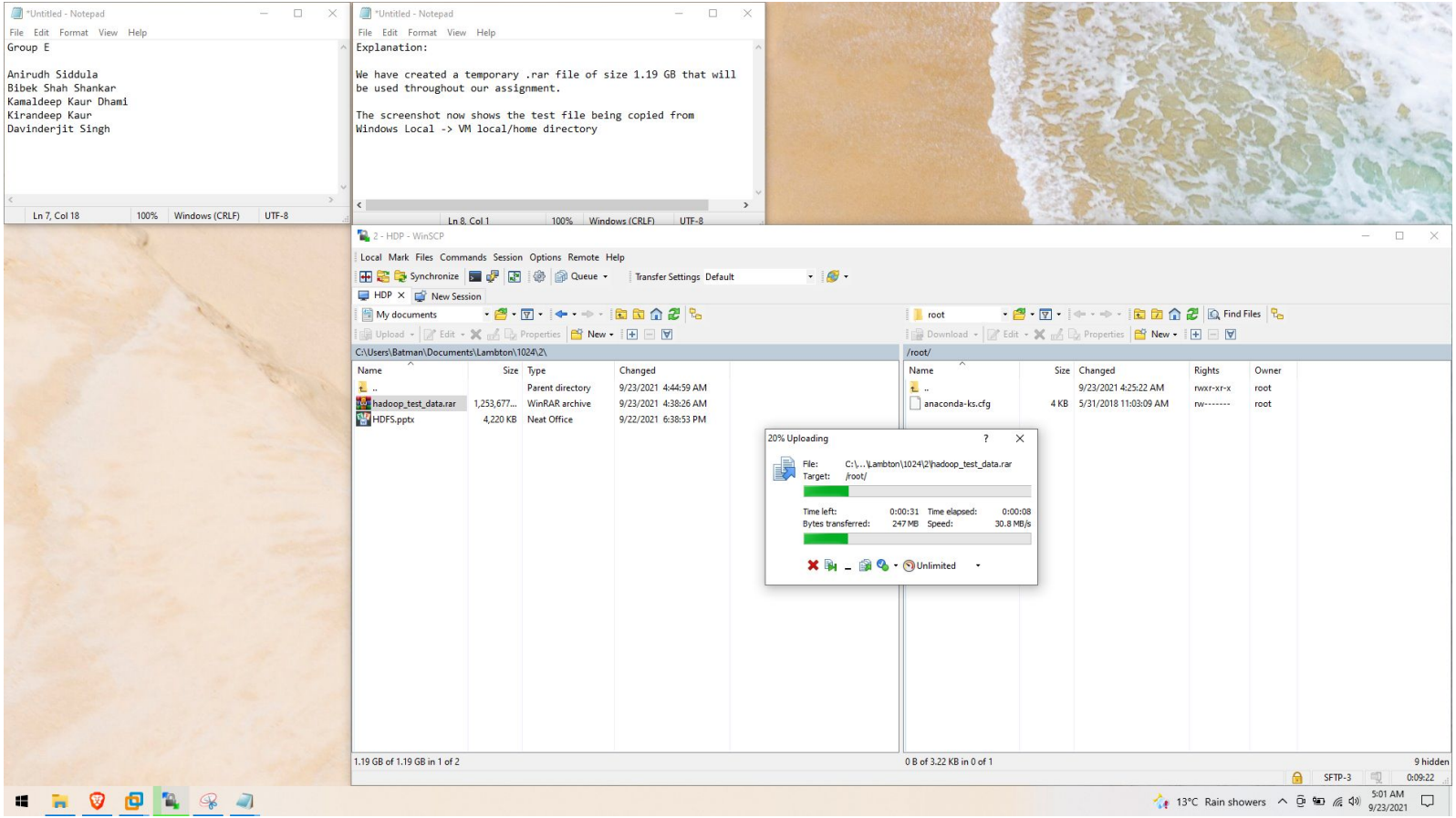
BDM 1024 - Data Technology Solutions: Assignment 2

Anirudh Siddula
Bibek Shah Shankhar
Kamaldeep Kaur Dhami
Davinderjit Singh
Kirandeep Kaur

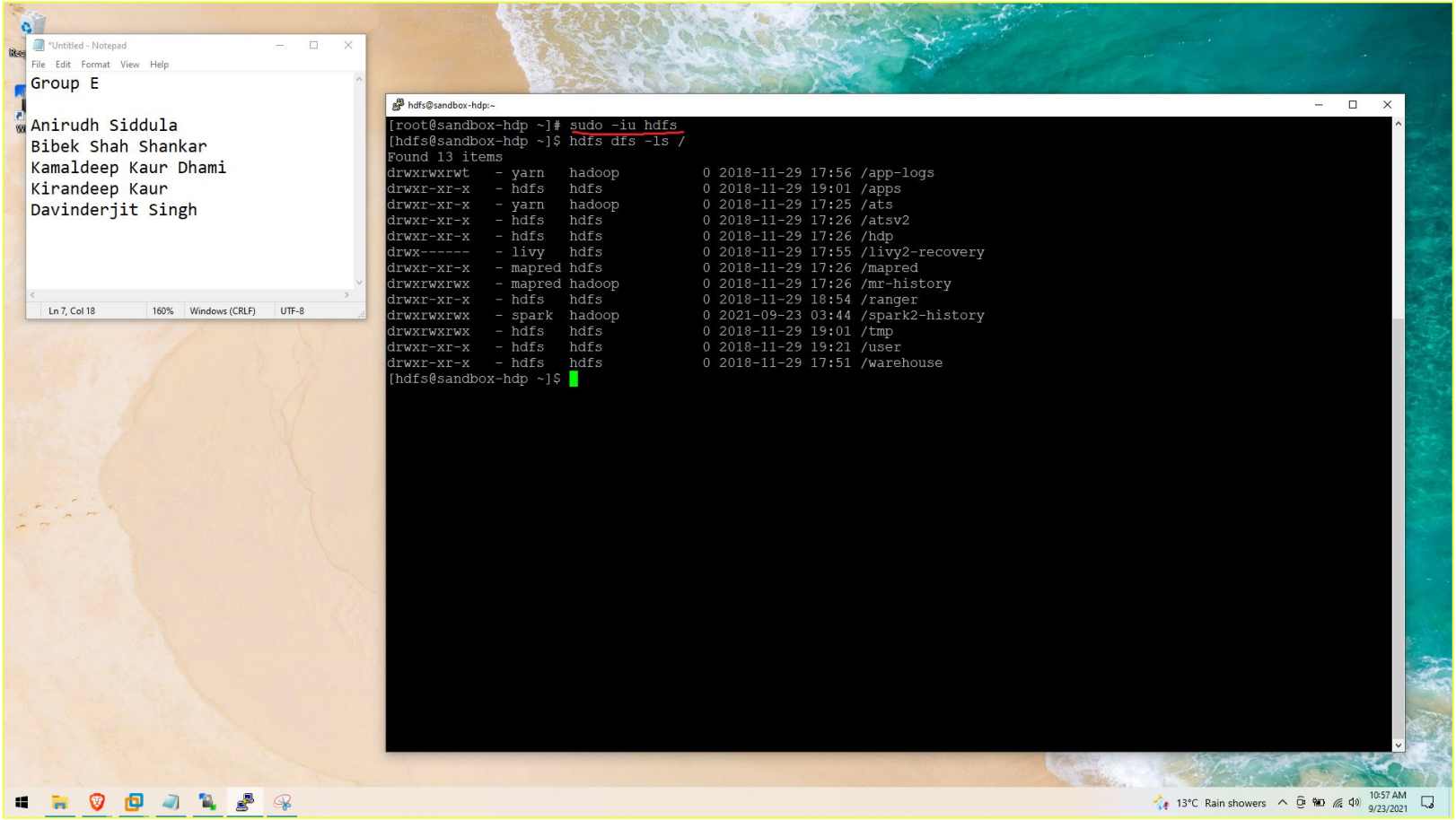
Lab1b-Part1: We have created a temporary archive file of size 1.1GB and transferred the file from Windows Local to HDP VM's local using WinScp as shown in the attached screenshot below.



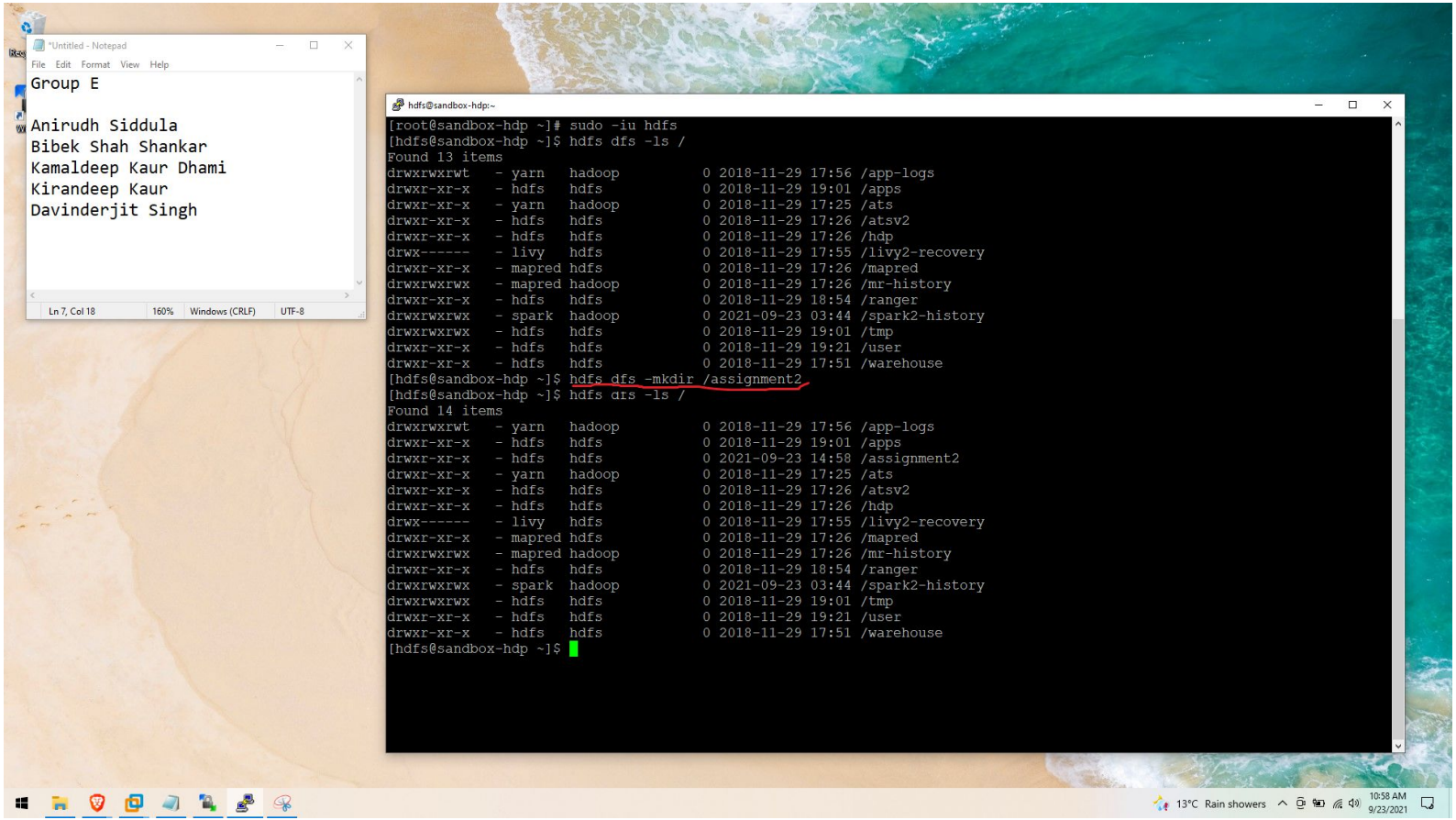
Lab1b-Part2:

DFS:

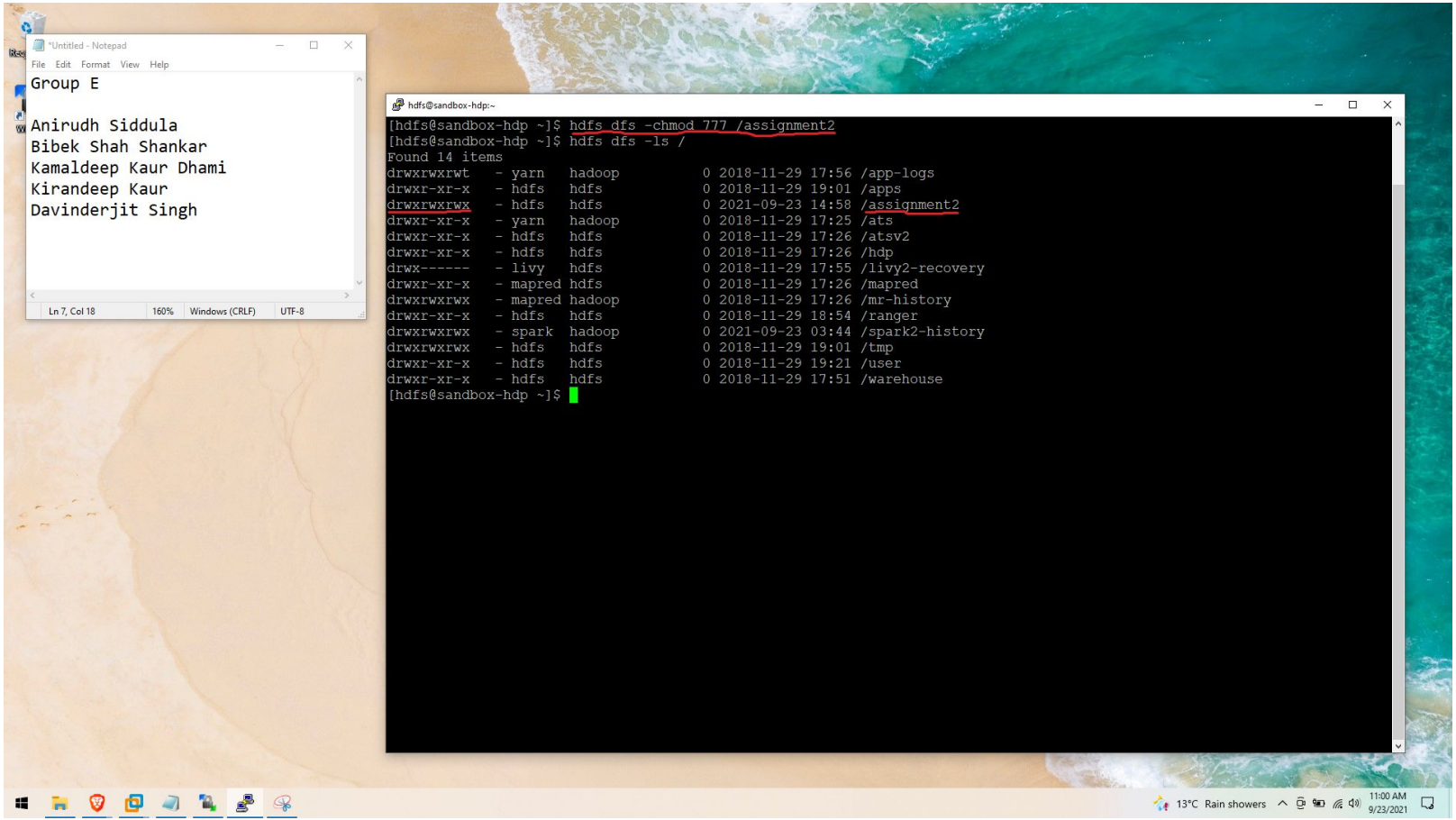
Logged in as hdfs user using super user do command and listed all the directories in the root folder of hadoop using “hdfs dfs -ls” command



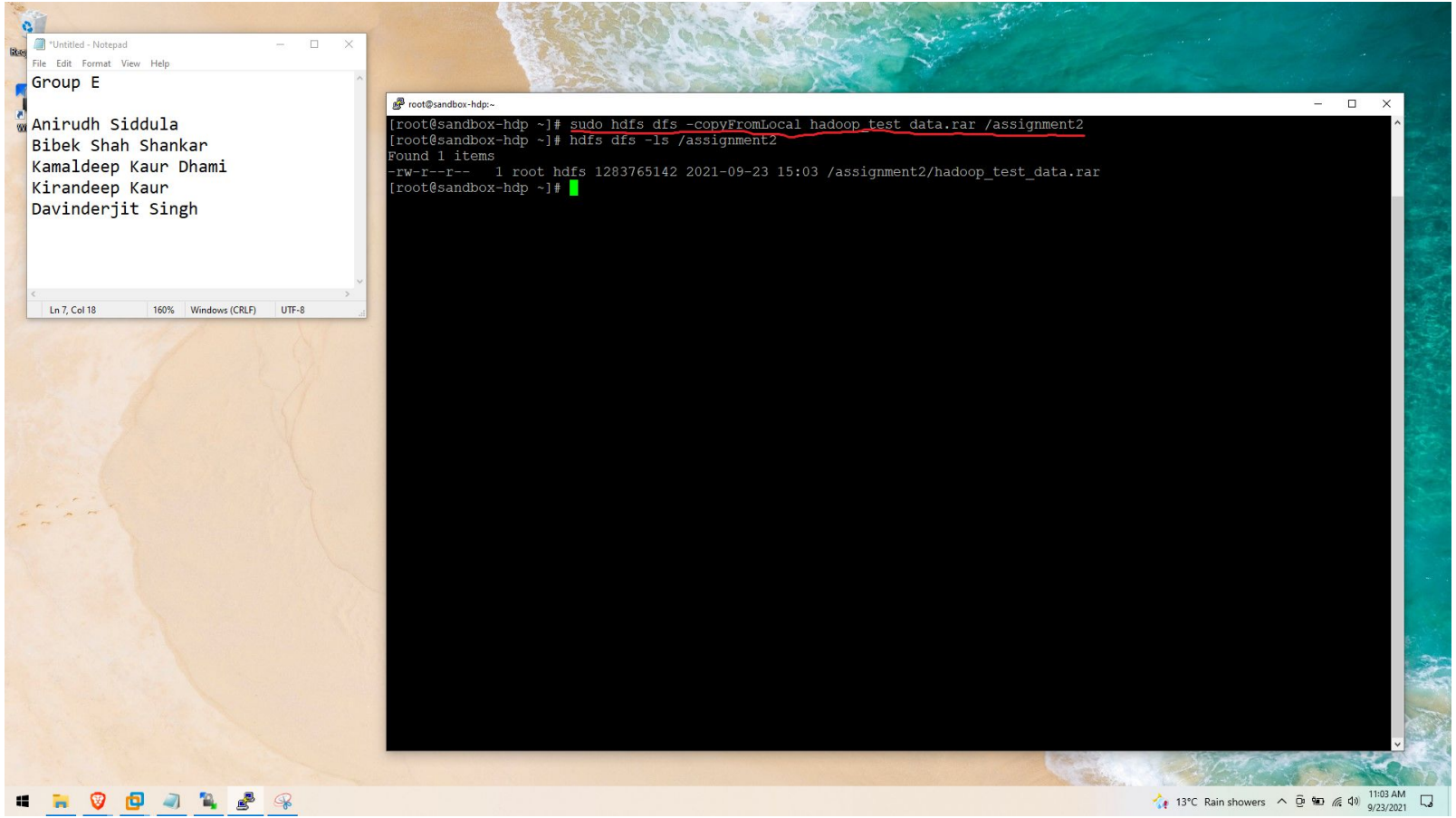
Created a directory in hadoop distributed file system using mkdir command where -mkdir being the generic unix command. Please refer the screenshot attached below.



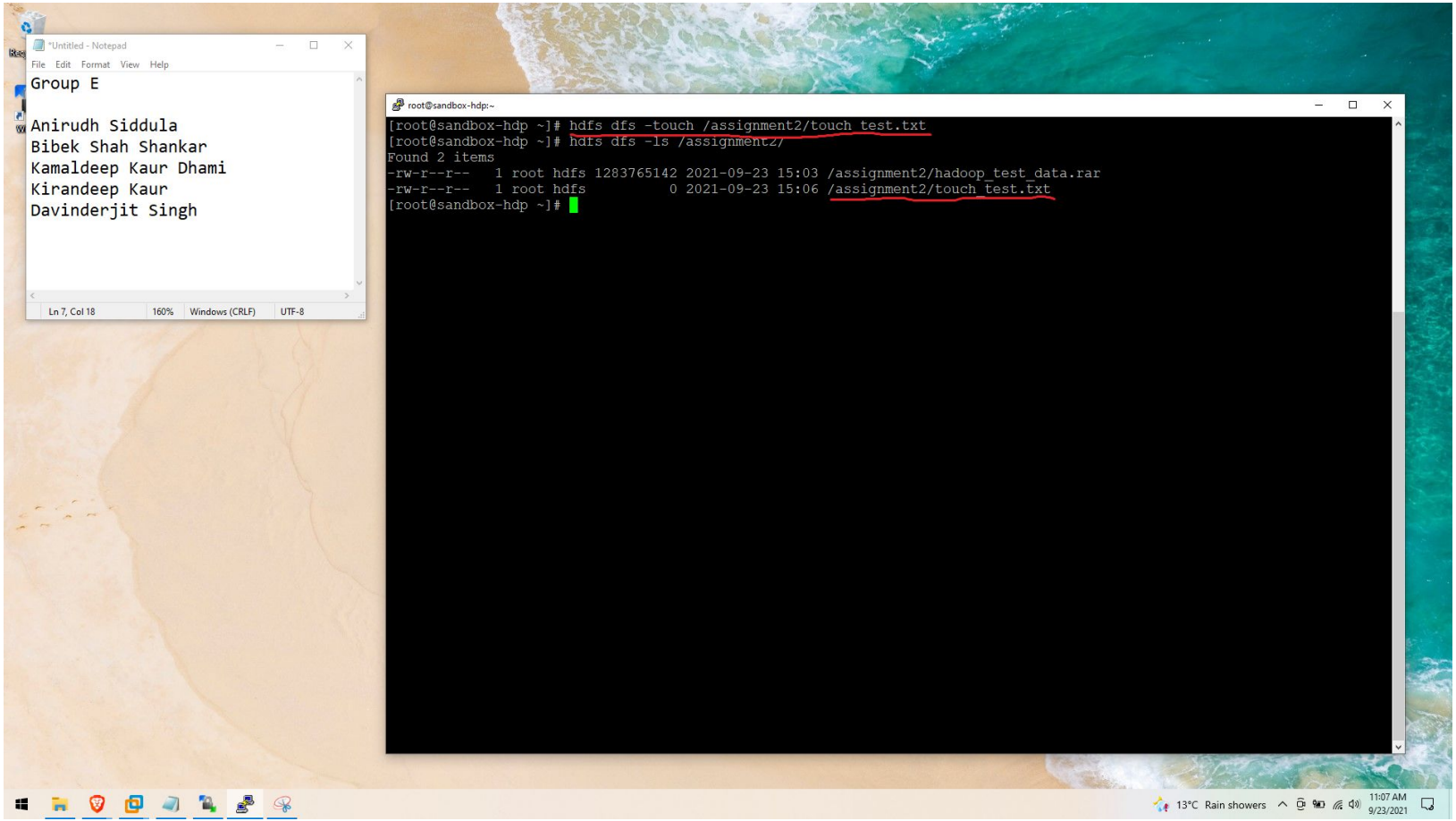
We have modified the directory permissions using “hdfs dfs -chmod” command to enable write permission when we copy the test archive file into hadoop. Please refer the underlined in screenshot for the updated permissions on the directory.



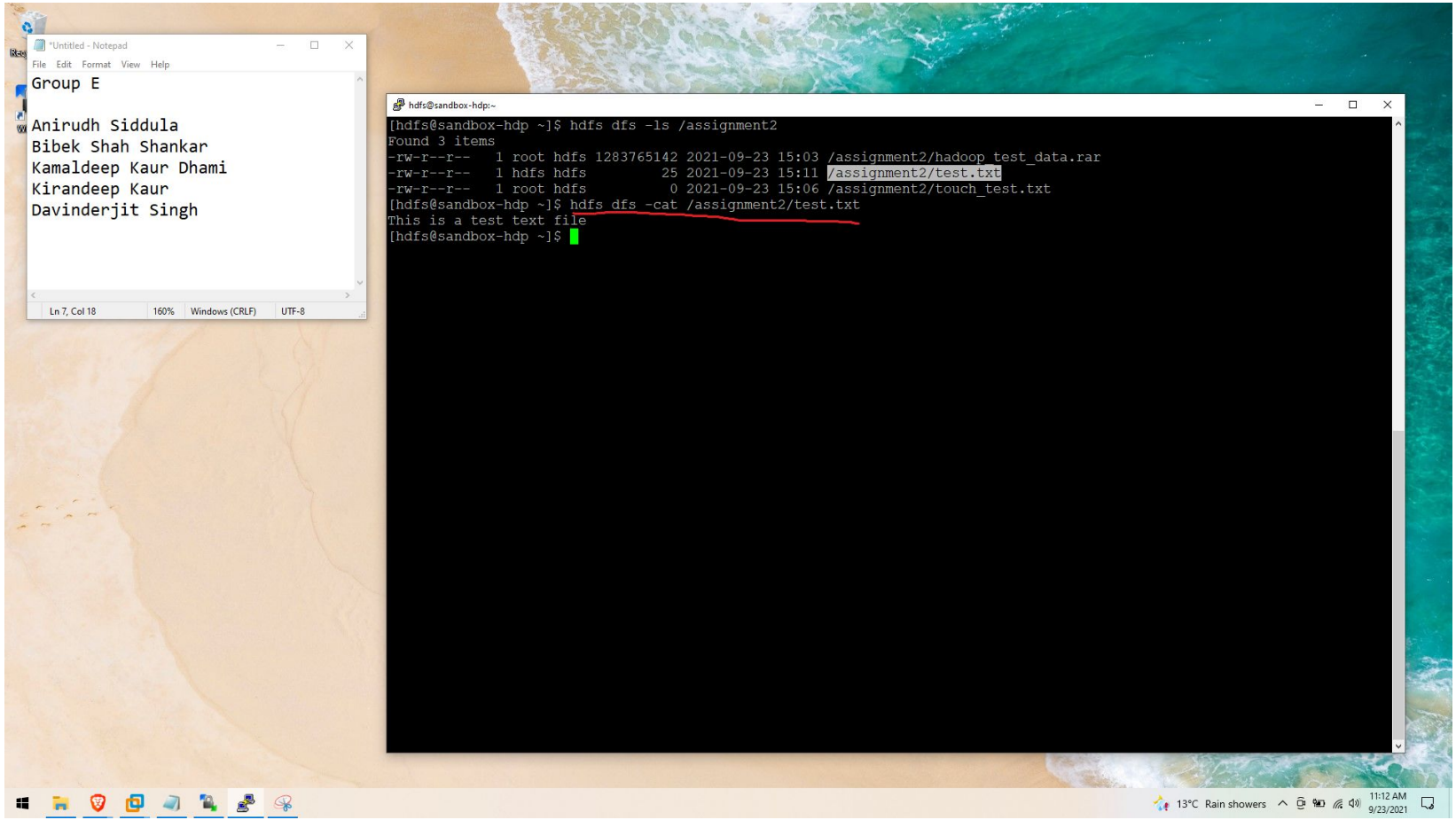
Used “hdfs dfs -copyFromLocal” command to copy file from VM’s local to hdfs’s assignment2 directory.



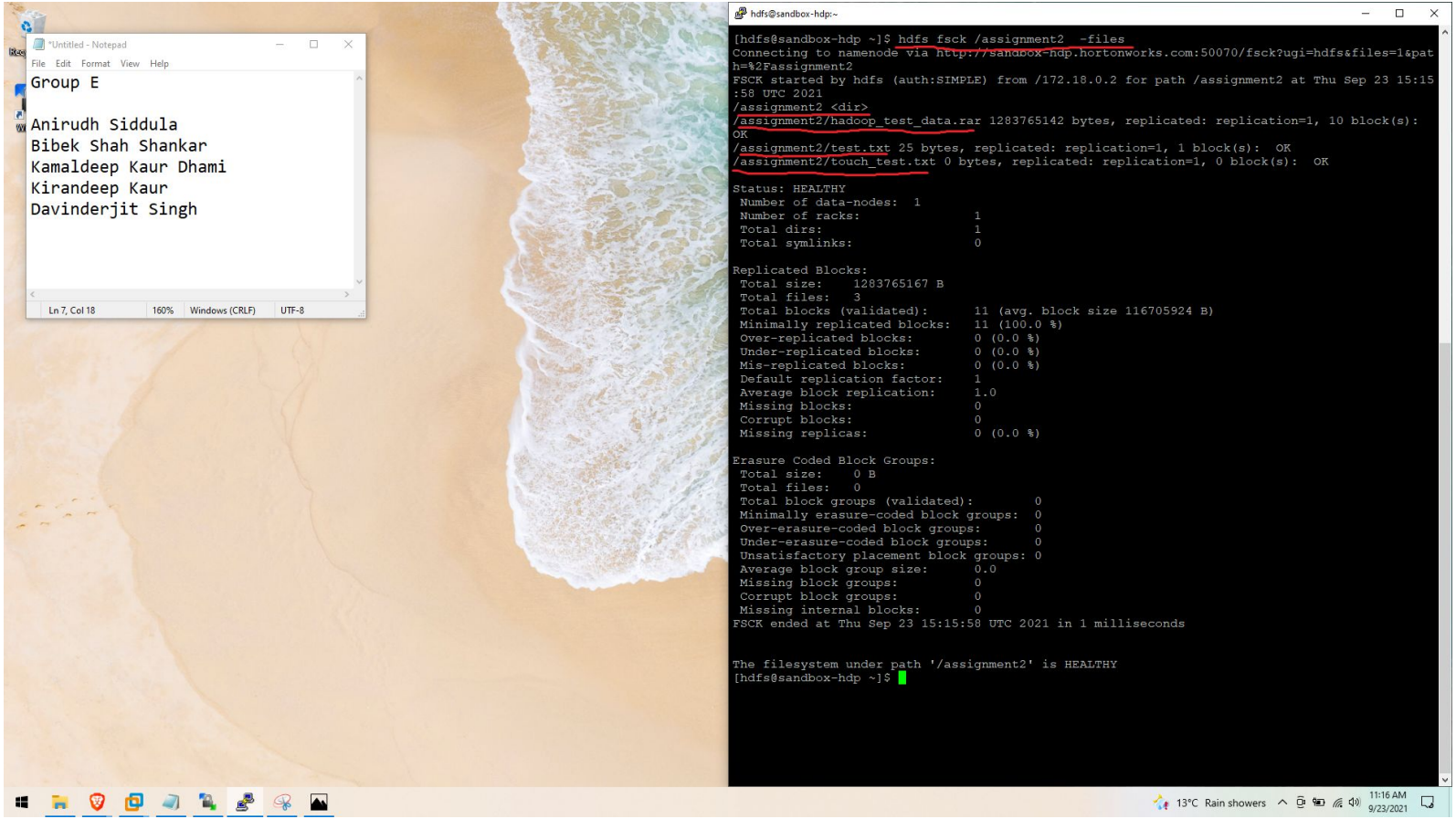
Used “hdfs dfs -touch” command to create a 0 byte file.



“hdfs dfs -cat” command to show the contents of the test file we have in our hdfs directory.



FSCK: We have run “hdfs fsck <path> -files” command to list the file present in our test directory(3 files as underlined in the screenshot).



Additional argument for -files in fsck command called “blocks” will show us the blocks the large file is divided into. Please refer the next slide for the detailed analysis.

*Untitled - Notepad

File Edit Format View Help

Group E

Anirudh Siddula
Bibek Shah Shankar
Kamaldeep Kaur Dhami
Kirandeep Kaur
Davinderjit Singh

Ln 7, Col 18 | 160% | Windows (CRLF) | UTF-8

hdfs@sandbox-hdp:/root

```
[hdfs@sandbox-hdp root]$ hdfs fsck /assignment2/ -files -blocks
Connecting to namenode via http://sandbox-hdp.hortonworks.com:50070/fsck?ugi=hdfs&files=1&blocks=1&path=%2Fassig
FSCK started by hdfs (auth:SIMPLE) from /172.18.0.2 for path /assignment2 at Thu Sep 23 15:36:50 UTC 2021
/assignment2 <dir>
/assignment2/hadoop_test_data.rar 1283765142 bytes, replicated: replication=1, 10 block(s):  OK
0. BP-1419118625-172.17.0.2-1543512323726:blk_1073748251_7450 len=134217728 Live_repl=1
1. BP-1419118625-172.17.0.2-1543512323726:blk_1073748252_7451 len=134217728 Live_repl=1
2. BP-1419118625-172.17.0.2-1543512323726:blk_1073748253_7452 len=134217728 Live_repl=1
3. BP-1419118625-172.17.0.2-1543512323726:blk_1073748254_7453 len=134217728 Live_repl=1
4. BP-1419118625-172.17.0.2-1543512323726:blk_1073748255_7454 len=134217728 Live_repl=1
5. BP-1419118625-172.17.0.2-1543512323726:blk_1073748256_7455 len=134217728 Live_repl=1
6. BP-1419118625-172.17.0.2-1543512323726:blk_1073748257_7456 len=134217728 Live_repl=1
7. BP-1419118625-172.17.0.2-1543512323726:blk_1073748258_7457 len=134217728 Live_repl=1
8. BP-1419118625-172.17.0.2-1543512323726:blk_1073748259_7458 len=134217728 Live_repl=1
9. BP-1419118625-172.17.0.2-1543512323726:blk_1073748260_7459 len=75805590 Live_repl=1

/assignment2/test.txt 25 bytes, replicated: replication=1, 1 block(s):  OK
0. BP-1419118625-172.17.0.2-1543512323726:blk_1073748263_7462 len=25 Live_repl=1

/assignment2/touch_test.txt 0 bytes, replicated: replication=1, 0 block(s):  OK


Status: HEALTHY
 Number of data-nodes:  1
 Number of racks:               1
 Total dirs:                    1
 Total symlinks:                0

Replicated Blocks:
 Total size:    1283765167 B
 Total files:   3
 Total blocks (validated):      11 (avg. block size 116705924 B)
 Minimally replicated blocks:   11 (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
 Missing blocks:                0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)

Erasure Coded Block Groups:
 Total size:    0 B
 Total files:   0
 Total block groups (validated):        0
 Minimally erasure-coded block groups:  0
 Over-erasure-coded block groups:       0
 Under-erasure-coded block groups:      0
 Unsatisfactory placement block groups: 0
 Average block group size:      0.0
 Missing block groups:          0
 Corrupt block groups:          0
 Missing internal blocks:       0
FSCK ended at Thu Sep 23 15:36:50 UTC 2021 in 10 milliseconds
```

13°C Rain showers 11:37 AM 9/23/2021

From the screenshot in previous slide we can see that our large file has been divided into 9 blocks.

We have confirmed each block size of our sandbox configuration from conf file as 128mb using the command

“hdfs getconf -confKey dfs.blocksize” its output being 134217728 bytes => 134217728/1024 = 131072KB = 128MB

We can see in our screenshot that our file has occupied 8 blocks and 9th block containing 75805590 bytes = 72MB

Also, we could see that for the zero byte file there is no blocks assigned.

The fsck command with “-locations” will show the location of each block in the assigned storage as underlined in the screenshot.
We have noted that there is no location for zero byte file due to no blocks being assigned for it.

*Untitled - Notepad

File Edit Format View Help

Group E

Anirudh Siddula
Bibek Shah Shankar
Kamaldeep Kaur Dhami
Kirandeep Kaur
Davinderjit Singh

Ln 7, Col 18 | 160% | Windows (CRLF) | UTF-8

hdfs@sandbox-hdp:/root

```
[hdfs@sandbox-hdp root]$ hdfs fsck /assignment2/ -files -blocks  -locations
Connecting to namenode via http://sandbox-hdp.hortonworks.com:50070/fsck?ugi=hdfs&files=1&blocks=1&locations=1&path=%2Fassignment2
FSCK started by hdfs (auth:SIMPLE) from /172.18.0.2 for path /assignment2 at Thu Sep 23 18:37:55 UTC 2021
/assignment2 <dir>
/assignment2/hadoop_test_data.rar 1283765142 bytes, replicated: replication=1, 10 block(s):  OK
0. BP-1419118625-172.17.0.2-1543512323726:blk_1073748251_7450 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[172.18.0.2:50010,DS-6c34
ba72-0587-4927-88a1-781ba7d444d9,DISK]]
1. BP-1419118625-172.17.0.2-1543512323726:blk_1073748252_7451 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[172.18.0.2:50010,DS-6c34
ba72-0587-4927-88a1-781ba7d444d9,DISK]]
2. BP-1419118625-172.17.0.2-1543512323726:blk_1073748253_7452 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[172.18.0.2:50010,DS-6c34
ba72-0587-4927-88a1-781ba7d444d9,DISK]]
3. BP-1419118625-172.17.0.2-1543512323726:blk_1073748254_7453 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[172.18.0.2:50010,DS-6c34
ba72-0587-4927-88a1-781ba7d444d9,DISK]]
4. BP-1419118625-172.17.0.2-1543512323726:blk_1073748255_7454 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[172.18.0.2:50010,DS-6c34
ba72-0587-4927-88a1-781ba7d444d9,DISK]]
5. BP-1419118625-172.17.0.2-1543512323726:blk_1073748256_7455 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[172.18.0.2:50010,DS-6c34
ba72-0587-4927-88a1-781ba7d444d9,DISK]]
6. BP-1419118625-172.17.0.2-1543512323726:blk_1073748257_7456 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[172.18.0.2:50010,DS-6c34
ba72-0587-4927-88a1-781ba7d444d9,DISK]]
7. BP-1419118625-172.17.0.2-1543512323726:blk_1073748258_7457 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[172.18.0.2:50010,DS-6c34
ba72-0587-4927-88a1-781ba7d444d9,DISK]]
8. BP-1419118625-172.17.0.2-1543512323726:blk_1073748259_7458 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[172.18.0.2:50010,DS-6c34
ba72-0587-4927-88a1-781ba7d444d9,DISK]]
9. BP-1419118625-172.17.0.2-1543512323726:blk_1073748260_7459 len=75805590 Live_repl=1  [DatanodeInfoWithStorage[172.18.0.2:50010,DS-6c34b
a72-0587-4927-88a1-781ba7d444d9,DISK]]

/assignment2/test.txt 25 bytes, replicated: replication=1, 1 block(s):  OK
0. BP-1419118625-172.17.0.2-1543512323726:blk_1073748263_7462 len=25 Live_repl=1  [DatanodeInfoWithStorage[172.18.0.2:50010,DS-6c34ba72-05
87-4927-88a1-781ba7d444d9,DISK]]

/assignment2/touch_test.txt 0 bytes, replicated: replication=1, 0 block(s):  OK


Status: HEALTHY
 Number of data-nodes:  1
 Number of racks:               1
 Total dirs:                    1
 Total symlinks:                0

Replicated Blocks:
 Total size:    1283765167 B
 Total files:   3
 Total blocks (validated):      11 (avg. block size 116705924 B)
 Minimally replicated blocks:   11 (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
 Missing blocks:                0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)

Erasure Coded Block Groups:
 Total size:    0 B
 Total files:   0
```

13°C Rain showers 2:38 PM 9/23/2021

The fsck command with additional “racks” argument will show the racks of each block. The rack number is same for all files as we are having our test environment in an single node configuration.

*Untitled - Notepad

File Edit Format View Help

Group E

Anirudh Siddula
Bibek Shah Shankar
Kamaldeep Kaur Dhami
Kirandeep Kaur
Davinderjit Singh

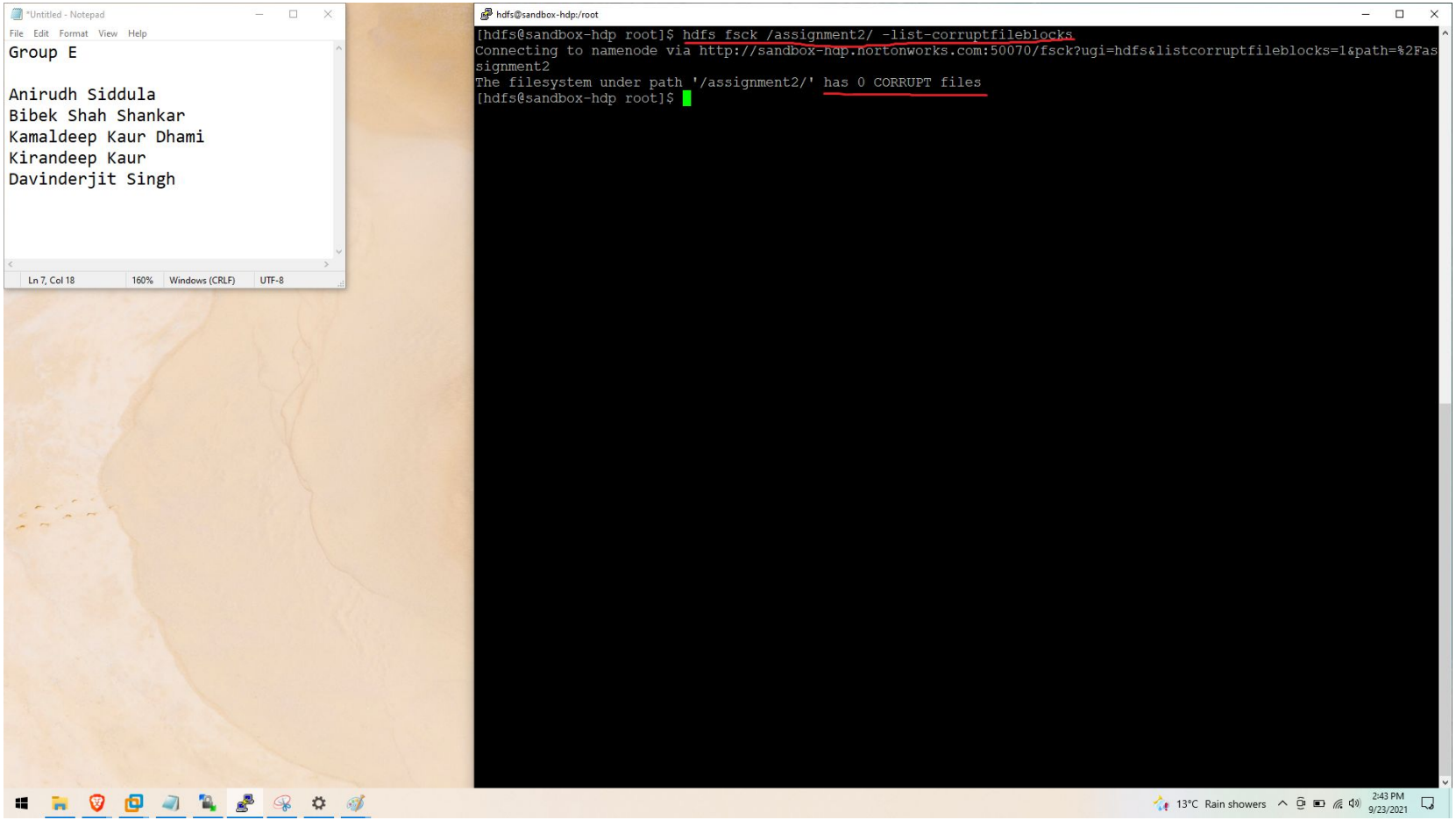
Ln 7, Col 18 | 160% | Windows (CRLF) | UTF-8

hdfs@sandbox-hdp:/root

```
[hdfs@sandbox-hdp root]$ hdfs fsck /assignment2/ -files -blocks -racks
Connecting to namenode via http://sandbox-hdp.hortonworks.com:50070/fsck?ugi=hdfs&files=1&blocks=1&racks=1&path=%2Fassignment
FSCK started by hdfs (auth:SIMPLE) from /172.18.0.2 for path /assignment2 at Thu Sep 23 18:21:39 UTC 2021
/assignment2 <dir>
/assignment2/hadoop_test_data.rar 1283765142 bytes, replicated: replication=1, 10 block(s):  OK
0. BP-1419118625-172.17.0.2-1543512323726:blk_1073748251_7450 len=134217728 Live_repl=1  [/default-rack/172.18.0.2:50010]
1. BP-1419118625-172.17.0.2-1543512323726:blk_1073748252_7451 len=134217728 Live_repl=1  [/default-rack/172.18.0.2:50010]
2. BP-1419118625-172.17.0.2-1543512323726:blk_1073748253_7452 len=134217728 Live_repl=1  [/default-rack/172.18.0.2:50010]
3. BP-1419118625-172.17.0.2-1543512323726:blk_1073748254_7453 len=134217728 Live_repl=1  [/default-rack/172.18.0.2:50010]
4. BP-1419118625-172.17.0.2-1543512323726:blk_1073748255_7454 len=134217728 Live_repl=1  [/default-rack/172.18.0.2:50010]
5. BP-1419118625-172.17.0.2-1543512323726:blk_1073748256_7455 len=134217728 Live_repl=1  [/default-rack/172.18.0.2:50010]
6. BP-1419118625-172.17.0.2-1543512323726:blk_1073748257_7456 len=134217728 Live_repl=1  [/default-rack/172.18.0.2:50010]
7. BP-1419118625-172.17.0.2-1543512323726:blk_1073748258_7457 len=134217728 Live_repl=1  [/default-rack/172.18.0.2:50010]
8. BP-1419118625-172.17.0.2-1543512323726:blk_1073748259_7458 len=134217728 Live_repl=1  [/default-rack/172.18.0.2:50010]
9. BP-1419118625-172.17.0.2-1543512323726:blk_1073748260_7459 len=75805590 Live_repl=1  [/default-rack/172.18.0.2:50010]

/assignment2/test.txt 25 bytes, replicated: replication=1, 1 block(s):  OK
0. BP-1419118625-172.17.0.2-1543512323726:blk_1073748263_7462 len=25 Live_repl=1  [/default-rack/172.18.0.2:50010]

/assignment2/touch_test.txt 0 bytes, replicated: replication=1, 0 block(s):  OK


Status: HEALTHY
 Number of data-nodes:  1
 Number of racks:               1
 Total dirs:                    1
 Total symlinks:                0

Replicated Blocks:
 Total size:    1283765167 B
 Total files:   3
 Total blocks (validated):      11 (avg. block size 116705924 B)
 Minimally replicated blocks:   11 (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
 Missing blocks:                0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)

Erasure Coded Block Groups:
 Total size:    0 B
 Total files:   0
 Total block groups (validated):        0
 Minimally erasure-coded block groups:  0
 Over-erasure-coded block groups:       0
 Under-erasure-coded block groups:      0
 Unsatisfactory placement block groups: 0
 Average block group size:      0.0
 Missing block groups:          0
 Corrupt block groups:          0
 Missing internal blocks:       0
FSCK ended at Thu Sep 23 18:21:39 UTC 2021 in 1 milliseconds
```

13°C Rain showers 2:23 PM 9/23/2021

The fsck -list-corruptfileblocks will get us all the blocks that have been corrupted and we having our sandbox as a fresh Install do not have any corrupt files in our hdfs.



References:

https://hadoop.apache.org/docs/r2.6.0/hadoop-project-dist/hadoop-hdfs/HDFSCommands.html

Thank you