59:04 INFO mapreduce.Job: map 0% reduce 0%
18/04/21 20:59:16 INFO mapreduce.Job: map 15% reduce 0%
18/04/21 20:59:19 INFO mapreduce.Job: map 19% reduce 0%
18/04/21 20:59:26 INFO mapreduce.Job: map 29% reduce 0%
18/04/21 20:59:32 INFO mapreduce.Job: map 34% reduce 0%
18/04/21 20:59:33 INFO mapreduce.Job: map 38% reduce 0%
18/04/21 20:59:39 INFO mapreduce.Job: map 43% reduce 0%
18/04/21 21:00:16 INFO mapreduce.Job: map 48% reduce 0%
18/04/21 21:00:22 INFO mapreduce.Job: map 53% reduce 0%
18/04/21 21:00:25 INFO mapreduce.Job: map 67% reduce 0%
18/04/21 21:00:35 INFO mapreduce.Job: map 60% reduce 0%
```

```
cc -- @quickstart/arc/data/output/cc/wcooc -- docker run --hostname=quickstart.cloudiera --privileged=true -t -i -v ~/Documents/DockerMR/ur...
..88 cloudera/quickstart /usr/bin/docker-quickstart ..ICLab2/data/cc -- jupyter-notebook - python ... ..ts/DockerMR/data/output/cc/wcooc -- -bash +
19/04/21 21:01:30 INFO mapreduce.Job: map 100% reduce 72%
19/04/21 21:01:31 INFO mapreduce.Job: map 100% reduce 75%
19/04/21 21:01:34 INFO mapreduce.Job: map 100% reduce 78%
19/04/21 21:01:37 INFO mapreduce.Job: map 100% reduce 80%
19/04/21 21:01:40 INFO mapreduce.Job: map 100% reduce 83%
19/04/21 21:01:43 INFO mapreduce.Job: map 100% reduce 85%
19/04/21 21:01:46 INFO mapreduce.Job: map 100% reduce 88%
19/04/21 21:01:49 INFO mapreduce.Job: map 100% reduce 91%
19/04/21 21:01:52 INFO mapreduce.Job: map 100% reduce 93%
19/04/21 21:01:55 INFO mapreduce.Job: map 100% reduce 96%
19/04/21 21:01:58 INFO mapreduce.Job: map 100% reduce 99%
19/04/21 21:02:00 INFO mapreduce.Job: map 100% reduce 100%
19/04/21 21:02:00 INFO mapreduce.Job: Job job_5555677373975_0004 completed successfully
19/04/21 21:02:00 INFO mapreduce.Job: Counters: 49
  File System Counters
    FILE: Number of bytes read=1135051286
    FILE: Number of bytes written=1797361600
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=1011733
    HDFS: Number of bytes written=82404056
    HDFS: Number of read operations=0
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
```

```
cc -- @quickstart/arc/data/output/cc/wcooc -- docker run --hostname=quickstart.cloudiera --privileged=true -t -i -v ~/Documents/DockerMR/ur...
..88 cloudera/quickstart /usr/bin/docker-quickstart ..ICLab2/data/cc -- jupyter-notebook - python ... ..ts/DockerMR/data/output/cc/wcooc -- -bash +
  Combine Input records=0
  Combine output records=0
  Reduce input groups=4727426
  Reduce shuffle bytes=567525610
  Reduce input records=31626624
  Reduce output records=4727426
  Spilled Records=94879872
  Shuffled Maps=42
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=2081
  CPU time spent (ms)=363860
  Physical memory (bytes) snapshot=1189269584
  Virtual memory (bytes) snapshot=4009200640
  Total committed heap usage (bytes)=1285862400
  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=1811485
  File Output Format Counters
    Bytes Written=82404056
```

```
wcooc -- -bash -- Solarized Dark ansi -- 134x26
88 cloudera/quickstart /usr/bin/docker-quickstart
..IC/..lab2/data/cc -- jupyter-notebook - python ...
..ts/DockerMR/data/output/cc/wcooc -- -bash

presid|trump 3080
trump|herder 3023
maeller|report 3021
presid|us 2790
investig|presid 2214
presid|wall 2021
internet|trump 1852
wall|trump 1888
state|us 1668
report|trump 1595
presid|state 1577
company|the 1577
internet|trump 1561
report|us 1541
trump|trump 1538
plan|trump 1402
company|internet 1404
govern|us 1488
company|state 1430
browser|company 1408
plan|presid 1400
govern|trump 1352
people|trump 1227
news|trump 1510
presid|report 1298
govern|presid 1280
```