Big Data Lab 2

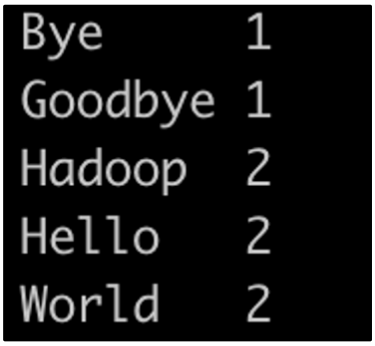
1. **Word Count**

   Total input paths to process : 2

   number of splits:3

```bash
hadoop fs -cat /user/beju1022/wc/out/part-00000
```

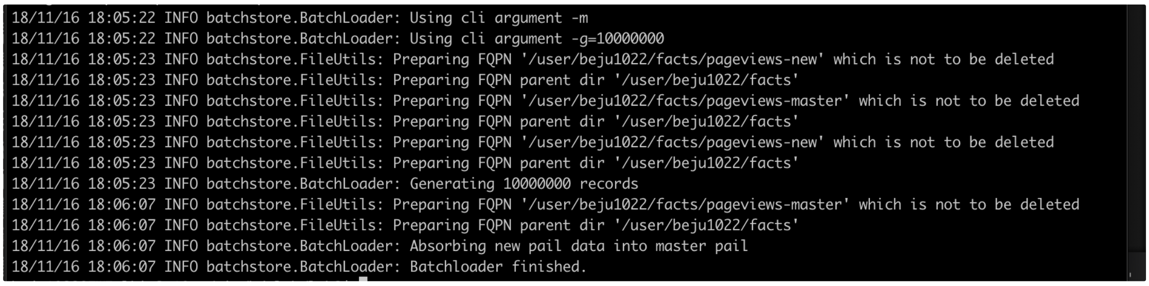
   Output:



```
Bye     1
Goodbye 1
Hadoop  2
Hello   2
World   2
```

2. **PageViews:**

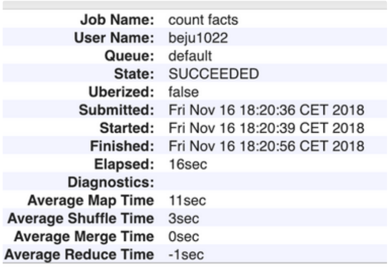
   - Generate new PageViews:



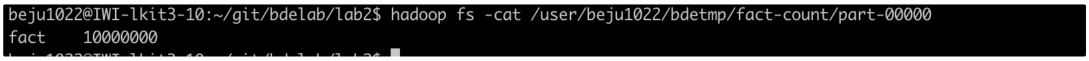
```
18/11/16 18:05:22 INFO batchstore.BatchLoader: Using cli argument -m
18/11/16 18:05:22 INFO batchstore.BatchLoader: Using cli argument -g=10000000
18/11/16 18:05:23 INFO batchstore.FileUtils: Preparing FQPN '/user/beju1022/facts/pageviews-new' which is not to be deleted
18/11/16 18:05:23 INFO batchstore.FileUtils: Preparing FQPN parent dir '/user/beju1022/facts'
18/11/16 18:05:23 INFO batchstore.FileUtils: Preparing FQPN '/user/beju1022/facts/pageviews-master' which is not to be deleted
18/11/16 18:05:23 INFO batchstore.FileUtils: Preparing FQPN parent dir '/user/beju1022/facts'
18/11/16 18:05:23 INFO batchstore.FileUtils: Preparing FQPN '/user/beju1022/facts/pageviews-new' which is not to be deleted
18/11/16 18:05:23 INFO batchstore.FileUtils: Preparing FQPN parent dir '/user/beju1022/facts'
18/11/16 18:05:23 INFO batchstore.BatchLoader: Generating 10000000 records
18/11/16 18:06:07 INFO batchstore.FileUtils: Preparing FQPN '/user/beju1022/facts/pageviews-master' which is not to be deleted
18/11/16 18:06:07 INFO batchstore.FileUtils: Preparing FQPN parent dir '/user/beju1022/facts'
18/11/16 18:06:07 INFO batchstore.BatchLoader: Absorbing new pail data into master pail
18/11/16 18:06:07 INFO batchstore.BatchLoader: Batchloader finished.
```

   - Ausführen des Map Reduce Jobs:

     - 7 splits



Job Name:	count facts
User Name:	beju1022
Queue:	default
State:	SUCCEEDED
Uberized:	false
Submitted:	Fri Nov 16 18:20:36 CET 2018
Started:	Fri Nov 16 18:20:39 CET 2018
Finished:	Fri Nov 16 18:20:56 CET 2018
Elapsed:	16sec
Diagnostics:	
Average Map Time	11sec
Average Shuffle Time	3sec
Average Merge Time	0sec
Average Reduce Time	-1sec



```
beju1022@IWI-lkit3-10:~/git/bdelab/lab2$ hadoop fs -cat /user/beju1022/bdetmp/fact-count/part-00000
fact    10000000
```

Aufgabe 2.1.2: Untersuchen sie den Pageview Index Job

▼ Output Map Reduce

```
18/11/19 12:21:22 INFO batchstore.FileUtils: Preparing FQPN
'/user/beju1022/facts/pageviews-master' which is not to be
deleted 18/11/19 12:21:22 INFO batchstore.FileUtils:
Preparing FQPN parent dir '/user/beju1022/facts' 18/11/19
12:21:23 INFO batchstore.FileUtils: Preparing FQPN
'/user/beju1022/bdetmp/page-count' which is to be deleted
18/11/19 12:21:23 INFO batchstore.FileUtils: Preparing FQPN
parent dir '/user/beju1022/bdetmp' 18/11/19 12:21:23 INFO
client.RMProxy: Connecting to ResourceManager at IWI-lkit-
ux-06.HS-Karlsruhe.DE/193.196.105.68:8050 18/11/19 12:21:23
INFO client.RMProxy: Connecting to ResourceManager at IWI-
lkit-ux-06.HS-Karlsruhe.DE/193.196.105.68:8050 18/11/19
12:21:23 WARN mapreduce.JobResourceUploader: Hadoop
command-line option parsing not performed. Implement the
Tool interface and execute your application with ToolRunner
to remedy this. 18/11/19 12:21:23 INFO net.NetworkTopology:
Adding a new node: /default-rack/193.196.105.110:50010
18/11/19 12:21:23 INFO net.NetworkTopology: Adding a new
node: /default-rack/193.196.105.103:50010 18/11/19 12:21:23
INFO net.NetworkTopology: Adding a new node: /default-
rack/193.196.105.113:50010 18/11/19 12:21:23 INFO
net.NetworkTopology: Adding a new node: /default-
rack/193.196.105.101:50010 18/11/19 12:21:23 INFO
net.NetworkTopology: Adding a new node: /default-
rack/193.196.105.109:50010 18/11/19 12:21:23 INFO
net.NetworkTopology: Adding a new node: /default-
rack/193.196.105.106:50010 18/11/19 12:21:26 INFO
hdfs.DFSClient: Exception in createBlockOutputStream
java.io.IOException: Got error, status message , ack with
firstBadLink as 193.196.105.120:50010 at
org.apache.hadoop.hdfs.protocol.datatransfer.DataTransferPr
otoUtil.checkBlockOpStatus(DataTransferProtoUtil.java:140)
at
org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.createB
lockOutputStream(DFSOutputStream.java:1359) at
org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.nextBlo
ckOutputStream(DFSOutputStream.java:1262) at
org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFS
OutputStream.java:448) 18/11/19 12:21:26 INFO
```

```
hdfs.DFSClient: Abandoning BP-1368902507-193.196.105.68-
1489054816217:blk_1073829941_89132 18/11/19 12:21:26 INFO
hdfs.DFSClient: Excluding datanode
DatanodeInfoWithStorage[193.196.105.120:50010,DS-128259d4-
23e5-4ff1-9488-daead26a0f11,DISK] 18/11/19 12:21:29 INFO
hdfs.DFSClient: Exception in createBlockOutputStream
java.io.IOException: Got error, status message , ack with
firstBadLink as 193.196.105.120:50010 at
org.apache.hadoop.hdfs.protocol.datatransfer.DataTransferPr
otoUtil.checkBlockOpStatus(DataTransferProtoUtil.java:140)
at
org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.createB
lockOutputStream(DFSOutputStream.java:1359) at
org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.nextBlo
ckOutputStream(DFSOutputStream.java:1262) at
org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFS
OutputStream.java:448) 18/11/19 12:21:29 INFO
hdfs.DFSClient: Abandoning BP-1368902507-193.196.105.68-
1489054816217:blk_1073829944_89135 18/11/19 12:21:29 INFO
hdfs.DFSClient: Excluding datanode
DatanodeInfoWithStorage[193.196.105.120:50010,DS-128259d4-
23e5-4ff1-9488-daead26a0f11,DISK] 18/11/19 12:21:29 INFO
mapreduce.JobSubmitter: number of splits:7 18/11/19
12:21:29 INFO mapreduce.JobSubmitter: Submitting tokens for
job: job_1541429278487_0035 18/11/19 12:21:29 INFO
impl.YarnClientImpl: Submitted application
application_1541429278487_0035 18/11/19 12:21:29 INFO
mapreduce.Job: The url to track the job: http://IWI-lkit-
ux-06.HS-
Karlsruhe.DE:8088/proxy/application_1541429278487_0035/
18/11/19 12:21:29 INFO mapreduce.Job: Running job:
job_1541429278487_0035 18/11/19 12:21:34 INFO
mapreduce.Job: Job job_1541429278487_0035 running in uber
mode : false 18/11/19 12:21:34 INFO mapreduce.Job: map 0%
reduce 0% 18/11/19 12:21:44 INFO mapreduce.Job: map 25%
reduce 0% 18/11/19 12:21:47 INFO mapreduce.Job: map 40%
reduce 0% 18/11/19 12:21:50 INFO mapreduce.Job: map 55%
reduce 0% 18/11/19 12:21:52 INFO mapreduce.Job: map 65%
reduce 0% 18/11/19 12:21:53 INFO mapreduce.Job: map 75%
reduce 0% 18/11/19 12:21:55 INFO mapreduce.Job: map 85%
reduce 0% 18/11/19 12:21:56 INFO mapreduce.Job: map 100%
reduce 0% 18/11/19 12:21:57 INFO mapreduce.Job: map 100%
reduce 100% 18/11/19 12:21:57 INFO mapreduce.Job: Job
job_1541429278487_0035 completed successfully 18/11/19
```

```
12:21:57 INFO mapreduce.Job: Counters: 50 File System
Counters FILE: Number of bytes read=7331770 FILE: Number of
bytes written=15619593 FILE: Number of read operations=0
FILE: Number of large read operations=0 FILE: Number of
write operations=0 HDFS: Number of bytes read=940878526
HDFS: Number of bytes written=1000196 HDFS: Number of read
operations=31 HDFS: Number of large read operations=0 HDFS:
Number of write operations=2 Job Counters Launched map
tasks=7 Launched reduce tasks=1 Data-local map tasks=6
Rack-local map tasks=1 Total time spent by all maps in
occupied slots (ms)=128334 Total time spent by all reduces
in occupied slots (ms)=2759 Total time spent by all map
tasks (ms)=128334 Total time spent by all reduce tasks
(ms)=2759 Total vcore-milliseconds taken by all map
tasks=128334 Total vcore-milliseconds taken by all reduce
tasks=2759 Total megabyte-milliseconds taken by all map
tasks=131414016 Total megabyte-milliseconds taken by all
reduce tasks=2825216 Map-Reduce Framework Map input
records=10000000 Map output records=10000000 Map output
bytes=408960000 Map output materialized bytes=7331806 Input
split bytes=1897 Combine input records=10000000 Combine
output records=170835 Reduce input groups=24434 Reduce
shuffle bytes=7331806 Reduce input records=170835 Reduce
output records=24434 Spilled Records=341670 Shuffled Maps
=7 Failed Shuffles=0 Merged Map outputs=7 GC time elapsed
(ms)=2527 CPU time spent (ms)=150510 Physical memory
(bytes) snapshot=2355933184 Virtual memory (bytes)
snapshot=15535398912 Total committed heap usage
(bytes)=1572339712 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=940876629 File Output
Format Counters Bytes Written=1000196
```

JavaScript

▼ URLS Beispiel ausgabe:

```
02.11.2016 - 00 http://ft.com/ 401 02.11.2016 - 00
http://ftc.gov/ 188 02.11.2016 - 00 http://gizmodo.com/ 232
02.11.2016 - 00 http://globo.com/ 220 02.11.2016 - 00
http://goo.ne.jp/ 217 02.11.2016 - 00 http://goodreads.com/
30 02.11.2016 - 00 http://google.co.jp/ 213 02.11.2016 - 00
http://google.com.br/ 214 02.11.2016 - 00 http://google.fr/
205 02.11.2016 - 00 http://google.ru/ 246 02.11.2016 - 00
```

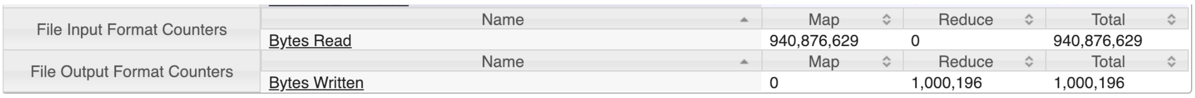
```
http://gov.uk/ 196 02.11.2016 - 00 http://gravatar.com/ 198
02.11.2016 - 00 http://guardian.co.uk/ 230 02.11.2016 - 00
http://hao123.com/ 230 02.11.2016 - 00 http://hc360.com/
352 02.11.2016 - 00 http://hhs.gov/ 225 02.11.2016 - 00
http://hibu.com/ 51 02.11.2016 - 00 http://histats.com/ 262
02.11.2016 - 00 http://hostgator.com/ 210 02.11.2016 - 00
http://house.gov/ 326 02.11.2016 - 00
http://howstuffworks.com/ 199 02.11.2016 - 00
http://hubpages.com/ 219 02.11.2016 - 00
http://hugedomains.com/ 52 02.11.2016 - 00 http://icio.us/
218 02.11.2016 - 00 http://icq.com/ 201 02.11.2016 - 00
http://infoseek.co.jp/ 212 02.11.2016 - 00
http://instagram.com/ 49 02.11.2016 - 00 http://issuu.com/
218 02.11.2016 - 00 http://istockphoto.com/ 212 02.11.2016
- 00 http://java.com/ 187 02.11.2016 - 00
http://kickstarter.com/ 196 02.11.2016 - 00 http://last.fm/
450 02.11.2016 - 00 http://latimes.com/ 7 02.11.2016 - 00
http://lycos.com/ 172 02.11.2016 - 00 http://mapquest.com/
200 02.11.2016 - 00 http://mayoclinic.com/ 205 02.11.2016 -
00 http://miibeian.gov.cn/ 204 02.11.2016 - 00
http://mit.edu/ 188 02.11.2016 - 00 http://mtv.com/ 149
02.11.2016 - 00 http://mysql.com/ 210 02.11.2016 - 00
http://naver.com/ 218 02.11.2016 - 00 http://nbcnews.com/
184 02.11.2016 - 00 http://netlog.com/ 216 02.11.2016 - 00
http://netscape.com/ 26 02.11.2016 - 00
http://networksolutions.com/ 236 02.11.2016 - 00
http://newsvine.com/ 212 02.11.2016 - 00 http://nifty.com/
114 02.11.2016 - 00 http://nih.gov/ 236 02.11.2016 - 00
http://nymag.com/ 194 02.11.2016 - 00 http://oracle.com/
407 02.11.2016 - 00 http://pagesperso-orange.fr/ 230
02.11.2016 - 00 http://parallels.com/ 388
```

JavaScript

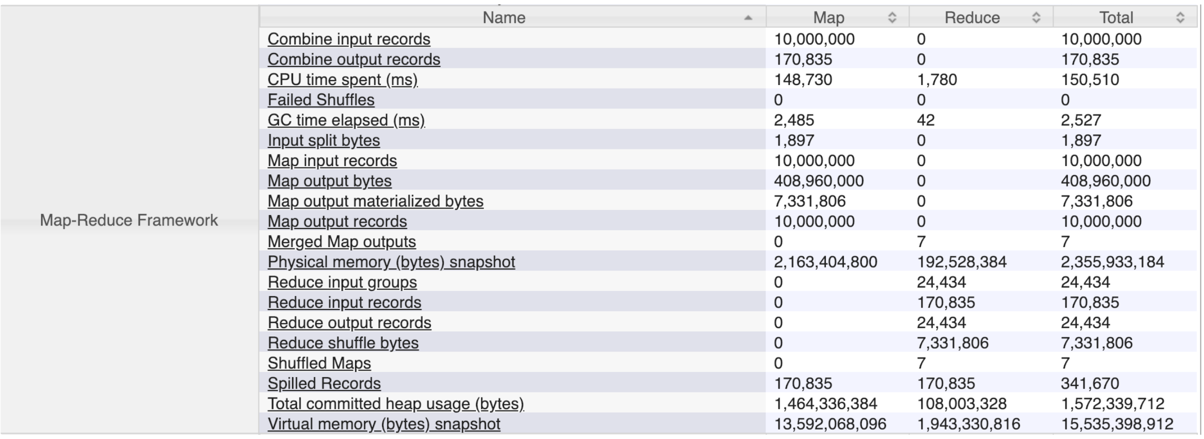
- Wie lange ist der Job gelaufen?
    - 25 sec
- Wieviele Mapper sind zum EINSATZ gekommen und wie lange haben diese im Schnitt gearbeitet?
    - Splits: 7
- Wieviele Daten haben Mapper und Reducer jeweils erzeugt?

http://iwi-lkit-hadoop.hs-karlsruhe.de:19888/jobhistory/jobcounters/job_1541429278487_0035

Siehe Counter für den Job:



	Name	Map	Reduce	Total
File Input Format Counters	Bytes Read	940,876,629	0	940,876,629
File Output Format Counters	Name	Map	Reduce	Total
	Bytes Written	0	1,000,196	1,000,196



	Name	Map	Reduce	Total
Map-Reduce Framework	Combine input records	10,000,000	0	10,000,000
	Combine output records	170,835	0	170,835
	CPU time spent (ms)	148,730	1,780	150,510
	Failed Shuffles	0	0	0
	GC time elapsed (ms)	2,485	42	2,527
	Input split bytes	1,897	0	1,897
	Map input records	10,000,000	0	10,000,000
	Map output bytes	408,960,000	0	408,960,000
	Map output materialized bytes	7,331,806	0	7,331,806
	Map output records	10,000,000	0	10,000,000
	Merged Map outputs	0	7	7
	Physical memory (bytes) snapshot	2,163,404,800	192,528,384	2,355,933,184
	Reduce input groups	0	24,434	24,434
	Reduce input records	0	170,835	170,835
	Reduce output records	0	24,434	24,434
	Reduce shuffle bytes	0	7,331,806	7,331,806
	Shuffled Maps	0	7	7
	Spilled Records	170,835	170,835	341,670
	Total committed heap usage (bytes)	1,464,336,384	108,003,328	1,572,339,712
	Virtual memory (bytes) snapshot	13,592,068,096	1,943,330,816	15,535,398,912