# MapReduce with Hadoop

## GREP:

Counting total number of ‘https’ word occurence in the hadoop homepage using built-in hadoop-mapreduce-example:

We run the hadoop job and then copy the results from HDFS to the local file system

```
[04/08/19]seed@VM:~/.../hadoop-2.7.6$ ls
bin  etc  hadoop_home_page.html  include  lib  libexec  LICENSE.txt  logs  NOTICE.txt  README.txt  sbin  share
[04/08/19]seed@VM:~/.../hadoop-2.7.6$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.6.jar grep /user/d
emo/hadoop_home_page.html /user/demo/hadoop_home_page.html_OUTPUT_2 'https'
Java HotSpot(TM) Client VM warning: You have loaded library /home/seed/Downloads/hadoop-2.7.6/lib/native/libhadoop.so.1.0.0
which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
19/04/08 19:05:02 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java c
lasses where applicable
19/04/08 19:05:02 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
19/04/08 19:05:02 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
19/04/08 19:05:03 INFO input.FileInputFormat: Total input paths to process : 1
19/04/08 19:05:03 INFO mapreduce.JobSubmitter: number of splits:1
19/04/08 19:05:03 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local501571242_0001
19/04/08 19:05:04 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
19/04/08 19:05:04 INFO mapreduce.Job: Running job: job_local501571242_0001
19/04/08 19:05:04 INFO mapred.LocalJobRunner: OutputCommitter set in config null
19/04/08 19:05:04 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
19/04/08 19:05:04 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
19/04/08 19:05:04 INFO mapred.LocalJobRunner: Waiting for map tasks
19/04/08 19:05:04 INFO mapred.LocalJobRunner: Starting task: attempt_local501571242_0001_m_000000_0
19/04/08 19:05:04 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
19/04/08 19:05:04 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
19/04/08 19:05:04 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/demo/hadoop_home_page.html:0+21721
19/04/08 19:05:04 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
19/04/08 19:05:04 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
19/04/08 19:05:04 INFO mapred.MapTask: soft limit at 83886080
19/04/08 19:05:04 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
19/04/08 19:05:04 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
```

We get 32 occurences of ‘https’

```
-rw-r--r--   1 seed supergroup      21721 2019-04-08 18:56 /user/demo/hadoop_home_page.html
drwxr-xr-x   - seed supergroup          0 2019-04-08 19:05 /user/demo/hadoop_home_page.html_OUTPUT_2
[04/08/19]seed@VM:~/.../hadoop-2.7.6$ bin/hdfs dfs -get /user/demo/hadoop_home_page.html_OUTPUT_2 hadoop_home_page.html_OUTP
UT_2
Java HotSpot(TM) Client VM warning: You have loaded library /home/seed/Downloads/hadoop-2.7.6/lib/native/libhadoop.so.1.0.0
which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
19/04/08 19:09:33 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java c
lasses where applicable
[04/08/19]seed@VM:~/.../hadoop-2.7.6$ cat hadoop_home_page.html_OUTPUT_2/part-r-00000
32      https
```

Using grep and checking the hadoop homepage directly, we also get 32 occurrences.

```
[04/08/19]seed@VM:~/.../hadoop-2.7.6$ cat hadoop_home_page.html | grep -o -w 'https' | wc -w
32
```

**WordCount:**

We are again using hadoop-mapreduce-examples to return the word count of each word in the document.

```
[04/08/19]seed@VM:~/.../hadoop-2.7.6$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.6.jar wordcount /u
ser/demo/hadoop_home_page.html /user/demo/hadoop_home_page.html_OUTPUT_1
Java HotSpot(TM) Client VM warning: You have loaded library /home/seed/Downloads/hadoop-2.7.6/lib/native/libhadoop.so.1.0.0
which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
19/04/08 19:18:29 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java c
lasses where applicable
19/04/08 19:18:30 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
19/04/08 19:18:30 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
19/04/08 19:18:30 INFO input.FileInputFormat: Total input paths to process : 1
19/04/08 19:18:31 INFO mapreduce.JobSubmitter: number of splits:1
19/04/08 19:18:31 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1129657162_0001
19/04/08 19:18:31 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
19/04/08 19:18:31 INFO mapreduce.Job: Running job: job_local1129657162_0001
19/04/08 19:18:31 INFO mapred.LocalJobRunner: OutputCommitter set in config null
19/04/08 19:18:31 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
19/04/08 19:18:31 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
19/04/08 19:18:31 INFO mapred.LocalJobRunner: Waiting for map tasks
19/04/08 19:18:31 INFO mapred.LocalJobRunner: Starting task: attempt_local1129657162_0001_m_000000_0
19/04/08 19:18:31 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
19/04/08 19:18:31 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
19/04/08 19:18:31 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/demo/hadoop_home_page.html:0+21721
19/04/08 19:18:32 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
19/04/08 19:18:32 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
19/04/08 19:18:32 INFO mapred.MapTask: soft limit at 83886080
19/04/08 19:18:32 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
19/04/08 19:18:32 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
19/04/08 19:18:32 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
19/04/08 19:18:32 INFO mapred.LocalJobRunner:
```

Returning the top 50 lines of the hadoop homepage, we see the list the words and their word count:

```
-rw-r--r--   1 seed supergroup      21721 2019-04-08 18:56 /user/demo/hadoop_home_page.html
drwxr-xr-x   - seed supergroup          0 2019-04-08 19:18 /user/demo/hadoop_home_page.html_OUTPUT_1
drwxr-xr-x   - seed supergroup          0 2019-04-08 19:05 /user/demo/hadoop_home_page.html_OUTPUT_2
[04/08/19]seed@VM:~/.../hadoop-2.7.6$ bin/hdfs dfs -get /user/demo/hadoop_home_page.html_OUTPUT_1 hadoop_home_page.html_OUTP
UT_1
Java HotSpot(TM) Client VM warning: You have loaded library /home/seed/Downloads/hadoop-2.7.6/lib/native/libhadoop.so.1.0.0
which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
19/04/08 19:22:17 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java c
lasses where applicable
[04/08/19]seed@VM:~/.../hadoop-2.7.6$ head -n 50 hadoop_home_page.html_OUTPUT_1/part-r-00000
">      4
"AS     8
"License");     8
$('table').addClass('table      1
$(function()    1
&raquo;</a>     3
&rarr;</a></p>  2
'UA-7453027-1', 1
'auto');        1
'pageview');    1
(!doNotTrack)   1
(HDFS™)</strong>:       1
(e.g.   1
(function(i,s,o,g,r,a,m){i['GoogleAnalyticsObject']=r;i[r]=i[r]||function(){    1
(i[r].q=i[r].q||[]).push(arguments)},i[r].l=1*new       1
(the    8
-->     8
/>      2
0.2.1-alpha     1
0.3.0-alpha     1
0.54.0" 1
1       1
1092    2
```

**Wordcount using Hadoop Streaming (Python):**

Mapper program (wordcount_map.py)

```
  GNU nano 2.5.3                          File: wordcount map.py

#!/usr/bin/env python
import sys
for line in sys.stdin:
        line = line.strip()
        words = line.split()
        for word in words:
                print '%s\t%s' % (word, 1)
```

Reducer Program (wordcount_red.py)

```
  GNU nano 2.5.3                          File: wordcount_red.py

#!/usr/bin/env python
import sys
tmp_word = None
total_count = 0
for line in sys.stdin:
        line = line.strip()
        word, count = line.split('\t')
        count = int(count)
        if tmp_word == word:
                total_count += count
        else:
                print '%s\t%s' % (tmp_word, total_count)
                total_count = count
                tmp_word = word
print '%s\t%s' % (tmp_word, total_count)
```

We run the mapper and reducer python files and copy the results to the local file system

```
[04/08/19]seed@VM:~/.../hadoop-2.7.6$ bin/hadoop jar share/hadoop/tools/lib/hadoop-streaming-2.7.6.jar -mapper ./wordcount_m
ap.py -reducer ./wordcount_red.py -input /user/demo/hadoop_home_page.html -output /user/demo/hadoop_home_page.html_OUTPUT_CO
UNT1
Java HotSpot(TM) Client VM warning: You have loaded library /home/seed/Downloads/hadoop-2.7.6/lib/native/libhadoop.so.1.0.0
which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
19/04/08 19:56:38 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java c
lasses where applicable
19/04/08 19:56:39 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
19/04/08 19:56:39 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
19/04/08 19:56:39 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
19/04/08 19:56:39 INFO mapred.FileInputFormat: Total input paths to process : 1
19/04/08 19:56:40 INFO mapreduce.JobSubmitter: number of splits:1
19/04/08 19:56:40 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local657839071_0001
19/04/08 19:56:40 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
19/04/08 19:56:40 INFO mapreduce.Job: Running job: job_local657839071_0001
19/04/08 19:56:40 INFO mapred.LocalJobRunner: OutputCommitter set in config null
19/04/08 19:56:40 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
19/04/08 19:56:40 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
19/04/08 19:56:40 INFO mapred.LocalJobRunner: Waiting for map tasks
19/04/08 19:56:40 INFO mapred.LocalJobRunner: Starting task: attempt_local657839071_0001_m_000000_0
19/04/08 19:56:40 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
19/04/08 19:56:40 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
19/04/08 19:56:40 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/demo/hadoop_home_page.html:0+21721
19/04/08 19:56:40 INFO mapred.MapTask: numReduceTasks: 1
19/04/08 19:56:41 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
19/04/08 19:56:41 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
19/04/08 19:56:41 INFO mapred.MapTask: soft limit at 83886080
19/04/08 19:56:41 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
19/04/08 19:56:41 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
```

We get the words and their word count similar to the built-in mapreduce example in hadoop

```
[04/08/19]seed@VM:~/.../hadoop-2.7.6$ bin/hdfs dfs -ls /user/demo
Java HotSpot(TM) Client VM warning: You have loaded library /home/seed/Downloads/hadoop-2.7.6/lib/native/libhadoop.so.1.0.0
which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
19/04/08 19:56:51 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java c
lasses where applicable
Found 5 items
-rw-r--r--   1 seed supergroup      21721 2019-04-08 18:56 /user/demo/hadoop_home_page.html
drwxr-xr-x   - seed supergroup          0 2019-04-08 19:18 /user/demo/hadoop_home_page.html_OUTPUT_1
drwxr-xr-x   - seed supergroup          0 2019-04-08 19:05 /user/demo/hadoop_home_page.html_OUTPUT_2
drwxr-xr-x   - seed supergroup          0 2019-04-08 19:47 /user/demo/hadoop_home_page.html_OUTPUT_COUNT
drwxr-xr-x   - seed supergroup          0 2019-04-08 19:56 /user/demo/hadoop_home_page.html_OUTPUT_COUNT1
[04/08/19]seed@VM:~/.../hadoop-2.7.6$ bin/hdfs dfs -get /user/demo/hadoop_home_page.html_OUTPUT_COUNT1 hadoop_home_page.html
_OUTPUT_COUNT1
Java HotSpot(TM) Client VM warning: You have loaded library /home/seed/Downloads/hadoop-2.7.6/lib/native/libhadoop.so.1.0.0
which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
19/04/08 19:59:29 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java c
lasses where applicable
[04/08/19]seed@VM:~/.../hadoop-2.7.6$ head hadoop_home_page.html_OUTPUT_COUNT1/part-00000
None    0
">      4
"AS     8
"License");     8
$('table').addClass('table      1
$(function()    1
&raquo;</a>     3
&rarr;</a></p>  2
'UA-7453027-1', 1
'auto');        1
[04/08/19]seed@VM:~/.../hadoop-2.7.6$
```