

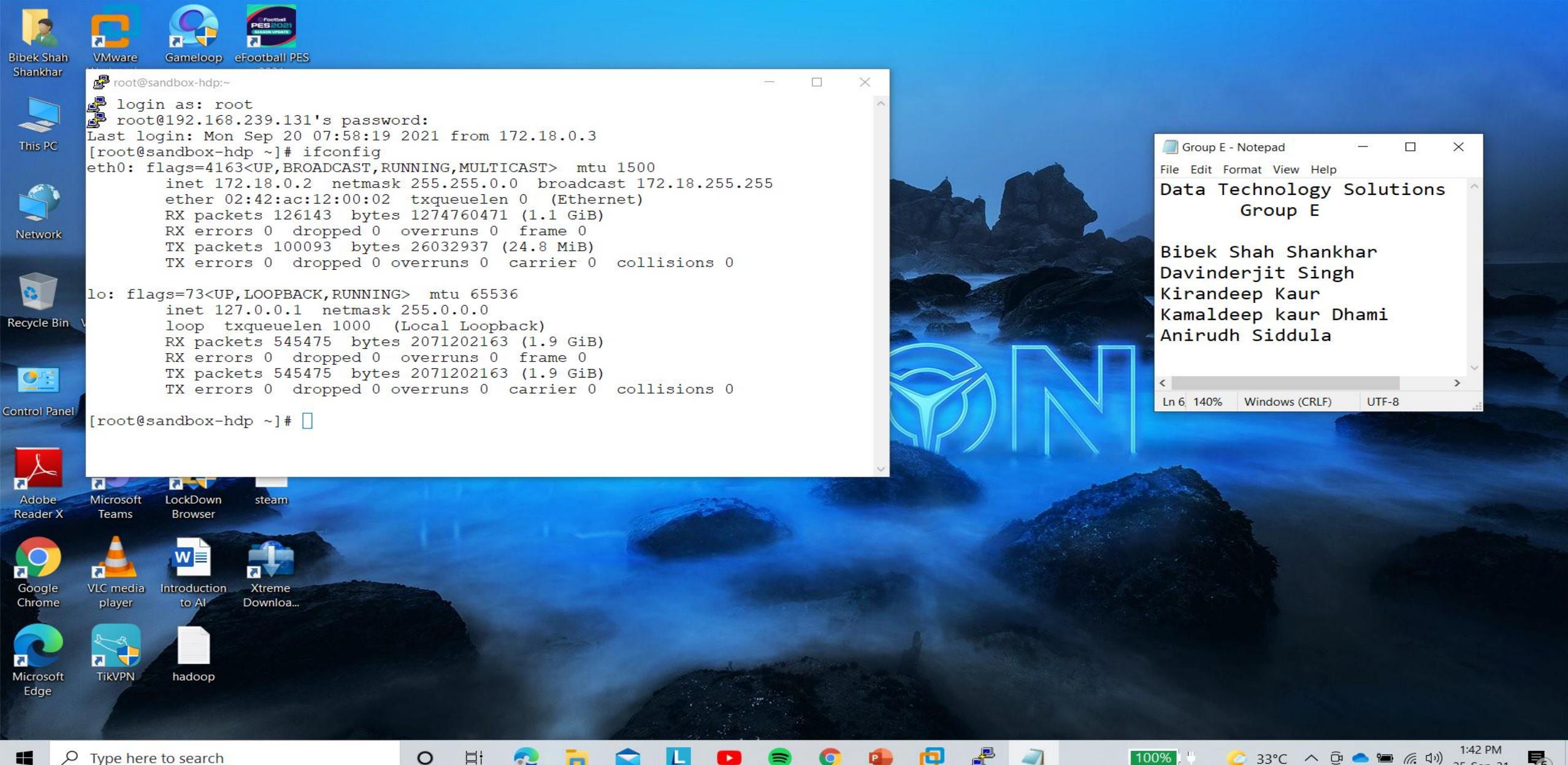
Group E

BDM 1024 - Data Technology Solutions

Assignment: Lab 1d -MapReduce

Submitted by:

1. Bibek Shah Shankhar
2. Anirudh Siddula
3. Kirandeep Kaur
4. Kamaldeep kaur Dhami
5. Davinderjit Singh



Login To Putty and Checking Ip address.

Ambari Sandbox 0 ops 14 alerts

Dashboard Services Hosts Alerts Admin raj_ops

Metrics Heatmaps Config History

Metric Actions Last 1 hour

HDFS Disk Usage: 28% (1/1)

DataNodes Live: 1/1

HDFS Links: NameNode, Secondary NameNode, 1 DataNodes

Memory Usage: No Data Available

Network Usage: No Data Available

CPU Usage: No Data Available

Cluster Load: No Data Available

NameNode Heap: 28%

NameNode RPC: 0.09 ms

NameNode CPU WIO: n/a

NameNode Uptime: 5.0 d

HBase Master Heap: n/a

HBase Links: No Active Master, 1 RegionServers, n/a

HBase Ave Load: n/a

HBase Master Uptime: n/a

ResourceManager Heap: 17%

ResourceManager Uptime: 5.0 d

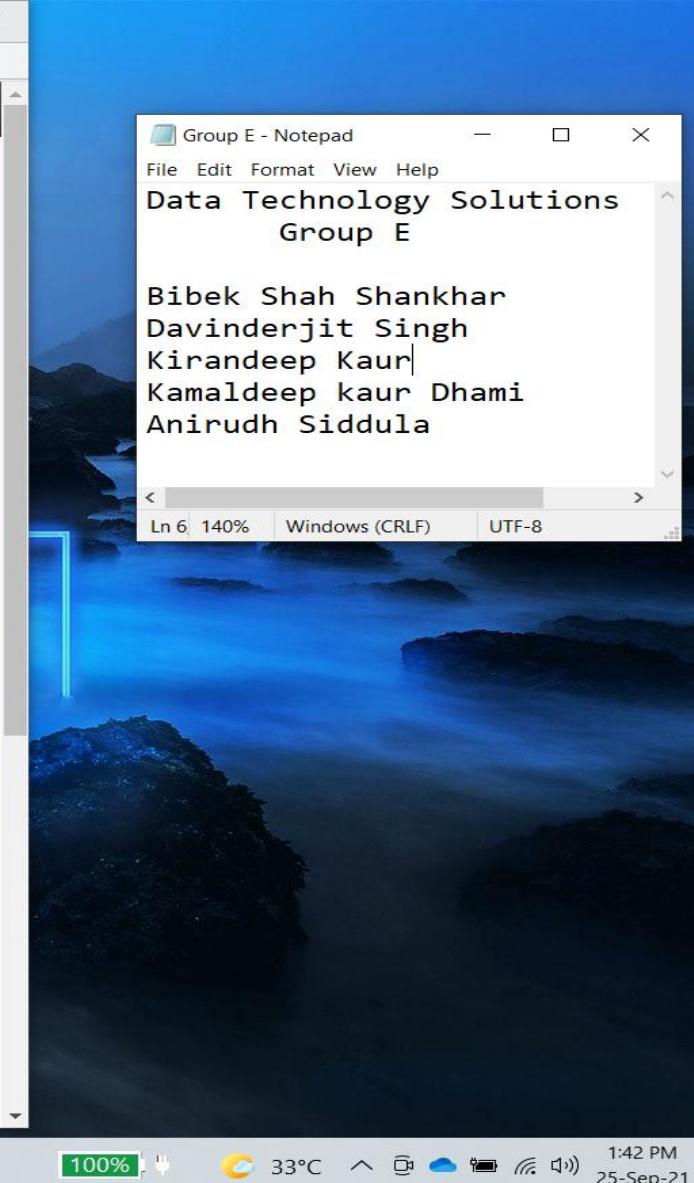
YARN Memory: 0%

NodeManagers Live: 1/1

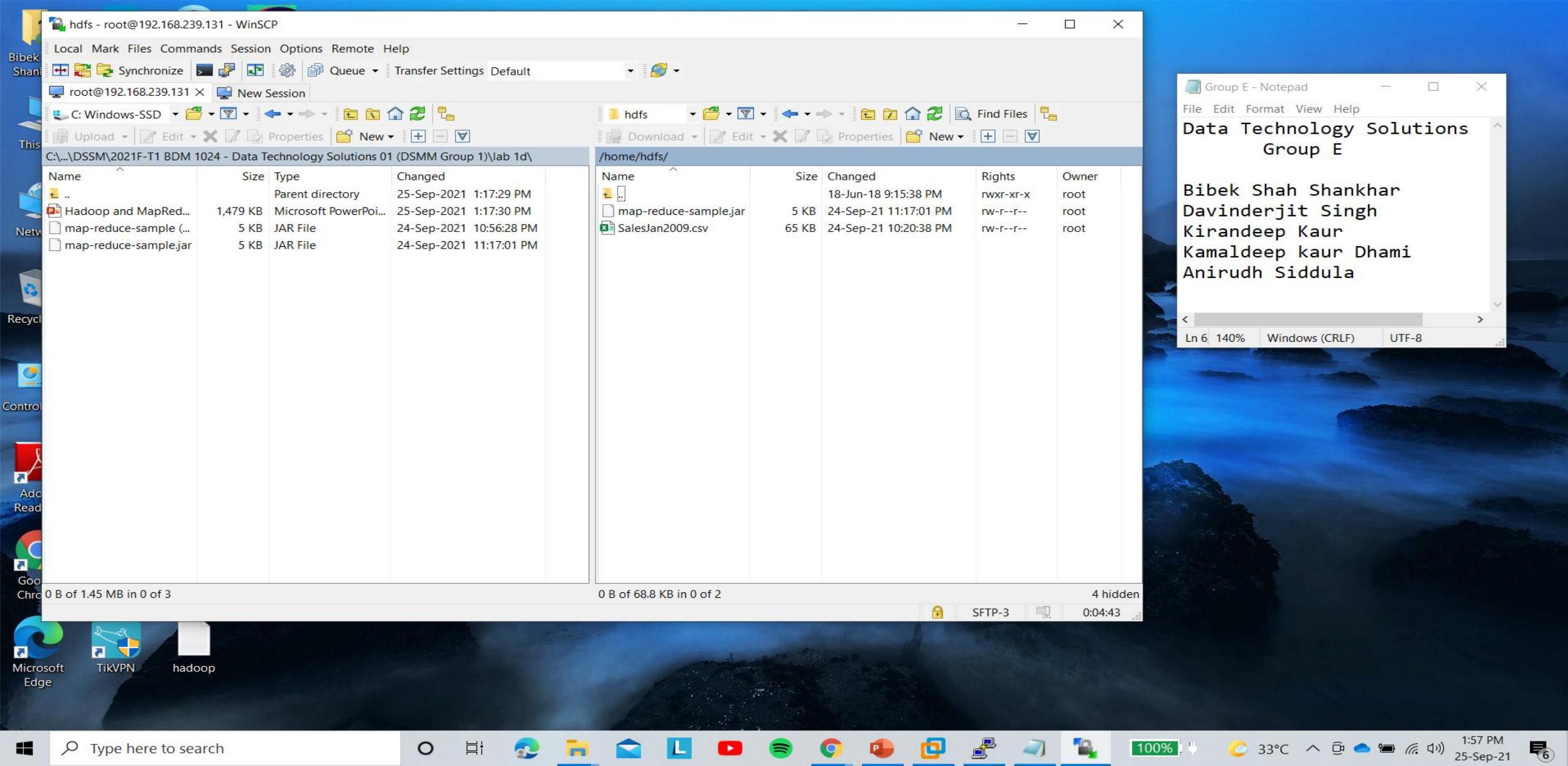
YARN Links: ResourceManager, 1 NodeManagers

Type here to search

100% 33°C 1:42 PM 25-Sep-21



Stopping Unused Services From Dashboard.



Transferring Local File From System To Hdfs Directory Through WinSCP.

The screenshot shows a Windows desktop environment with several open windows:

- WinSCP Session Window:** Title bar: "hdfs - root@192.168.239.131 - WinSCP". Menus: Local, Mark, Files, Commands, Session, Options, Remote, Help. Buttons: Synchronize, Queue, Transfer Settings, Default. Session details: root@192.168.239.131 X New Session. Left sidebar: This PC, Network, Recycle Bin, Control Panel, Adobe Reader X, Google Chrome, Microsoft Edge, TikVPN, hadoop. Main area: Two file lists. Left list: C:\...\DSSM\2021F-T1 BDM 1024 - Data Technology Solutions 01 (DSMM Group 1)\lab 1d\, showing files: Hadoop and MapRed..., map-reduce-sample..., map-reduce-samplejar. Right list: /home/hdfs/, showing files: .., map-reduce-sample.jar, SalesJan2009.csv.
- File Explorer Window:** Title bar: "Group E - Notepad". Menus: File, Edit, Format, View, Help. Content: "Data Technology Solutions Group E" followed by a list of names: Bibek Shah, Shankhar, Davinderjit Singh, Kirandeep Kaur, Kamaldeep kaur Dhami, Anirudh Siddula.
- Terminal Window:** Title bar: "root@sandbox-hdp:/home/hdfs". Command history:

```
[root@sandbox-hdp hdfs]# cd /home/hdfs
[root@sandbox-hdp hdfs]# ls
map-reduce-sample.jar  SalesJan2009.csv
[root@sandbox-hdp hdfs]#
```
- Taskbar:** Shows the Start button, a search bar, and icons for various system and application tasks.

Checking Transferred File from Putty
Commands.

The image shows a Windows desktop environment with a blue-themed wallpaper featuring a mountain landscape. On the left side, there is a vertical column of pinned icons for various applications: Bibek Shah Shankhar, VMware Workstation..., Gameloop, eFootball PES 2021, This PC, WinSCP, Network, WhatsApp, Recycle Bin, Visual Studio Code, Control Panel, Group E, Adobe Reader X, Microsoft Teams, Google Chrome, VLC media player, Microsoft Edge, TikVPN, and hadoop.

A terminal window titled "root@sandbox-hdp:" is open, displaying the following command-line session:

```
[root@sandbox-hdp hdfs]# cd /home/hdfs  
[root@sandbox-hdp hdfs]# ls  
map-reduce-sample.jar SalesJan2009.csv  
[root@sandbox-hdp hdfs]# cd /  
[root@sandbox-hdp /]# ls  
anaconda-post.log hadoop mysql-connector-java-5.1.45 sandbox-flavour  
apps home opt sbin  
bin kafka-logs packer-files srv  
boot lib proc sys  
cgroups_test lib64 root tmp  
dev media run usr  
etc mnt sandbox var  
[root@sandbox-hdp /]#
```

To the right of the terminal, a Notepad window titled "Group E - Notepad" is open, containing the following text:

Data Technology Solutions
Group E

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

The Notepad window also shows the status bar with "Ln 6, 140% Windows (CRLF) UTF-8".

The taskbar at the bottom of the screen includes the Start button, a search bar with the placeholder "Type here to search", and several pinned icons: File Explorer, Task View, Mail, LinkedIn, YouTube, Spotify, Google Chrome, Microsoft Edge, TikTok, and a lock icon. The system tray shows the battery level at 100%, the temperature at 33°C, and the date and time as 1:59 PM, 25-Sep-21.

Listing Files in The Directories.

Bibek Shah
Shankhar

VMware Workstation...
Gameloop eFootball PES 2021

This PC WinSCP

Network WhatsApp

Recycle Bin Visual Studio 2022 Code

Control Panel Group E

Adobe Reader X Microsoft Teams

Google Chrome VLC media player

Microsoft Edge TikVPN hadoop

Football PES 2021

hdfs@sandbox-hdp:~

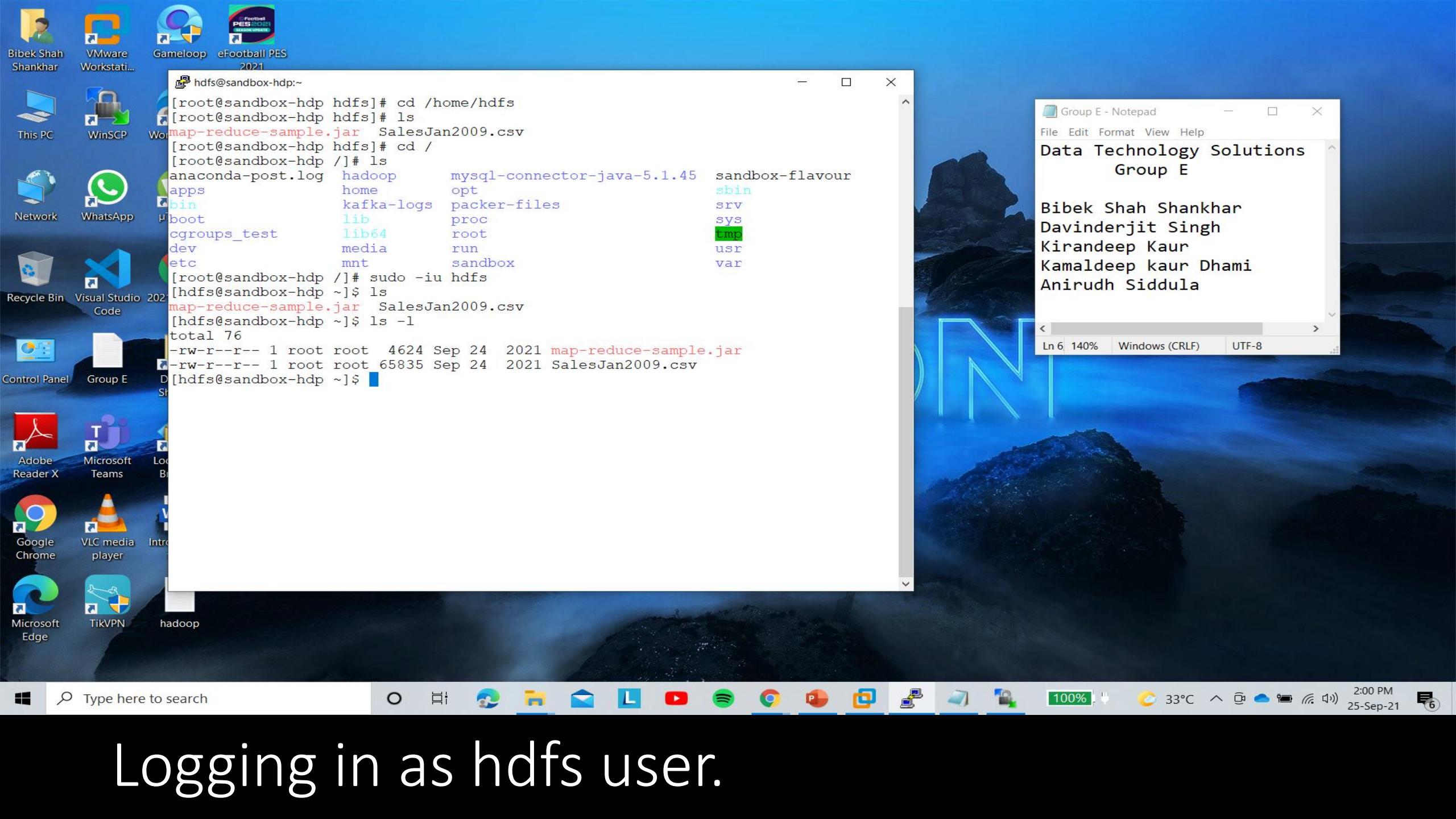
```
[root@sandbox-hdp hdfs]# cd /home/hdfs
[root@sandbox-hdp hdfs]# ls
map-reduce-sample.jar  SalesJan2009.csv
[root@sandbox-hdp hdfs]# cd /
[root@sandbox-hdp /]# ls
anaconda-post.log  hadoop  mysql-connector-java-5.1.45  sandbox-flavour
apps  home  opt
boot  kafka-logs  packer-files
cgroups_test  lib  proc
dev  lib64  root
etc  media  run
etc  mnt  sandbox
[root@sandbox-hdp /]# sudo -iu hdfs
[hdfs@sandbox-hdp ~]$ ls
map-reduce-sample.jar  SalesJan2009.csv
[hdfs@sandbox-hdp ~]$ ls -l
total 76
-rw-r--r-- 1 root root 4624 Sep 24 2021 map-reduce-sample.jar
-rw-r--r-- 1 root root 65835 Sep 24 2021 SalesJan2009.csv
[hdfs@sandbox-hdp ~]$
```

Group E - Notepad

Data Technology Solutions Group E

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

Ln 6 140% Windows (CRLF) UTF-8



Type here to search

2:00 PM 25-Sep-21

Logging in as hdfs user.

Bibek Shah
Shankhar

VMware Workstation

Game

Football PES 2021

This PC

WinSCP

Work -

Network

WhatsApp

uTorrent

Recycle Bin

Visual Studio 2021F-
Code

Control Panel

Group E

DSShort

Adobe Reader X

Microsoft Teams

LockD Brow

Google Chrome

VLC media player

Introdu to A

Microsoft Edge

TikVPN

hadoop

hdfs@sandbox-hdp:~

```
[root@sandbox-hdp hdfs]# cd /home/hdfs
[root@sandbox-hdp hdfs]# ls
map-reduce-sample.jar  SalesJan2009.csv
[root@sandbox-hdp hdfs]# cd /
[root@sandbox-hdp /]# ls
anaconda-post.log      hadoop      mysql-connector-java-5.1.45  sandbox-flavour
apps                  home       opt                           sbin
bin                   kafka-logs  packer-files
boot                  lib        proc
cgrouptest           lib64      root
dev                   media     run
etc                   mnt       sandbox
[root@sandbox-hdp /]# sudo -iu hdfs
[hdfs@sandbox-hdp ~]$ ls
map-reduce-sample.jar  SalesJan2009.csv
[hdfs@sandbox-hdp ~]$ ls -l
total 76
-rw-r--r-- 1 root root 4624 Sep 24 2021 map-reduce-sample.jar
-rw-r--r-- 1 root root 65835 Sep 24 2021 SalesJan2009.csv
[hdfs@sandbox-hdp ~]$ hdfs dfs -copyFromLocal SalesJan2009.csv /TestData
[hdfs@sandbox-hdp ~]$ hdfs dfs -ls /TestData/SalesJan2009.csv
-rw-r--r-- 1 hdfs hdfs 65835 2021-09-20 09:12 /TestData/SalesJan2009.csv
[hdfs@sandbox-hdp ~]$
```

Group E - Notepad

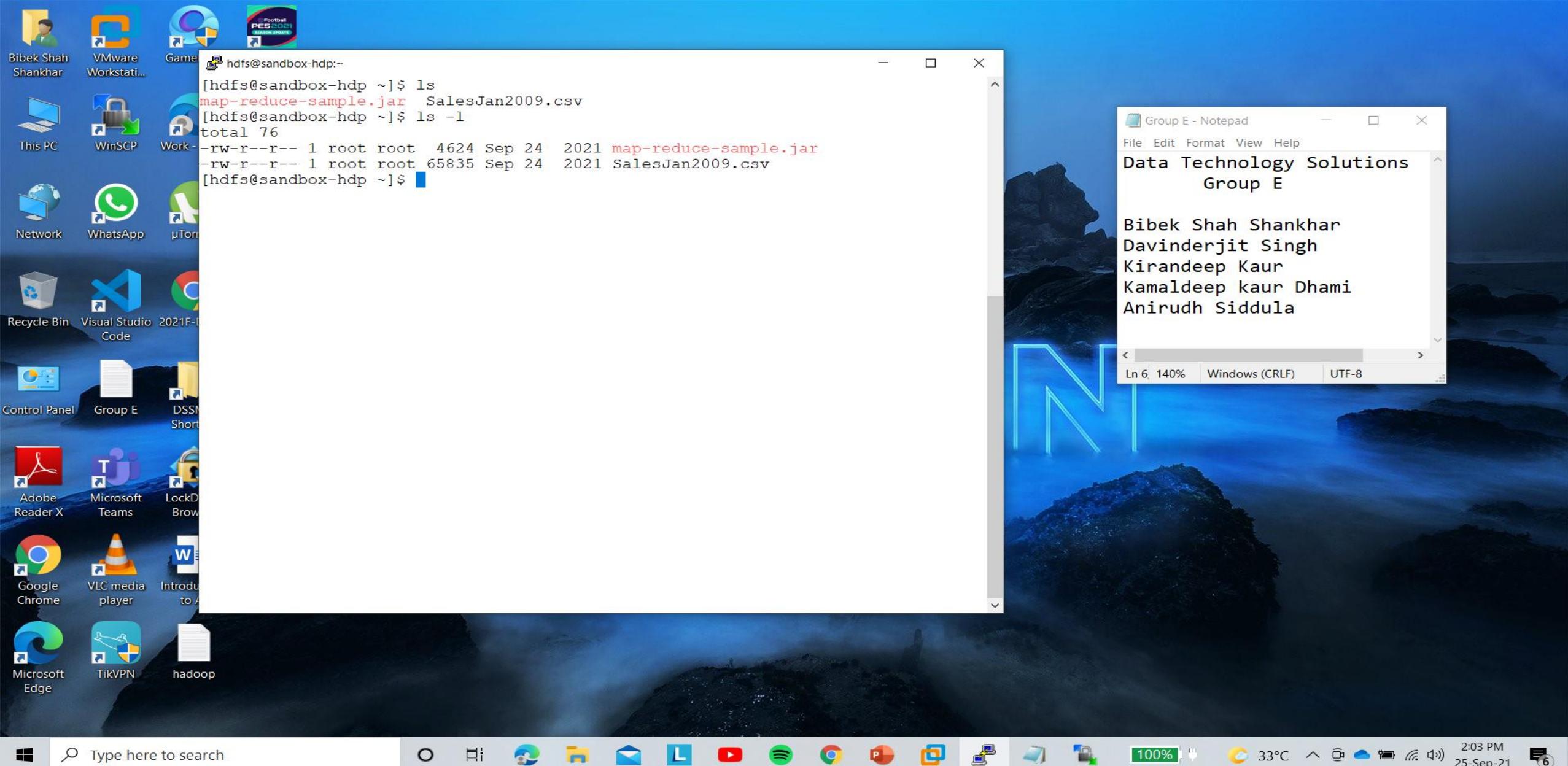
File Edit Format View Help

Data Technology Solutions
Group E

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

Ln 6 140% Windows (CRLF) UTF-8

Copying Files from Local System to HDFS and Listing.



Listing the Copied Files.

Bibek Shah
Shankhar

VMware Workstation

Game

Football PES 2021

This PC

WinSCP

Work

Network

WhatsApp

muTorr

Recycle Bin

Visual Studio 2021 F-Code

Control Panel

Group E

DSSM Short

Adobe Reader X

Microsoft Teams

LockD Brow

Google Chrome

VLC media player

Introdu to A

Microsoft Edge

TikVPN

hadoop

hdfs@sandbox-hdp:~

```
[hdfs@sandbox-hdp ~]$ ls
map-reduce-sample.jar  SalesJan2009.csv
[hdfs@sandbox-hdp ~]$ ls -l
total 76
-rw-r--r-- 1 root root 4624 Sep 24 2021 map-reduce-sample.jar
-rw-r--r-- 1 root root 65835 Sep 24 2021 SalesJan2009.csv
[hdfs@sandbox-hdp ~]$ hdfs dfs -ls /
Found 13 items
drwxr-xr-x  - root  hdfs          0 2021-09-20 09:12 /TestData
drwxr-xr-x  - root  hdfs          0 2021-09-20 08:13 /Testfile
drwxrwxrwx  - yarn  hadoop        0 2018-06-18 15:18 /app-logs
drwxr-xr-x  - hdfs  hdfs          0 2018-06-18 16:13 /apps
drwxr-xr-x  - yarn  hadoop        0 2018-06-18 14:52 /ats
drwxr-xr-x  - hdfs  hdfs          0 2018-06-18 14:52 /hdp
drwxr-----  - livy  hdfs          0 2018-06-18 15:11 /livy2-recovery
drwxr-xr-x  - mapred hdfs        0 2018-06-18 14:52 /mapred
drwxrwxrwx  - mapred hadoop      0 2018-06-18 14:52 /mr-history
drwxr-xr-x  - hdfs  hdfs          0 2018-06-18 15:59 /ranger
drwxrwxrwx  - spark  hadoop      0 2021-09-20 08:43 /spark2-history
drwxrwxrwx  - hdfs  hdfs          0 2018-06-18 16:06 /tmp
drwxr-xr-x  - hdfs  hdfs          0 2018-06-18 16:08 /user
[hdfs@sandbox-hdp ~]$ hdfs dfs -ls /TestData
Found 2 items
-rw-r--r--  1 hdfs  hdfs   65835 2021-09-20 09:12 /TestData/SalesJan2009.csv
-rw-r--r--  1 root  hdfs 1261316392 2021-09-20 07:34 /TestData/Test.rar
[hdfs@sandbox-hdp ~]$
```

Group E - Notepad

File Edit Format View Help

Data Technology Solutions
Group E

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

Ln 6, 140% Windows (CRLF) UTF-8

Type here to search

2:05 PM 25-Sep-21

100% 33°C

Listing files from Hdfs and From TestData Directory Which includes our Project Files.

Bibek Shah
Shankhar

VMware Workstation

This PC

WinSCP

Work -

Network

WhatsApp

µTorrent

Recycle Bin

Visual Studio 2021 F-
Code

Control Panel

Group E

DSSM

Adobe Reader X

Microsoft Teams

LockDown Browser

Google Chrome

VLC media player

Microsoft Edge

TikVPN

hadoop

FootBall

hdfs@sandbox-hdp:~

```
[hdfs@sandbox-hdp ~]$ ls
map-reduce-sample.jar  SalesJan2009.csv
[hdfs@sandbox-hdp ~]$ ls -l
total 76
-rw-r--r-- 1 root root 4624 Sep 24 2021 map-reduce-sample.jar
-rw-r--r-- 1 root root 65835 Sep 24 2021 SalesJan2009.csv
[hdfs@sandbox-hdp ~]$ hdfs dfs -ls /
Found 13 items
drwxr-xr-x  - root  hdfs          0 2021-09-20 09:12 /TestData
drwxr-xr-x  - root  hdfs          0 2021-09-20 08:13 /Testfile
drwxrwxrwx  - yarn  hadoop        0 2018-06-18 15:18 /app-logs
drwxr-xr-x  - yarn  hadoop        0 2018-06-18 16:13 /apps
drwxr-xr-x  - hdfs  hdfs          0 2018-06-18 14:52 /ats
drwxr-xr-x  - hdfs  hdfs          0 2018-06-18 14:52 /hdp
drwx-----  - livy  hdfs          0 2018-06-18 15:11 /livy2-recovery
drwxr-xr-x  - mapred hdfs        0 2018-06-18 14:52 /mapred
drwxrwxrwx  - mapred hadoop      0 2018-06-18 14:52 /mr-history
drwxr-xr-x  - hdfs  hdfs          0 2018-06-18 15:59 /ranger
drwxrwxrwx  - spark  hadoop      0 2021-09-20 08:43 /spark2-history
drwxrwxrwx  - hdfs  hdfs          0 2018-06-18 16:06 /tmp
drwxr-xr-x  - hdfs  hdfs          0 2018-06-18 16:08 /user
[hdfs@sandbox-hdp ~]$ hdfs dfs -ls /TestData
Found 2 items
-rw-r--r--  1 hdfs hdfs  65835 2021-09-20 09:12 /TestData/SalesJan2009.csv
-rw-r--r--  1 root hdfs 1261316392 2021-09-20 07:34 /TestData/Test.rar
[hdfs@sandbox-hdp ~]$ clear
[hdfs@sandbox-hdp ~]$ hadoop jar map-reduce-sample.jar /TestData/SalesJan2009.cs
v /TestData/mp_output
21/09/20 09:17:55 INFO client.RMProxy: Connecting to ResourceManager at sandbox-
hdp.hortonworks.com/172.18.0.2:8032
21/09/20 09:17:55 INFO client.AHSProxy: Connecting to Application History server
at sandbox-hdp.hortonworks.com/172.18.0.2:10200
21/09/20 09:17:55 INFO client.RMProxy: Connecting to ResourceManager at sandbox-
hdp.hortonworks.com/172.18.0.2:8032
21/09/20 09:17:55 INFO client.AHSProxy: Connecting to Application History server
at sandbox-hdp.hortonworks.com/172.18.0.2:10200
21/09/20 09:17:55 WARN mapreduce.JobResourceUploader: Hadoop command-line option
parsing not performed. Implement the Tool interface and execute your applicatio
```

Group E - Notepad

File Edit Format View Help

Data Technology Solutions
Group E

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

Ln 6 140% Windows (CRLF) UTF-8

We run the Jar file Which Contains The mapreduce code using Hadoop to execute SalesJan2009.csv and Trying to store Results in TestData/mp_output Directory.

```
[hdfs@sandbox-hdp ~]$ hadoop jar map-reduce-sample.jar /TestData/SalesJan2009.cs  
v /TestData/mp_output  
21/09/20 09:17:55 INFO client.RMProxy: Connecting to ResourceManager at sandbox-  
hdp.hortonworks.com/172.18.0.2:8032  
21/09/20 09:17:55 INFO client.AHSProxy: Connecting to Application History server  
at sandbox-hdp.hortonworks.com/172.18.0.2:10200  
21/09/20 09:17:55 INFO client.RMProxy: Connecting to ResourceManager at sandbox-  
hdp.hortonworks.com/172.18.0.2:8032  
21/09/20 09:17:55 INFO client.AHSProxy: Connecting to Application History server  
at sandbox-hdp.hortonworks.com/172.18.0.2:10200  
21/09/20 09:17:55 WARN mapreduce.JobResourceUploader: Hadoop command-line option  
parsing not performed. Implement the Tool interface and execute your applicatio  
n with ToolRunner to remedy this.  
21/09/20 09:17:56 INFO mapred.FileInputFormat: Total input paths to process : 1  
21/09/20 09:17:56 INFO mapreduce.JobSubmitter: number of splits:2  
21/09/20 09:17:56 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_16  
32122556196_0001  
21/09/20 09:17:57 INFO impl.YarnClientImpl: Submitted application application_16  
32122556196_0001  
21/09/20 09:17:57 INFO mapreduce.Job: The url to track the job: http://sandbox-h  
dp.hortonworks.com:8088/proxy/application_1632122556196_0001/  
21/09/20 09:17:57 INFO mapreduce.Job: Running job: job_1632122556196_0001  
  
21/09/20 09:20:24 INFO mapreduce.Job: Job job_1632122556196_0001 running in uber mode : fa  
21/09/20 09:20:24 INFO mapreduce.Job: map 0% reduce 0%  
21/09/20 09:20:31 INFO mapreduce.Job: map 100% reduce 0%  
21/09/20 09:20:36 INFO mapreduce.Job: map 100% reduce 100%  
21/09/20 09:20:37 INFO mapreduce.Job: Job job_1632122556196_0001 completed successfully  
21/09/20 09:20:37 INFO mapreduce.Job: Counters: 49  
File System Counters  
FILE: Number of bytes read=12824  
FILE: Number of bytes written=485640  
FILE: Number of read operations=0  
FILE: Number of large read operations=0  
FILE: Number of write operations=0  
HDFS: Number of bytes read=98987  
HDFS: Number of bytes written=43  
HDFS: Number of read operations=9
```

Group E - Notepad
File Edit Format View Help
Data Technology Solutions
Group E

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

< Ln 6, 140% Windows (CRLF) UTF-8 >

Type here to search 2:10 PM
25-Sep-21

Our map-reduce Code is running and executed Sucessfully
100%.

```
hdfs@sandbox-hdp:~  
21/09/20 09:17:57 INFO mapreduce.Job: The url to track the job: http://sandbox-h  
dp.hortonworks.com:8088/proxy/application_1632122556196_0001/  
21/09/20 09:17:57 INFO mapreduce.Job: Running job: job_1632122556196_0001  
  
21/09/20 09:20:24 INFO mapreduce.Job: Job job_1632122556196_0001 running in uber mode : fa  
21/09/20 09:20:24 INFO mapreduce.Job: map 0% reduce 0%  
21/09/20 09:20:31 INFO mapreduce.Job: map 100% reduce 0%  
21/09/20 09:20:36 INFO mapreduce.Job: map 100% reduce 100%  
21/09/20 09:20:37 INFO mapreduce.Job: Job job_1632122556196_0001 completed successfully  
21/09/20 09:20:37 INFO mapreduce.Job: Counters: 49  
File System Counters  
FILE: Number of bytes read=12824  
FILE: Number of bytes written=485640  
FILE: Number of read operations=0  
FILE: Number of large read operations=0  
FILE: Number of write operations=0  
HDFS: Number of bytes read=98987  
HDFS: Number of bytes written=43  
HDFS: Number of read operations=9  
HDFS: Number of large read operations=0  
HDFS: Number of write operations=2  
Job Counters  
Launched map tasks=2  
Launched reduce tasks=1  
Data-local map tasks=2  
Total time spent by all maps in occupied slots (ms)=10401  
Total time spent by all reduces in occupied slots (ms)=2349  
Total time spent by all map tasks (ms)=10401  
Total time spent by all reduce tasks (ms)=2349  
Total vcore-milliseconds taken by all map tasks=10401  
Total vcore-milliseconds taken by all reduce tasks=2349  
Total megabyte-milliseconds taken by all map tasks=2600250  
Total megabyte-milliseconds taken by all reduce tasks=587250  
Map-Reduce Framework  
Map input records=998  
Map output records=998  
Map output bytes=10822  
Map output materialized bytes=12830
```

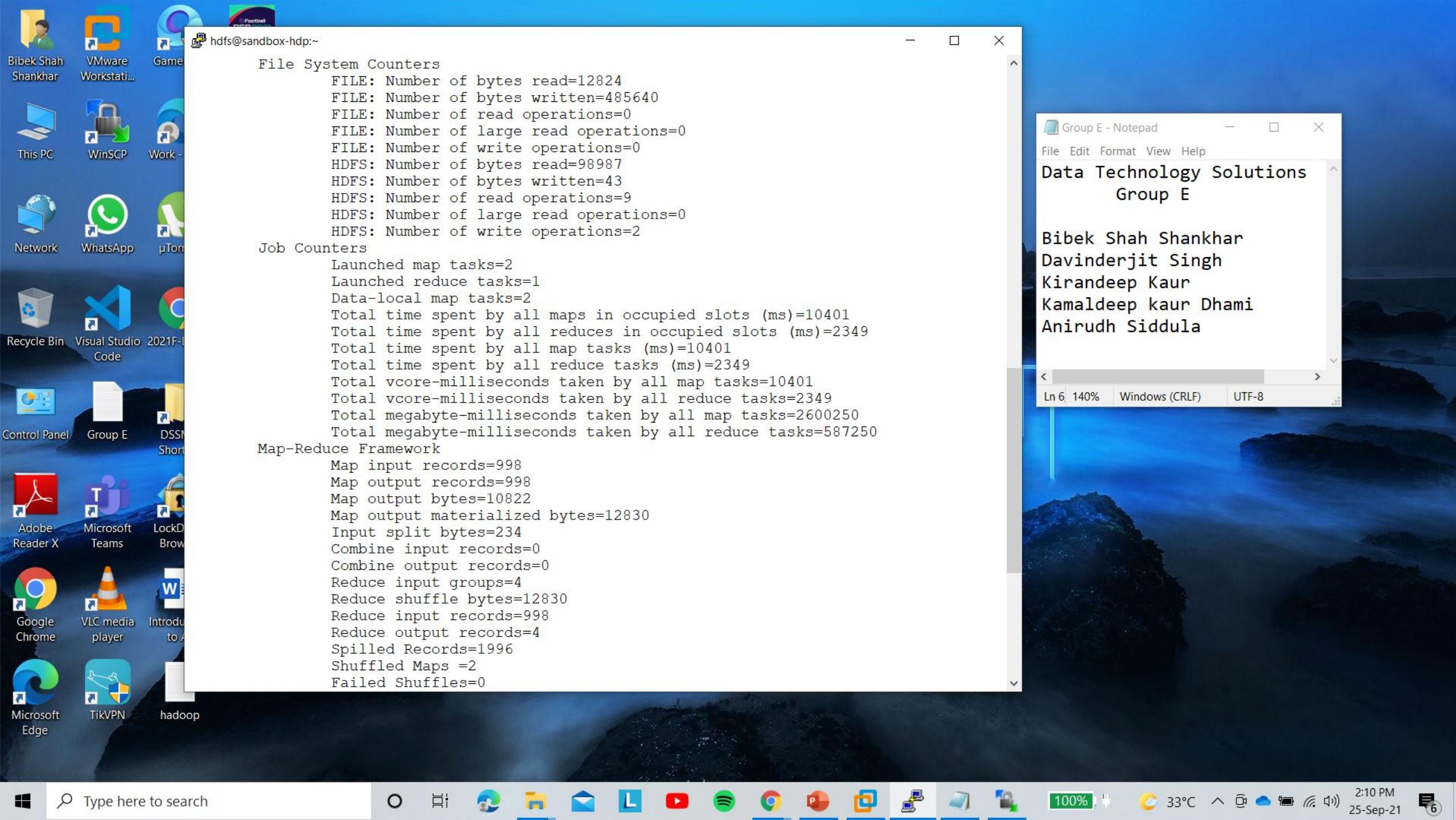
Group E - Notepad
File Edit Format View Help
Data Technology Solutions
Group E

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

Ln 6 140% Windows (CRLF) UTF-8



We Have 49 Counters.

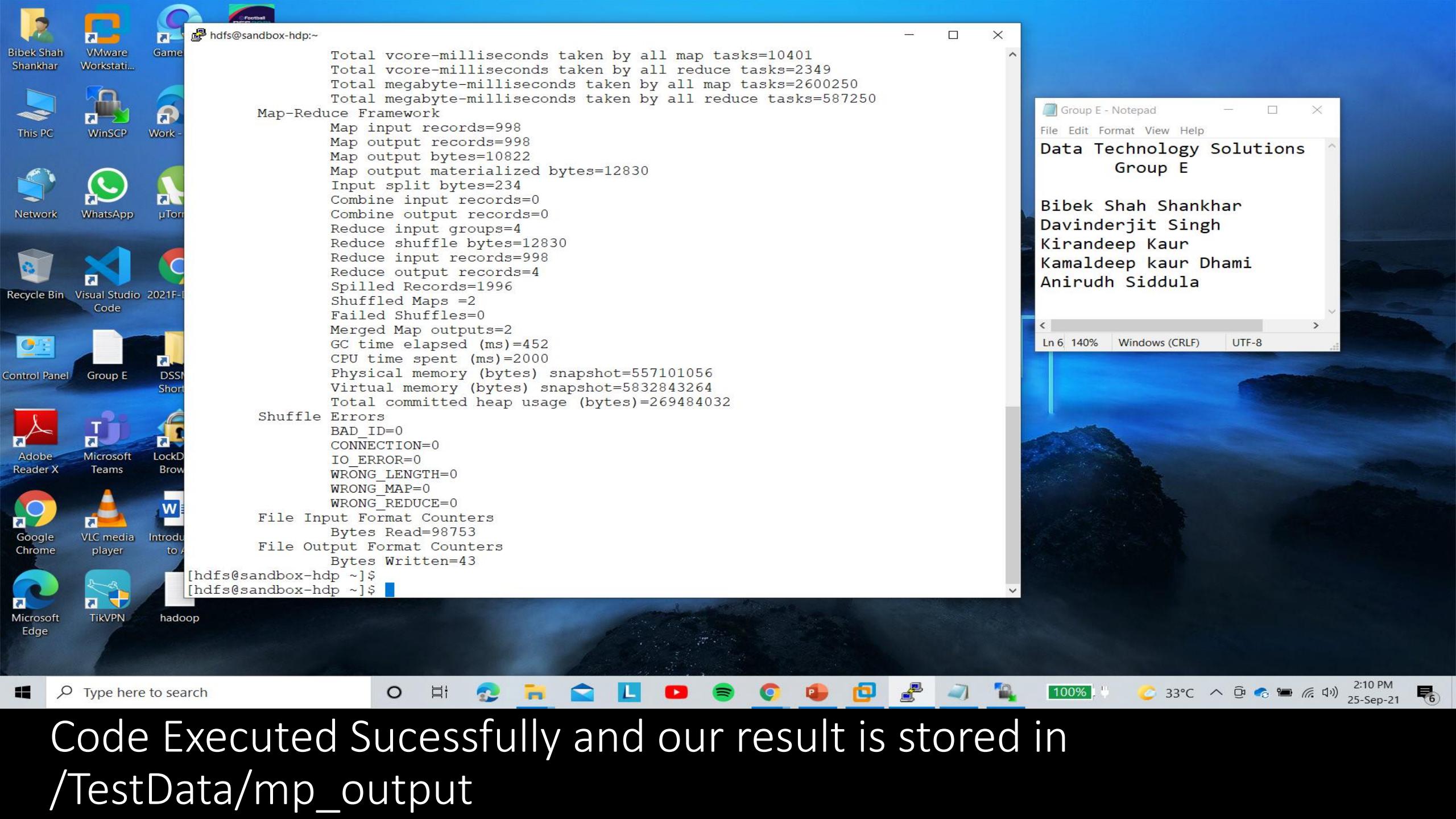


```
Total vcore-milliseconds taken by all map tasks=10401
Total vcore-milliseconds taken by all reduce tasks=2349
Total megabyte-milliseconds taken by all map tasks=2600250
Total megabyte-milliseconds taken by all reduce tasks=587250
Map-Reduce Framework
  Map input records=998
  Map output records=998
  Map output bytes=10822
  Map output materialized bytes=12830
  Input split bytes=234
  Combine input records=0
  Combine output records=0
  Reduce input groups=4
  Reduce shuffle bytes=12830
  Reduce input records=998
  Reduce output records=4
  Spilled Records=1996
  Shuffled Maps =2
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=452
  CPU time spent (ms)=2000
  Physical memory (bytes) snapshot=557101056
  Virtual memory (bytes) snapshot=5832843264
  Total committed heap usage (bytes)=269484032
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=98753
File Output Format Counters
  Bytes Written=43
[hdfs@sandbox-hdp ~]$
[hdfs@sandbox-hdp ~]$
```

Group E - Notepad
File Edit Format View Help
Data Technology Solutions
Group E

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

Ln 6 140% Windows (CRLF) UTF-8



Type here to search 2:10 PM 25-Sep-21 6

Code Executed Sucessfully and our result is stored in
/TestData/mp_output

Bibek Shah
Shankhar

VMware
Workstation

Gameloop eFootball PES

This PC

Network

Recycle Bin

Visual Studio Code

Control Panel

Adobe Reader X

Microsoft Team

Google Chrome

VLC media player

Microsoft Edge

[hdfs@sandbox-hdp ~]\$ hdfs dfs -ls /TestData/
Found 3 items
-rw-r--r-- 1 hdfs hdfs 65835 2021-09-20 09:12 /TestData/SalesJan2009.csv
-rw-r--r-- 1 root hdfs 1261316392 2021-09-20 07:34 /TestData/Test.rar
drwxr-xr-x - hdfs hdfs 0 2021-09-20 09:20 /TestData/mp_output
[hdfs@sandbox-hdp ~]\$ hdfs dfs -ls /TestData/mp_output
Found 2 items
-rw-r--r-- 1 hdfs hdfs 0 2021-09-20 09:20 /TestData/mp_output/_SUCCESS
-rw-r--r-- 1 hdfs hdfs 43 2021-09-20 09:20 /TestData/mp_output/part-00000
[hdfs@sandbox-hdp ~]\$ hdfs dfs -cat /TestData/mp_output/_SUCCESS
[hdfs@sandbox-hdp ~]\$ hdfs dfs -cat /TestData/mp_output/part-00000
Amex 110
Diners 89
Mastercard 277
Visa 522
[hdfs@sandbox-hdp ~]\$

Group E - Notepad

File Edit Format View Help

Data Technology Solutions
Group E

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

< >
Ln 6 140% Windows (CRLF) UTF-8

We can see we have a result and Analysis of our mapreduce code.