

HDFS

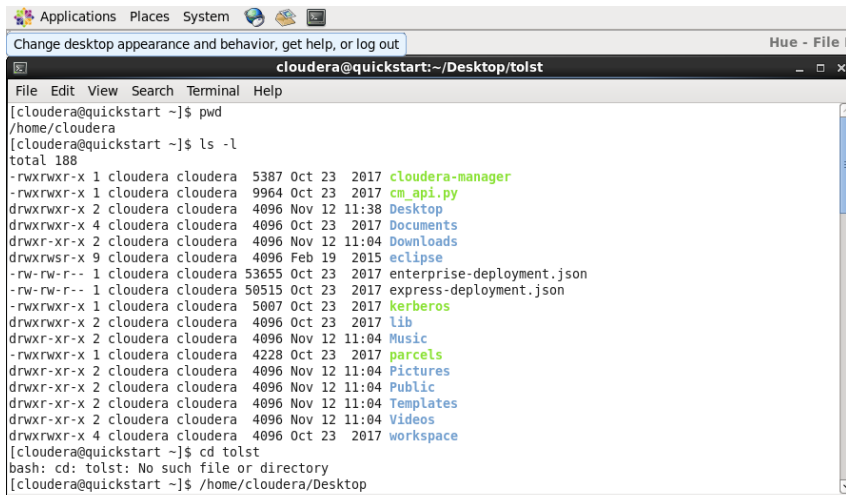
1. Когда мы перетащили файлы с произведением Льва Толстого – мы перетащили их в файловую систему виртуальной машины, но не в HDFS, соответственно, в первую очередь нам нужно перенести их в папку нашего пользователя именно на HDFS.

Определим, где мы находимся:

```
pwd
```

Выведем содержание директории:

```
ls -l
```



```
cloudera@quickstart: ~/Desktop/tolst
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ pwd
/home/cloudera
[cloudera@quickstart ~]$ ls -l
total 188
-rwxrwxr-x 1 cloudera cloudera 5387 Oct 23 2017 cloudera-manager
-rwxrwxr-x 1 cloudera cloudera 9964 Oct 23 2017 cm_api.py
drwxrwxr-x 2 cloudera cloudera 4096 Nov 12 11:38 Desktop
drwxrwxr-x 4 cloudera cloudera 4096 Oct 23 2017 Documents
drwxr-xr-x 2 cloudera cloudera 4096 Nov 12 11:04 Downloads
drwxrwsr-x 9 cloudera cloudera 4096 Feb 19 2015 eclipse
-rw-rw-r-- 1 cloudera cloudera 53655 Oct 23 2017 enterprise-deployment.json
-rw-rw-r-- 1 cloudera cloudera 50515 Oct 23 2017 express-deployment.json
-rwxrwxr-x 1 cloudera cloudera 5007 Oct 23 2017 kerberos
drwxrwxr-x 2 cloudera cloudera 4096 Oct 23 2017 lib
drwxr-xr-x 2 cloudera cloudera 4096 Nov 12 11:04 Music
-rwxrwxr-x 1 cloudera cloudera 4228 Oct 23 2017 parcels
drwxr-xr-x 2 cloudera cloudera 4096 Nov 12 11:04 Pictures
drwxr-xr-x 2 cloudera cloudera 4096 Nov 12 11:04 Public
drwxr-xr-x 2 cloudera cloudera 4096 Nov 12 11:04 Templates
drwxr-xr-x 2 cloudera cloudera 4096 Nov 12 11:04 Videos
drwxrwxr-x 4 cloudera cloudera 4096 Oct 23 2017 workspace
[cloudera@quickstart ~]$ cd tolst
bash: cd: tolst: No such file or directory
[cloudera@quickstart ~]$ /home/cloudera/Desktop
```

Томы Льва Толстого лежат на рабочем столе в заранее созданной папке *tolst*.

Проверим это с помощью команд:

```
cd /home/cloudera/Desktop/tolst
```

```
ls -l
```

Скопируем их из нашей папки */home/Downloads/* в *hdfs*:

```
hdfs dfs -copyFromLocal /home/cloudera/Downloads/voina-i-mir-tom-1.txt
```

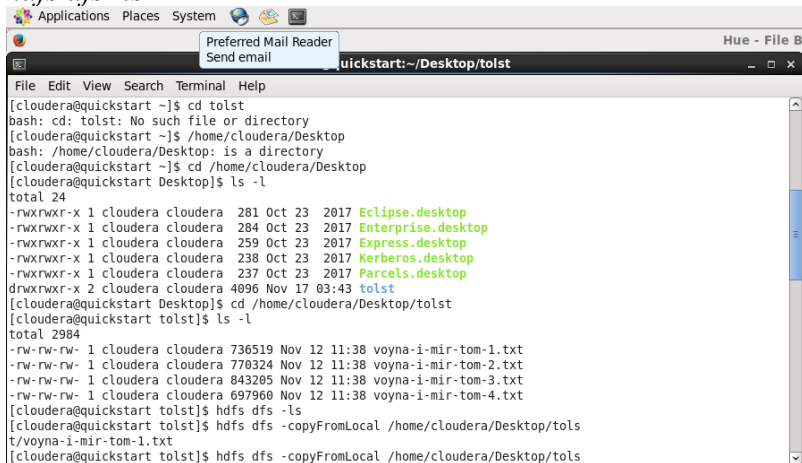
```
hdfs dfs -copyFromLocal /home/cloudera/Downloads/voina-i-mir-tom-2.txt
```

```
hdfs dfs -copyFromLocal /home/cloudera/Downloads/voina-i-mir-tom-3.txt
```

```
hdfs dfs -copyFromLocal /home/cloudera/Downloads/voina-i-mir-tom-4.txt
```

Проверим, что файлы скопированы в *hdfs*:

```
hdfs dfs -ls
```



```
cloudera@quickstart: ~/Desktop/tolst
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ cd tolst
bash: cd: tolst: No such file or directory
[cloudera@quickstart ~]$ /home/cloudera/Desktop
bash: /home/cloudera/Desktop: is a directory
[cloudera@quickstart ~]$ cd /home/cloudera/Desktop
[cloudera@quickstart Desktop]$ ls -l
total 24
-rwxrwxr-x 1 cloudera cloudera 281 Oct 23 2017 Eclipse.desktop
-rwxrwxr-x 1 cloudera cloudera 284 Oct 23 2017 Enterprise.desktop
-rwxrwxr-x 1 cloudera cloudera 259 Oct 23 2017 Express.desktop
-rwxrwxr-x 1 cloudera cloudera 238 Oct 23 2017 Kerberos.desktop
-rwxrwxr-x 1 cloudera cloudera 237 Oct 23 2017 Parcels.desktop
drwxrwxr-x 2 cloudera cloudera 4096 Nov 17 03:43 tolst
[cloudera@quickstart Desktop]$ cd /home/cloudera/Desktop/tolst
[cloudera@quickstart tolst]$ ls -l
total 2984
-rw-rw-rw- 1 cloudera cloudera 736519 Nov 12 11:38 voyna-i-mir-tom-1.txt
-rw-rw-rw- 1 cloudera cloudera 770324 Nov 12 11:38 voyna-i-mir-tom-2.txt
-rw-rw-rw- 1 cloudera cloudera 843205 Nov 12 11:38 voyna-i-mir-tom-3.txt
-rw-rw-rw- 1 cloudera cloudera 697960 Nov 12 11:38 voyna-i-mir-tom-4.txt
[cloudera@quickstart tolst]$ hdfs dfs -ls
[cloudera@quickstart tolst]$ hdfs dfs -copyFromLocal /home/cloudera/Desktop/tolst
t/voyna-i-mir-tom-1.txt
[cloudera@quickstart tolst]$ hdfs dfs -copyFromLocal /home/cloudera/Desktop/tolst
```

```
cloudera@quickstart: ~/Desktop/tolst
total 2984
-rw-rw-rw- 1 cloudera cloudera 736519 Nov 12 11:38 voyna-i-mir-tom-1.txt
-rw-rw-rw- 1 cloudera cloudera 770324 Nov 12 11:38 voyna-i-mir-tom-2.txt
-rw-rw-rw- 1 cloudera cloudera 843205 Nov 12 11:38 voyna-i-mir-tom-3.txt
-rw-rw-rw- 1 cloudera cloudera 697960 Nov 12 11:38 voyna-i-mir-tom-4.txt
[cloudera@quickstart tolst]$ hdfs dfs -ls
[cloudera@quickstart tolst]$ hdfs dfs -copyFromLocal /home/cloudera/Desktop/tols
t/voyna-i-mir-tom-1.txt
[cloudera@quickstart tolst]$ hdfs dfs -copyFromLocal /home/cloudera/Desktop/tols
t/voyna-i-mir-tom-2.txt
[cloudera@quickstart tolst]$ hdfs dfs -copyFromLocal /home/cloudera/Desktop/tols
t/voyna-i-mir-tom-3.txt
[cloudera@quickstart tolst]$ hdfs dfs -copyFromLocal /home/cloudera/Desktop/tols
t/voyna-i-mir-tom-4.txt
[cloudera@quickstart tolst]$ hdfs dfs -ls
Found 4 items
-rw-r--r-- 1 cloudera cloudera 736519 2022-11-17 03:53 voyna-i-mir-tom-1.t
xt
-rw-r--r-- 1 cloudera cloudera 770324 2022-11-17 03:53 voyna-i-mir-tom-2.t
xt
-rw-r--r-- 1 cloudera cloudera 843205 2022-11-17 03:54 voyna-i-mir-tom-3.t
xt
-rw-r--r-- 1 cloudera cloudera 697960 2022-11-17 03:54 voyna-i-mir-tom-4.t
xt
```

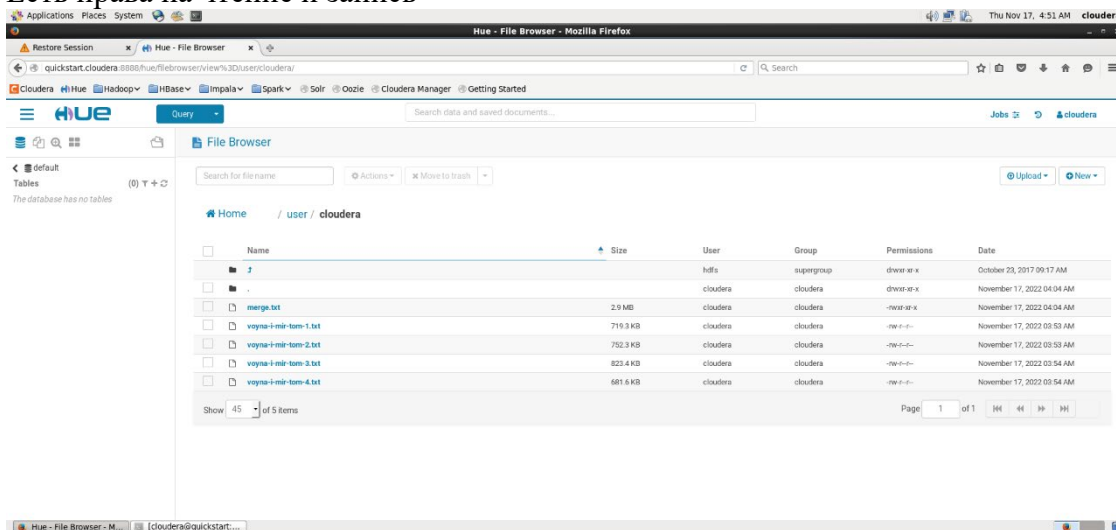
2. После того, как файлы окажутся на HDFS попробуйте выполнить команду, которая выводит содержимое папки. Особенно обратите внимание на права доступа к вашим файлам.

Вот они наши права доступа:

hdfs dfs -ls

```
cloudera@quickstart: ~/Desktop/tolst
total 2984
-rw-rw-rw- 1 cloudera cloudera 736519 Nov 12 11:38 voyna-i-mir-tom-1.txt
-rw-rw-rw- 1 cloudera cloudera 770324 Nov 12 11:38 voyna-i-mir-tom-2.txt
-rw-rw-rw- 1 cloudera cloudera 843205 Nov 12 11:38 voyna-i-mir-tom-3.txt
-rw-rw-rw- 1 cloudera cloudera 697960 Nov 12 11:38 voyna-i-mir-tom-4.txt
[cloudera@quickstart tolst]$ hdfs dfs -ls
[cloudera@quickstart tolst]$ hdfs dfs -copyFromLocal /home/cloudera/Desktop/tols
t/voyna-i-mir-tom-1.txt
[cloudera@quickstart tolst]$ hdfs dfs -copyFromLocal /home/cloudera/Desktop/tols
t/voyna-i-mir-tom-2.txt
[cloudera@quickstart tolst]$ hdfs dfs -copyFromLocal /home/cloudera/Desktop/tols
t/voyna-i-mir-tom-3.txt
[cloudera@quickstart tolst]$ hdfs dfs -copyFromLocal /home/cloudera/Desktop/tols
t/voyna-i-mir-tom-4.txt
[cloudera@quickstart tolst]$ hdfs dfs -ls
Found 4 items
-rw-r--r-- 1 cloudera cloudera 736519 2022-11-17 03:53 voyna-i-mir-tom-1.t
xt
-rw-r--r-- 1 cloudera cloudera 770324 2022-11-17 03:53 voyna-i-mir-tom-2.t
xt
-rw-r--r-- 1 cloudera cloudera 843205 2022-11-17 03:54 voyna-i-mir-tom-3.t
xt
-rw-r--r-- 1 cloudera cloudera 697960 2022-11-17 03:54 voyna-i-mir-tom-4.t
xt
```

Есть права на чтение и запись



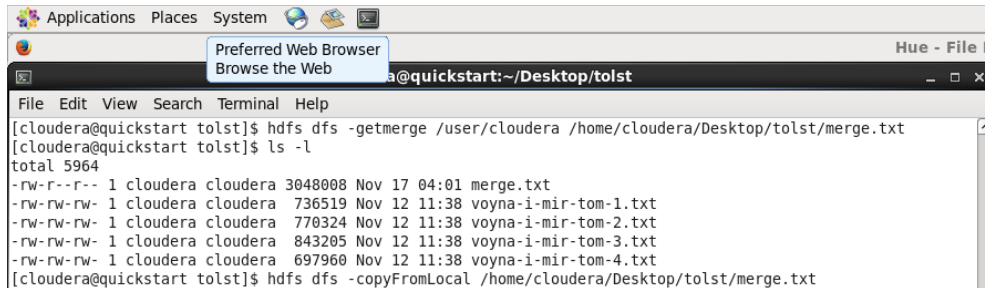
3. Далее сожмим все 4 тома в 1 файл.

Для сжатия воспользуемся командой:

```
hdfs dfs -getmerge /user/cloudera /home/cloudera/Desktop/tolst/merge.txt
```

И скопируем этот файл в hdfs:

```
hdfs dfs -copyFromLocal /home/cloudera/Desktop/tolst/merge.txt
```



```
[cloudera@quickstart tolst]$ hdfs dfs -getmerge /user/cloudera /home/cloudera/Desktop/tolst/merge.txt
[cloudera@quickstart tolst]$ ls -l
total 5964
-rw-r--r-- 1 cloudera cloudera 3048008 Nov 17 04:01 merge.txt
-rw-rw-rw- 1 cloudera cloudera 736519 Nov 12 11:38 voyna-i-mir-tom-1.txt
-rw-rw-rw- 1 cloudera cloudera 770324 Nov 12 11:38 voyna-i-mir-tom-2.txt
-rw-rw-rw- 1 cloudera cloudera 843205 Nov 12 11:38 voyna-i-mir-tom-3.txt
-rw-rw-rw- 1 cloudera cloudera 697960 Nov 12 11:38 voyna-i-mir-tom-4.txt
[cloudera@quickstart tolst]$ hdfs dfs -copyFromLocal /home/cloudera/Desktop/tolst/merge.txt
```

4. Теперь давайте изменим права доступа к нашему файлу. Чтобы с нашим файлом могли взаимодействовать коллеги, установите режим доступа, который дает полный доступ для владельца файла, а для сторонних пользователей возможность читать и выполнять.

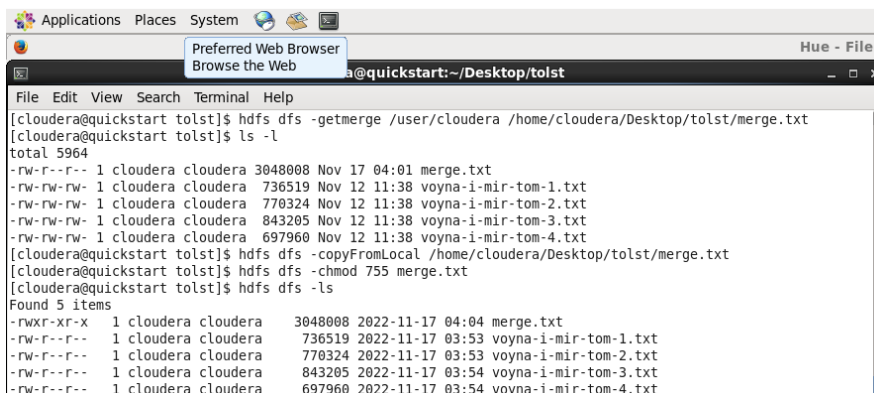
Воспользуемся командой:

```
hdfs dfs -chmod 755 merge.txt
```

5. Попробуйте заново использовать команду для вывода содержимого папки и обратите внимание как изменились права доступа к файлу.

И посмотрим содержимое папки hdfs:

```
hdfs dfs -ls
```



```
[cloudera@quickstart tolst]$ hdfs dfs -getmerge /user/cloudera /home/cloudera/Desktop/tolst/merge.txt
[cloudera@quickstart tolst]$ ls -l
total 5964
-rw-r--r-- 1 cloudera cloudera 3048008 Nov 17 04:01 merge.txt
-rw-rw-rw- 1 cloudera cloudera 736519 Nov 12 11:38 voyna-i-mir-tom-1.txt
-rw-rw-rw- 1 cloudera cloudera 770324 Nov 12 11:38 voyna-i-mir-tom-2.txt
-rw-rw-rw- 1 cloudera cloudera 843205 Nov 12 11:38 voyna-i-mir-tom-3.txt
-rw-rw-rw- 1 cloudera cloudera 697960 Nov 12 11:38 voyna-i-mir-tom-4.txt
[cloudera@quickstart tolst]$ hdfs dfs -copyFromLocal /home/cloudera/Desktop/tolst/merge.txt
[cloudera@quickstart tolst]$ hdfs dfs -chmod 755 merge.txt
[cloudera@quickstart tolst]$ hdfs dfs -ls
Found 5 items
-rwxr-xr-x 1 cloudera cloudera 3048008 2022-11-17 04:04 merge.txt
-rw-r--r-- 1 cloudera cloudera 736519 2022-11-17 03:53 voyna-i-mir-tom-1.txt
-rw-r--r-- 1 cloudera cloudera 770324 2022-11-17 03:53 voyna-i-mir-tom-2.txt
-rw-r--r-- 1 cloudera cloudera 843205 2022-11-17 03:54 voyna-i-mir-tom-3.txt
-rw-r--r-- 1 cloudera cloudera 697960 2022-11-17 03:54 voyna-i-mir-tom-4.txt
```

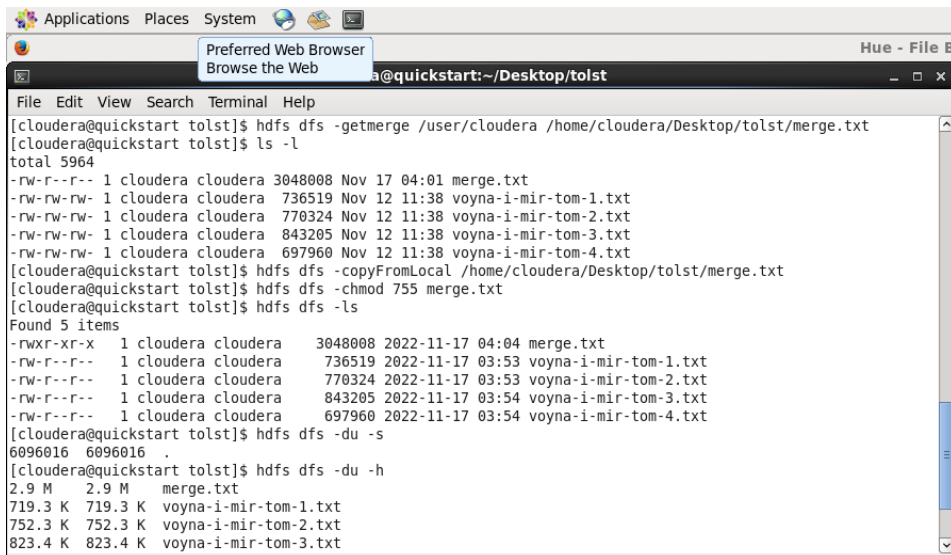
6. Теперь попробуем вывести на экран информацию о том, сколько места на диске занимает наш файл. Желательно, чтобы размер файла был удобночитаемым.

Команда для вывода размера файла:

```
hdfs dfs -du -s
```

Для удобства чтения возьмем команду:

```
hdfs dfs -du -h
```



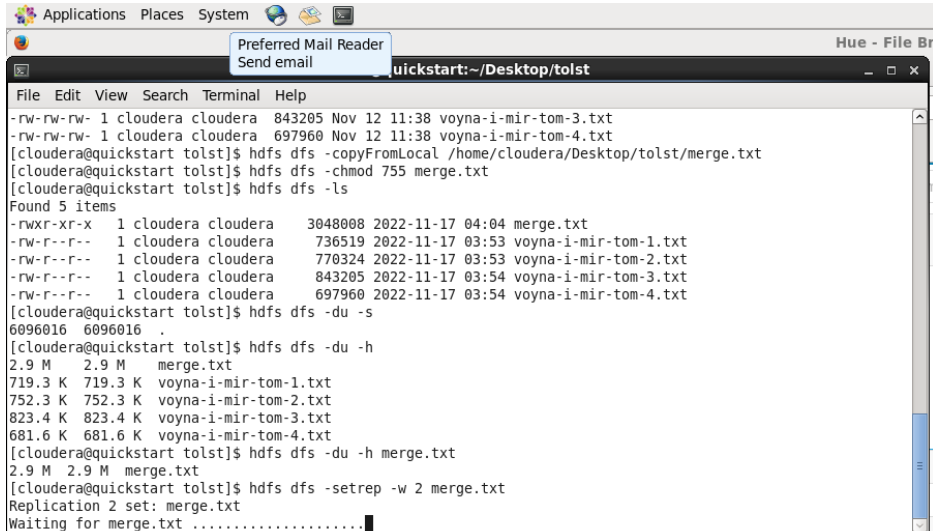
```
[cloudera@quickstart tolst]$ hdfs dfs -getmerge /user/cloudera /home/cloudera/Desktop/tolst/merge.txt
[cloudera@quickstart tolst]$ ls -l
total 5964
-rw-r--r-- 1 cloudera cloudera 3048008 Nov 17 04:01 merge.txt
-rw-rw-rw- 1 cloudera cloudera 736519 Nov 12 11:38 voyna-i-mir-tom-1.txt
-rw-rw-rw- 1 cloudera cloudera 770324 Nov 12 11:38 voyna-i-mir-tom-2.txt
-rw-rw-rw- 1 cloudera cloudera 843205 Nov 12 11:38 voyna-i-mir-tom-3.txt
-rw-rw-rw- 1 cloudera cloudera 697960 Nov 12 11:38 voyna-i-mir-tom-4.txt
[cloudera@quickstart tolst]$ hdfs dfs -copyFromLocal /home/cloudera/Desktop/tolst/merge.txt
[cloudera@quickstart tolst]$ hdfs dfs -chmod 755 merge.txt
[cloudera@quickstart tolst]$ hdfs dfs -ls
Found 5 items
-rwxr-xr-x 1 cloudera cloudera 3048008 2022-11-17 04:04 merge.txt
-rw-r--r-- 1 cloudera cloudera 736519 2022-11-17 03:53 voyna-i-mir-tom-1.txt
-rw-r--r-- 1 cloudera cloudera 770324 2022-11-17 03:53 voyna-i-mir-tom-2.txt
-rw-r--r-- 1 cloudera cloudera 843205 2022-11-17 03:54 voyna-i-mir-tom-3.txt
-rw-r--r-- 1 cloudera cloudera 697960 2022-11-17 03:54 voyna-i-mir-tom-4.txt
[cloudera@quickstart tolst]$ hdfs dfs -du -s
6096016 6096016 .
[cloudera@quickstart tolst]$ hdfs dfs -du -h
2.9 M 2.9 M merge.txt
719.3 K 719.3 K voyna-i-mir-tom-1.txt
752.3 K 752.3 K voyna-i-mir-tom-2.txt
823.4 K 823.4 K voyna-i-mir-tom-3.txt
```

Видим, что размер файла 2.9 Мб. Занимаемое файлом место на диске также 2.9 Мб.

7. На экране вы можете заметить 2 числа. Первое число – это фактический размер файла, а второе – это занимаемое файлом место на диске с учетом репликации. По умолчанию в данной версии HDFS эти числа будут одинаковы – это означает, что никакой репликации нет – нас это не очень устраивает, мы хотели бы, чтобы у наших файлов существовали резервные копии, поэтому напишите команду, которая изменит фактор репликации на 2.

Изменим число репликаций на 2:

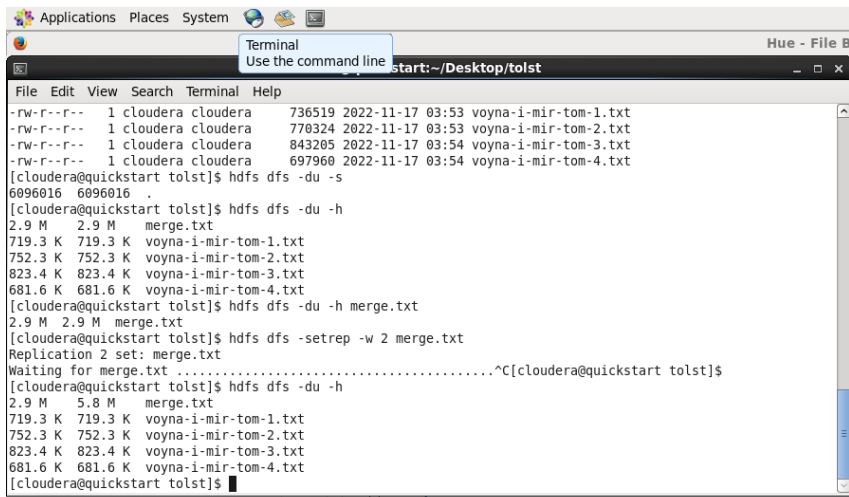
hdfs dfs -setrep -w 2 merge.txt



```
[cloudera@quickstart tolst]$ hdfs dfs -copyFromLocal /home/cloudera/Desktop/tolst/merge.txt
[cloudera@quickstart tolst]$ hdfs dfs -chmod 755 merge.txt
[cloudera@quickstart tolst]$ hdfs dfs -ls
Found 5 items
-rwxr-xr-x 1 cloudera cloudera 3048008 2022-11-17 04:04 merge.txt
-rw-r--r-- 1 cloudera cloudera 736519 2022-11-17 03:53 voyna-i-mir-tom-1.txt
-rw-r--r-- 1 cloudera cloudera 770324 2022-11-17 03:53 voyna-i-mir-tom-2.txt
-rw-r--r-- 1 cloudera cloudera 843205 2022-11-17 03:54 voyna-i-mir-tom-3.txt
-rw-r--r-- 1 cloudera cloudera 697960 2022-11-17 03:54 voyna-i-mir-tom-4.txt
[cloudera@quickstart tolst]$ hdfs dfs -du -s
6096016 6096016 .
[cloudera@quickstart tolst]$ hdfs dfs -du -h
2.9 M 2.9 M merge.txt
719.3 K 719.3 K voyna-i-mir-tom-1.txt
752.3 K 752.3 K voyna-i-mir-tom-2.txt
823.4 K 823.4 K voyna-i-mir-tom-3.txt
681.6 K 681.6 K voyna-i-mir-tom-4.txt
[cloudera@quickstart tolst]$ hdfs dfs -du -h merge.txt
2.9 M 2.9 M merge.txt
[cloudera@quickstart tolst]$ hdfs dfs -setrep -w 2 merge.txt
Replication 2 set: merge.txt
Waiting for merge.txt .....
```

8. Повторите команду, которая выводит информацию о том, какое место на диске занимает файл и убедитесь, что изменения произошли.

hdfs dfs -du -h



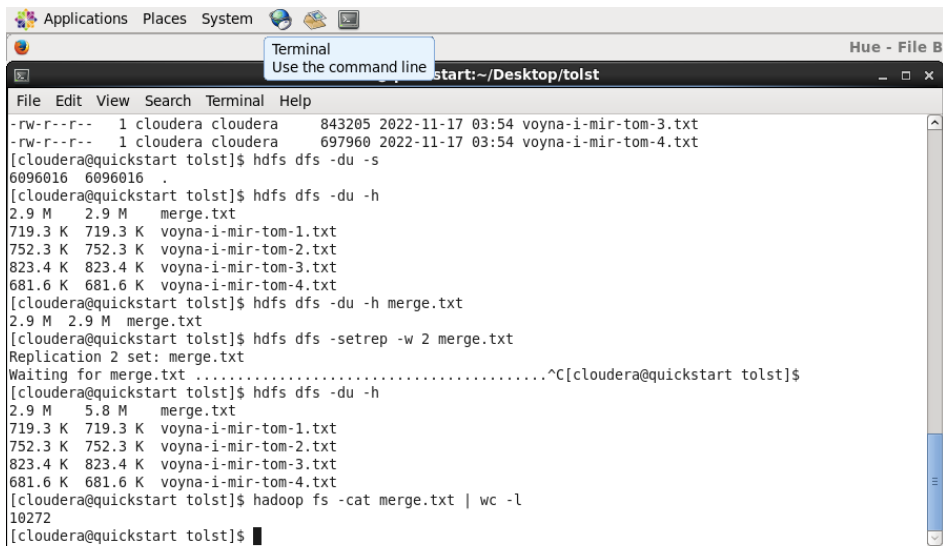
```
File Edit View Search Terminal Help
-rw-r--r-- 1 cloudera cloudera 736519 2022-11-17 03:53 voyna-i-mir-tom-1.txt
-rw-r--r-- 1 cloudera cloudera 770324 2022-11-17 03:53 voyna-i-mir-tom-2.txt
-rw-r--r-- 1 cloudera cloudera 843205 2022-11-17 03:54 voyna-i-mir-tom-3.txt
-rw-r--r-- 1 cloudera cloudera 697960 2022-11-17 03:54 voyna-i-mir-tom-4.txt
[cloudera@quickstart tolst]$ hdfs dfs -du -s
6096016 6096016 .
[cloudera@quickstart tolst]$ hdfs dfs -du -h
2.9 M 2.9 M merge.txt
719.3 K 719.3 K voyna-i-mir-tom-1.txt
752.3 K 752.3 K voyna-i-mir-tom-2.txt
823.4 K 823.4 K voyna-i-mir-tom-3.txt
681.6 K 681.6 K voyna-i-mir-tom-4.txt
[cloudera@quickstart tolst]$ hdfs dfs -du -h merge.txt
2.9 M 2.9 M merge.txt
[cloudera@quickstart tolst]$ hdfs dfs -setrep -w 2 merge.txt
Replication 2 set: merge.txt
Waiting for merge.txt .....^C[cloudera@quickstart tolst]$
[cloudera@quickstart tolst]$ hdfs dfs -du -h
2.9 M 5.8 M merge.txt
719.3 K 719.3 K voyna-i-mir-tom-1.txt
752.3 K 752.3 K voyna-i-mir-tom-2.txt
823.4 K 823.4 K voyna-i-mir-tom-3.txt
681.6 K 681.6 K voyna-i-mir-tom-4.txt
[cloudera@quickstart tolst]$
```

Видим, что теперь занимаемое файлом место на диске в два раза увеличилось и стало 5.8 Мб.

9. Напишите команду, которая подсчитывает количество строк в вашем файле

Команда:

`hadoop fs -cat user/cloudera/merge.txt | wc -l`



```
File Edit View Search Terminal Help
-rw-r--r-- 1 cloudera cloudera 843205 2022-11-17 03:54 voyna-i-mir-tom-3.txt
-rw-r--r-- 1 cloudera cloudera 697960 2022-11-17 03:54 voyna-i-mir-tom-4.txt
[cloudera@quickstart tolst]$ hdfs dfs -du -s
6096016 6096016 .
[cloudera@quickstart tolst]$ hdfs dfs -du -h
2.9 M 2.9 M merge.txt
719.3 K 719.3 K voyna-i-mir-tom-1.txt
752.3 K 752.3 K voyna-i-mir-tom-2.txt
823.4 K 823.4 K voyna-i-mir-tom-3.txt
681.6 K 681.6 K voyna-i-mir-tom-4.txt
[cloudera@quickstart tolst]$ hdfs dfs -du -h merge.txt
2.9 M 2.9 M merge.txt
[cloudera@quickstart tolst]$ hdfs dfs -setrep -w 2 merge.txt
Replication 2 set: merge.txt
Waiting for merge.txt .....^C[cloudera@quickstart tolst]$
[cloudera@quickstart tolst]$ hdfs dfs -du -h
2.9 M 5.8 M merge.txt
719.3 K 719.3 K voyna-i-mir-tom-1.txt
752.3 K 752.3 K voyna-i-mir-tom-2.txt
823.4 K 823.4 K voyna-i-mir-tom-3.txt
681.6 K 681.6 K voyna-i-mir-tom-4.txt
[cloudera@quickstart tolst]$ hadoop fs -cat merge.txt | wc -l
10272
[cloudera@quickstart tolst]$
```

В нашем файле 10272 строки.

10. В качестве результатов вашей работы, запишите ваши команды и вывод этих команд в отдельный файл и выложите его на github.

Данный файл будет запущен в репозитории: https://github.com/piskovo4ka/1T_Sprint_3.1