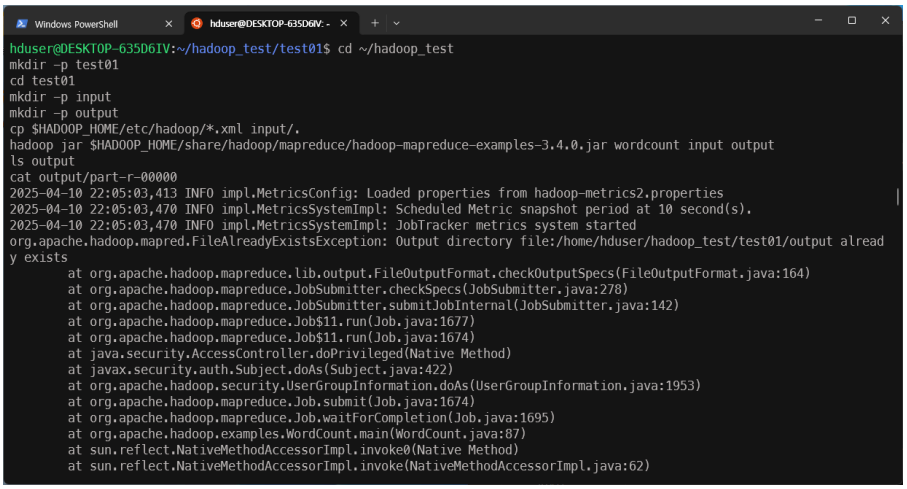
분산시스템 과제 #3

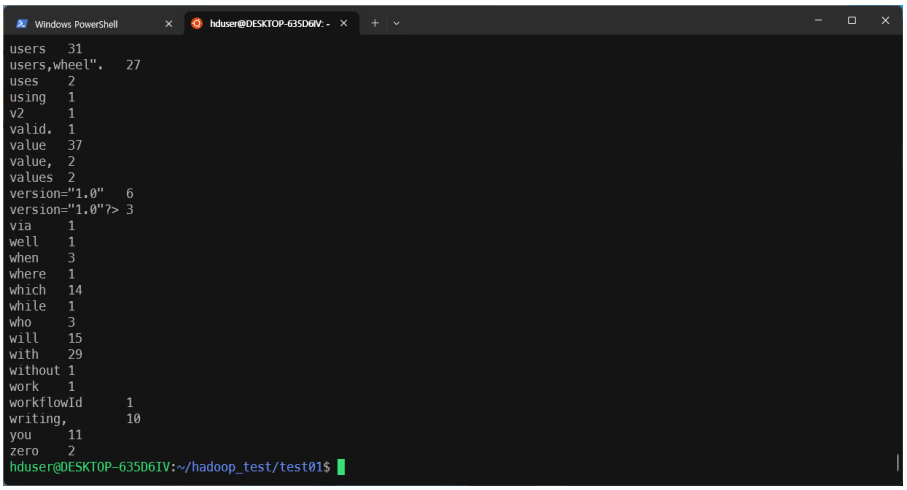
박종현, 2025-04-10
공과대학 컴퓨터정보통신공학과

과제 목표

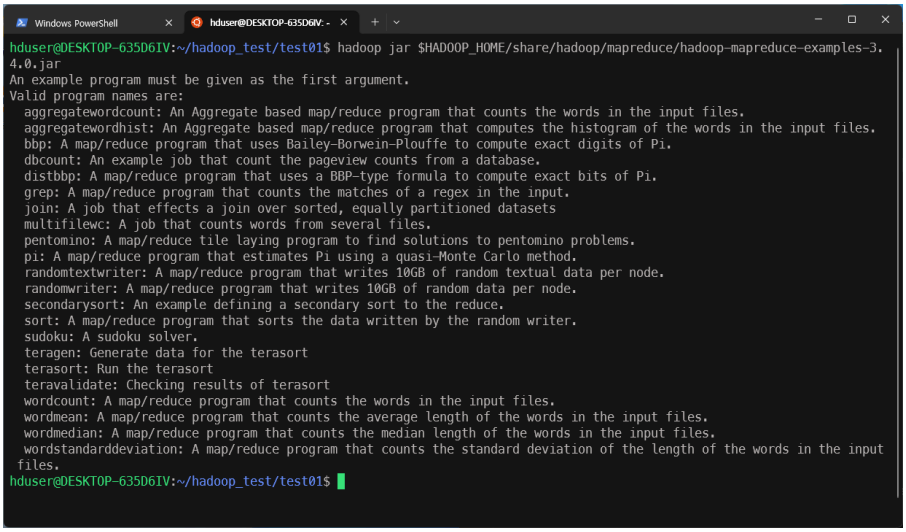
1. 2-4 자료 : 14, 15쪽
2. 2-5 자료 : 9, 16, 17쪽
3. 2-6 자료 : 25, 26쪽
4. 2-7 자료 : 10, 13, 24쪽

[Ch 2-4 p.14]

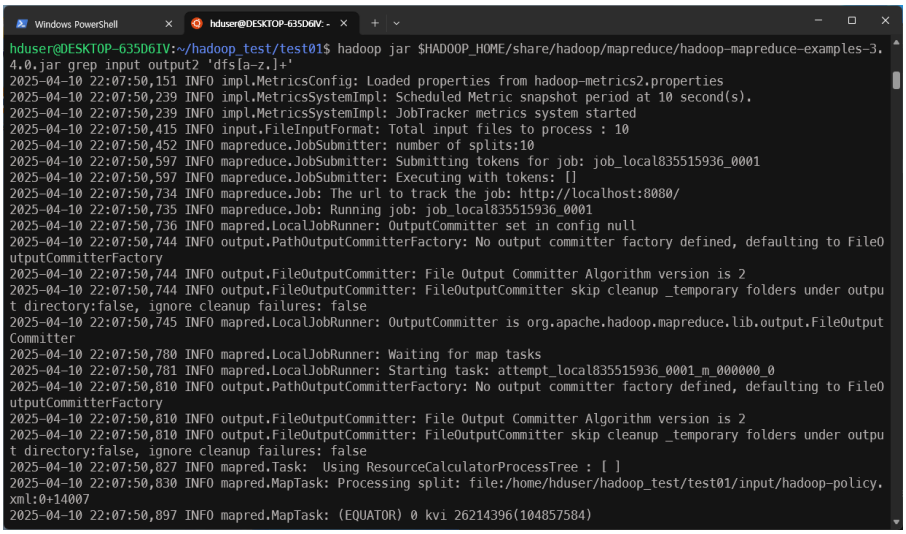




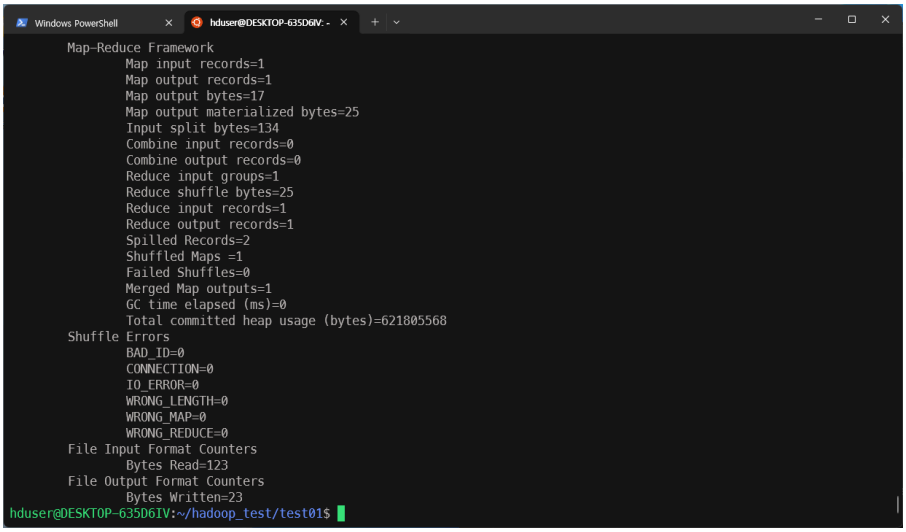
[Ch 2-4 p.15]



```
hduser@DESKTOP-635D6IV:~/hadoop_test/test01$ hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.
4.0.jar
An example program must be given as the first argument.
Valid program names are:
  aggregatewordcount: An Aggregate based map/reduce program that counts the words in the input files.
  aggregatewordhist: An Aggregate based map/reduce program that computes the histogram of the words in the input files.
  bbp: A map/reduce program that uses Bailey-Borwein-Plouffe to compute exact digits of Pi.
  dbcount: An example job that count the pageview counts from a database.
  distbbp: A map/reduce program that uses a BBP-type formula to compute exact bits of Pi.
  grep: A map/reduce program that counts the matches of a regex in the input.
  join: A job that effects a join over sorted, equally partitioned datasets
  multifilewc: A job that counts words from several files.
  pentomino: A map/reduce tile laying program to find solutions to pentomino problems.
  pi: A map/reduce program that estimates Pi using a quasi-Monte Carlo method.
  randomtextwriter: A map/reduce program that writes 10GB of random textual data per node.
  randomwriter: A map/reduce program that writes 10GB of random data per node.
  secondarysort: An example defining a secondary sort to the reduce.
  sort: A map/reduce program that sorts the data written by the random writer.
  sudoku: A sudoku solver.
  teragen: Generate data for the terasort
  terasort: Run the terasort
  teravalidate: Checking results of terasort
  wordcount: A map/reduce program that counts the words in the input files.
  wordmean: A map/reduce program that counts the average length of the words in the input files.
  wordmedian: A map/reduce program that counts the median length of the words in the input files.
  wordstandarddeviation: A map/reduce program that counts the standard deviation of the length of the words in the input
 files.
hduser@DESKTOP-635D6IV:~/hadoop_test/test01$
```

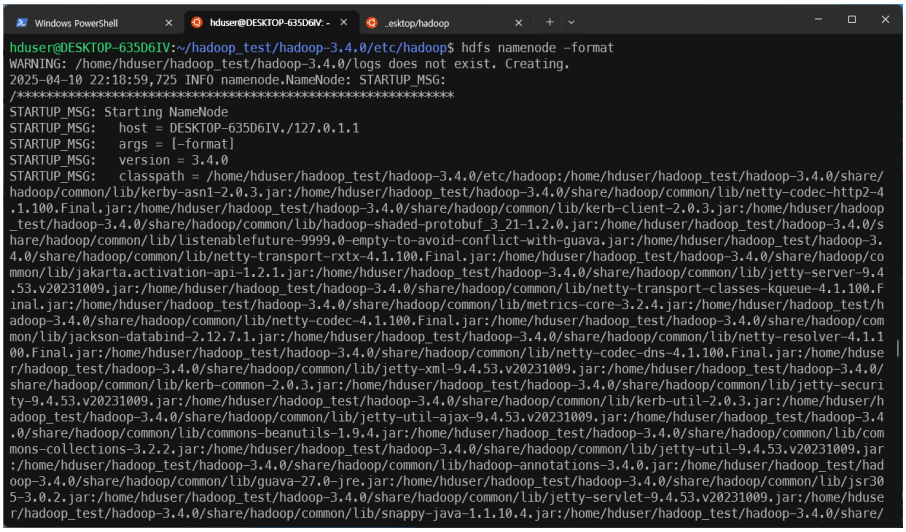


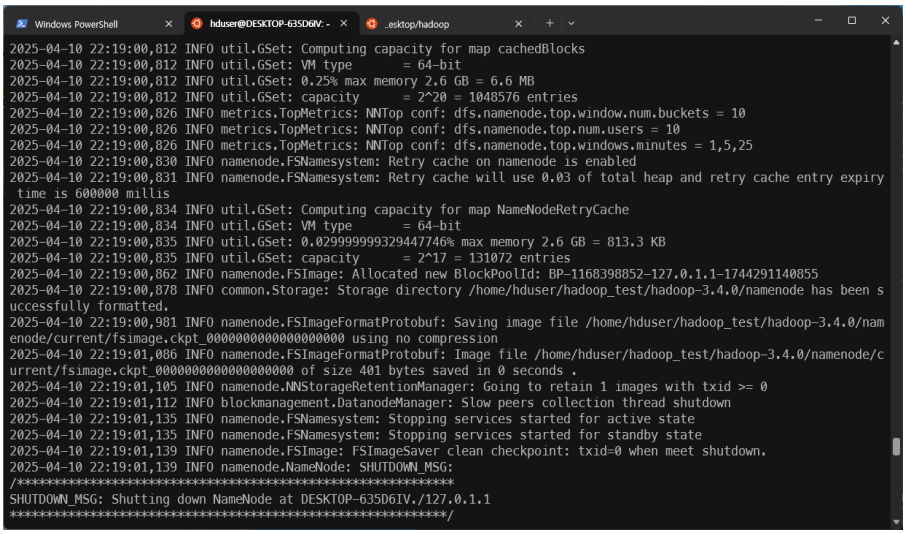
```
hduser@DESKTOP-635D6IV:~/hadoop_test/test01$ hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.
4.0.jar grep input output2 'dfs[a-z.]+'
2025-04-10 22:07:50,151 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-04-10 22:07:50,239 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-04-10 22:07:50,239 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-04-10 22:07:50,415 INFO input.FileInputFormat: Total input files to process : 10
2025-04-10 22:07:50,452 INFO mapreduce.JobSubmitter: number of splits:10
2025-04-10 22:07:50,597 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local835515936_0001
2025-04-10 22:07:50,597 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-10 22:07:50,734 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-04-10 22:07:50,735 INFO mapreduce.Job: Running job: job_local835515936_0001
2025-04-10 22:07:50,736 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2025-04-10 22:07:50,744 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileO
utputCommitterFactory
2025-04-10 22:07:50,744 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-04-10 22:07:50,744 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under outpu
t directory:false, ignore cleanup failures: false
2025-04-10 22:07:50,745 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutput
Committer
2025-04-10 22:07:50,780 INFO mapred.LocalJobRunner: Waiting for map tasks
2025-04-10 22:07:50,781 INFO mapred.LocalJobRunner: Starting task: attempt_local835515936_0001_m_000000_0
2025-04-10 22:07:50,810 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileO
utputCommitterFactory
2025-04-10 22:07:50,810 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-04-10 22:07:50,810 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under outpu
t directory:false, ignore cleanup failures: false
2025-04-10 22:07:50,827 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2025-04-10 22:07:50,830 INFO mapred.MapTask: Processing split: file:/home/hduser/hadoop_test/test01/input/hadoop-policy.
xml:0+14007
2025-04-10 22:07:50,897 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
```

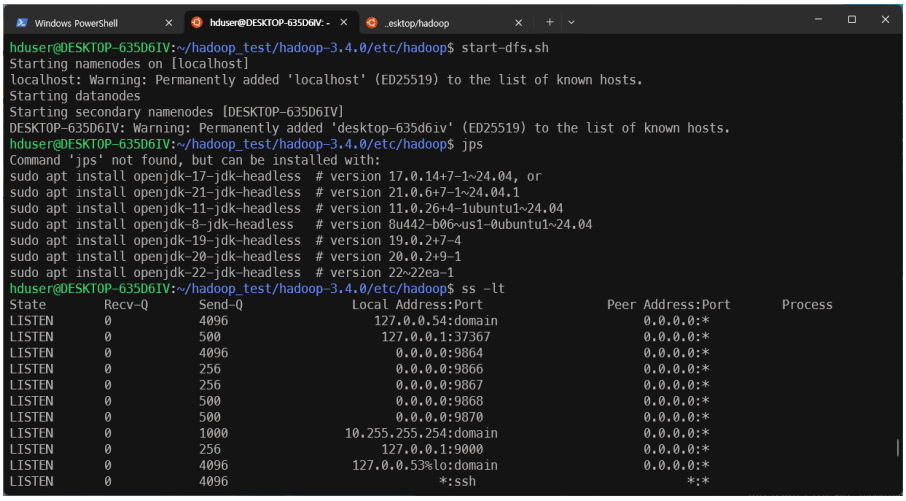


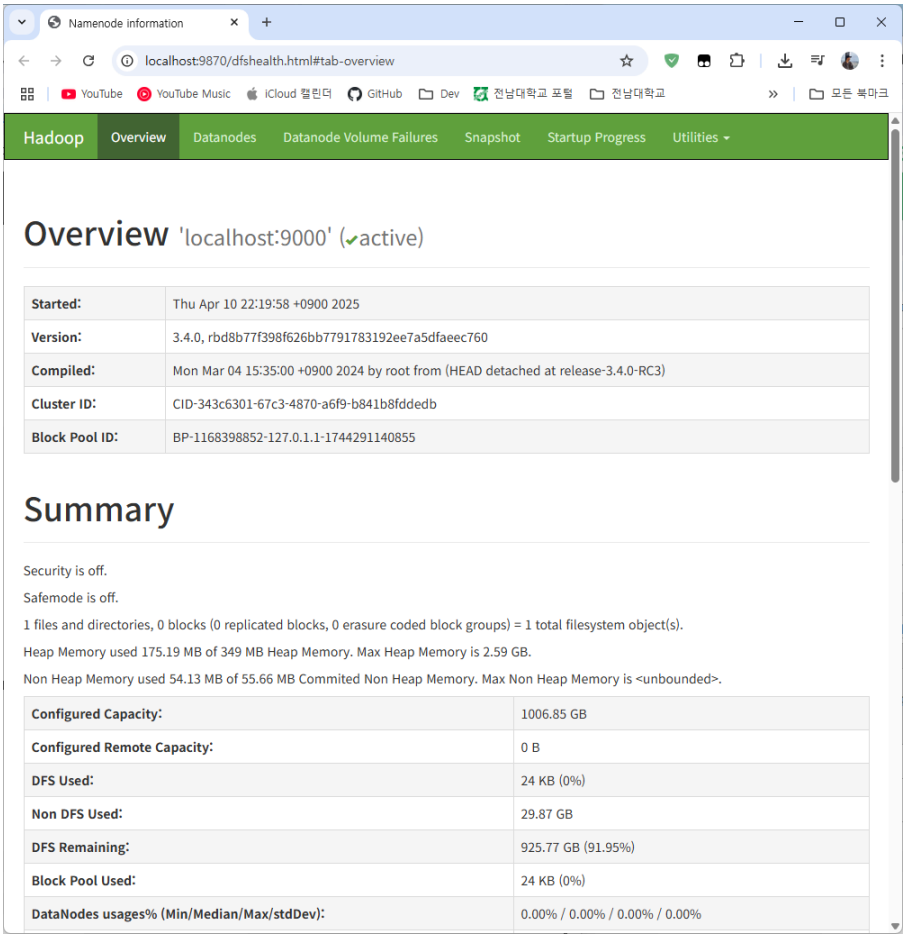
```
        Map-Reduce Framework
                Map input records=1
                Map output records=1
                Map output bytes=17
                Map output materialized bytes=25
                Input split bytes=134
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=25
                Reduce input records=1
                Reduce output records=1
                Spilled Records=2
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=621805568
        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=123
        File Output Format Counters
                Bytes Written=23
hduser@DESKTOP-635D6IV:~/hadoop_test/test01$
```

[Ch 2-5 p.9]

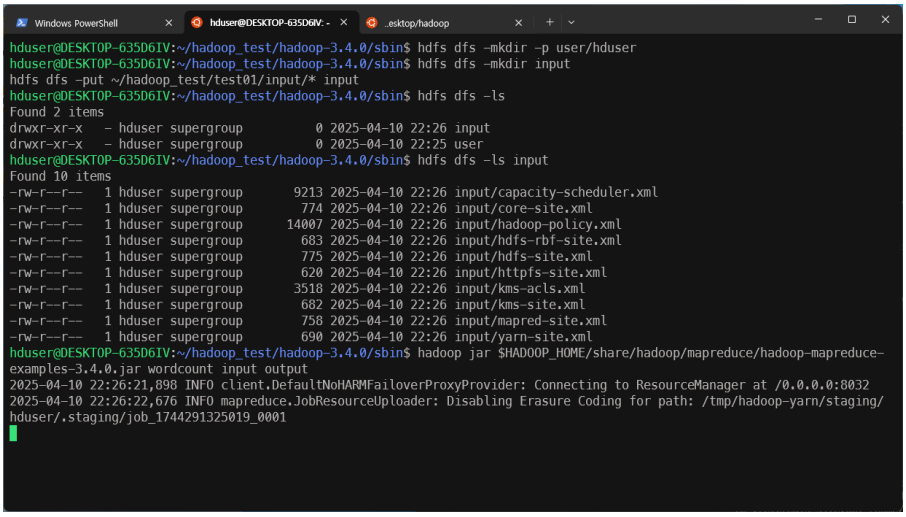

Windows PowerShell
hduser@DESKTOP-635D6IV: ~
~esktop/hadoop
hduser@DESKTOP-635D6IV:~/hadoop_test/hadoop-3.4.0/etc/hadoop$ hdfs namenode -format
WARNING: /home/hduser/hadoop_test/hadoop-3.4.0/logs does not exist. Creating.
2025-04-10 22:18:59,725 INFO namenode.NameNode: STARTUP_MSG:
/**
STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = DESKTOP-635D6IV./127.0.1.1
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 3.4.0
STARTUP_MSG: classpath = /home/hduser/hadoop_test/hadoop-3.4.0/etc/hadoop:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/kerby-asn1-2.0.3.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/netty-codec-http2-4.1.100.Final.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/kerb-client-2.0.3.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/hadoop-shaded-protobuf_3_21-1.2.0.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/listenablefuture-9999.0-empty-to-avoid-conflict-with-guava.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/netty-transport-rxtx-4.1.100.Final.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/jakarta.activation-api-1.2.1.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/jetty-server-9.4.53.v20231009.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/netty-transport-classes-kqueue-4.1.100.Final.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/metrics-core-3.2.4.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/netty-codec-4.1.100.Final.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/jackson-databind-2.12.7.1.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/netty-resolver-4.1.100.Final.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/netty-codec-dns-4.1.100.Final.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/jetty-xml-9.4.53.v20231009.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/kerb-common-2.0.3.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/jetty-security-9.4.53.v20231009.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/kerb-util-2.0.3.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/jetty-util-ajax-9.4.53.v20231009.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/commons-beanutils-1.9.4.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/commons-collections-3.2.2.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/jetty-util-9.4.53.v20231009.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/hadoop-annotations-3.4.0.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/guava-27.0-jre.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/jsr305-3.0.2.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/jetty-servlet-9.4.53.v20231009.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/hadoop/common/lib/snappy-java-1.1.10.4.jar:/home/hduser/hadoop_test/hadoop-3.4.0/share/

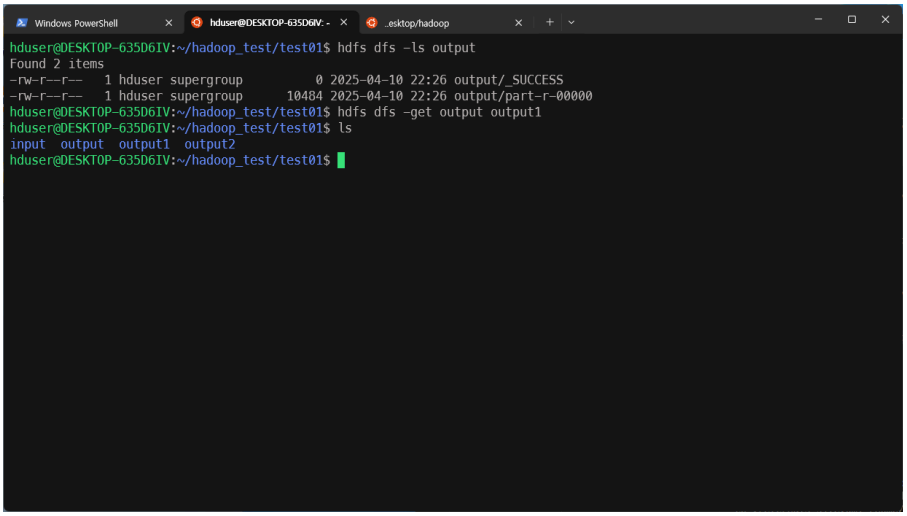

Windows PowerShell
hduser@DESKTOP-635D6IV: ~
~esktop/hadoop
2025-04-10 22:19:00,812 INFO util.GSet: Computing capacity for map cachedBlocks
2025-04-10 22:19:00,812 INFO util.GSet: VM type = 64-bit
2025-04-10 22:19:00,812 INFO util.GSet: 0.25% max memory 2.6 GB = 6.6 MB
2025-04-10 22:19:00,812 INFO util.GSet: capacity = 2^20 = 1048576 entries
2025-04-10 22:19:00,826 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
2025-04-10 22:19:00,826 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2025-04-10 22:19:00,826 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
2025-04-10 22:19:00,830 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2025-04-10 22:19:00,831 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 600000 millis
2025-04-10 22:19:00,834 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2025-04-10 22:19:00,834 INFO util.GSet: VM type = 64-bit
2025-04-10 22:19:00,835 INFO util.GSet: 0.029999999329447746% max memory 2.6 GB = 813.3 KB
2025-04-10 22:19:00,835 INFO util.GSet: capacity = 2^17 = 131072 entries
2025-04-10 22:19:00,862 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1168398852-127.0.1.1-1744291140855
2025-04-10 22:19:00,878 INFO common.Storage: Storage directory /home/hduser/hadoop_test/hadoop-3.4.0/namenode has been successfully formatted.
2025-04-10 22:19:00,981 INFO namenode.FSImageFormatProtobuf: Saving image file /home/hduser/hadoop_test/hadoop-3.4.0/namenode/current/fsimage.ckpt_0000000000000000000 using no compression
2025-04-10 22:19:01,086 INFO namenode.FSImageFormatProtobuf: Image file /home/hduser/hadoop_test/hadoop-3.4.0/namenode/current/fsimage.ckpt_0000000000000000000 of size 401 bytes saved in 0 seconds .
2025-04-10 22:19:01,105 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2025-04-10 22:19:01,112 INFO blockmanagement.DatanodeManager: Slow peers collection thread shutdown
2025-04-10 22:19:01,135 INFO namenode.FSNamesystem: Stopping services started for active state
2025-04-10 22:19:01,135 INFO namenode.FSNamesystem: Stopping services started for standby state
2025-04-10 22:19:01,139 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2025-04-10 22:19:01,139 INFO namenode.NameNode: SHUTDOWN_MSG:
/**
SHUTDOWN_MSG: Shutting down NameNode at DESKTOP-635D6IV./127.0.1.1
**/


Windows PowerShell
hduser@DESKTOP-635D6IV: ~
~esktop/hadoop
hduser@DESKTOP-635D6IV:~/hadoop_test/hadoop-3.4.0/etc/hadoop$ start-dfs.sh
Starting namenodes on [localhost]
localhost: Warning: Permanently added 'localhost' (ED25519) to the list of known hosts.
Starting datanodes
Starting secondary namenodes [DESKTOP-635D6IV]
DESKTOP-635D6IV: Warning: Permanently added 'desktop-635d6iv' (ED25519) to the list of known hosts.
hduser@DESKTOP-635D6IV:~/hadoop_test/hadoop-3.4.0/etc/hadoop$ jps
Command 'jps' not found, but can be installed with:
sudo apt install openjdk-17-jdk-headless # version 17.0.14+7-1~24.04, or
sudo apt install openjdk-21-jdk-headless # version 21.0.6+7-1~24.04.1
sudo apt install openjdk-11-jdk-headless # version 11.0.26+4-1ubuntu1~24.04
sudo apt install openjdk-8-jdk-headless # version 8u442-b06~us1-0ubuntu1~24.04
sudo apt install openjdk-19-jdk-headless # version 19.0.2+7-4
sudo apt install openjdk-20-jdk-headless # version 20.0.2+9-1
sudo apt install openjdk-22-jdk-headless # version 22~22ea-1
hduser@DESKTOP-635D6IV:~/hadoop_test/hadoop-3.4.0/etc/hadoop$ ss -lt
State Recv-Q Send-Q Local Address:Port Peer Address:Port Process
LISTEN 0 4096 127.0.0.54:domain 0.0.0.0:*
LISTEN 0 500 127.0.0.1:37367 0.0.0.0:*
LISTEN 0 4096 0.0.0.0:9864 0.0.0.0:*
LISTEN 0 256 0.0.0.0:9866 0.0.0.0:*
LISTEN 0 256 0.0.0.0:9867 0.0.0.0:*
LISTEN 0 500 0.0.0.0:9868 0.0.0.0:*
LISTEN 0 500 0.0.0.0:9870 0.0.0.0:*
LISTEN 0 1000 10.255.255.254:domain 0.0.0.0:*
LISTEN 0 256 127.0.0.1:9000 0.0.0.0:*
LISTEN 0 4096 127.0.0.53%lo:domain 0.0.0.0:*
LISTEN 0 4096 *:ssh *:*

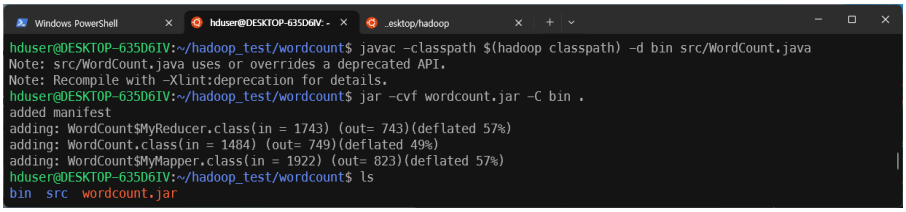

Namenode information
localhost:9870/dfshealth.html#tab-overview
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Overview 'localhost:9000' (active)
Started: Thu Apr 10 22:19:58 +0900 2025
Version: 3.4.0, rbd8b77f398f626bb7791783192ee7a5dfaeec760
Compiled: Mon Mar 04 15:35:00 +0900 2024 by root from (HEAD detached at release-3.4.0-RC3)
Cluster ID: CID-343c6301-67c3-4870-a6f9-b841b8fddedb
Block Pool ID: BP-1168398852-127.0.1.1-1744291140855
Summary
Security is off.
Safemode is off.
1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).
Heap Memory used 175.19 MB of 349 MB Heap Memory. Max Heap Memory is 2.59 GB.
Non Heap Memory used 54.13 MB of 55.66 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.
Configured Capacity: 1006.85 GB
Configured Remote Capacity: 0 B
DFS Used: 24 KB (0%)
Non DFS Used: 29.87 GB
DFS Remaining: 925.77 GB (91.95%)
Block Pool Used: 24 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev): 0.00% / 0.00% / 0.00% / 0.00%

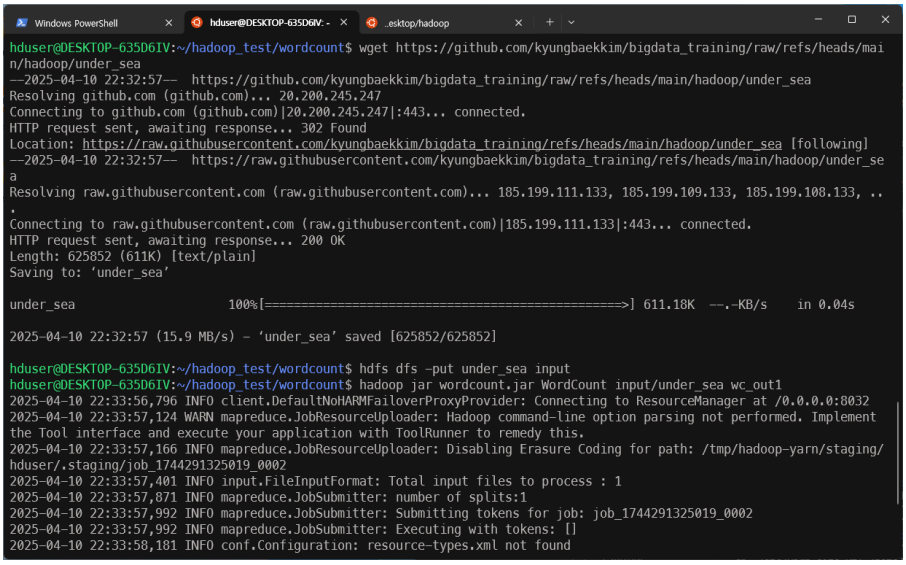
[Ch 2-5 p.16-17]

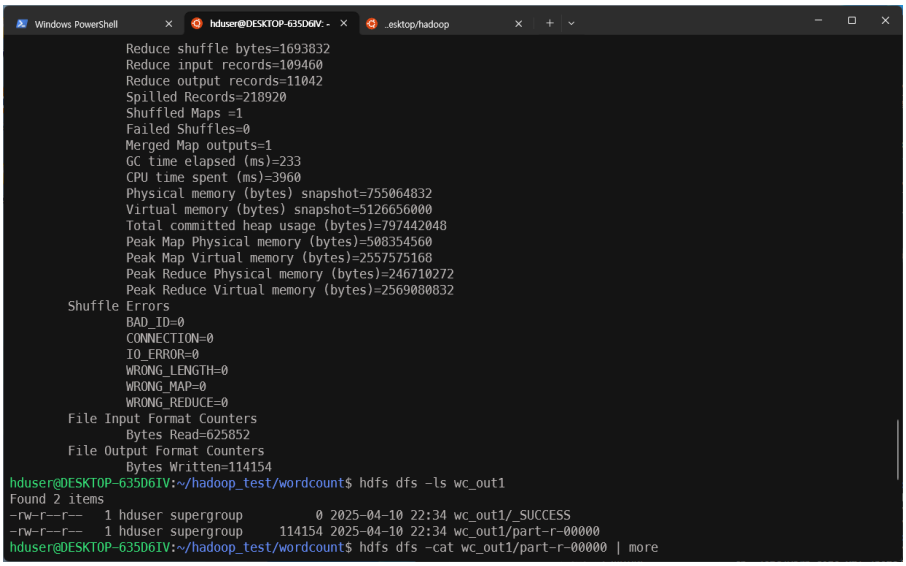

Windows PowerShell
hduser@DESKTOP-635D6IV: ~
~esktop/hadoop
hduser@DESKTOP-635D6IV:~/hadoop_test/hadoop-3.4.0/sbin$ hdfs dfs -mkdir -p user/hduser
hduser@DESKTOP-635D6IV:~/hadoop_test/hadoop-3.4.0/sbin$ hdfs dfs -mkdir input
hdfs dfs -put ~/hadoop_test/test01/input/* input
hduser@DESKTOP-635D6IV:~/hadoop_test/hadoop-3.4.0/sbin$ hdfs dfs -ls
Found 2 items
drwxr-xr-x - hduser supergroup 0 2025-04-10 22:26 input
drwxr-xr-x - hduser supergroup 0 2025-04-10 22:25 user
hduser@DESKTOP-635D6IV:~/hadoop_test/hadoop-3.4.0/sbin$ hdfs dfs -ls input
Found 10 items
-rw-r--r-- 1 hduser supergroup 9213 2025-04-10 22:26 input/capacity-scheduler.xml
-rw-r--r-- 1 hduser supergroup 774 2025-04-10 22:26 input/core-site.xml
-rw-r--r-- 1 hduser supergroup 14007 2025-04-10 22:26 input/hadoop-policy.xml
-rw-r--r-- 1 hduser supergroup 683 2025-04-10 22:26 input/hdfs-rbf-site.xml
-rw-r--r-- 1 hduser supergroup 775 2025-04-10 22:26 input/hdfs-site.xml
-rw-r--r-- 1 hduser supergroup 620 2025-04-10 22:26 input/httpfs-site.xml
-rw-r--r-- 1 hduser supergroup 3518 2025-04-10 22:26 input/kms-acls.xml
-rw-r--r-- 1 hduser supergroup 682 2025-04-10 22:26 input/kms-site.xml
-rw-r--r-- 1 hduser supergroup 758 2025-04-10 22:26 input/mapred-site.xml
-rw-r--r-- 1 hduser supergroup 690 2025-04-10 22:26 input/yarn-site.xml
hduser@DESKTOP-635D6IV:~/hadoop_test/hadoop-3.4.0/sbin$ hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.4.0.jar wordcount input output
2025-04-10 22:26:21,898 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-04-10 22:26:22,676 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hduser/.staging/job_1744291325019_0001

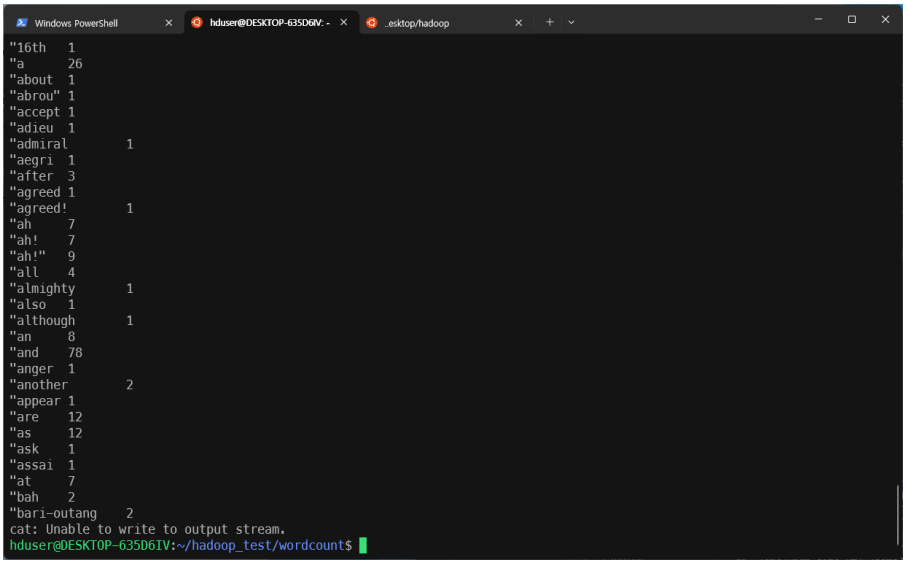

Windows PowerShell
hduser@DESKTOP-635D6IV: ~
~esktop/hadoop
hduser@DESKTOP-635D6IV:~/hadoop_test/test01$ hdfs dfs -ls output
Found 2 items
-rw-r--r--   1 hduser supergroup   0 2025-04-10 22:26 output/_SUCCESS
-rw-r--r--   1 hduser supergroup   10484 2025-04-10 22:26 output/part-r-00000
hduser@DESKTOP-635D6IV:~/hadoop_test/test01$ hdfs dfs -get output output1
hduser@DESKTOP-635D6IV:~/hadoop_test/test01$ ls
input output output1 output2
hduser@DESKTOP-635D6IV:~/hadoop_test/test01$

[Ch 2-6 p.25-26]

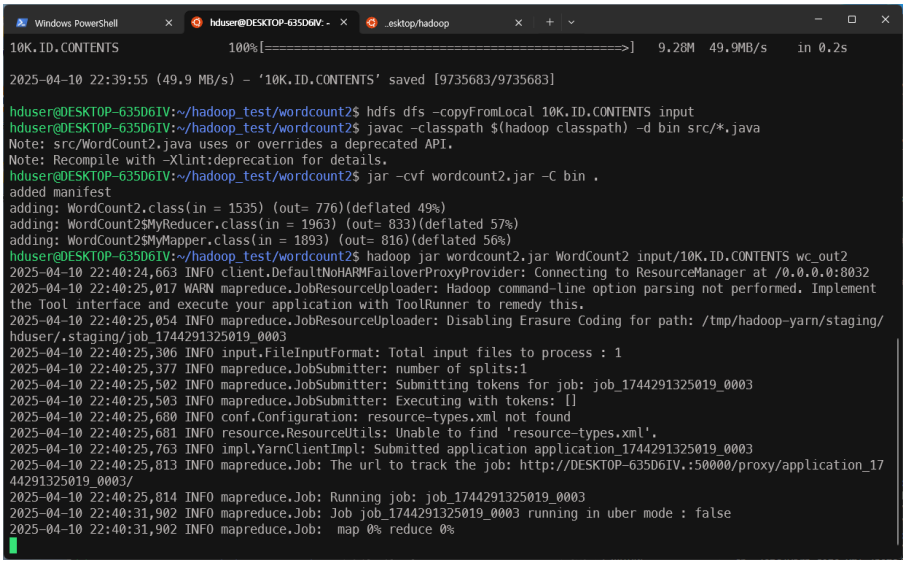

Windows PowerShell
hduser@DESKTOP-635D6IV: ~
~esktop/hadoop
hduser@DESKTOP-635D6IV:~/hadoop_test/wordcount$ javac -classpath $(hadoop classpath) -d bin src/WordCount.java
Note: src/WordCount.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
hduser@DESKTOP-635D6IV:~/hadoop_test/wordcount$ jar -cvf wordcount.jar -C bin .
added manifest
adding: WordCount$MyReducer.class(in = 1743) (out= 743)(deflated 57%)
adding: WordCount.class(in = 1484) (out= 749)(deflated 49%)
adding: WordCount$MyMapper.class(in = 1922) (out= 823)(deflated 57%)
hduser@DESKTOP-635D6IV:~/hadoop_test/wordcount$ ls
bin src wordcount.jar

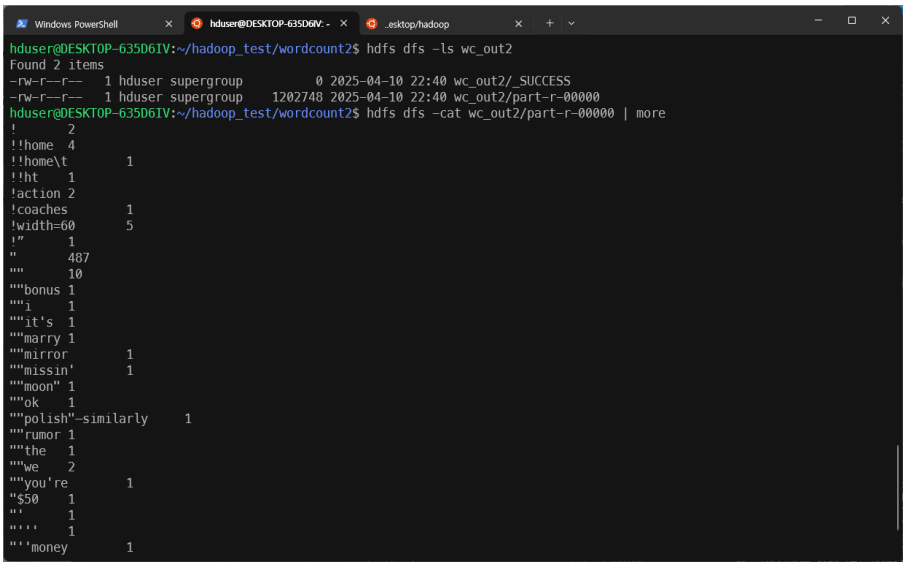

Windows PowerShell
hduser@DESKTOP-635D6IV: ~
~esktop/hadoop
hduser@DESKTOP-635D6IV:~/hadoop_test/wordcount$ wget https://github.com/kyungbaekkim/bigdata_training/raw/refs/heads/main/hadoop/under_sea
--2025-04-10 22:32:57-- https://github.com/kyungbaekkim/bigdata_training/raw/refs/heads/main/hadoop/under_sea
Resolving github.com (github.com)... 20.200.245.247
Connecting to github.com (github.com)|20.200.245.247|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://raw.githubusercontent.com/kyungbaekkim/bigdata_training/refs/heads/main/hadoop/under_sea [following]
--2025-04-10 22:32:57-- https://raw.githubusercontent.com/kyungbaekkim/bigdata_training/refs/heads/main/hadoop/under_sea
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.111.133, 185.199.109.133, 185.199.108.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.111.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 625852 (611K) [text/plain]
Saving to: 'under_sea'
under_sea 100%[=====>] 611.18K --.-KB/s in 0.04s
2025-04-10 22:32:57 (15.9 MB/s) - 'under_sea' saved [625852/625852]
hduser@DESKTOP-635D6IV:~/hadoop_test/wordcount$ hdfs dfs -put under_sea input
hduser@DESKTOP-635D6IV:~/hadoop_test/wordcount$ hadoop jar wordcount.jar WordCount input/under_sea wc_out1
2025-04-10 22:33:56,796 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-04-10 22:33:57,124 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-04-10 22:33:57,166 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hduser/.staging/job_1744291325019_0002
2025-04-10 22:33:57,401 INFO input.FileInputFormat: Total input files to process : 1
2025-04-10 22:33:57,871 INFO mapreduce.JobSubmitter: number of splits:1
2025-04-10 22:33:57,992 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744291325019_0002
2025-04-10 22:33:57,992 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-10 22:33:58,181 INFO conf.Configuration: resource-types.xml not found


Windows PowerShell
hduser@DESKTOP-635D6IV: ~
~esktop/hadoop
Reduce shuffle bytes=1693832
Reduce input records=109460
Reduce output records=11042
Spilled Records=218920
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=233
CPU time spent (ms)=3960
Physical memory (bytes) snapshot=755064832
Virtual memory (bytes) snapshot=5126656000
Total committed heap usage (bytes)=797442048
Peak Map Physical memory (bytes)=508354560
Peak Map Virtual memory (bytes)=2557575168
Peak Reduce Physical memory (bytes)=246710272
Peak Reduce Virtual memory (bytes)=2569080832
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=625852
File Output Format Counters
Bytes Written=114154
hduser@DESKTOP-635D6IV:~/hadoop_test/wordcount$ hdfs dfs -ls wc_out1
Found 2 items
-rw-r--r-- 1 hduser supergroup 0 2025-04-10 22:34 wc_out1/_SUCCESS
-rw-r--r-- 1 hduser supergroup 114154 2025-04-10 22:34 wc_out1/part-r-00000
hduser@DESKTOP-635D6IV:~/hadoop_test/wordcount$ hdfs dfs -cat wc_out1/part-r-00000 | more

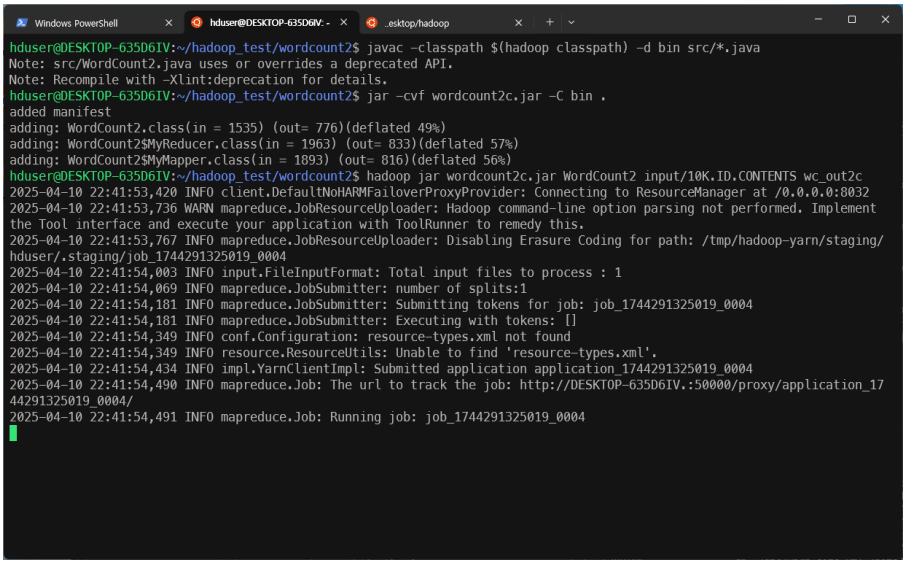

Windows PowerShell
hduser@DESKTOP-635D6IV: ~
..esktop/hadoop
"16th 1
"a 26
"about 1
"abrou" 1
"accept 1
"adieu 1
"admiral 1
"aegri 1
"after 3
"agreed 1
"agreed! 1
"ah 7
"ah! 7
"ah!" 9
"all 4
"almighty 1
"also 1
"although 1
"an 8
"and 78
"anger 1
"another 2
"appear 1
"are 12
"as 12
"ask 1
"assai 1
"at 7
"bah 2
"bari-outang 2
cat: Unable to write to output stream.
hduser@DESKTOP-635D6IV:~/hadoop_test/wordcount$

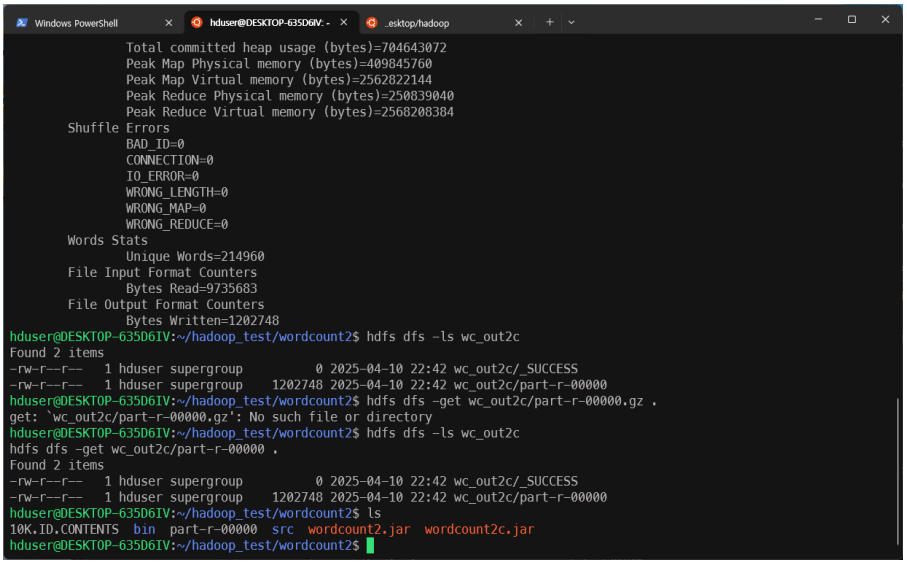
[Ch 2-7 p.10]



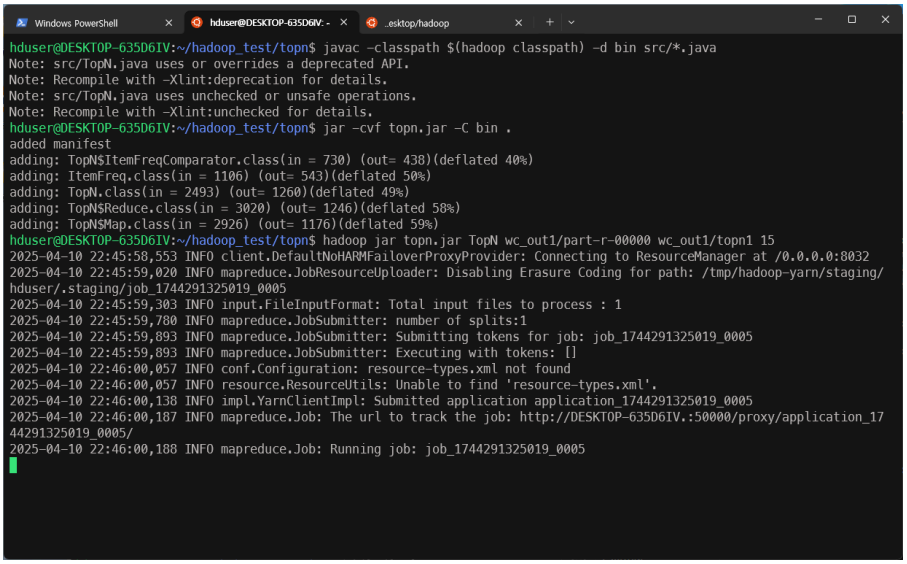


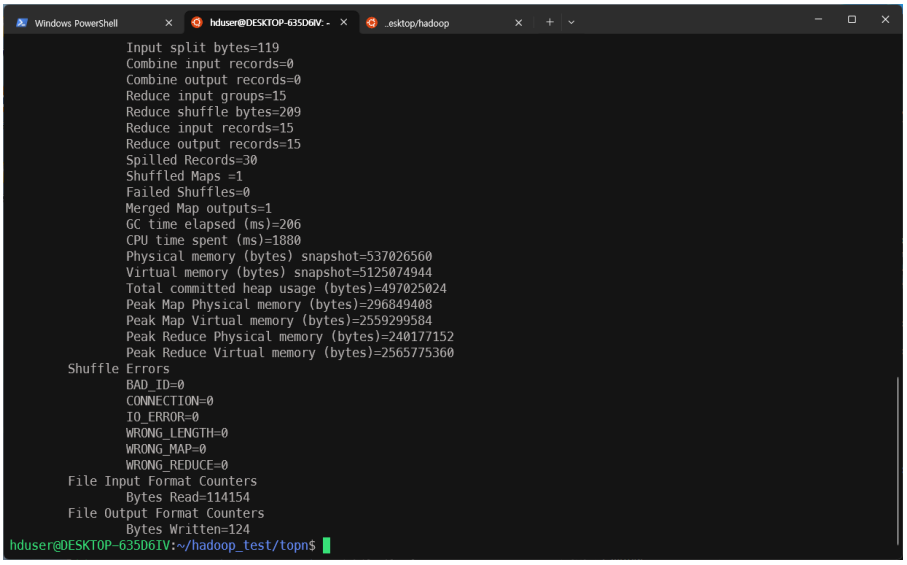
[Ch 2-7 p.13]

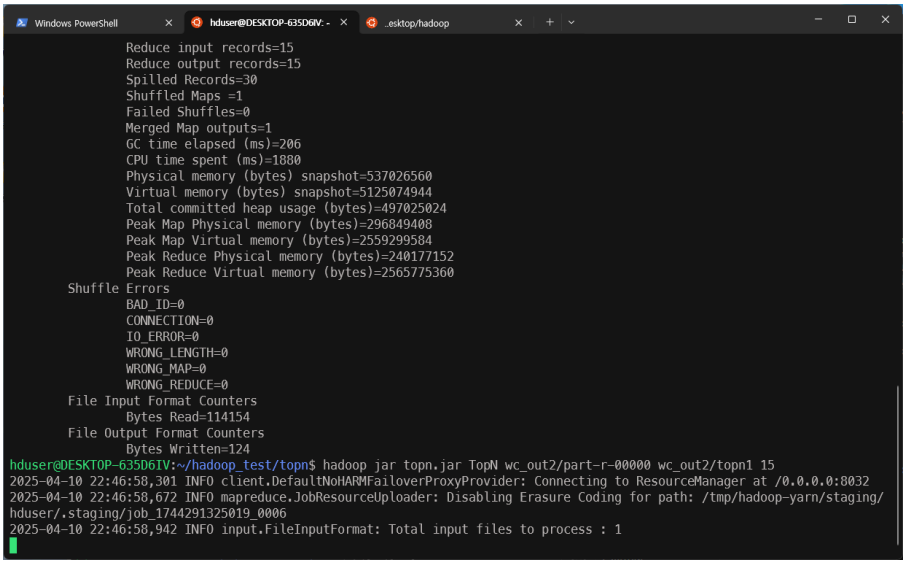


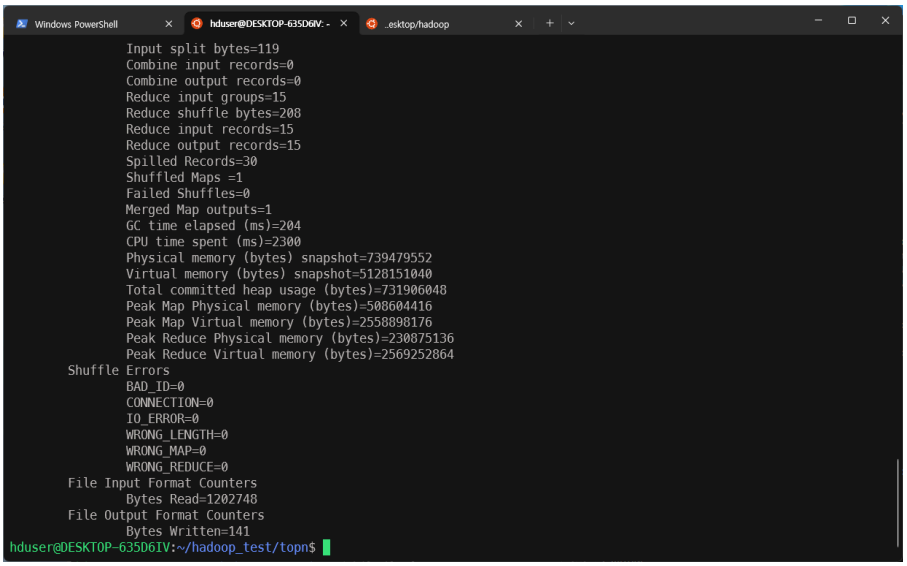

Windows PowerShell
hduser@DESKTOP-635D6IV: ~
~esktop/hadoop
Total committed heap usage (bytes)=704643072
Peak Map Physical memory (bytes)=409845760
Peak Map Virtual memory (bytes)=2562822144
Peak Reduce Physical memory (bytes)=250839040
Peak Reduce Virtual memory (bytes)=2568208384
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
Words Stats
Unique Words=214960
File Input Format Counters
Bytes Read=9735683
File Output Format Counters
Bytes Written=1202748
hduser@DESKTOP-635D6IV:~/hadoop_test/wordcount2$ hdfs dfs -ls wc_out2c
Found 2 items
-rw-r--r-- 1 hduser supergroup 0 2025-04-10 22:42 wc_out2c/_SUCCESS
-rw-r--r-- 1 hduser supergroup 1202748 2025-04-10 22:42 wc_out2c/part-r-00000
hduser@DESKTOP-635D6IV:~/hadoop_test/wordcount2$ hdfs dfs -get wc_out2c/part-r-00000.gz .
get: `wc_out2c/part-r-00000.gz': No such file or directory
hduser@DESKTOP-635D6IV:~/hadoop_test/wordcount2$ hdfs dfs -ls wc_out2c
hdfs dfs -get wc_out2c/part-r-00000 .
Found 2 items
-rw-r--r-- 1 hduser supergroup 0 2025-04-10 22:42 wc_out2c/_SUCCESS
-rw-r--r-- 1 hduser supergroup 1202748 2025-04-10 22:42 wc_out2c/part-r-00000
hduser@DESKTOP-635D6IV:~/hadoop_test/wordcount2$ ls
10K.ID.CONTENTS bin part-r-00000 src wordcount2.jar wordcount2c.jar
hduser@DESKTOP-635D6IV:~/hadoop_test/wordcount2$

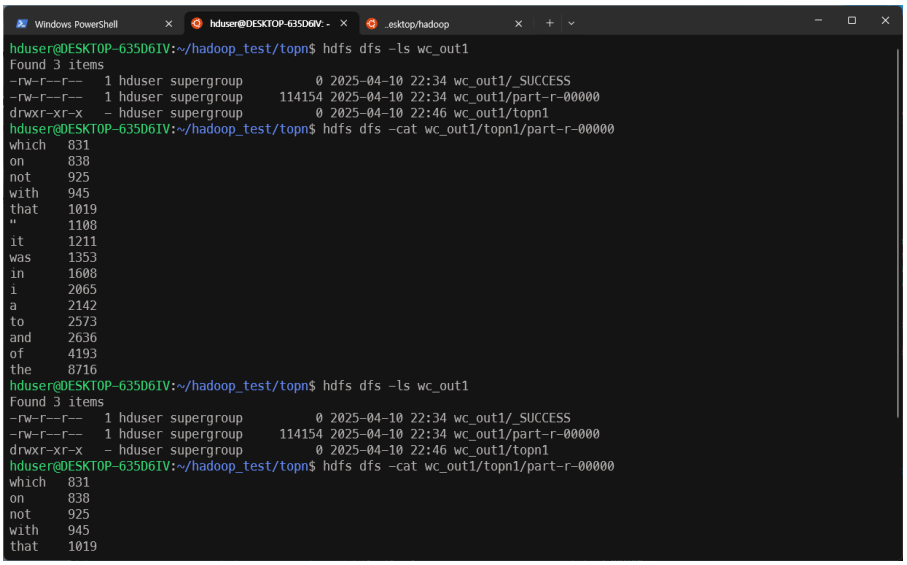
[Ch 2-7 p.24]


Windows PowerShell
hduser@DESKTOP-635D6IV: ~
~esktop/hadoop
hduser@DESKTOP-635D6IV:~/hadoop_test/topn$ javac -classpath $(hadoop classpath) -d bin src/*.java
Note: src/TopN.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
Note: src/TopN.java uses unchecked or unsafe operations.
Note: Recompile with -Xlint:unchecked for details.
hduser@DESKTOP-635D6IV:~/hadoop_test/topn$ jar -cvf topn.jar -C bin .
added manifest
adding: TopN$ItemFreqComparator.class(in = 730) (out= 438)(deflated 40%)
adding: ItemFreq.class(in = 1106) (out= 543)(deflated 50%)
adding: TopN.class(in = 2493) (out= 1260)(deflated 49%)
adding: TopN$Reduce.class(in = 3020) (out= 1246)(deflated 58%)
adding: TopN$Map.class(in = 2926) (out= 1176)(deflated 59%)
hduser@DESKTOP-635D6IV:~/hadoop_test/topn$ hadoop jar topn.jar TopN wc_out1/part-r-00000 wc_out1/topn1 15
2025-04-10 22:45:58,553 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-04-10 22:45:59,020 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hduser/.staging/job_1744291325019_0005
2025-04-10 22:45:59,303 INFO input.FileInputFormat: Total input files to process : 1
2025-04-10 22:45:59,780 INFO mapreduce.JobSubmitter: number of splits:1
2025-04-10 22:45:59,893 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744291325019_0005
2025-04-10 22:45:59,893 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-04-10 22:46:00,057 INFO conf.Configuration: resource-types.xml not found
2025-04-10 22:46:00,057 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-04-10 22:46:00,138 INFO impl.YarnClientImpl: Submitted application application_1744291325019_0005
2025-04-10 22:46:00,187 INFO mapreduce.Job: The url to track the job: http://DESKTOP-635D6IV.:50000/proxy/application_1744291325019_0005/
2025-04-10 22:46:00,188 INFO mapreduce.Job: Running job: job_1744291325019_0005


Windows PowerShell
hduser@DESKTOP-635D6IV: ~
~esktop/hadoop
Input split bytes=119
Combine input records=0
Combine output records=0
Reduce input groups=15
Reduce shuffle bytes=209
Reduce input records=15
Reduce output records=15
Spilled Records=30
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=206
CPU time spent (ms)=1880
Physical memory (bytes) snapshot=537026560
Virtual memory (bytes) snapshot=5125074944
Total committed heap usage (bytes)=497025024
Peak Map Physical memory (bytes)=296849408
Peak Map Virtual memory (bytes)=2559299584
Peak Reduce Physical memory (bytes)=240177152
Peak Reduce Virtual memory (bytes)=2565775360
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=114154
File Output Format Counters
Bytes Written=124
hduser@DESKTOP-635D6IV:~/hadoop_test/topn$


Windows PowerShell
hduser@DESKTOP-635D6IV: ~
~esktop/hadoop
Reduce input records=15
Reduce output records=15
Spilled Records=30
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=206
CPU time spent (ms)=1880
Physical memory (bytes) snapshot=537026560
Virtual memory (bytes) snapshot=5125074944
Total committed heap usage (bytes)=497025024
Peak Map Physical memory (bytes)=296849408
Peak Map Virtual memory (bytes)=2559299584
Peak Reduce Physical memory (bytes)=240177152
Peak Reduce Virtual memory (bytes)=2565775360
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=114154
File Output Format Counters
Bytes Written=124
hduser@DESKTOP-635D6IV:~/hadoop_test/topn$ hadoop jar topn.jar TopN wc_out2/part-r-00000 wc_out2/topn1 15
2025-04-10 22:46:58,301 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-04-10 22:46:58,672 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hduser/.staging/job_1744291325019_0006
2025-04-10 22:46:58,942 INFO input.FileInputFormat: Total input files to process : 1


Windows PowerShell
hduser@DESKTOP-635D6IV: ~
~esktop/hadoop
Input split bytes=119
Combine input records=0
Combine output records=0
Reduce input groups=15
Reduce shuffle bytes=208
Reduce input records=15
Reduce output records=15
Spilled Records=30
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=204
CPU time spent (ms)=2300
Physical memory (bytes) snapshot=739479552
Virtual memory (bytes) snapshot=5128151040
Total committed heap usage (bytes)=731906048
Peak Map Physical memory (bytes)=508604416
Peak Map Virtual memory (bytes)=2558898176
Peak Reduce Physical memory (bytes)=230875136
Peak Reduce Virtual memory (bytes)=2569252864
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=1202748
File Output Format Counters
Bytes Written=141
hduser@DESKTOP-635D6IV:~/hadoop_test/topn$


Windows PowerShell
hduser@DESKTOP-635D6IV: ~
~esktop/hadoop
hduser@DESKTOP-635D6IV:~/hadoop_test/topn$ hdfs dfs -ls wc_out1
Found 3 items
-rw-r--r-- 1 hduser supergroup 0 2025-04-10 22:34 wc_out1/_SUCCESS
-rw-r--r-- 1 hduser supergroup 114154 2025-04-10 22:34 wc_out1/part-r-00000
drwxr-xr-x - hduser supergroup 0 2025-04-10 22:46 wc_out1/topn1
hduser@DESKTOP-635D6IV:~/hadoop_test/topn$ hdfs dfs -cat wc_out1/topn1/part-r-00000
which 831
on 838
not 925
with 945
that 1019
" 1108
it 1211
was 1353
in 1608
i 2065
a 2142
to 2573
and 2636
of 4193
the 8716
hduser@DESKTOP-635D6IV:~/hadoop_test/topn$ hdfs dfs -ls wc_out1
Found 3 items
-rw-r--r-- 1 hduser supergroup 0 2025-04-10 22:34 wc_out1/_SUCCESS
-rw-r--r-- 1 hduser supergroup 114154 2025-04-10 22:34 wc_out1/part-r-00000
drwxr-xr-x - hduser supergroup 0 2025-04-10 22:46 wc_out1/topn1
hduser@DESKTOP-635D6IV:~/hadoop_test/topn$ hdfs dfs -cat wc_out1/topn1/part-r-00000
which 831
on 838
not 925
with 945
that 1019