**Big Data Analytics-Lab-CSE6034\_Lab-Assessment - 1:**

**Implementation of Hadoop Commands and Map-Reduce Programs ( Word Count and Market Analysis) using java**

**Submitted By: 20MAI0001 - NIHARIKA MAITRA**

**Git Repo link :**

<https://github.com/Niharika-20-MAI-01/Winter-2021-Big-Data-Analytics-LabCSE6034-Niharika-20MAI01>

**Activity - 1 : Implementation of Hadoop Commands**

**Hadoop Commands and their corresponding Outputs obtained on executing each command in the Terminal window followed by Screenshot of the Hadoop Commands and their corresponding Outputs obtained on executing each command in the Terminal window :**

**Description :- 1. To Print the Hadoop version :**

D:\>hadoop version

Hadoop 2.8.0

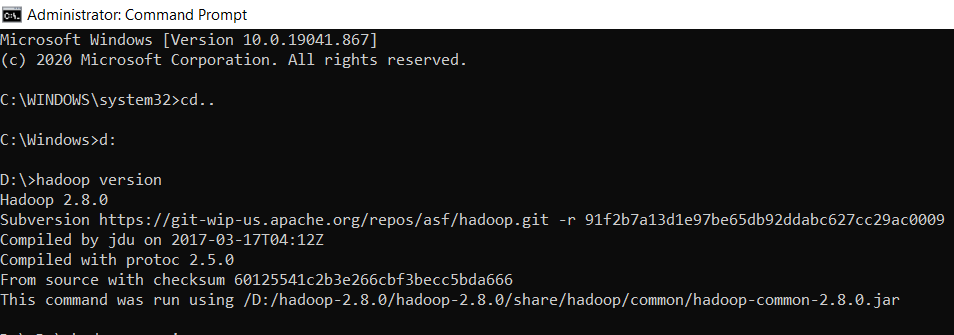
Subversion https://git-wip-us.apache.org/repos/asf/hadoop.git -r 91f2b7a13d1e97be65db92ddabc627cc29ac0009

Compiled by jdu on 2017-03-17T04:12Z

Compiled with protoc 2.5.0

From source with checksum 60125541c2b3e266cbf3becc5bda666

This command was run using /D:/hadoop-2.8.0/hadoop-2.8.0/share/hadoop/common/hadoop-common-2.8.0.j**ar**



**Description :- 2. To List the contents of the root directory in HDFS :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -ls /

Found 8 items

drwxr-xr-x - NRika supergroup 0 2021-02-28 21:06 /input1\_dir

-rw-r--r-- 1 NRika supergroup 1888 2021-03-01 02:45 /input\_dir

drwxr-xr-x - NRika supergroup 0 2021-02-18 12:35 /input\_dir1

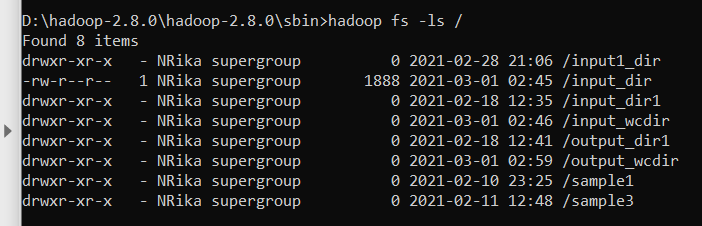
drwxr-xr-x - NRika supergroup 0 2021-03-01 02:46 /input\_wcdir

drwxr-xr-x - NRika supergroup 0 2021-02-18 12:41 /output\_dir1

drwxr-xr-x - NRika supergroup 0 2021-03-01 02:59 /output\_wcdir

drwxr-xr-x - NRika supergroup 0 2021-02-10 23:25 /sample1

drwxr-xr-x - NRika supergroup 0 2021-02-11 12:48 /sample3



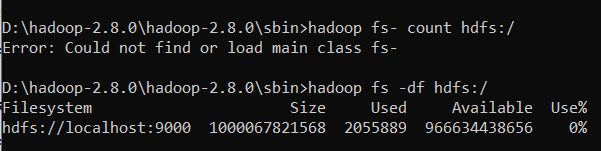
**Description :- 3. To Report the amount of space used and**

**available on currently mounted filesystem :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -df hdfs:/

Filesystem Size Used Available Use%

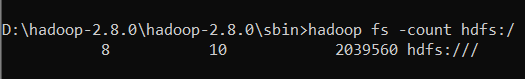
hdfs://localhost:9000 1000067821568 2055889 966634438656 0%



**Description :- 4. To Count the number of directories,files and bytes under the paths that match the specified file pattern :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -count hdfs:/

8 10 2039560 hdfs:///



**Description :- 5. To Run a DFS filesystem checking utility :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fsck - /

DEPRECATED: Use of this script to execute hdfs command is deprecated.

Instead use the hdfs command for it.

Connecting to namenode via http://localhost:50070/fsck?ugi=NRika&path=%2F

FSCK started by NRika (auth:SIMPLE) from /127.0.0.1 for path / at Sat Apr 03 18:42:31 IST 2021

..........Status: HEALTHY

Total size: 2039560 B

Total dirs: 8

Total files: 10

Total symlinks: 0

Total blocks (validated): 8 (avg. block size 254945 B)

Minimally replicated blocks: 8 (100.0 %)

Over-replicated blocks: 0 (0.0 %)

Under-replicated blocks: 0 (0.0 %)

Mis-replicated blocks: 0 (0.0 %)

Default replication factor: 1

Average block replication: 1.0

Corrupt blocks: 0

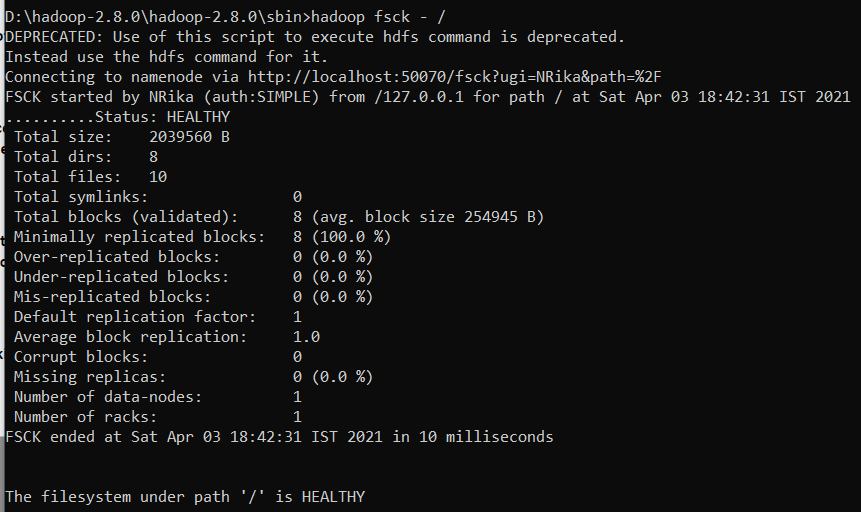
Missing replicas: 0 (0.0 %)

Number of data-nodes: 1

Number of racks: 1

FSCK ended at Sat Apr 03 18:42:31 IST 2021 in 10 milliseconds

The filesystem under path '/' is HEALTHY



**Description :- 6. To Run a cluster balancing utility :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop balancer

DEPRECATED: Use of this script to execute hdfs command is deprecated.

Instead use the hdfs command for it.

21/04/03 18:49:03 INFO balancer.Balancer: namenodes = [hdfs://localhost:9000]

21/04/03 18:49:03 INFO balancer.Balancer: parameters = Balancer.BalancerParameters [BalancingPolicy.Node, threshold = 10.0, max idle iteration = 5, #excluded nodes = 0, #included nodes = 0, #source nodes = 0, #blockpools = 0, run during upgrade = false]

21/04/03 18:49:03 INFO balancer.Balancer: included nodes = []

21/04/03 18:49:03 INFO balancer.Balancer: excluded nodes = []

21/04/03 18:49:03 INFO balancer.Balancer: source nodes = []

Time Stamp Iteration# Bytes Already Moved Bytes Left To Move Bytes Being Moved

21/04/03 18:49:04 INFO balancer.Balancer: dfs.balancer.movedWinWidth = 5400000 (default=5400000)

21/04/03 18:49:04 INFO balancer.Balancer: dfs.balancer.moverThreads = 1000 (default=1000)

21/04/03 18:49:04 INFO balancer.Balancer: dfs.balancer.dispatcherThreads = 200 (default=200)

21/04/03 18:49:04 INFO balancer.Balancer: dfs.datanode.balance.max.concurrent.moves = 50 (default=50)

21/04/03 18:49:04 INFO balancer.Balancer: dfs.balancer.getBlocks.size = 2147483648 (default=2147483648)

21/04/03 18:49:04 INFO balancer.Balancer: dfs.balancer.getBlocks.min-block-size = 10485760 (default=10485760)

21/04/03 18:49:04 INFO balancer.Balancer: dfs.balancer.max-size-to-move = 10737418240 (default=10737418240)

21/04/03 18:49:04 INFO balancer.Balancer: dfs.blocksize = 134217728 (default=134217728)

21/04/03 18:49:04 INFO net.NetworkTopology: Adding a new node: /default-rack/127.0.0.1:50010

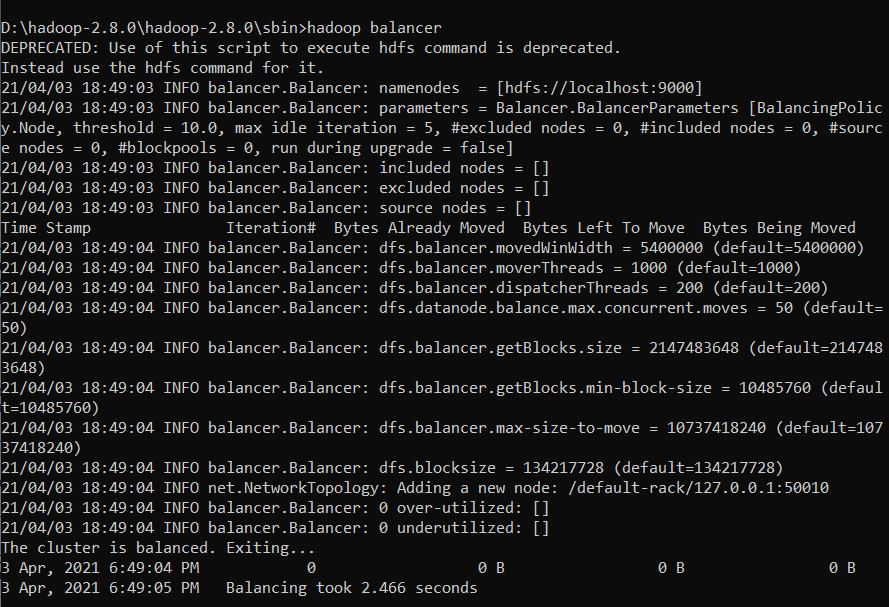
21/04/03 18:49:04 INFO balancer.Balancer: 0 over-utilized: []

21/04/03 18:49:04 INFO balancer.Balancer: 0 underutilized: []

The cluster is balanced. Exiting...

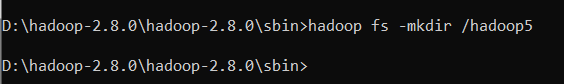
3 Apr, 2021 6:49:04 PM 0 0 B 0 B 0 B

3 Apr, 2021 6:49:05 PM Balancing took 2.466 seconds



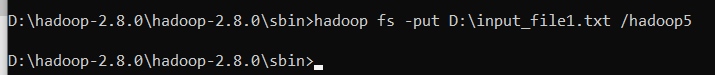
**Description :- 7. To Create a new directory named “hadoop5” in HDFS. Since you’re currently logged in your home directory in HDFS :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -mkdir /hadoop5



**Description :- 8. To Add a sample text file : “input\_file1.txt” from the local directory : “D” to the new directory : “hadoop5” created in HDFS during the previous step :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -put D:\input\_file1.txt /hadoop5

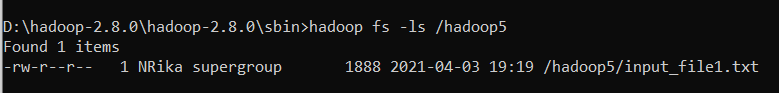


**Description :- 9. To List the contents of this new directory : “hadoop5” created in HDFS :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -ls /hadoop5

Found 1 items

-rw-r--r-- 1 NRika supergroup 1888 2021-04-03 19:19 /hadoop5/input\_file1.txt



**Description :- 10. To Add the entire local directory called “tmp” to the directory : “hadoop5” created in HDFS :**

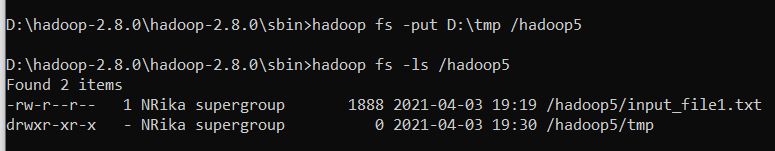
D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -put D:\tmp /hadoop5

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -ls /hadoop5

Found 2 items

-rw-r--r-- 1 NRika supergroup 1888 2021-04-03 19:19 /hadoop5/input\_file1.txt

drwxr-xr-x - NRika supergroup 0 2021-04-03 19:30 /hadoop5/tmp



**Description :- 11. To list and show the items added in the directory : “hadoop5” in HDFS (any command that does not have an absolute path is**

**interpreted as relative to that directory) :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -ls /

Found 10 items

drwxr-xr-x - NRika supergroup 0 2021-04-03 19:30 /hadoop5

drwxr-xr-x - NRika supergroup 0 2021-02-28 21:06 /input1\_dir

-rw-r--r-- 1 NRika supergroup 1888 2021-03-01 02:45 /input\_dir

drwxr-xr-x - NRika supergroup 0 2021-02-18 12:35 /input\_dir1

drwxr-xr-x - NRika supergroup 0 2021-03-01 02:46 /input\_wcdir

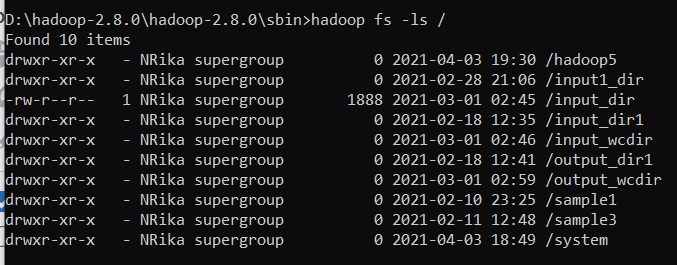
drwxr-xr-x - NRika supergroup 0 2021-02-18 12:41 /output\_dir1

drwxr-xr-x - NRika supergroup 0 2021-03-01 02:59 /output\_wcdir

drwxr-xr-x - NRika supergroup 0 2021-02-10 23:25 /sample1

drwxr-xr-x - NRika supergroup 0 2021-02-11 12:48 /sample3

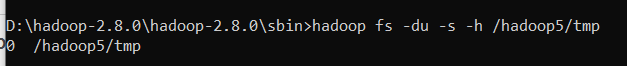
drwxr-xr-x - NRika supergroup 0 2021-04-03 18:49 /system

****

**Description :- 12. To See / check how much space this directory : “tmp”, occupies in HDFS:**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -du -s -h /hadoop5/tmp

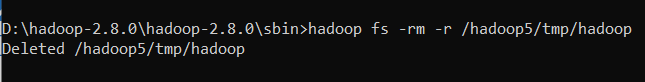
0 /hadoop5/tmp



**Description :- 13. To Delete a directory “hadoop” from the “tmp” directory inside “hadoop5” directory in HDFS :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -rm -r /hadoop5/tmp/hadoop

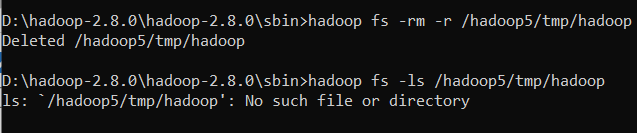
Deleted /hadoop5/tmp/hadoop



**Description :- 14. To Ensure this directory : “hadoop” is no longer in HDFS :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -ls /hadoop5/tmp/hadoop

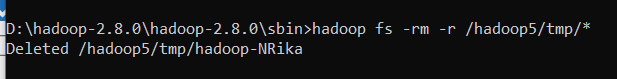
ls: `/hadoop5/tmp/hadoop': No such file or directory



**Description :- 15. To Delete all directories from the “tmp” directory using a wildcard :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -rm -r /hadoop5/tmp/\*

Deleted /hadoop5/tmp/hadoop-NRika



**Description :- 16. To empty the trash :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -expunge

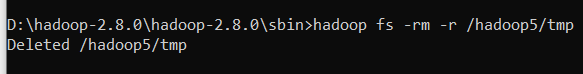


**Description :- 17. To finally remove the entire “tmp” directory and all**

**of its contents in HDFS :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -rm -r /hadoop5/tmp

Deleted /hadoop5/tmp



**Description :- 18. To List the “hadoop5” directory present in HDFS again :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -ls hadoop5 /

ls: `hadoop5': No such file or directory

Found 10 items

drwxr-xr-x - NRika supergroup 0 2021-04-03 20:21 /hadoop5

drwxr-xr-x - NRika supergroup 0 2021-02-28 21:06 /input1\_dir

-rw-r--r-- 1 NRika supergroup 1888 2021-03-01 02:45 /input\_dir

drwxr-xr-x - NRika supergroup 0 2021-02-18 12:35 /input\_dir1

drwxr-xr-x - NRika supergroup 0 2021-03-01 02:46 /input\_wcdir

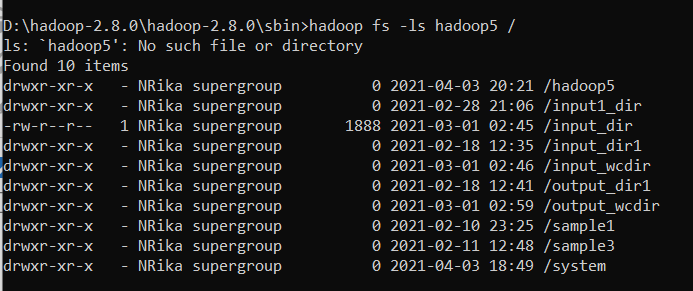
drwxr-xr-x - NRika supergroup 0 2021-02-18 12:41 /output\_dir1

drwxr-xr-x - NRika supergroup 0 2021-03-01 02:59 /output\_wcdir

drwxr-xr-x - NRika supergroup 0 2021-02-10 23:25 /sample1

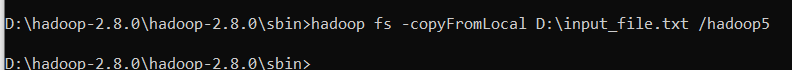
drwxr-xr-x - NRika supergroup 0 2021-02-11 12:48 /sample3

drwxr-xr-x - NRika supergroup 0 2021-04-03 18:49 /system



**Description :- 19. To Add a sample text file “input\_file.txt” from the local directory : “D” to the directory : “hadoop5” created in HDFS :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -copyFromLocal D:\input\_file.txt /hadoop5



**Description :- 20. To view the contents of the text file “input\_file.txt”**

**which is present in the “hadoop5” directory in HDFS :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -cat /hadoop5/input\_file.txt

23 23 27 43 24 25 26 26 26 26 25 26 25

26 27 28 28 28 30 31 31 31 30 30 30 29

31 32 32 32 33 34 35 36 36 34 34 34 34

39 38 39 39 39 41 42 43 40 39 38 38 40

38 39 39 39 39 41 41 41 28 40 39 39 45

23 23 27 43 24 25 26 26 26 26 25 26 25

26 27 28 28 28 30 31 31 31 30 30 30 29

31 32 32 32 33 34 35 36 36 34 34 34 34

39 38 39 39 39 41 42 43 40 39 38 38 40

38 39 39 39 39 41 41 41 28 40 39 39 45

23 23 27 43 24 25 26 26 26 26 25 26 25

26 27 28 28 28 30 31 31 31 30 30 30 29

31 32 32 32 33 34 35 36 36 34 34 34 34

39 38 39 39 39 41 42 43 40 39 38 38 40

38 39 39 39 39 41 41 41 28 40 39 39 45

23 23 27 43 24 25 26 26 26 26 25 26 25

26 27 28 28 28 30 31 31 31 30 30 30 29

31 32 32 32 33 34 35 36 36 34 34 34 34

39 38 39 39 39 41 42 43 40 39 38 38 40

38 39 39 39 39 41 41 41 28 40 39 39 45

23 23 27 43 24 25 26 26 26 26 25 26 25

26 27 28 28 28 30 31 31 31 30 30 30 29

31 32 32 32 33 34 35 36 36 34 34 34 34

39 38 39 39 39 41 42 43 40 39 38 38 40

38 39 39 39 39 41 41 41 28 40 39 39 45

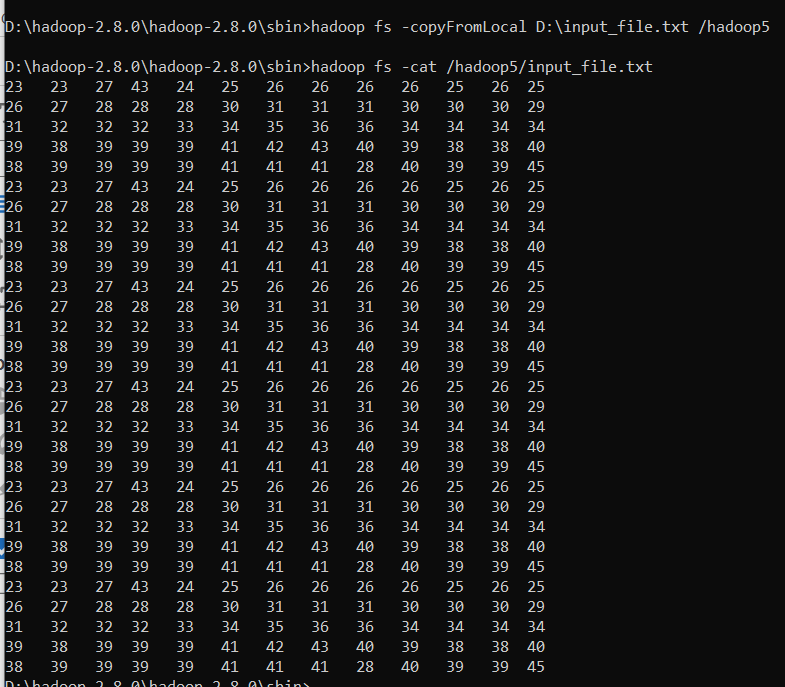
23 23 27 43 24 25 26 26 26 26 25 26 25

26 27 28 28 28 30 31 31 31 30 30 30 29

31 32 32 32 33 34 35 36 36 34 34 34 34

39 38 39 39 39 41 42 43 40 39 38 38 40

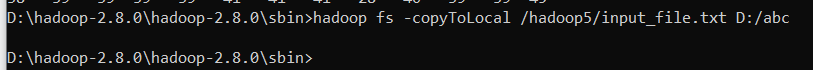
38 39 39 39 39 41 41 41 28 40 39 39 45



**Description :- 21. To Add the text file “input\_file.txt” from “hadoop5” directory which is present in HDFS directory**

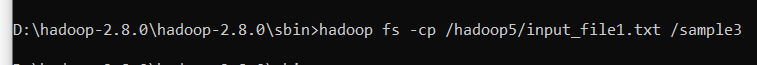
**to the directory “abc” which is present in local directory : “D” :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -copyToLocal /hadoop5/input\_file.txt D:/abc



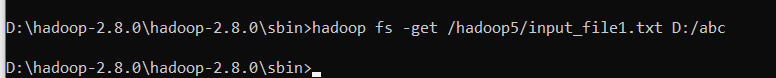
**Description :- 22. To copy files between directories present in HDFS using ‘cp’ command : “input\_file1.txt” text file copying from “hadoop5” directory in HDFS to the “sample3” directory in HDFS :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -cp /hadoop5/input\_file1.txt /sample3



**Description :- 23. To use the ‘-get’ command alternatively to the ‘-copyToLocal’ command :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -get /hadoop5/input\_file1.txt D:/abc



**Description :- 24. To Display last kilobyte of the file “input\_file.txt” to stdout :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -tail /hadoop5/input\_file.txt

9 38 38 40

38 39 39 39 39 41 41 41 28 40 39 39 45

23 23 27 43 24 25 26 26 26 26 25 26 25

26 27 28 28 28 30 31 31 31 30 30 30 29

31 32 32 32 33 34 35 36 36 34 34 34 34

39 38 39 39 39 41 42 43 40 39 38 38 40

38 39 39 39 39 41 41 41 28 40 39 39 45

23 23 27 43 24 25 26 26 26 26 25 26 25

26 27 28 28 28 30 31 31 31 30 30 30 29

31 32 32 32 33 34 35 36 36 34 34 34 34

39 38 39 39 39 41 42 43 40 39 38 38 40

38 39 39 39 39 41 41 41 28 40 39 39 45

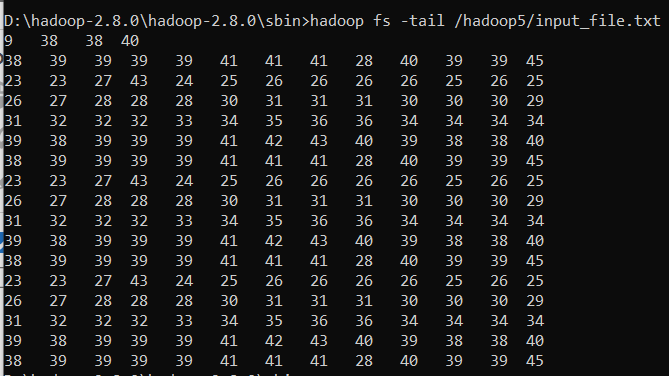
23 23 27 43 24 25 26 26 26 26 25 26 25

26 27 28 28 28 30 31 31 31 30 30 30 29

31 32 32 32 33 34 35 36 36 34 34 34 34

39 38 39 39 39 41 42 43 40 39 38 38 40

38 39 39 39 39 41 41 41 28 40 39 39 45

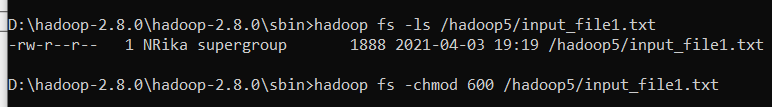


**Description :- 25. To Use ‘-chmod’ command to change permissions of the file “input\_file1.txt” in HDFS (Default file permissions are 666 in HDFS) :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -ls /hadoop5/input\_file1.txt

-rw-r--r-- 1 NRika supergroup 1888 2021-04-03 19:19 /hadoop5/input\_file1.txt

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -chmod 600 /hadoop5/input\_file1.txt

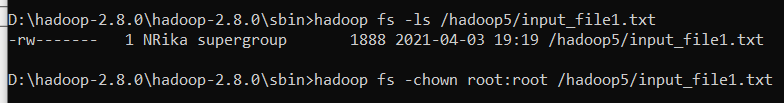


**Description :- 26. To use ‘-chown’ to change owner name and group name of the file “input\_file1.txt” in HDFS simultaneously :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -ls /hadoop5/input\_file1.txt

-rw------- 1 NRika supergroup 1888 2021-04-03 19:19 /hadoop5/input\_file1.txt

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -chown root:root /hadoop5/input\_file1.txt

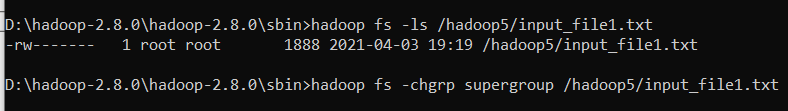


**Description :- 27. To use ‘-chgrp’ command to change group name of the file “input\_file1.txt” in HDFS :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -ls /hadoop5/input\_file1.txt

-rw------- 1 root root 1888 2021-04-03 19:19 /hadoop5/input\_file1.txt

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -chgrp supergroup /hadoop5/input\_file1.txt



**Description :- 28. To move a directory from one location to other in HDFS : to move directory : “hadoop5” from HDFS into the directory : “sample3” in HDFS :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -mv /hadoop5 /sample3

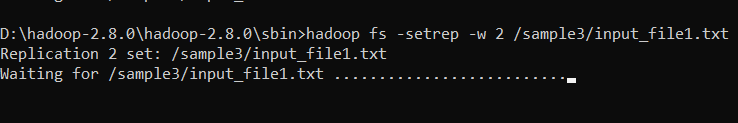


**Description :- 29. To Use ‘-setrep’ command to change replication factor of the file “input\_file1.txt” in HDFS (Default replication factor to a file is 3) :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -setrep -w 2 /sample3/input\_file1.txt

Replication 2 set: /sample3/input\_file1.txt

Waiting for /sample3/input\_file1.txt .....................



**Description :- 30. To copy a directory from one node in the cluster to another using ‘-distcp’ command, ‘-overwrite’ command to overwrite in an existing files**

**‘-update’ command to synchronize both directories : (to copy directory : “sample3” from namenodeA to namenodeB) in HDFS :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -distcp hdfs://namenodeA/sample3 hdfs://namenodeB/sample6

-distcp: Unknown command

Usage: hadoop fs [generic options]

[-appendToFile <localsrc> ... <dst>]

[-cat [-ignoreCrc] <src> ...]

[-checksum <src> ...]

[-chgrp [-R] GROUP PATH...]

[-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]

[-chown [-R] [OWNER][:[GROUP]] PATH...]

[-copyFromLocal [-f] [-p] [-l] [-d] <localsrc> ... <dst>]

[-copyToLocal [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]

[-count [-q] [-h] [-v] [-t [<storage type>]] [-u] [-x] <path> ...]

[-cp [-f] [-p | -p[topax]] [-d] <src> ... <dst>]

[-createSnapshot <snapshotDir> [<snapshotName>]]

[-deleteSnapshot <snapshotDir> <snapshotName>]

[-df [-h] [<path> ...]]

[-du [-s] [-h] [-x] <path> ...]

[-expunge]

[-find <path> ... <expression> ...]

[-get [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]

[-getfacl [-R] <path>]

[-getfattr [-R] {-n name | -d} [-e en] <path>]

[-getmerge [-nl] [-skip-empty-file] <src> <localdst>]

[-help [cmd ...]]

[-ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] [<path> ...]]

[-mkdir [-p] <path> ...]

[-moveFromLocal <localsrc> ... <dst>]

[-moveToLocal <src> <localdst>]

[-mv <src> ... <dst>]

[-put [-f] [-p] [-l] [-d] <localsrc> ... <dst>]

[-renameSnapshot <snapshotDir> <oldName> <newName>]

[-rm [-f] [-r|-R] [-skipTrash] [-safely] <src> ...]

[-rmdir [--ignore-fail-on-non-empty] <dir> ...]

[-setfacl [-R] [{-b|-k} {-m|-x <acl\_spec>} <path>]|[--set <acl\_spec> <path>]]

[-setfattr {-n name [-v value] | -x name} <path>]

[-setrep [-R] [-w] <rep> <path> ...]

[-stat [format] <path> ...]

[-tail [-f] <file>]

[-test -[defsz] <path>]

[-text [-ignoreCrc] <src> ...]

[-touchz <path> ...]

[-truncate [-w] <length> <path> ...]

[-usage [cmd ...]]

Generic options supported are

-conf <configuration file> specify an application configuration file

-D <property=value> use value for given property

-fs <file:///|hdfs://namenode:port> specify default filesystem URL to use, overrides 'fs.defaultFS' property from configurations.

-jt <local|resourcemanager:port> specify a ResourceManager

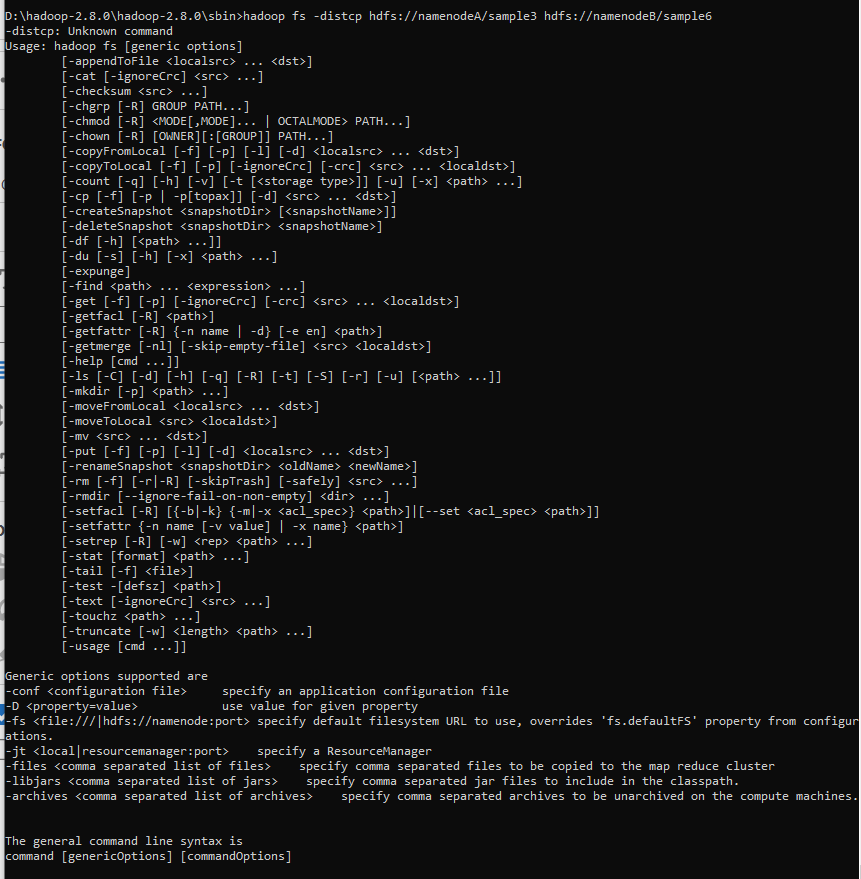
-files <comma separated list of files> specify comma separated files to be copied to the map reduce cluster

-libjars <comma separated list of jars> specify comma separated jar files to include in the classpath.

-archives <comma separated list of archives> specify comma separated archives to be unarchived on the compute machines.

The general command line syntax is

command [genericOptions] [commandOptions]

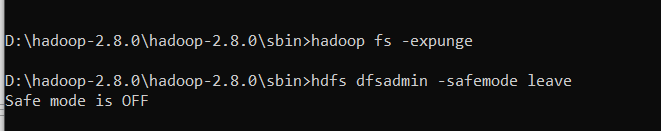


**Description :- 31. To make the name node leave safe mode :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -expunge

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hdfs dfsadmin -safemode leave

Safe mode is OFF



**Description :- 32. To List all the hadoop file system shell commands :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs

Usage: hadoop fs [generic options]

[-appendToFile <localsrc> ... <dst>]

[-cat [-ignoreCrc] <src> ...]

[-checksum <src> ...]

[-chgrp [-R] GROUP PATH...]

[-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]

[-chown [-R] [OWNER][:[GROUP]] PATH...]

[-copyFromLocal [-f] [-p] [-l] [-d] <localsrc> ... <dst>]

[-copyToLocal [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]

[-count [-q] [-h] [-v] [-t [<storage type>]] [-u] [-x] <path> ...]

[-cp [-f] [-p | -p[topax]] [-d] <src> ... <dst>]

[-createSnapshot <snapshotDir> [<snapshotName>]]

[-deleteSnapshot <snapshotDir> <snapshotName>]

[-df [-h] [<path> ...]]

[-du [-s] [-h] [-x] <path> ...]

[-expunge]

[-find <path> ... <expression> ...]

[-get [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]

[-getfacl [-R] <path>]

[-getfattr [-R] {-n name | -d} [-e en] <path>]

[-getmerge [-nl] [-skip-empty-file] <src> <localdst>]

[-help [cmd ...]]

[-ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] [<path> ...]]

[-mkdir [-p] <path> ...]

[-moveFromLocal <localsrc> ... <dst>]

[-moveToLocal <src> <localdst>]

[-mv <src> ... <dst>]

[-put [-f] [-p] [-l] [-d] <localsrc> ... <dst>]

[-renameSnapshot <snapshotDir> <oldName> <newName>]

[-rm [-f] [-r|-R] [-skipTrash] [-safely] <src> ...]

[-rmdir [--ignore-fail-on-non-empty] <dir> ...]

[-setfacl [-R] [{-b|-k} {-m|-x <acl\_spec>} <path>]|[--set <acl\_spec> <path>]]

[-setfattr {-n name [-v value] | -x name} <path>]

[-setrep [-R] [-w] <rep> <path> ...]

[-stat [format] <path> ...]

[-tail [-f] <file>]

[-test -[defsz] <path>]

[-text [-ignoreCrc] <src> ...]

[-touchz <path> ...]

[-truncate [-w] <length> <path> ...]

[-usage [cmd ...]]

Generic options supported are

-conf <configuration file> specify an application configuration file

-D <property=value> use value for given property

-fs <file:///|hdfs://namenode:port> specify default filesystem URL to use, overrides 'fs.defaultFS' property from configurations.

-jt <local|resourcemanager:port> specify a ResourceManager

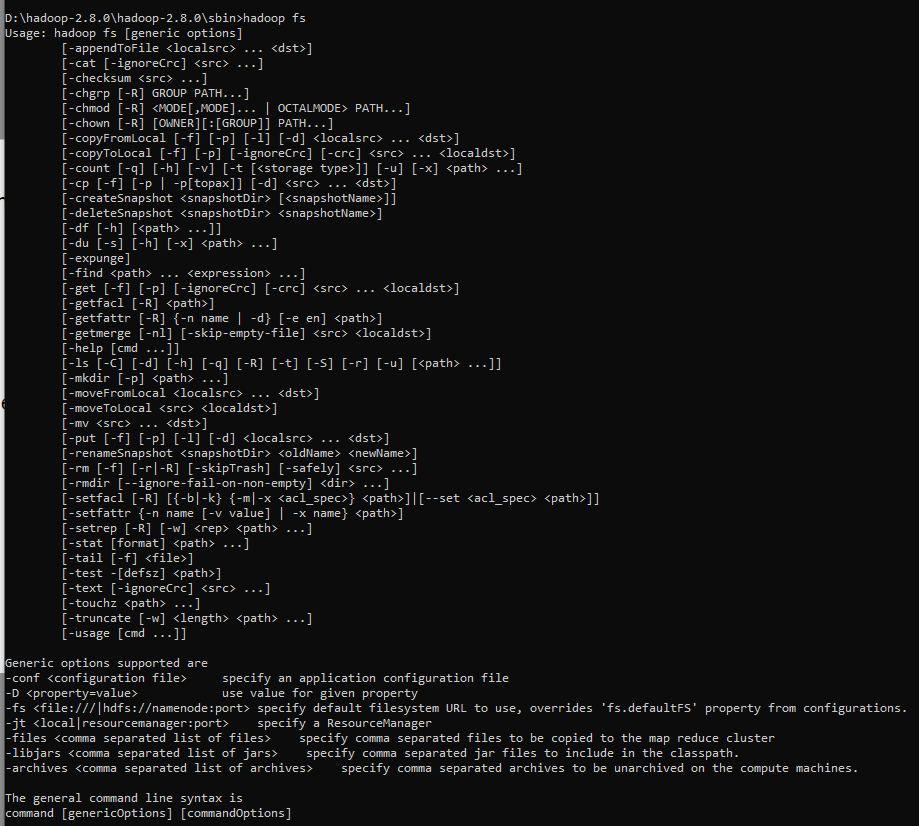
-files <comma separated list of files> specify comma separated files to be copied to the map reduce cluster

-libjars <comma separated list of jars> specify comma separated jar files to include in the classpath.

-archives <comma separated list of archives> specify comma separated archives to be unarchived on the compute machines.

The general command line syntax is

command [genericOptions] [commandOptions]



**Description :- 33. To execute the hadoop “help!” command :**

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -help

Usage: hadoop fs [generic options]

[-appendToFile <localsrc> ... <dst>]

[-cat [-ignoreCrc] <src> ...]

[-checksum <src> ...]

[-chgrp [-R] GROUP PATH...]

[-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]

[-chown [-R] [OWNER][:[GROUP]] PATH...]

[-copyFromLocal [-f] [-p] [-l] [-d] <localsrc> ... <dst>]

[-copyToLocal [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]

[-count [-q] [-h] [-v] [-t [<storage type>]] [-u] [-x] <path> ...]

[-cp [-f] [-p | -p[topax]] [-d] <src> ... <dst>]

[-createSnapshot <snapshotDir> [<snapshotName>]]

[-deleteSnapshot <snapshotDir> <snapshotName>]

[-df [-h] [<path> ...]]

[-du [-s] [-h] [-x] <path> ...]

[-expunge]

[-find <path> ... <expression> ...]

[-get [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]

[-getfacl [-R] <path>]

[-getfattr [-R] {-n name | -d} [-e en] <path>]

[-getmerge [-nl] [-skip-empty-file] <src> <localdst>]

[-help [cmd ...]]

[-ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] [<path> ...]]

[-mkdir [-p] <path> ...]

[-moveFromLocal <localsrc> ... <dst>]

[-moveToLocal <src> <localdst>]

[-mv <src> ... <dst>]

[-put [-f] [-p] [-l] [-d] <localsrc> ... <dst>]

[-renameSnapshot <snapshotDir> <oldName> <newName>]

[-rm [-f] [-r|-R] [-skipTrash] [-safely] <src> ...]

[-rmdir [--ignore-fail-on-non-empty] <dir> ...]

[-setfacl [-R] [{-b|-k} {-m|-x <acl\_spec>} <path>]|[--set <acl\_spec> <path>]]

[-setfattr {-n name [-v value] | -x name} <path>]

[-setrep [-R] [-w] <rep> <path> ...]

[-stat [format] <path> ...]

[-tail [-f] <file>]

[-test -[defsz] <path>]

[-text [-ignoreCrc] <src> ...]

[-touchz <path> ...]

[-truncate [-w] <length> <path> ...]

[-usage [cmd ...]]

-appendToFile <localsrc> ... <dst> :

Appends the contents of all the given local files to the given dst file. The dst

file will be created if it does not exist. If <localSrc> is -, then the input is

read from stdin.

-cat [-ignoreCrc] <src> ... :

Fetch all files that match the file pattern <src> and display their content on

stdout.

-checksum <src> ... :

Dump checksum information for files that match the file pattern <src> to stdout.

Note that this requires a round-trip to a datanode storing each block of the

file, and thus is not efficient to run on a large number of files. The checksum

of a file depends on its content, block size and the checksum algorithm and

parameters used for creating the file.

-chgrp [-R] GROUP PATH... :

This is equivalent to -chown ... :GROUP ...

-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH... :

Changes permissions of a file. This works similar to the shell's chmod command

with a few exceptions.

-R modifies the files recursively. This is the only option currently

supported.

<MODE> Mode is the same as mode used for the shell's command. The only

letters recognized are 'rwxXt', e.g. +t,a+r,g-w,+rwx,o=r.

<OCTALMODE> Mode specifed in 3 or 4 digits. If 4 digits, the first may be 1 or

0 to turn the sticky bit on or off, respectively. Unlike the

shell command, it is not possible to specify only part of the

mode, e.g. 754 is same as u=rwx,g=rx,o=r.

If none of 'augo' is specified, 'a' is assumed and unlike the shell command, no

umask is applied.

-chown [-R] [OWNER][:[GROUP]] PATH... :

Changes owner and group of a file. This is similar to the shell's chown command

with a few exceptions.

-R modifies the files recursively. This is the only option currently

supported.

If only the owner or group is specified, then only the owner or group is

modified. The owner and group names may only consist of digits, alphabet, and

any of [-\_./@a-zA-Z0-9 ]. The names are case sensitive.

WARNING: Avoid using '.' to separate user name and group though Linux allows it.

If user names have dots in them and you are using local file system, you might

see surprising results since the shell command 'chown' is used for local files.

-copyFromLocal [-f] [-p] [-l] [-d] <localsrc> ... <dst> :

Identical to the -put command.

-copyToLocal [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst> :

Identical to the -get command.

-count [-q] [-h] [-v] [-t [<storage type>]] [-u] [-x] <path> ... :

Count the number of directories, files and bytes under the paths

that match the specified file pattern. The output columns are:

DIR\_COUNT FILE\_COUNT CONTENT\_SIZE PATHNAME

or, with the -q option:

QUOTA REM\_QUOTA SPACE\_QUOTA REM\_SPACE\_QUOTA

DIR\_COUNT FILE\_COUNT CONTENT\_SIZE PATHNAME

The -h option shows file sizes in human readable format.

The -v option displays a header line.

The -x option excludes snapshots from being calculated.

The -t option displays quota by storage types.

It should be used with -q or -u option, otherwise it will be ignored.

If a comma-separated list of storage types is given after the -t option,

it displays the quota and usage for the specified types.

Otherwise, it displays the quota and usage for all the storage

types that support quota. The list of possible storage types(case insensitive):

ram\_disk, ssd, disk and archive.

It can also pass the value '', 'all' or 'ALL' to specify all the storage types.

The -u option shows the quota and

the usage against the quota without the detailed content summary.

-cp [-f] [-p | -p[topax]] [-d] <src> ... <dst> :

Copy files that match the file pattern <src> to a destination. When copying

multiple files, the destination must be a directory. Passing -p preserves status

[topax] (timestamps, ownership, permission, ACLs, XAttr). If -p is specified

with no <arg>, then preserves timestamps, ownership, permission. If -pa is

specified, then preserves permission also because ACL is a super-set of

permission. Passing -f overwrites the destination if it already exists. raw

namespace extended attributes are preserved if (1) they are supported (HDFS

only) and, (2) all of the source and target pathnames are in the /.reserved/raw

hierarchy. raw namespace xattr preservation is determined solely by the presence

(or absence) of the /.reserved/raw prefix and not by the -p option. Passing -d

will skip creation of temporary file(<dst>.\_COPYING\_).

-createSnapshot <snapshotDir> [<snapshotName>] :

Create a snapshot on a directory

-deleteSnapshot <snapshotDir> <snapshotName> :

Delete a snapshot from a directory

-df [-h] [<path> ...] :

Shows the capacity, free and used space of the filesystem. If the filesystem has

multiple partitions, and no path to a particular partition is specified, then

the status of the root partitions will be shown.

-h Formats the sizes of files in a human-readable fashion rather than a number

of bytes.

-du [-s] [-h] [-x] <path> ... :

Show the amount of space, in bytes, used by the files that match the specified

file pattern. The following flags are optional:

-s Rather than showing the size of each individual file that matches the

pattern, shows the total (summary) size.

-h Formats the sizes of files in a human-readable fashion rather than a number

of bytes.

-x Excludes snapshots from being counted.

Note that, even without the -s option, this only shows size summaries one level

deep into a directory.

The output is in the form

size name(full path)

-expunge :

Delete files from the trash that are older than the retention threshold

-find <path> ... <expression> ... :

Finds all files that match the specified expression and

applies selected actions to them. If no <path> is specified

then defaults to the current working directory. If no

expression is specified then defaults to -print.

The following primary expressions are recognised:

-name pattern

-iname pattern

Evaluates as true if the basename of the file matches the

pattern using standard file system globbing.

If -iname is used then the match is case insensitive.

-print

-print0

Always evaluates to true. Causes the current pathname to be

written to standard output followed by a newline. If the -print0

expression is used then an ASCII NULL character is appended rather

than a newline.

The following operators are recognised:

expression -a expression

expression -and expression

expression expression

Logical AND operator for joining two expressions. Returns

true if both child expressions return true. Implied by the

juxtaposition of two expressions and so does not need to be

explicitly specified. The second expression will not be

applied if the first fails.

-get [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst> :

Copy files that match the file pattern <src> to the local name. <src> is kept.

When copying multiple files, the destination must be a directory. Passing -f

overwrites the destination if it already exists and -p preserves access and

modification times, ownership and the mode.

-getfacl [-R] <path> :

Displays the Access Control Lists (ACLs) of files and directories. If a

directory has a default ACL, then getfacl also displays the default ACL.

-R List the ACLs of all files and directories recursively.

<path> File or directory to list.

-getfattr [-R] {-n name | -d} [-e en] <path> :

Displays the extended attribute names and values (if any) for a file or

directory.

-R Recursively list the attributes for all files and directories.

-n name Dump the named extended attribute value.

-d Dump all extended attribute values associated with pathname.

-e <encoding> Encode values after retrieving them.Valid encodings are "text",

"hex", and "base64". Values encoded as text strings are enclosed

in double quotes ("), and values encoded as hexadecimal and

base64 are prefixed with 0x and 0s, respectively.

<path> The file or directory.

-getmerge [-nl] [-skip-empty-file] <src> <localdst> :

Get all the files in the directories that match the source file pattern and

merge and sort them to only one file on local fs. <src> is kept.

-nl Add a newline character at the end of each file.

-skip-empty-file Do not add new line character for empty file.

-help [cmd ...] :

Displays help for given command or all commands if none is specified.

-ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] [<path> ...] :

List the contents that match the specified file pattern. If path is not

specified, the contents of /user/<currentUser> will be listed. For a directory a

list of its direct children is returned (unless -d option is specified).

Directory entries are of the form:

permissions - userId groupId sizeOfDirectory(in bytes)

modificationDate(yyyy-MM-dd HH:mm) directoryName

and file entries are of the form:

permissions numberOfReplicas userId groupId sizeOfFile(in bytes)

modificationDate(yyyy-MM-dd HH:mm) fileName

-C Display the paths of files and directories only.

-d Directories are listed as plain files.

-h Formats the sizes of files in a human-readable fashion

rather than a number of bytes.

-q Print ? instead of non-printable characters.

-R Recursively list the contents of directories.

-t Sort files by modification time (most recent first).

-S Sort files by size.

-r Reverse the order of the sort.

-u Use time of last access instead of modification for

display and sorting.

-mkdir [-p] <path> ... :

Create a directory in specified location.

-p Do not fail if the directory already exists

-moveFromLocal <localsrc> ... <dst> :

Same as -put, except that the source is deleted after it's copied.

-moveToLocal <src> <localdst> :

Not implemented yet

-mv <src> ... <dst> :

Move files that match the specified file pattern <src> to a destination <dst>.

When moving multiple files, the destination must be a directory.

-put [-f] [-p] [-l] [-d] <localsrc> ... <dst> :

Copy files from the local file system into fs. Copying fails if the file already

exists, unless the -f flag is given.

Flags:

-p Preserves access and modification times, ownership and the mode.

-f Overwrites the destination if it already exists.

-l Allow DataNode to lazily persist the file to disk. Forces

replication factor of 1. This flag will result in reduced

durability. Use with care.

-d Skip creation of temporary file(<dst>.\_COPYING\_).

-renameSnapshot <snapshotDir> <oldName> <newName> :

Rename a snapshot from oldName to newName

-rm [-f] [-r|-R] [-skipTrash] [-safely] <src> ... :

Delete all files that match the specified file pattern. Equivalent to the Unix

command "rm <src>"

-f If the file does not exist, do not display a diagnostic message or

modify the exit status to reflect an error.

-[rR] Recursively deletes directories.

-skipTrash option bypasses trash, if enabled, and immediately deletes <src>.

-safely option requires safety confirmation, if enabled, requires

confirmation before deleting large directory with more than

<hadoop.shell.delete.limit.num.files> files. Delay is expected when

walking over large directory recursively to count the number of

files to be deleted before the confirmation.

-rmdir [--ignore-fail-on-non-empty] <dir> ... :

Removes the directory entry specified by each directory argument, provided it is

empty.

-setfacl [-R] [{-b|-k} {-m|-x <acl\_spec>} <path>]|[--set <acl\_spec> <path>] :

Sets Access Control Lists (ACLs) of files and directories.

Options:

-b Remove all but the base ACL entries. The entries for user, group

and others are retained for compatibility with permission bits.

-k Remove the default ACL.

-R Apply operations to all files and directories recursively.

-m Modify ACL. New entries are added to the ACL, and existing entries

are retained.

-x Remove specified ACL entries. Other ACL entries are retained.

--set Fully replace the ACL, discarding all existing entries. The

<acl\_spec> must include entries for user, group, and others for

compatibility with permission bits.

<acl\_spec> Comma separated list of ACL entries.

<path> File or directory to modify.

-setfattr {-n name [-v value] | -x name} <path> :

Sets an extended attribute name and value for a file or directory.

-n name The extended attribute name.

-v value The extended attribute value. There are three different encoding

methods for the value. If the argument is enclosed in double quotes,

then the value is the string inside the quotes. If the argument is

prefixed with 0x or 0X, then it is taken as a hexadecimal number. If

the argument begins with 0s or 0S, then it is taken as a base64

encoding.

-x name Remove the extended attribute.

<path> The file or directory.

-setrep [-R] [-w] <rep> <path> ... :

Set the replication level of a file. If <path> is a directory then the command

recursively changes the replication factor of all files under the directory tree

rooted at <path>.

-w It requests that the command waits for the replication to complete. This

can potentially take a very long time.

-R It is accepted for backwards compatibility. It has no effect.

-stat [format] <path> ... :

Print statistics about the file/directory at <path>

in the specified format. Format accepts filesize in

blocks (%b), type (%F), group name of owner (%g),

name (%n), block size (%o), replication (%r), user name

of owner (%u), modification date (%y, %Y).

%y shows UTC date as "yyyy-MM-dd HH:mm:ss" and

%Y shows milliseconds since January 1, 1970 UTC.

If the format is not specified, %y is used by default.

-tail [-f] <file> :

Show the last 1KB of the file.

-f Shows appended data as the file grows.

-test -[defsz] <path> :

Answer various questions about <path>, with result via exit status.

-d return 0 if <path> is a directory.

-e return 0 if <path> exists.

-f return 0 if <path> is a file.

-s return 0 if file <path> is greater than zero bytes in size.

-w return 0 if file <path> exists and write permission is granted.

-r return 0 if file <path> exists and read permission is granted.

-z return 0 if file <path> is zero bytes in size, else return 1.

-text [-ignoreCrc] <src> ... :

Takes a source file and outputs the file in text format.

The allowed formats are zip and TextRecordInputStream and Avro.

-touchz <path> ... :

Creates a file of zero length at <path> with current time as the timestamp of

that <path>. An error is returned if the file exists with non-zero length

-truncate [-w] <length> <path> ... :

Truncate all files that match the specified file pattern to the specified

length.

-w Requests that the command wait for block recovery to complete, if

necessary.

-usage [cmd ...] :

Displays the usage for given command or all commands if none is specified.

Generic options supported are

-conf <configuration file> specify an application configuration file

-D <property=value> use value for given property

-fs <file:///|hdfs://namenode:port> specify default filesystem URL to use, overrides 'fs.defaultFS' property from configurations.

-jt <local|resourcemanager:port> specify a ResourceManager

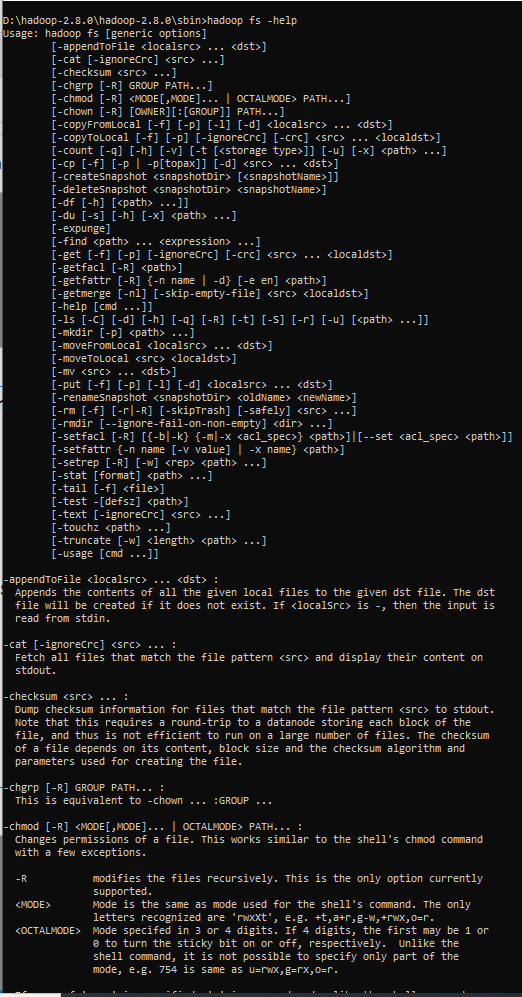
-files <comma separated list of files> specify comma separated files to be copied to the map reduce cluster

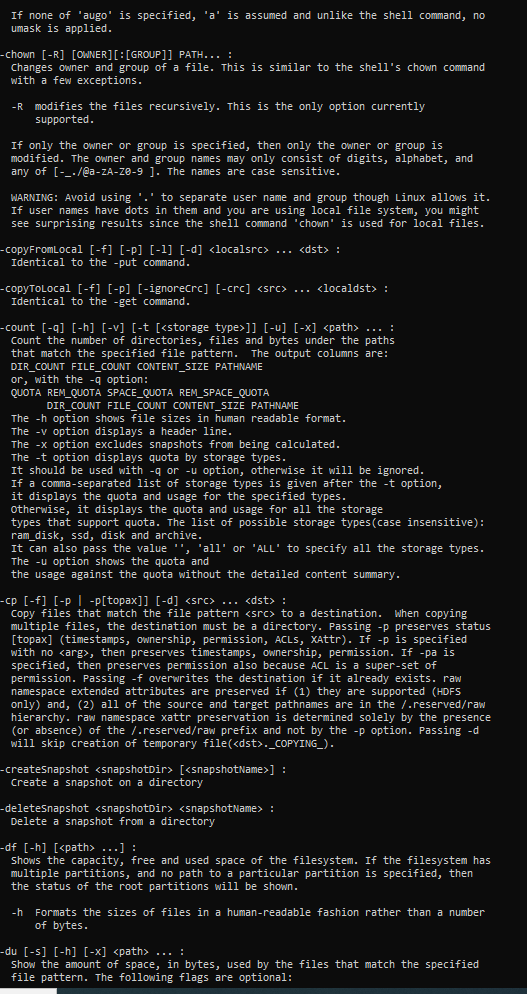
-libjars <comma separated list of jars> specify comma separated jar files to include in the classpath.

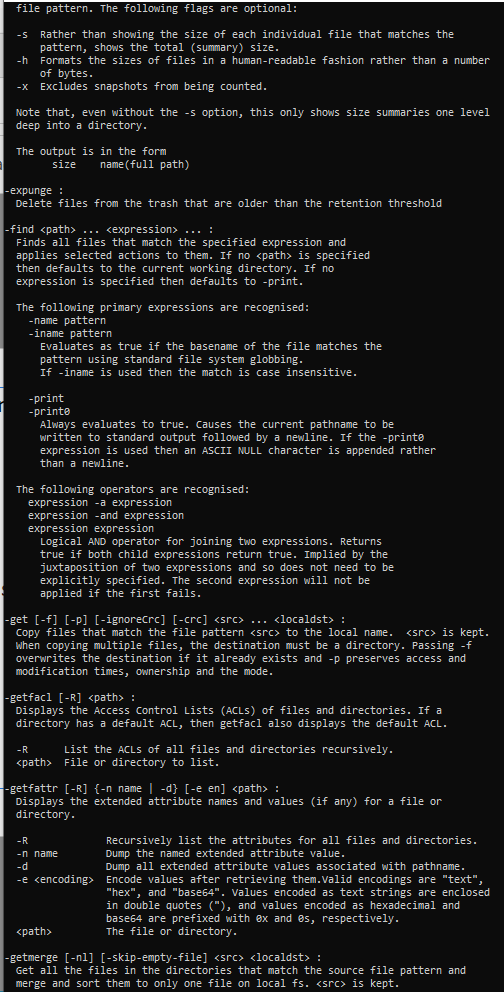
-archives <comma separated list of archives> specify comma separated archives to be unarchived on the compute machines.

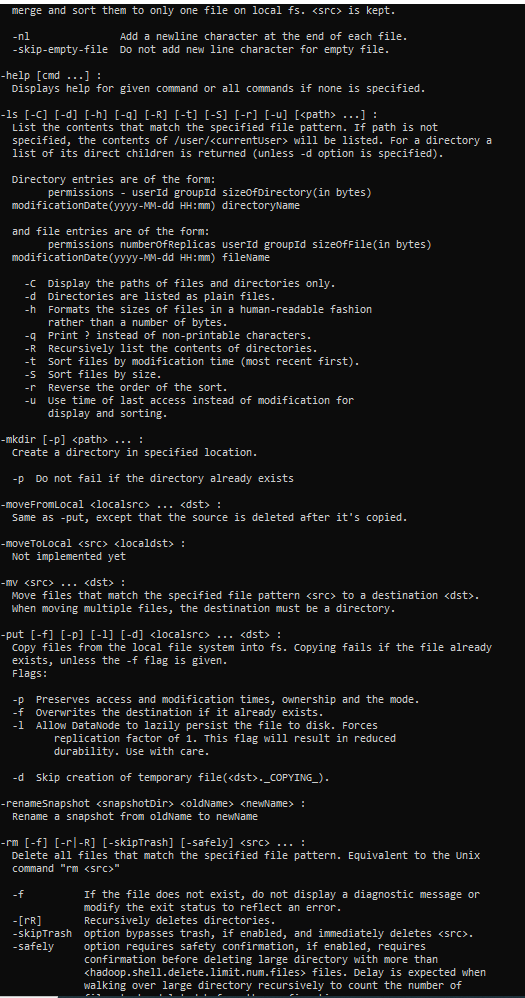
The general command line syntax is

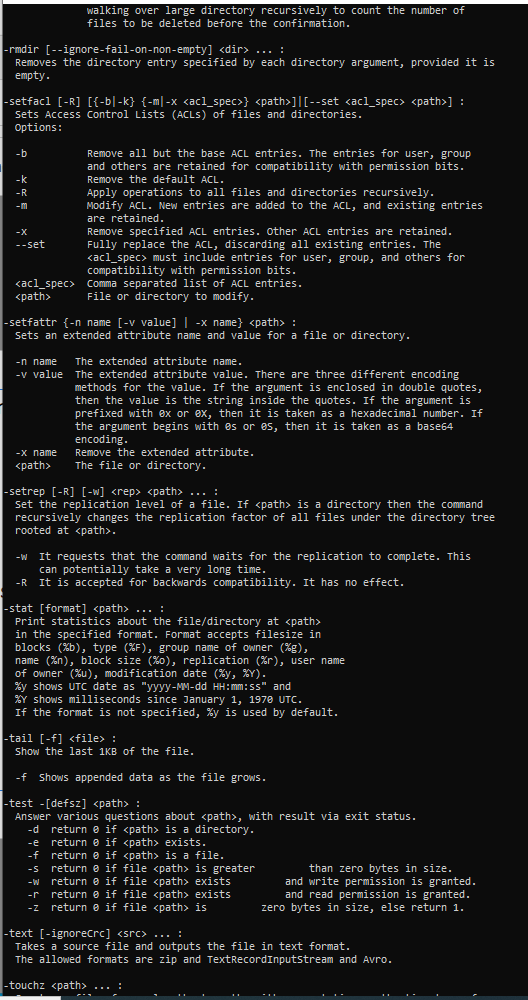
command [genericOptions] [commandOptions]













**Activity - 2 : Implementation of the Map-Reduce Program : “ Word Count”**

Hadoop Commands and their corresponding Outputs obtained on executing the Map-Reduce :- “Word Count” Program in Terminal window:

Microsoft Windows [Version 10.0.18363.1379]

(c) 2019 Microsoft Corporation. All rights reserved.

C:\WINDOWS\system32>cd..

C:\Windows>d:

D:\>cd D:\hadoop-2.8.0\hadoop-2.8.0\sbin

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>start-all.cmd

This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd

starting yarn daemons

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -mkdir /input\_dir

mkdir: Cannot create directory /input\_dir. Name node is in safe mode.

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -mkdir /input\_wcdir

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -put D:/input\_file.txt /input\_dir

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -put D:/input\_file.txt /input\_wcdir

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -ls /input\_wcdir/

Found 1 items

-rw-r--r-- 1 NRika supergroup 1888 2021-03-01 02:46 /input\_wcdir/input\_file.txt

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop dfs -cat /input\_wcdir/input\_file.txt

DEPRECATED: Use of this script to execute hdfs command is deprecated.

Instead use the hdfs command for it.

23 23 27 43 24 25 26 26 26 26 25 26 25

26 27 28 28 28 30 31 31 31 30 30 30 29

31 32 32 32 33 34 35 36 36 34 34 34 34

39 38 39 39 39 41 42 43 40 39 38 38 40

38 39 39 39 39 41 41 41 28 40 39 39 45

23 23 27 43 24 25 26 26 26 26 25 26 25

26 27 28 28 28 30 31 31 31 30 30 30 29

31 32 32 32 33 34 35 36 36 34 34 34 34

39 38 39 39 39 41 42 43 40 39 38 38 40

38 39 39 39 39 41 41 41 28 40 39 39 45

23 23 27 43 24 25 26 26 26 26 25 26 25

26 27 28 28 28 30 31 31 31 30 30 30 29

31 32 32 32 33 34 35 36 36 34 34 34 34

39 38 39 39 39 41 42 43 40 39 38 38 40

38 39 39 39 39 41 41 41 28 40 39 39 45

23 23 27 43 24 25 26 26 26 26 25 26 25

26 27 28 28 28 30 31 31 31 30 30 30 29

31 32 32 32 33 34 35 36 36 34 34 34 34

39 38 39 39 39 41 42 43 40 39 38 38 40

38 39 39 39 39 41 41 41 28 40 39 39 45

23 23 27 43 24 25 26 26 26 26 25 26 25

26 27 28 28 28 30 31 31 31 30 30 30 29

31 32 32 32 33 34 35 36 36 34 34 34 34

39 38 39 39 39 41 42 43 40 39 38 38 40

38 39 39 39 39 41 41 41 28 40 39 39 45

23 23 27 43 24 25 26 26 26 26 25 26 25

26 27 28 28 28 30 31 31 31 30 30 30 29

31 32 32 32 33 34 35 36 36 34 34 34 34

39 38 39 39 39 41 42 43 40 39 38 38 40

38 39 39 39 39 41 41 41 28 40 39 39 45

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop jar D:/MapReduceClient.jar wordcount /input\_wcdir/output\_wcdir

Usage: wordcount <in> [<in>...] <out>

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop jar D:/MapReduceClient.jar wordcount /input\_wcdir /output\_wcdir

21/03/01 02:59:17 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id

21/03/01 02:59:17 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=

21/03/01 02:59:18 INFO input.FileInputFormat: Total input files to process : 1

21/03/01 02:59:18 INFO mapreduce.JobSubmitter: number of splits:1

21/03/01 02:59:18 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_local859052974\_0001

21/03/01 02:59:19 INFO mapreduce.Job: The url to track the job: http://localhost:8080/

21/03/01 02:59:19 INFO mapreduce.Job: Running job: job\_local859052974\_0001

21/03/01 02:59:19 INFO mapred.LocalJobRunner: OutputCommitter set in config null

21/03/01 02:59:19 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1

21/03/01 02:59:19 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup \_temporary folders under output directory:false, ignore cleanup failures: false

21/03/01 02:59:19 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter

21/03/01 02:59:19 INFO mapred.LocalJobRunner: Waiting for map tasks

21/03/01 02:59:19 INFO mapred.LocalJobRunner: Starting task: attempt\_local859052974\_0001\_m\_000000\_0

21/03/01 02:59:19 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1

21/03/01 02:59:19 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup \_temporary folders under output directory:false, ignore cleanup failures: false

21/03/01 02:59:19 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.

21/03/01 02:59:19 INFO mapred.Task: Using ResourceCalculatorProcessTree : org.apache.hadoop.yarn.util.WindowsBasedProcessTree@6af4bcc7

21/03/01 02:59:19 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/input\_wcdir/input\_file.txt:0+1888

21/03/01 02:59:19 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)

21/03/01 02:59:19 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100

21/03/01 02:59:19 INFO mapred.MapTask: soft limit at 83886080

21/03/01 02:59:19 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600

21/03/01 02:59:19 INFO mapred.MapTask: kvstart = 26214396; length = 6553600

21/03/01 02:59:19 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer

21/03/01 02:59:19 INFO mapred.LocalJobRunner:

21/03/01 02:59:19 INFO mapred.MapTask: Starting flush of map output

21/03/01 02:59:19 INFO mapred.MapTask: Spilling map output

21/03/01 02:59:19 INFO mapred.MapTask: bufstart = 0; bufend = 2730; bufvoid = 104857600

21/03/01 02:59:19 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26212840(104851360); length = 1557/6553600

21/03/01 02:59:19 INFO mapred.MapTask: Finished spill 0

21/03/01 02:59:19 INFO mapred.Task: Task:attempt\_local859052974\_0001\_m\_000000\_0 is done. And is in the process of committing

21/03/01 02:59:19 INFO mapred.LocalJobRunner: map

21/03/01 02:59:19 INFO mapred.Task: Task 'attempt\_local859052974\_0001\_m\_000000\_0' done.

21/03/01 02:59:19 INFO mapred.LocalJobRunner: Finishing task: attempt\_local859052974\_0001\_m\_000000\_0

21/03/01 02:59:19 INFO mapred.LocalJobRunner: map task executor complete.

21/03/01 02:59:19 INFO mapred.LocalJobRunner: Waiting for reduce tasks

21/03/01 02:59:19 INFO mapred.LocalJobRunner: Starting task: attempt\_local859052974\_0001\_r\_000000\_0

21/03/01 02:59:19 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1

21/03/01 02:59:19 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup \_temporary folders under output directory:false, ignore cleanup failures: false

21/03/01 02:59:19 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.

21/03/01 02:59:19 INFO mapred.Task: Using ResourceCalculatorProcessTree : org.apache.hadoop.yarn.util.WindowsBasedProcessTree@24469f1a

21/03/01 02:59:19 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@2f66e4c2

21/03/01 02:59:19 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshold=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10

21/03/01 02:59:19 INFO reduce.EventFetcher: attempt\_local859052974\_0001\_r\_000000\_0 Thread started: EventFetcher for fetching Map Completion Events

21/03/01 02:59:19 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt\_local859052974\_0001\_m\_000000\_0 decomp: 191 len: 195 to MEMORY

21/03/01 02:59:19 INFO reduce.InMemoryMapOutput: Read 191 bytes from map-output for attempt\_local859052974\_0001\_m\_000000\_0

21/03/01 02:59:19 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 191, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->191

21/03/01 02:59:19 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning

21/03/01 02:59:19 INFO mapred.LocalJobRunner: 1 / 1 copied.

21/03/01 02:59:19 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs

21/03/01 02:59:19 INFO mapred.Merger: Merging 1 sorted segments

21/03/01 02:59:19 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 186 bytes

21/03/01 02:59:19 INFO reduce.MergeManagerImpl: Merged 1 segments, 191 bytes to disk to satisfy reduce memory limit

21/03/01 02:59:19 INFO reduce.MergeManagerImpl: Merging 1 files, 195 bytes from disk

21/03/01 02:59:19 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce

21/03/01 02:59:19 INFO mapred.Merger: Merging 1 sorted segments

21/03/01 02:59:19 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 186 bytes

21/03/01 02:59:19 INFO mapred.LocalJobRunner: 1 / 1 copied.

21/03/01 02:59:19 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords

21/03/01 02:59:19 INFO mapred.Task: Task:attempt\_local859052974\_0001\_r\_000000\_0 is done. And is in the process of committing

21/03/01 02:59:19 INFO mapred.LocalJobRunner: 1 / 1 copied.

21/03/01 02:59:19 INFO mapred.Task: Task attempt\_local859052974\_0001\_r\_000000\_0 is allowed to commit now

21/03/01 02:59:19 INFO output.FileOutputCommitter: Saved output of task 'attempt\_local859052974\_0001\_r\_000000\_0' to hdfs://localhost:9000/output\_wcdir/\_temporary/0/task\_local859052974\_0001\_r\_000000

21/03/01 02:59:19 INFO mapred.LocalJobRunner: reduce > reduce

21/03/01 02:59:19 INFO mapred.Task: Task 'attempt\_local859052974\_0001\_r\_000000\_0' done.

21/03/01 02:59:19 INFO mapred.LocalJobRunner: Finishing task: attempt\_local859052974\_0001\_r\_000000\_0

21/03/01 02:59:19 INFO mapred.LocalJobRunner: reduce task executor complete.

21/03/01 02:59:20 INFO mapreduce.Job: Job job\_local859052974\_0001 running in uber mode : false

21/03/01 02:59:20 INFO mapreduce.Job: map 100% reduce 100%

21/03/01 02:59:20 INFO mapreduce.Job: Job job\_local859052974\_0001 completed successfully

21/03/01 02:59:20 INFO mapreduce.Job: Counters: 35

File System Counters

FILE: Number of bytes read=604640

FILE: Number of bytes written=1252227

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=3776

HDFS: Number of bytes written=120

HDFS: Number of read operations=13

HDFS: Number of large read operations=0

HDFS: Number of write operations=4

Map-Reduce Framework

Map input records=30

Map output records=390

Map output bytes=2730

Map output materialized bytes=195

Input split bytes=113

Combine input records=390

Combine output records=21

Reduce input groups=21

Reduce shuffle bytes=195

Reduce input records=21

Reduce output records=21

Spilled Records=42

Shuffled Maps =1

Failed Shuffles=0

Merged Map outputs=1

GC time elapsed (ms)=4

Total committed heap usage (bytes)=463470592

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=1888

File Output Format Counters

Bytes Written=120

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop dfs -cat /output\_wcdir/\*

DEPRECATED: Use of this script to execute hdfs command is deprecated.

Instead use the hdfs command for it.

23 12

24 6

25 18

26 36

27 12

28 24

29 6

30 24

31 24

32 18

33 6

34 30

35 6

36 12

38 24

39 66

40 18

41 24

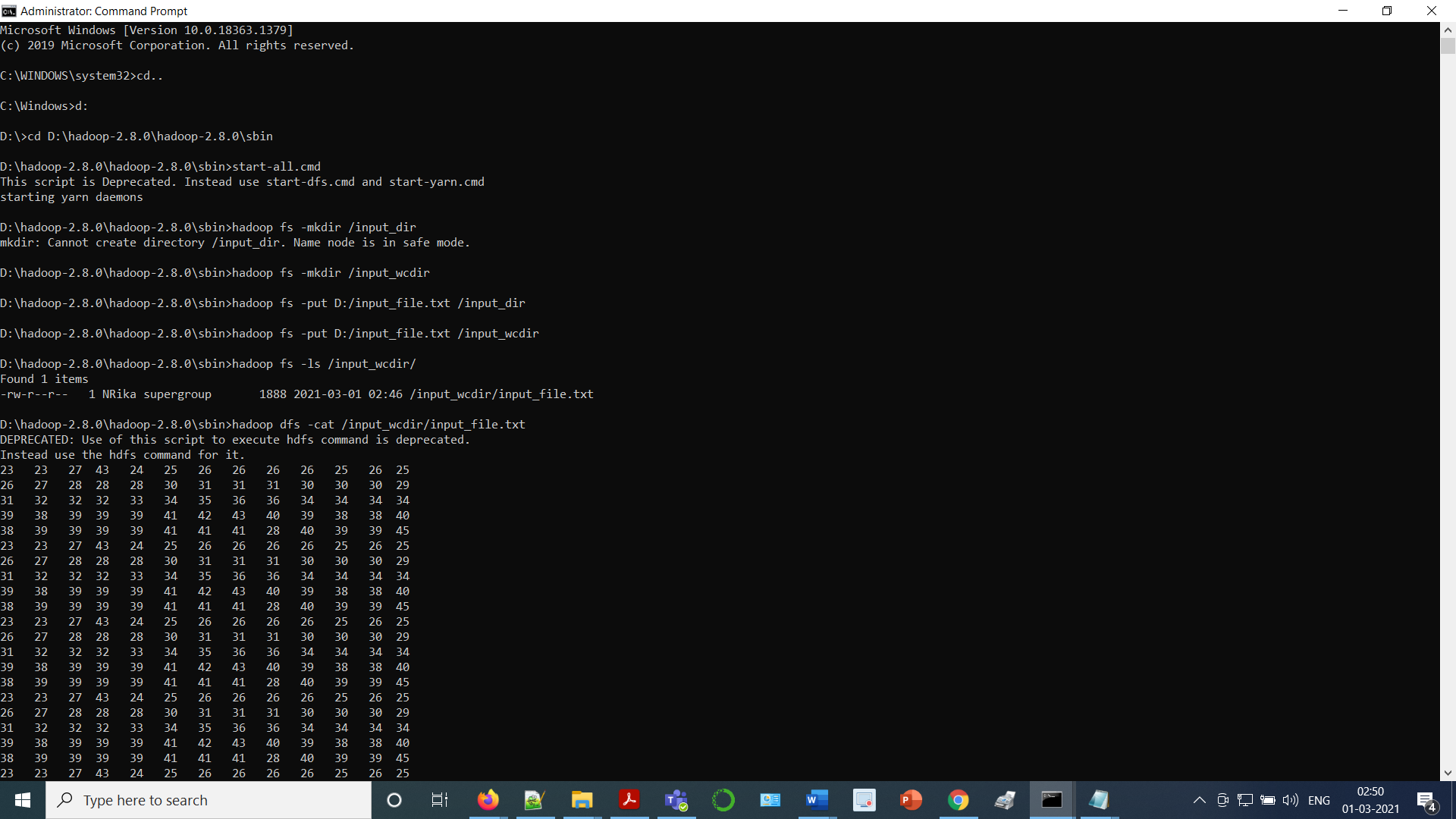
42 6

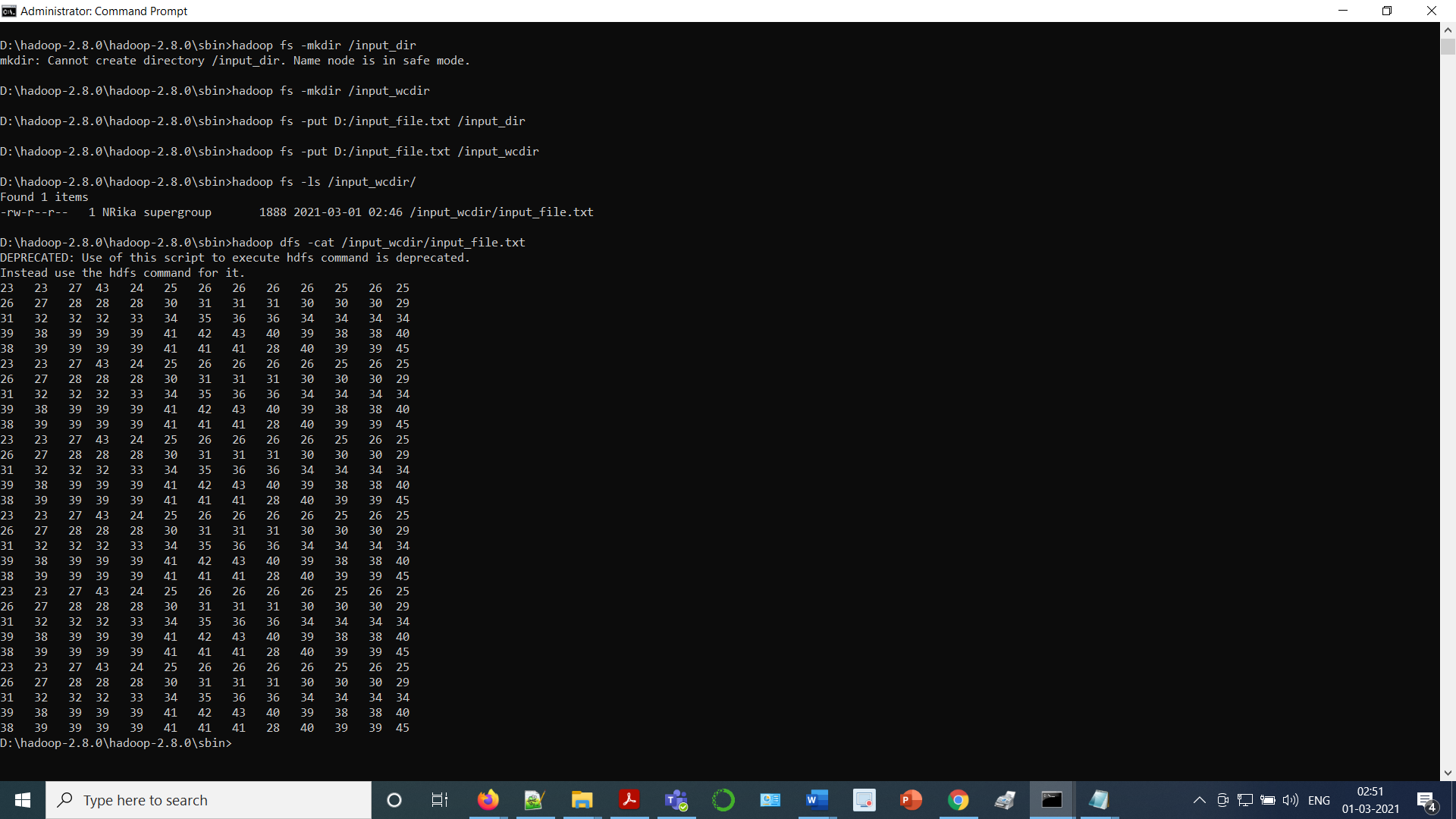
43 12

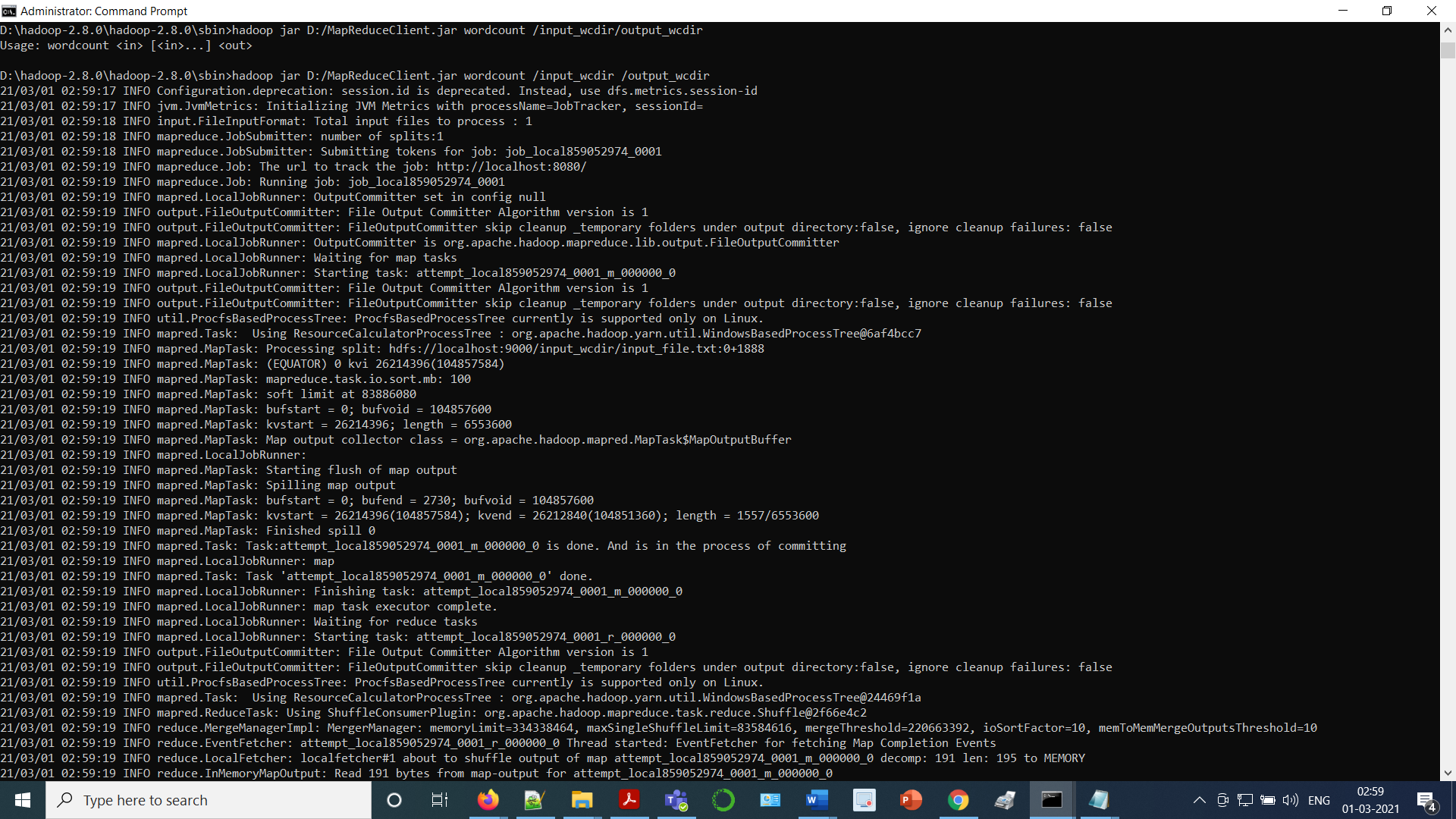
45 6

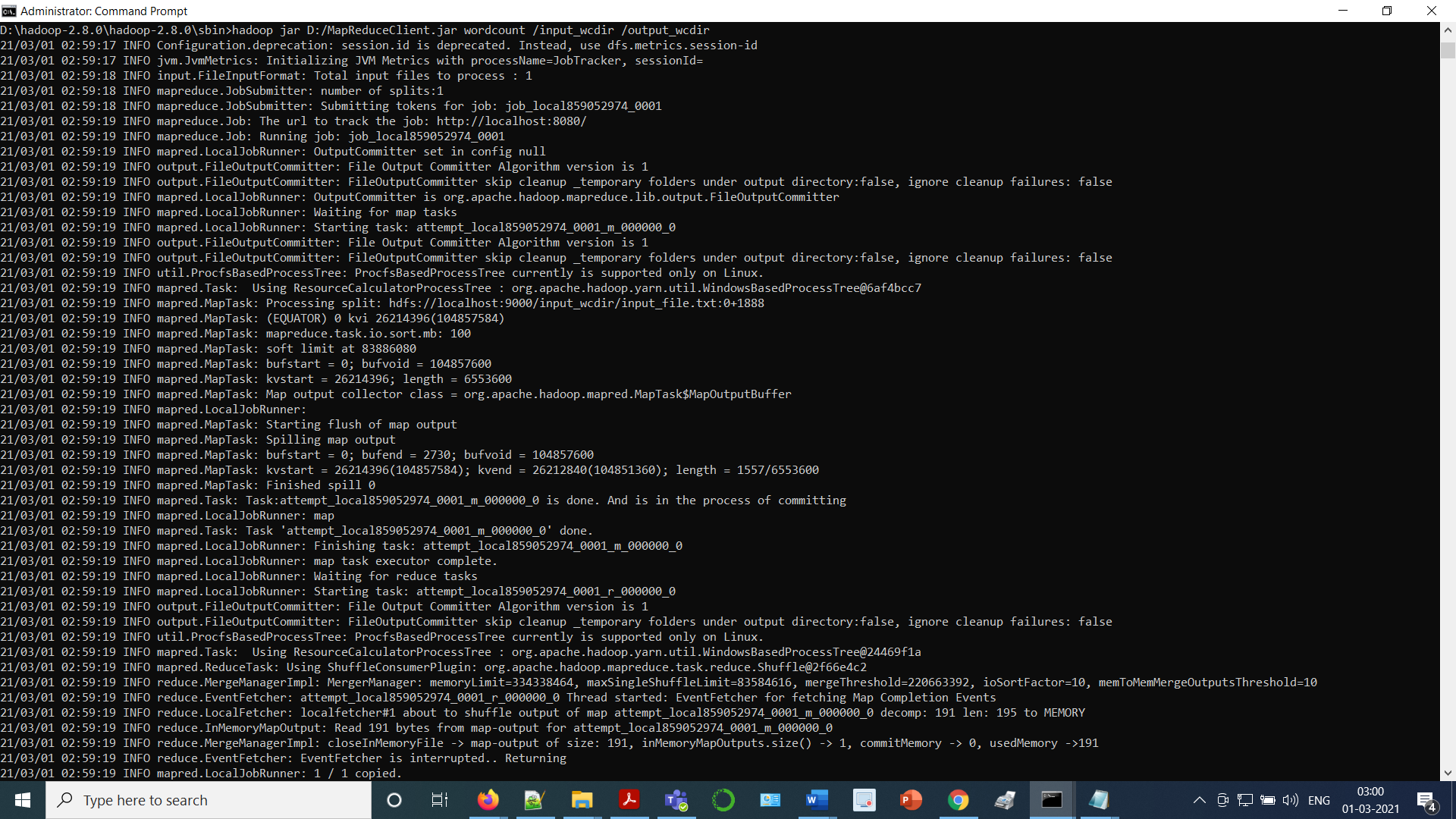
D:\hadoop-2.8.0\hadoop-2.8.0\sbin>

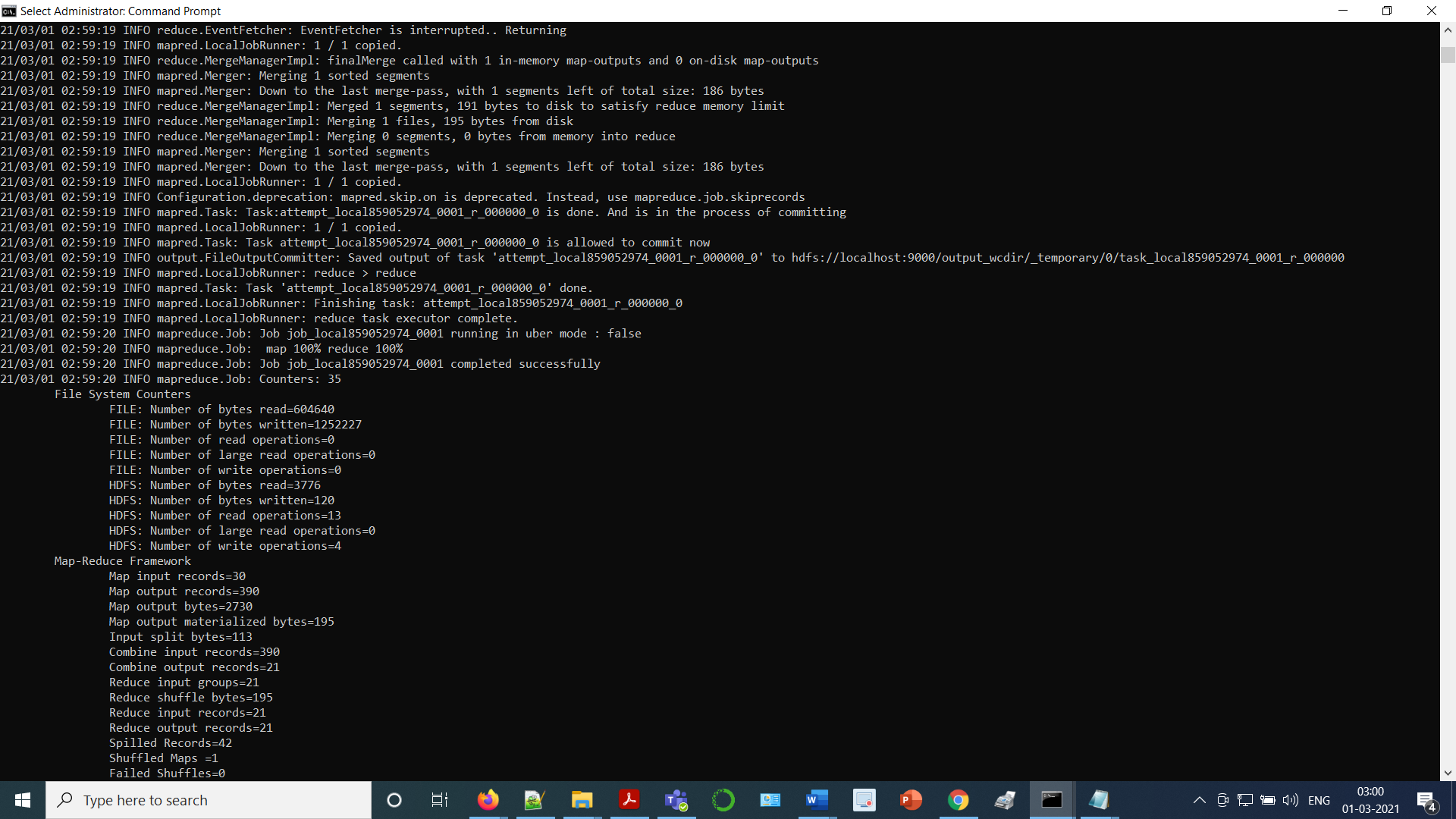
Screenshot of the Hadoop Commands and their corresponding Outputs obtained on executing the Map-Reduce :- Word Count Program in Terminal window:

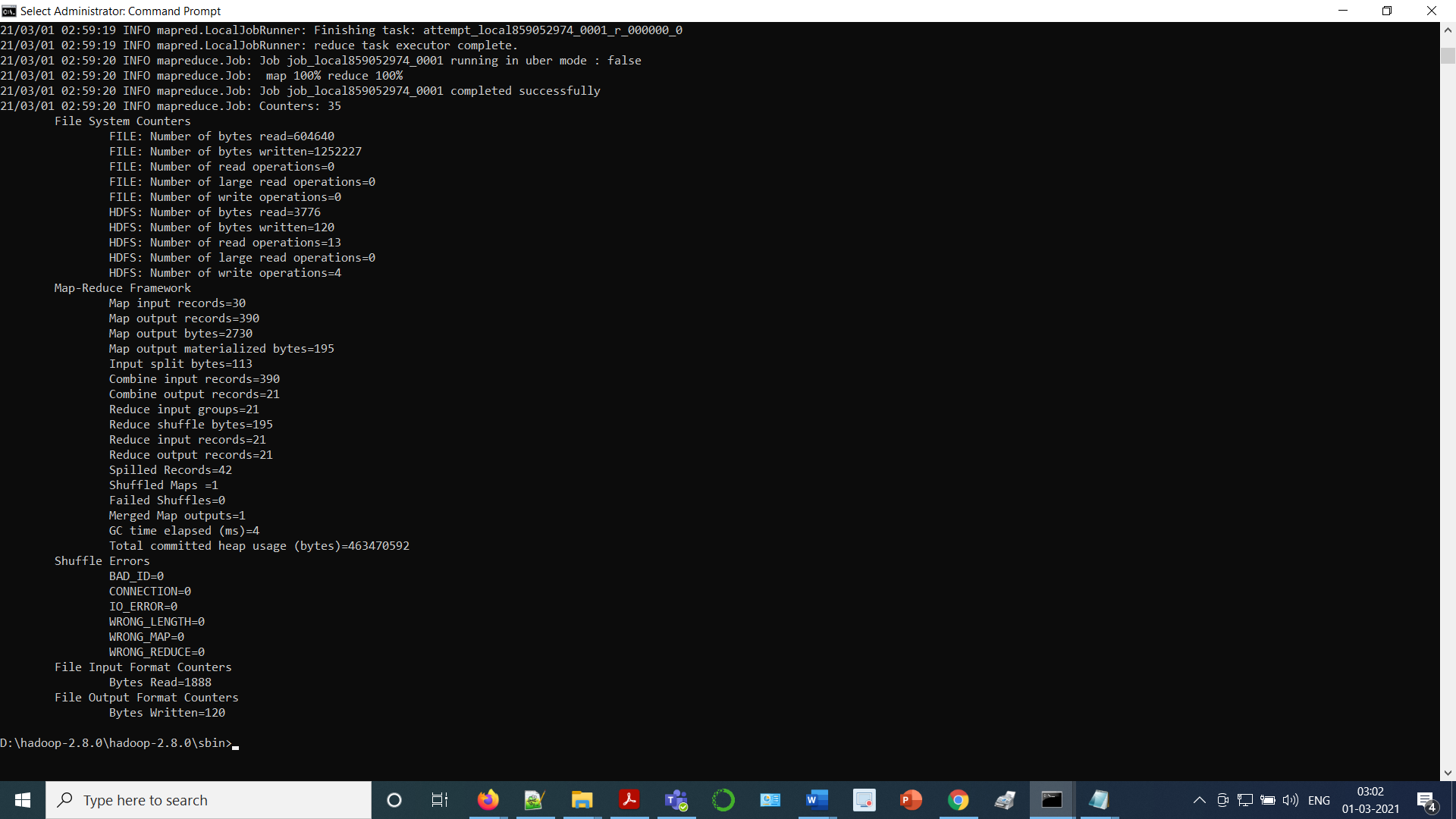


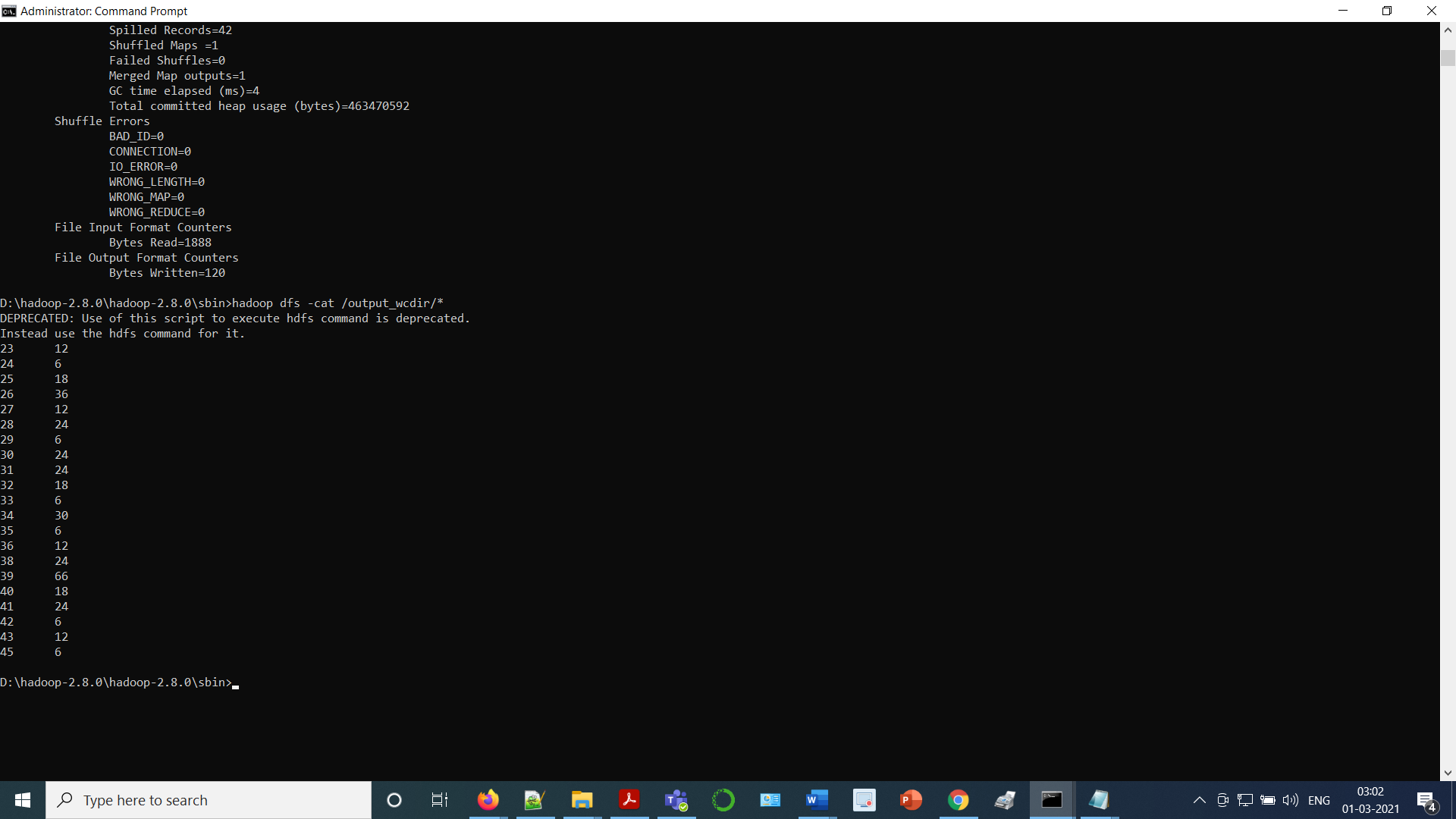












Java Code for the Map-Reduce :- Word Count Program :

import java.io.IOException;

import java.util.StringTokenizer;

import org.apache.hadoop.conf.Configuration;

import org.apache.hadoop.fs.Path;

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Job;

import org.apache.hadoop.mapreduce.Mapper;

import org.apache.hadoop.mapreduce.Reducer;

import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;

import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class WordCount {

public static class TokenizerMapper

extends Mapper<Object, Text, Text, IntWritable>{

private final static IntWritable one = new IntWritable(1);

private Text word = new Text();

public void map(Object key, Text value, Context context

) throws IOException, InterruptedException {

StringTokenizer itr = new StringTokenizer(value.toString());

while (itr.hasMoreTokens()) {

word.set(itr.nextToken());

context.write(word, one);

}

}

}

public static class IntSumReducer

extends Reducer<Text,IntWritable,Text,IntWritable> {

private IntWritable result = new IntWritable();

public void reduce(Text key, Iterable<IntWritable> values,

Context context

) throws IOException, InterruptedException {

int sum = 0;

for (IntWritable val : values) {

sum += val.get();

}

result.set(sum);

context.write(key, result);

}

}

public static void main(String[] args) throws Exception {

Configuration conf = new Configuration();

Job job = Job.getInstance(conf, "word count");

job.setJarByClass(WordCount.class);

job.setMapperClass(TokenizerMapper.class);

job.setCombinerClass(IntSumReducer.class);

job.setReducerClass(IntSumReducer.class);

job.setOutputKeyClass(Text.class);

job.setOutputValueClass(IntWritable.class);

FileInputFormat.addInputPath(job, new Path(args[0]));

FileOutputFormat.setOutputPath(job, new Path(args[1]));

System.exit(job.waitForCompletion(true) ? 0 : 1);

}

}

**Activity - 3 : Implementation of the Map-Reduce**

**Program : “Market Analysis”**

Hadoop Commands and their corresponding Outputs obtained on executing the Map-Reduce :- “Market Analysis” Program in Terminal window:

Microsoft Windows [Version 10.0.18363.1379]

(c) 2019 Microsoft Corporation. All rights reserved.

C:\WINDOWS\system32>cd..

C:\Windows>d:

D:\>cd D:\hadoop-2.8.0\hadoop-2.8.0\sbin

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>start-all.cmd

This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd

starting yarn daemons

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -mkdir input1\_dir

mkdir: `input1\_dir': No such file or directory

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -mkdir /input1\_dir

D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -put D:/markets\_2\_mrp.txt /input1\_dir

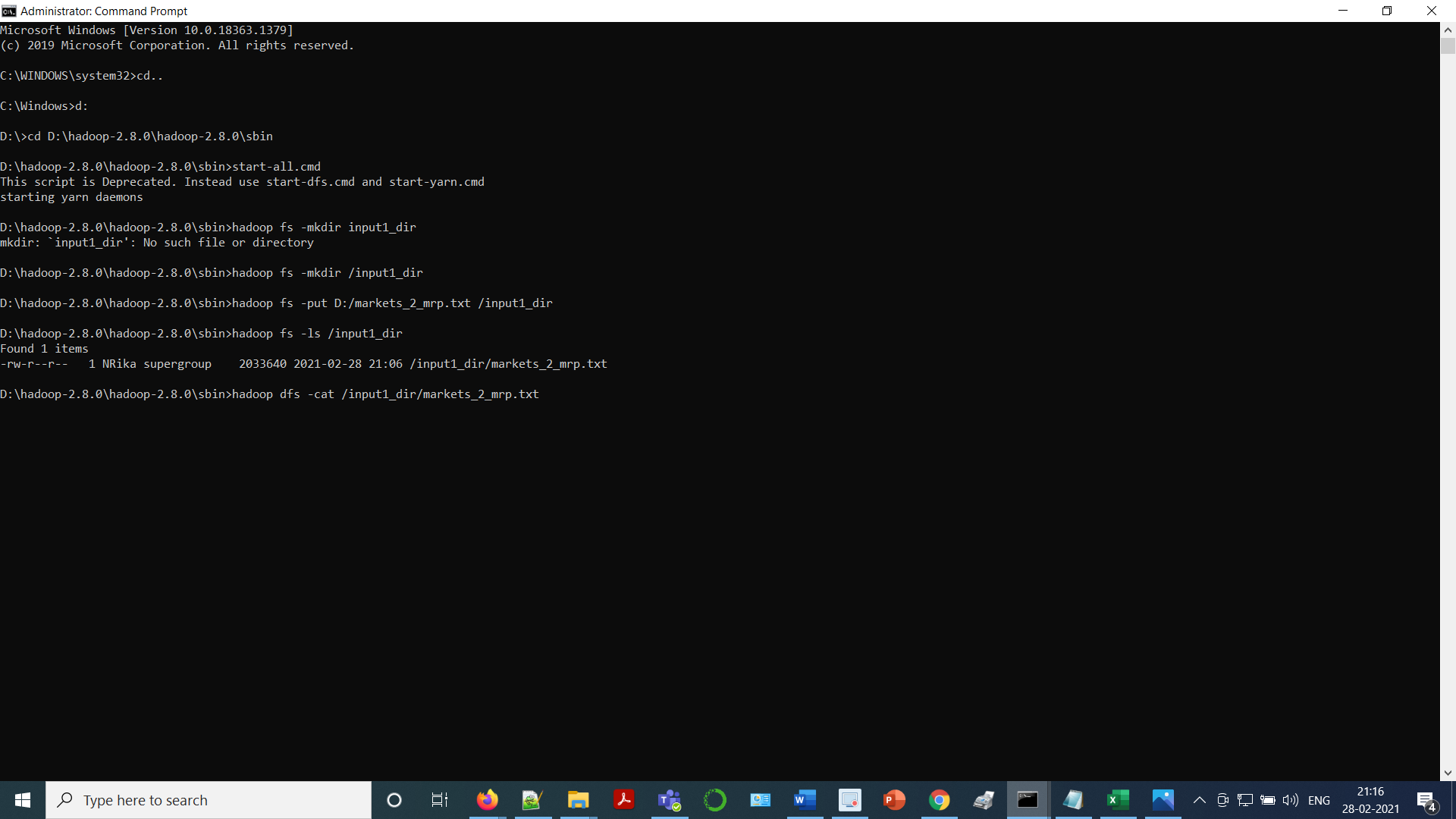
D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop fs -ls /input1\_dir

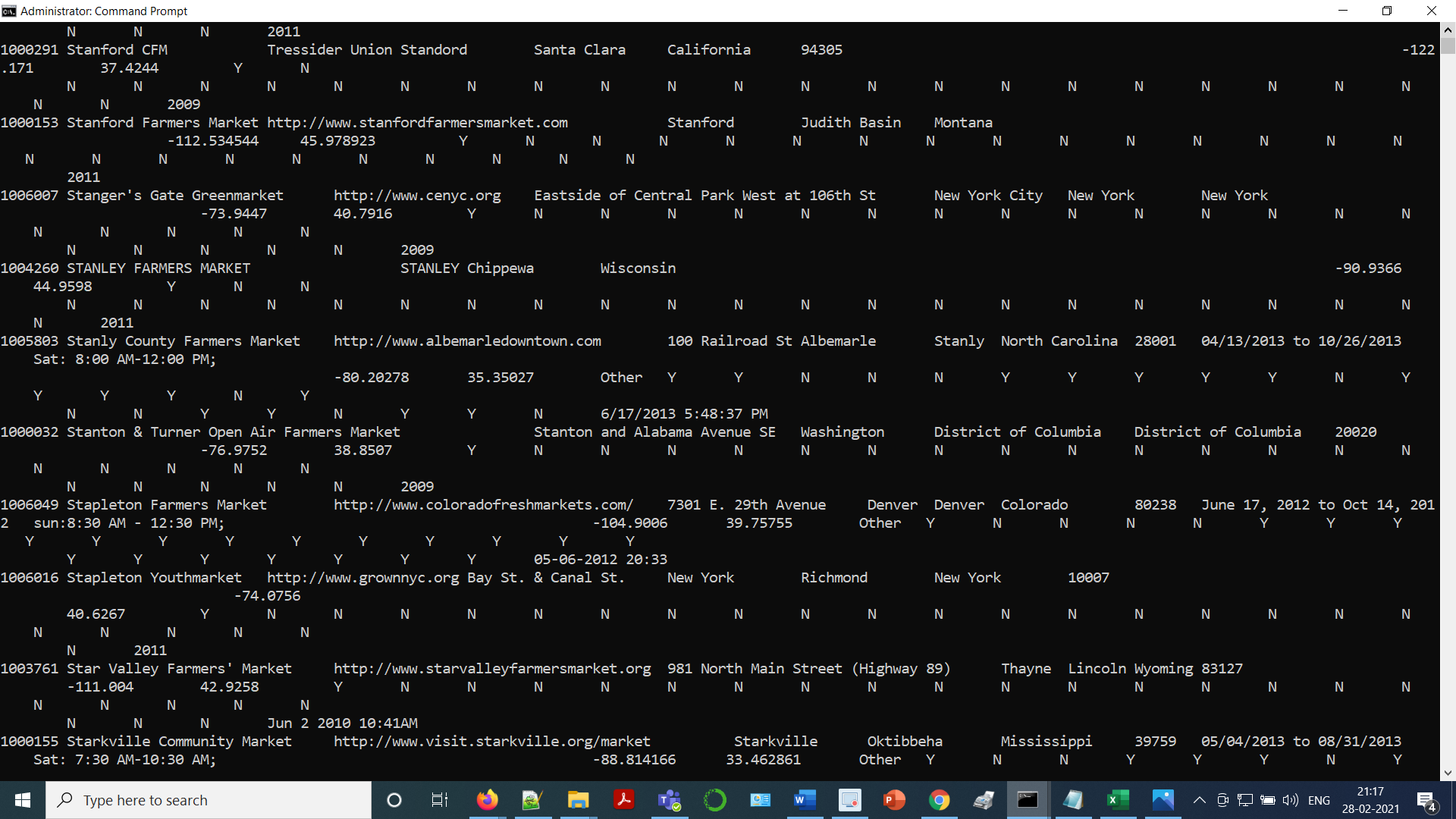
Found 1 items

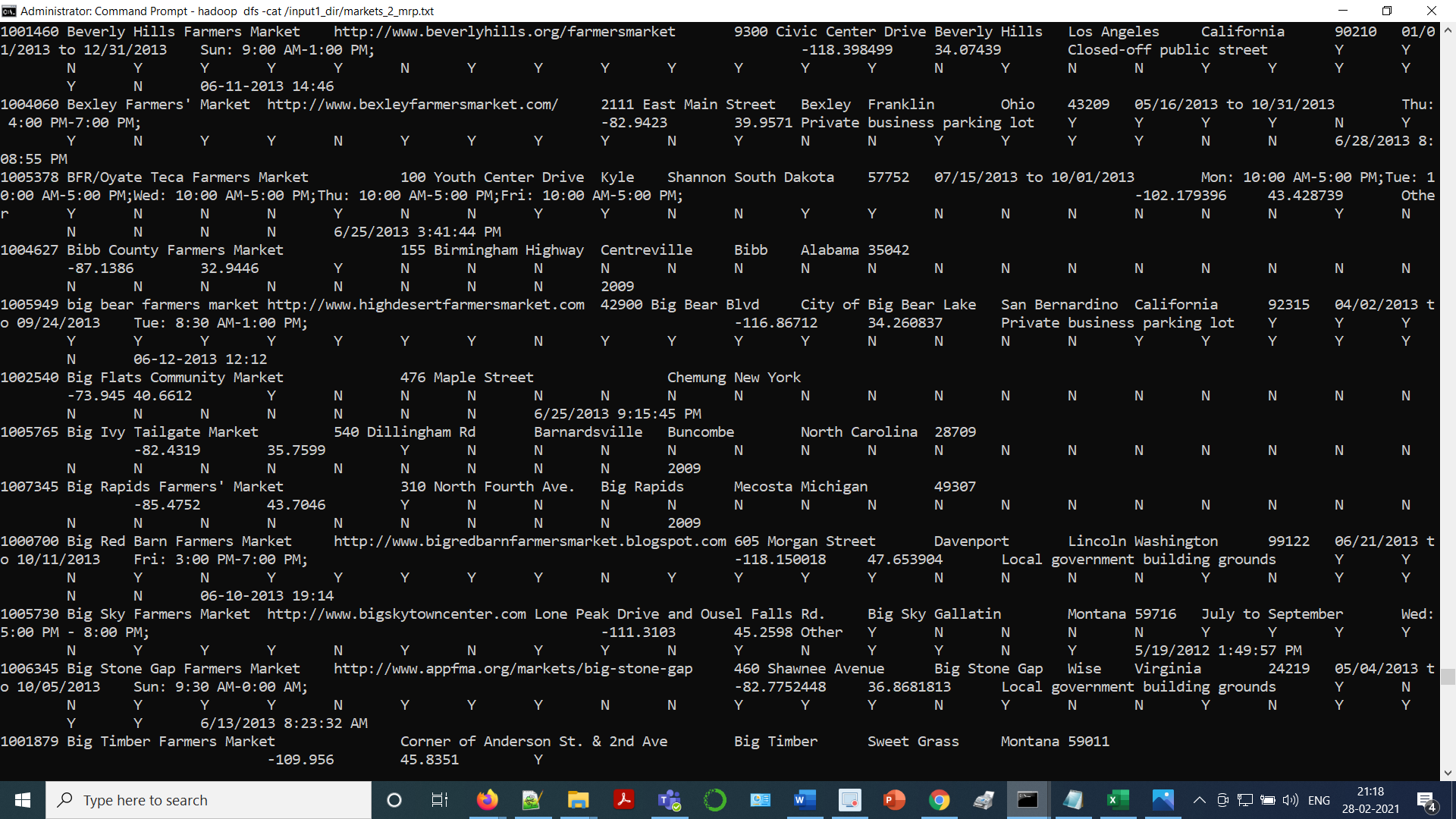
-rw-r--r-- 1 NRika supergroup 2033640 2021-02-28 21:06 /input1\_dir/markets\_2\_mrp.txt

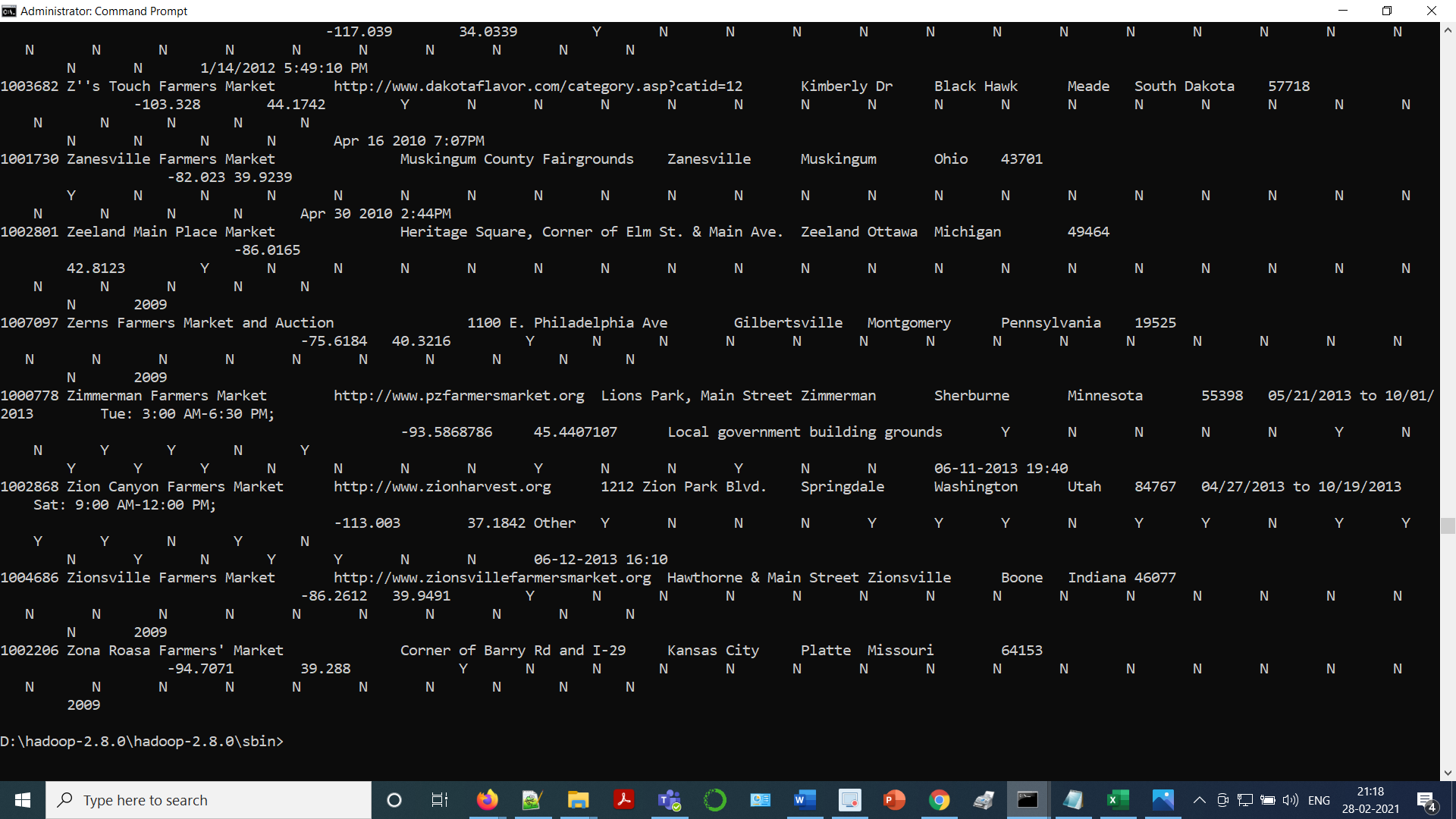
D:\hadoop-2.8.0\hadoop-2.8.0\sbin>hadoop dfs -cat /input1\_dir/markets\_2\_mrp.txt

Screenshot of the Hadoop Commands and their corresponding Outputs obtained on executing the Map-Reduce :- “Market Analysis” Program in Terminal window:









Java Code for the Map-Reduce :- Market Analysis Program :

import java.io.IOException;

import java.util.\*;

import org.apache.hadoop.fs.Path;

import org.apache.hadoop.conf.\*;

import org.apache.hadoop.io.\*;

import org.apache.hadoop.mapred.\*;

import org.apache.hadoop.util.\*;

public class MarketEvaluation extends Configured implements Tool{

public static class MapClass extends MapReduceBase implements Mapper<LongWritable, Text, Text, Text>{

private Text location = new Text();

private Text rating = new Text();

@Override

public void map(LongWritable key, Text value, OutputCollector<Text, Text> output, Reporter reporter) throws IOException{

String[] rows = value.toString().split(",");

if(rows.length > 28){ //Each row has 29 columns

String marketName = rows[0];

String website = rows[1];

String city = rows[2];

String state = rows[3];

int count = 0; // Initialized with Zero

String evaluation = "Not Evaluated";

for (int col=4; col<=28; col++) //columns 4-28 contain data about what the market offers in 'Y' or 'N' format

{

if(rows[col].equals("Y"))

count++;

}

count=(count\*100)/25; // There are 25 segments of market offerings, and it calculates %age of 'Y'

if (count > 59) { // Assigning 60-100% as High Value Market

evaluation = "High";

}

else if (count > 39 && count < 60) { //Assigning 40-59% as Medium Value Market

evaluation = "Medium";

}

else if (count > 0 && count < 40) { // Assigning 1-39% as Low Value Market

evaluation = "Low";

}

else if (count == 0 ) {

evaluation = "Zero";

}

location.set(marketName + ", " + website + ", " + city + ", " + state + ", ");

rating.set(evaluation + "\t" + count);

output.collect(location, rating);

}

}

}

public static class Reduce extends MapReduceBase implements Reducer<Text, Text, Text, Text>

{

@Override

public void reduce(Text key, Iterator<Text> values, OutputCollector<Text, Text> output, Reporter reporter) throws IOException{

int rating = 0; // Variable Initialized

String strEvaluation = "Not Evaluated"; //Variable Initialized

while(values.hasNext()){

String tokens[] = (values.next().toString()).split("\t");

String strEval = String.valueOf(tokens[0]); //gets string Value as High or Medium or Low or Zero

int val = Integer.parseInt(tokens[1]); //gets %age value

if(val > 0)

{

rating = val;

strEvaluation = strEval;

}

}

output.collect(key, new Text(strEvaluation + "," + rating + "%"));

}

}

@Override

public int run(String[] args) throws IOException

{

return 0;

}

public static void main(String[] args) throws IOException{

JobConf conf = new JobConf(MarketEvaluation.class);

conf.setJobName("MarketEvaluation");

conf.setOutputKeyClass(Text.class);

conf.setOutputValueClass(Text.class);

conf.setMapperClass(MapClass.class);

conf.setReducerClass(Reduce.class);

conf.setInputFormat(TextInputFormat.class);

conf.setOutputFormat(TextOutputFormat.class);

FileInputFormat.setInputPaths(conf, new Path(args[0]));

FileOutputFormat.setOutputPath(conf, new Path(args[1]));

JobClient.runJob(conf);

}

}

**A brief Description / Inference of the**

**Map-Reduce :- “Market Analysis” Program :**

This Map-Reduce Program :- “Market Analysis” basically gives the information about the markets and the products available in different regions based on the seasons.

The fields listed in the dataset used for performing / executing the Map-Reduce Program :- “Market Analysis” are as follows:

MID, MarketName, Website, street, city, County, State, zip, Season1Date, Season1Time, Season2Date, Season2Time, Season3Date, Season3Time, Season4Date, Season4Time, x, y, Location, Credit, WIC, WICcash, SFMNP, SNAP, Bakedgoods, Cheese, Crafts, Flowers, Eggs, Seafood, Herbs, Vegetables, Honey, Jams, Maple, Meat, Nursery, Nuts, Plants, Poultry, Prepared, Soap, Trees, Wine, updateTime.

In this problem we need to select any particular county and calculate the percentage of different products produced by each Market in that particular county.

Here, in the dataset used, we have a total of 24 products which consists of the values Y or N.

We need to count the products that a particular Market will produce, and which is denoted by : Y and also we need to calculate the percentage of that type of products as count%25.

We also need to divide and display the products into three categories : High, Medium and Low, keeping into account the following constraints for each category of products :

High – above 60 %

Medium – less than or equal to 60% and greater than 40%

Low – less than or equal to 40%

Also we need to find / determine the count of the Markets that come under the category : “HIGH”.