

Course: Big Data

Lab 02

HDFS

Fill answers of the questions below in the given tables.

Your screenshots must **contain commands** for required operations.

Question 1:

Download and unzip [lab02.zip](#) to obtain 365 text files.

Create the folder `/user/<username>/lab02` in your HDFS

Copy the entire `lab02/` from your local filesystem to `/user/<username>/lab02` in HDFS

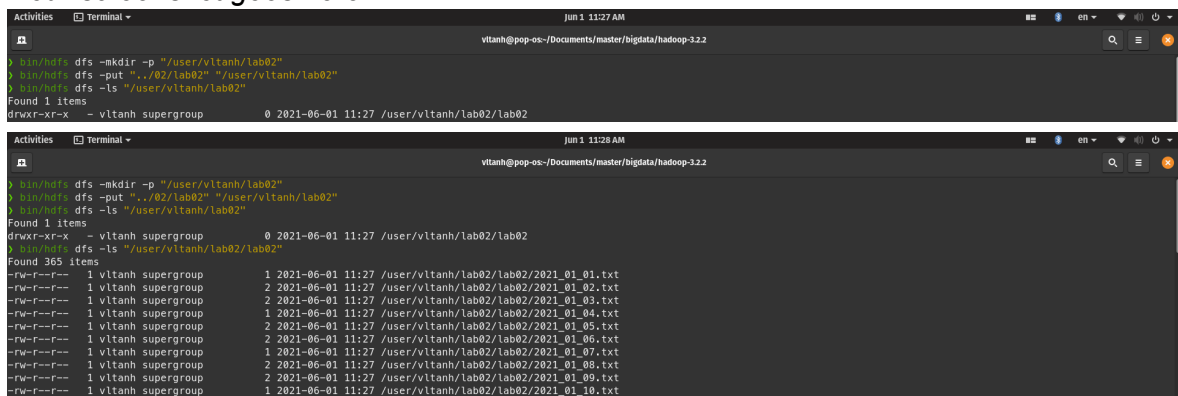
Take two screenshots to show your results.

For example,

```
Found 1 items
drwxr-xr-x  - ntan supergroup          0 2021-05-18 15:27 /user/ntan/lab02/lab02

Found 365 items
-rw-r--r--  1 ntan supergroup          1 2021-05-18 15:27 /user/ntan/lab02/lab02/2021_01_01.txt
-rw-r--r--  1 ntan supergroup          2 2021-05-18 15:26 /user/ntan/lab02/lab02/2021_01_02.txt
-rw-r--r--  1 ntan supergroup          2 2021-05-18 15:26 /user/ntan/lab02/lab02/2021_01_03.txt
-rw-r--r--  1 ntan supergroup          1 2021-05-18 15:26 /user/ntan/lab02/lab02/2021_01_04.txt
-rw-r--r--  1 ntan supergroup          2 2021-05-18 15:26 /user/ntan/lab02/lab02/2021_01_05.txt
-rw-r--r--  1 ntan supergroup          2 2021-05-18 15:26 /user/ntan/lab02/lab02/2021_01_06.txt
-rw-r--r--  1 ntan supergroup          1 2021-05-18 15:26 /user/ntan/lab02/lab02/2021_01_07.txt
```

Your screenshot goes here



```
Jun 1 11:27 AM
vltanh@pop-os:~/Documents/master/bigdata/hadoop-3.2.2

$ bin/hdfs dfs -mkdir -p "/user/vltanh/lab02"
$ bin/hdfs dfs -put "../02/lab02" "/user/vltanh/lab02"
$ bin/hdfs dfs -ls "/user/vltanh/lab02"
Found 1 items
drwxr-xr-x  - vltanh supergroup          0 2021-06-01 11:27 /user/vltanh/lab02/lab02

Jun 1 11:28 AM
vltanh@pop-os:~/Documents/master/bigdata/hadoop-3.2.2

$ bin/hdfs dfs -mkdir -p "/user/vltanh/lab02"
$ bin/hdfs dfs -put "../02/lab02" "/user/vltanh/lab02"
$ bin/hdfs dfs -ls "/user/vltanh/lab02"
Found 1 items
drwxr-xr-x  - vltanh supergroup          0 2021-06-01 11:27 /user/vltanh/lab02/lab02
$ bin/hdfs dfs -ls "/user/vltanh/lab02/lab02"
Found 365 items
-rw-r--r--  1 vltanh supergroup          1 2021-06-01 11:27 /user/vltanh/lab02/lab02/2021_01_01.txt
-rw-r--r--  1 vltanh supergroup          2 2021-06-01 11:27 /user/vltanh/lab02/lab02/2021_01_02.txt
-rw-r--r--  1 vltanh supergroup          2 2021-06-01 11:27 /user/vltanh/lab02/lab02/2021_01_03.txt
-rw-r--r--  1 vltanh supergroup          1 2021-06-01 11:27 /user/vltanh/lab02/lab02/2021_01_04.txt
-rw-r--r--  1 vltanh supergroup          2 2021-06-01 11:27 /user/vltanh/lab02/lab02/2021_01_05.txt
-rw-r--r--  1 vltanh supergroup          2 2021-06-01 11:27 /user/vltanh/lab02/lab02/2021_01_06.txt
-rw-r--r--  1 vltanh supergroup          1 2021-06-01 11:27 /user/vltanh/lab02/lab02/2021_01_07.txt
-rw-r--r--  1 vltanh supergroup          2 2021-06-01 11:27 /user/vltanh/lab02/lab02/2021_01_08.txt
-rw-r--r--  1 vltanh supergroup          2 2021-06-01 11:27 /user/vltanh/lab02/lab02/2021_01_09.txt
-rw-r--r--  1 vltanh supergroup          1 2021-06-01 11:27 /user/vltanh/lab02/lab02/2021_01_10.txt
```

Question 2:

Create 12 folders, corresponding to 12 months in a year, in `/user/<username>/lab02` in HDFS.

Notice: folder names must in the format like `_01`, `_02`, ..., `_12`

Take a screenshot to show your result.

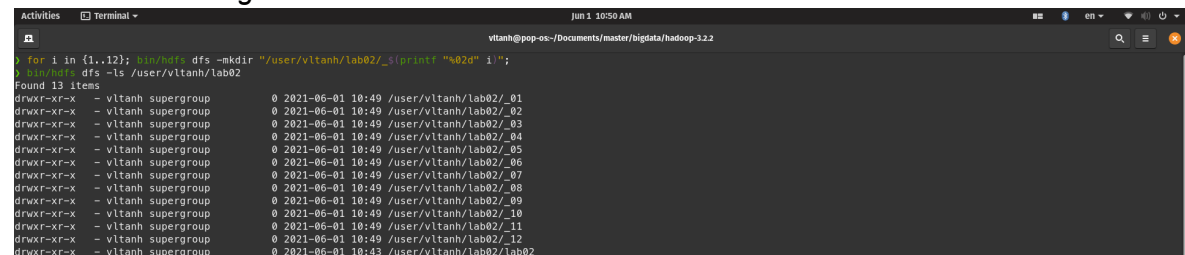
Hint:

- loop in bash scripts
- `$(printf ...)` for formatting strings in bash scripts.

For example,

```
Found 13 items
drwxr-xr-x  - ntan supergroup          0 2021-05-18 15:30 lab02/_01
drwxr-xr-x  - ntan supergroup          0 2021-05-18 15:30 lab02/_02
drwxr-xr-x  - ntan supergroup          0 2021-05-18 15:30 lab02/_03
drwxr-xr-x  - ntan supergroup          0 2021-05-18 15:30 lab02/_04
drwxr-xr-x  - ntan supergroup          0 2021-05-18 15:30 lab02/_05
drwxr-xr-x  - ntan supergroup          0 2021-05-18 15:30 lab02/_06
drwxr-xr-x  - ntan supergroup          0 2021-05-18 15:30 lab02/_07
drwxr-xr-x  - ntan supergroup          0 2021-05-18 15:30 lab02/_08
drwxr-xr-x  - ntan supergroup          0 2021-05-18 15:30 lab02/_09
drwxr-xr-x  - ntan supergroup          0 2021-05-18 15:30 lab02/_10
drwxr-xr-x  - ntan supergroup          0 2021-05-18 15:30 lab02/_11
drwxr-xr-x  - ntan supergroup          0 2021-05-18 15:30 lab02/_12
drwxr-xr-x  - ntan supergroup          0 2021-05-18 15:27 lab02/lab02
```

Your screenshot goes here



```
Activities Terminal Jun 1 10:50 AM vltanh@pop-os:~/Documents/master/bigdata/hadoop-3.2.2
$ for i in {1..12}; do bin/hdfs dfs -mkdir "/user/vltanh/lab02/$(printf "%02d" $i)"; done
$ bin/hdfs dfs -ls /user/vltanh/lab02
Found 13 items
drwxr-xr-x  - vltanh supergroup          0 2021-06-01 10:49 /user/vltanh/lab02/_01
drwxr-xr-x  - vltanh supergroup          0 2021-06-01 10:49 /user/vltanh/lab02/_02
drwxr-xr-x  - vltanh supergroup          0 2021-06-01 10:49 /user/vltanh/lab02/_03
drwxr-xr-x  - vltanh supergroup          0 2021-06-01 10:49 /user/vltanh/lab02/_04
drwxr-xr-x  - vltanh supergroup          0 2021-06-01 10:49 /user/vltanh/lab02/_05
drwxr-xr-x  - vltanh supergroup          0 2021-06-01 10:49 /user/vltanh/lab02/_06
drwxr-xr-x  - vltanh supergroup          0 2021-06-01 10:49 /user/vltanh/lab02/_07
drwxr-xr-x  - vltanh supergroup          0 2021-06-01 10:49 /user/vltanh/lab02/_08
drwxr-xr-x  - vltanh supergroup          0 2021-06-01 10:49 /user/vltanh/lab02/_09
drwxr-xr-x  - vltanh supergroup          0 2021-06-01 10:49 /user/vltanh/lab02/_10
drwxr-xr-x  - vltanh supergroup          0 2021-06-01 10:49 /user/vltanh/lab02/_11
drwxr-xr-x  - vltanh supergroup          0 2021-06-01 10:49 /user/vltanh/lab02/_12
drwxr-xr-x  - vltanh supergroup          0 2021-06-01 10:43 /user/vltanh/lab02/lab02
```

Question 3:

Text files in **lab02/** have a filename format like **YYYY_MM_DD.txt**

Move each text file in **lab02/** to the folder which is corresponding to the month in its filename.

E.g. **20201_01_01.txt** to **_01/**

Take a screenshot to show the content in **lab02/_01/**

Take a screenshot to show the sizes (in bytes) of the 12 folders.

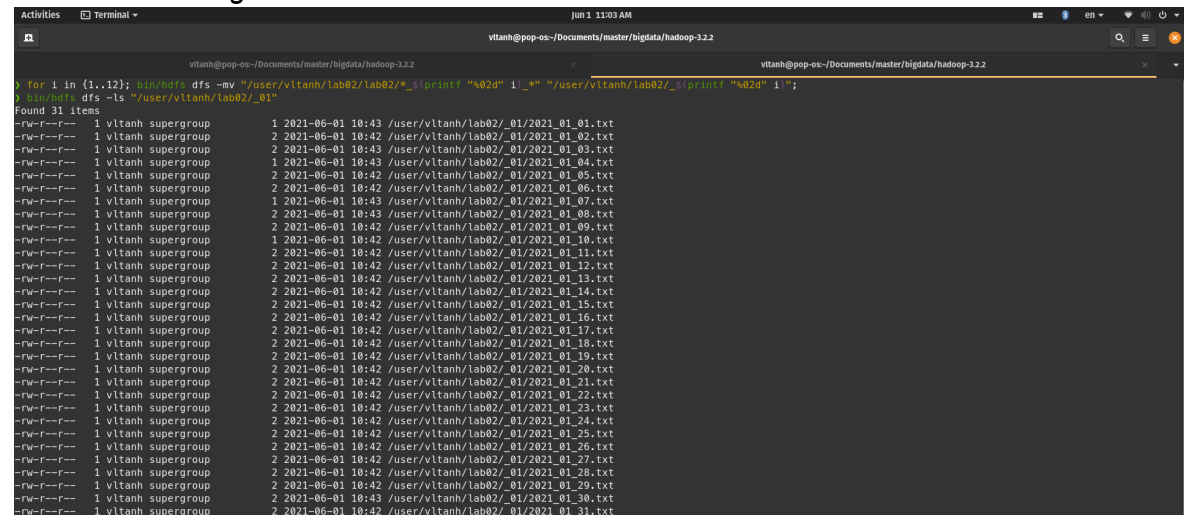
Hint: loop in bash scripts

For example,

```
Found 31 items
-rw-r--r-- 1 ntan supergroup 1 2021-05-18 15:27 lab02/_01/2021_01_01.txt
-rw-r--r-- 1 ntan supergroup 2 2021-05-18 15:26 lab02/_01/2021_01_02.txt
-rw-r--r-- 1 ntan supergroup 2 2021-05-18 15:26 lab02/_01/2021_01_03.txt
-rw-r--r-- 1 ntan supergroup 1 2021-05-18 15:26 lab02/_01/2021_01_04.txt
-rw-r--r-- 1 ntan supergroup 2 2021-05-18 15:26 lab02/_01/2021_01_05.txt
```

```
58 58 lab02/_01
52 52 lab02/_02
58 58 lab02/_03
56 56 lab02/_04
58 58 lab02/_05
56 56 lab02/_06
58 58 lab02/_07
58 58 lab02/_08
56 56 lab02/_09
58 58 lab02/_10
56 56 lab02/_11
58 58 lab02/_12
```

Your screenshot goes here



```
Activities Terminal v
vltanh@pop-os:~/Documents/master/bigdata/hadoop-3.2.2
vltanh@pop-os:~/Documents/master/bigdata/hadoop-3.2.2
vltanh@pop-os:~/Documents/master/bigdata/hadoop-3.2.2
vltanh@pop-os:~/Documents/master/bigdata/hadoop-3.2.2
for i in {1..12}; do mkdir -p "/user/vltanh/lab02/_$i"; find "/user/vltanh/lab02/" -type f -name "*$i*" -exec mv {} "/user/vltanh/lab02/_$i/" \; done
vltanh@pop-os:~/Documents/master/bigdata/hadoop-3.2.2
$ find -ls
Found 31 items
-rw-r--r-- 1 vltanh supergroup 1 2021-06-01 10:43 /user/vltanh/lab02/_01/2021_01_01.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_02.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:43 /user/vltanh/lab02/_01/2021_01_03.txt
-rw-r--r-- 1 vltanh supergroup 1 2021-06-01 10:43 /user/vltanh/lab02/_01/2021_01_04.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_05.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_06.txt
-rw-r--r-- 1 vltanh supergroup 1 2021-06-01 10:43 /user/vltanh/lab02/_01/2021_01_07.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:43 /user/vltanh/lab02/_01/2021_01_08.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_09.txt
-rw-r--r-- 1 vltanh supergroup 1 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_10.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_11.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_12.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_13.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_14.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_15.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_16.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_17.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_18.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_19.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_20.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_21.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_22.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_23.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_24.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_25.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_26.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_27.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_28.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_29.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:43 /user/vltanh/lab02/_01/2021_01_30.txt
-rw-r--r-- 1 vltanh supergroup 2 2021-06-01 10:42 /user/vltanh/lab02/_01/2021_01_31.txt
```

```
Activities Terminal Jun 1 11:06 AM vitanh@pop-os:~/Documents/master/bigdata/hadoop-3.2.2
y bin/hdfs dfs -du "/user/vltanh/lab02/"
58 58 /user/vltanh/lab02/_01
52 52 /user/vltanh/lab02/_02
58 58 /user/vltanh/lab02/_03
56 56 /user/vltanh/lab02/_04
58 58 /user/vltanh/lab02/_05
56 56 /user/vltanh/lab02/_06
58 58 /user/vltanh/lab02/_07
58 58 /user/vltanh/lab02/_08
56 56 /user/vltanh/lab02/_09
58 58 /user/vltanh/lab02/_10
56 56 /user/vltanh/lab02/_11
58 58 /user/vltanh/lab02/_12
0 0 /user/vltanh/lab02/lab02
```

Question 4:

Using **cat** to display the content of all files (in month order)

- day 10th
- day 01st
- day 07th
- day 04th

Take 4 screenshots of the corresponding results above.

For example,

```
2021-05-18 15:57:20,510 INFO s
se, remoteHostTrusted = false
e
c
2
1
2
1
a
9
3
7
f
e
```

Your screenshot goes here

```
Activities Terminal Jun 1 11:10 AM vitanh@pop-os:~/Documents/master/bigdata/hadoop-3.2.2
y bin/hdfs dfs -cat "/user/vltanh/lab02/*_10.txt"
<3__i__love__

Activities Terminal Jun 1 11:10 AM vitanh@pop-os:~/Documents/master/bigdata/hadoop-3.2.2
y bin/hdfs dfs -cat "/user/vltanh/lab02/*_01.txt"
__you__for__

Activities Terminal Jun 1 11:10 AM vitanh@pop-os:~/Documents/master/bigdata/hadoop-3.2.2
y bin/hdfs dfs -cat "/user/vltanh/lab02/*_07.txt"
a__thousand__

Activities Terminal Jun 1 11:11 AM vitanh@pop-os:~/Documents/master/bigdata/hadoop-3.2.2
y bin/hdfs dfs -cat "/user/vltanh/lab02/*_04.txt"
__years__<3__
```

Question 5:

Concatenate contents in the 4 screenshots of Question 4 to form the completed message. Finally, write it in the given table.

Your answer goes here

The 4 messages together read:

“_<3__i__love__you__for__a__thousand__years__<3_.”

Or:

“❤️ I love you for a thousand years ❤️.”

Submission Notice

- Export your answer file as pdf
- Rename the pdf following the format:

<student number>_HoTen.pdf

E.g: **123456_NguyenThanhAn.pdf**

If you have not been assigned a student number yet, then use 123456 instead.

- Careless mistakes in filename, format, question order, etc. are not accepted (0 pts).
- Submission URL: <https://forms.gle/bPfd5NXthKTTToHLx9>

● **Deadline: 17:00, 03/06/2021 (Thu).**