Урок 2. HDFS

ssh student3_10@37.139.32.56 -D localhost:8080

Создаем директорию **data** в HDFS: hdfs dfs -mkdir data

```
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ hdfs dfs -ls /
Found 13 items
drwxrwxrwt   - yarn    hadoop          0 2021-04-08 15:19 /app-logs
drwxr-xr-x   - hdfs    hdfs            0 2021-03-25 08:38 /apps
drwxr-xr-x   - yarn    hadoop          0 2021-03-20 14:24 /ats
drwxr-xr-x   - hdfs    hdfs            0 2021-03-20 14:25 /atsv2
drwxr-xr-x   - hdfs    hdfs            0 2021-03-20 14:24 /hdp
drwx------   - livy    hdfs            0 2021-03-20 14:31 /livy2-recovery
drwxr-xr-x   - mapred  hdfs            0 2021-03-20 14:24 /mapred
drwxrwxrwx   - mapred  hadoop          0 2021-03-31 21:30 /mr-history
drwxrwxrwx   - spark   hadoop          0 2021-04-08 21:07 /spark2-history
drwxr-xr-x   - hdfs    hdfs            0 2021-04-06 09:11 /system
drwxrwxrwx   - hdfs    hdfs            0 2021-04-07 13:46 /tmp
drwxr-xr-x   - hdfs    hdfs            0 2021-04-02 16:10 /user
drwxr-xr-x   - hdfs    hdfs            0 2021-04-02 08:21 /warehouse
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ ^C
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ pwd
/home/student3_10
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ hdfs dfs -mkdir data
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ ls
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ hdfs dfs -ls
Found 1 items
drwxr-xr-x   - student3_10 student3_10          0 2021-04-08 21:10 data
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$
```

Удаляем директорию **data**: hdfs dfs -rm -r data

```
drwxr-xr-x   - hdfs     hdfs          0 2021-03-20 14:25 /atsv2
drwxr-xr-x   - hdfs     hdfs          0 2021-03-20 14:24 /hdp
drwx------   - livy     hdfs          0 2021-03-20 14:31 /livy2-recovery
drwxr-xr-x   - mapred   hdfs          0 2021-03-20 14:24 /mapred
drwxrwxrwx   - mapred   hadoop        0 2021-03-31 21:30 /mr-history
drwxrwxrwx   - spark    hadoop        0 2021-04-08 21:07 /spark2-history
drwxr-xr-x   - hdfs     hdfs          0 2021-04-06 09:11 /system
drwxrwxrwx   - hdfs     hdfs          0 2021-04-07 13:46 /tmp
drwxr-xr-x   - hdfs     hdfs          0 2021-04-02 16:10 /user
drwxr-xr-x   - hdfs     hdfs          0 2021-04-02 08:21 /warehouse
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ ^C
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ pwd
/home/student3_10
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ hdfs dfs -mkdir data
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ ls
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ hdfs dfs -ls
Found 1 items
drwxr-xr-x   - student3_10 student3_10          0 2021-04-08 21:10 data
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ hdfs dfs -rm -r data
21/04/08 21:14:03 INFO fs.TrashPolicyDefault: Moved: 'hdfs://bigdataanalytics2-h
ead-shdpt-v31-1-0.novalocal:8020/user/student3_10/data' to trash at: hdfs://bigd
ataanalytics2-head-shdpt-v31-1-0.novalocal:8020/user/student3_10/.Trash/Current/
user/student3_10/data
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ 
```

Просмотрим директорию: hdfs dfs -ls

```
drwxrwxrwx   - mapred   hadoop        0 2021-03-31 21:30 /mr-history
drwxrwxrwx   - spark    hadoop        0 2021-04-08 21:07 /spark2-history
drwxr-xr-x   - hdfs     hdfs          0 2021-04-06 09:11 /system
drwxrwxrwx   - hdfs     hdfs          0 2021-04-07 13:46 /tmp
drwxr-xr-x   - hdfs     hdfs          0 2021-04-02 16:10 /user
drwxr-xr-x   - hdfs     hdfs          0 2021-04-02 08:21 /warehouse
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ ^C
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ pwd
/home/student3_10
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ hdfs dfs -mkdir data
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ ls
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ hdfs dfs -ls
Found 1 items
drwxr-xr-x   - student3_10 student3_10          0 2021-04-08 21:10 data
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ hdfs dfs -rm -r data
21/04/08 21:14:03 INFO fs.TrashPolicyDefault: Moved: 'hdfs://bigdataanalytics2-h
ead-shdpt-v31-1-0.novalocal:8020/user/student3_10/data' to trash at: hdfs://bigd
ataanalytics2-head-shdpt-v31-1-0.novalocal:8020/user/student3_10/.Trash/Current/
user/student3_10/data
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ ls
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ hdfs dfs -ls
Found 1 items
drwx------   - student3_10 student3_10          0 2021-04-08 21:14 .Trash
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ 
```

Удаляем директорию **.Trash** со всем содержимым: hdfs dfs -rm -r .Trash

Создаем директорию **data** в HDFS и в ней файл **1.txt**: hdfs dfs -touchz data/1.txt

Удаляем директорию **data** со всем содержимым: hdfs dfs -rm -r data

Создаем директорию **data** в HDFS: hdfs dfs -mkdir data

Удаляем директорию **data** со всем содержимым: hdfs dfs -rm -r data

Положить файл **urbandict.csv** в HDFS: scp urbandict.csv student3_10@37.139.32.56:.

Копируем файл **urbandict.csv** в HDFS: **hdfs dfs -put urbandict.csv**

Просмотреть размер: ls -al

Копируем файл в HDFS: hdfs dfs -put **urbandict.csv**

Смотрим размер файла: hdfs dfs -ls -h

фактор репликации 2: hdfs dfs -du -h urbandict.csv

Изменяем файл репликации 3: hdfs dfs -setrep 3 urbandict.csv

Посмотрим все папки в корне HDFS: hdfs dfs -ls -h /

Посмотрим файл: hdfs dfs -ls -h /user/student3_10

Находимся в директории: pwd

Смотрим что фал urbandict.csv хранится в home: ls -al

Скопировать файл urbandict.csv: hdfs dfs -get urbandict.csv urbandict2.csv

Смотрим файл: ls -alh

Удаляем файл urbandict2.csv: rm -rf urbandict2.csv

```
drwx------     4 student3_10 student3_10 4,0K апр 11 07:11 .
drwxr-xr-x. 213 root        root        4,0K апр  1 16:31 ..
-rw-------     1 student3_10 student3_10 3,0K апр 11 06:32 .bash_history
-rw-r--r--     1 student3_10 student3_10   18 авг  8  2019 .bash_logout
-rw-r--r--     1 student3_10 student3_10  193 авг  8  2019 .bash_profile
-rw-r--r--     1 student3_10 student3_10  231 авг  8  2019 .bashrc
drwxrwxr-x     2 student3_10 student3_10 4,0K апр  3 18:35 .oracle_jre_usage
drwx------     2 student3_10 student3_10 4,0K апр  3 18:40 .ssh
-rw-r--r--     1 student3_10 student3_10 514M апр 11 07:11 urbandict2.csv
-rw-rw-r--     1 student3_10 student3_10 514M апр 10 17:39 urbandict.csv
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ rm -rf urbandict2.csv
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ ls -alh
итого 514M
drwx------     4 student3_10 student3_10 4,0K апр 11 07:17 .
drwxr-xr-x. 213 root        root        4,0K апр  1 16:31 ..
-rw-------     1 student3_10 student3_10 3,0K апр 11 06:32 .bash_history
-rw-r--r--     1 student3_10 student3_10   18 авг  8  2019 .bash_logout
-rw-r--r--     1 student3_10 student3_10  193 авг  8  2019 .bash_profile
-rw-r--r--     1 student3_10 student3_10  231 авг  8  2019 .bashrc
drwxrwxr-x     2 student3_10 student3_10 4,0K апр  3 18:35 .oracle_jre_usage
drwx------     2 student3_10 student3_10 4,0K апр  3 18:40 .ssh
-rw-rw-r--     1 student3_10 student3_10 514M апр 10 17:39 urbandict.csv
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$
```

Посмотреть как файл хранится на файловой системе: hdfs fsck **urbandict.csv** -files -bloks -locations

Файл   Правка   Вид   Терминал   Вкладки   Справка

```
drwxrwxr-x    2 student3_10 student3_10 4,0K апр  3 18:35 .oracle_jre_usage
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ hdfs fsck urbandict.csv -files -bloks -locations
fsck: Illegal option '-bloks'
Usage: hdfs fsck <path> [-list-corruptfileblocks | [-move | -delete | -openforwrite] [-files [-blocks [-locat
ions | -racks | -replicaDetails | -upgradedomains]]]] [-includeSnapshots] [-showprogress] [-storagepolicies]
[-maintenance] [-blockId <blk_Id>]
        <path>   start checking from this path
        -move    move corrupted files to /lost+found
        -delete delete corrupted files
        -files   print out files being checked
        -openforwrite   print out files opened for write
        -includeSnapshots       include snapshot data if the given path indicates a snapshottable directory o
r there are snapshottable directories under it
        -list-corruptfileblocks print out list of missing blocks and files they belong to
        -files -blocks  print out block report
        -files -blocks -locations       print out locations for every block
        -files -blocks -racks   print out network topology for data-node locations
        -files -blocks -replicaDetails  print out each replica details
        -files -blocks -upgradedomains  print out upgrade domains for every block
        -storagepolicies        print out storage policy summary for the blocks
        -maintenance     print out maintenance state node details
        -showprogress   show progress in output. Default is OFF (no progress)
        -blockId        print out which file this blockId belongs to, locations (nodes, racks) of this block,
 and other diagnostics info (under replicated, corrupted or not, etc)

Please Note:
        1. By default fsck ignores files opened for write, use -openforwrite to report such files. They are u
sually  tagged CORRUPT or HEALTHY depending on their block allocation status
        2. Option -includeSnapshots should not be used for comparing stats, should be used only for HEALTH ch
eck, as this may contain duplicates if the same file present in both original fs tree and inside snapshots.

Generic options supported are:
-conf <configuration file>      specify an application configuration file
-D <property=value>             define a value for a 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 <file1,...>              specify a comma-separated list of files to be copied to the map reduce clus
ter
-libjars <jar1,...>             specify a comma-separated list of jar files to be included in the classpath
-archives <archive1,...>        specify a comma-separated list of archives to be unarchived on the compute
machines

The general command line syntax is:
command [genericOptions] [commandOptions]

[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$
```
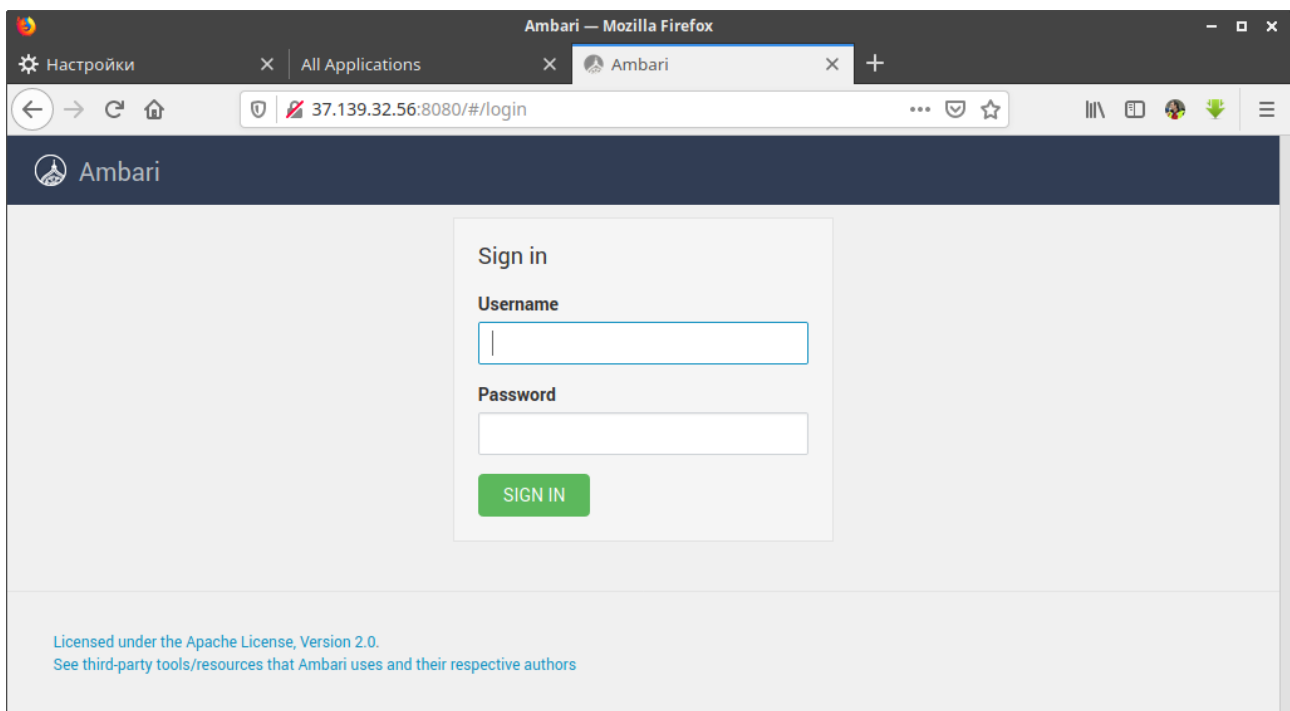
Вывести только начало файла: hdfs dfs -cat **urbandict.csv** | head

Файл   Правка   Вид   Терминал   Вкладки   Справка

```
-archives <archive1,...>                specify a comma-separated list of archives to be unarchived on the compute machines

The general command line syntax is:
command [genericOptions] [commandOptions]

[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ hdfs dfs -cat urbandict.csv | head
word_id,word,up_votes,down_votes,author,definition
0000007,Janky,296,255,dc397b2f,"Undesirable; less-than optimum."
0000008,slumpin',16,37,dc397b2f,"low down and funky, but [knee deep] enough to ride to."
0000009,yayeeyay,19,27,dc397b2f,"affirmation; suggestion of encouragement, approval, or interest."
0000012,hard-core,162,96,d1610749,"anything out of our league that can be good or bad."
0000013,brutal,12,45,40ece1ef,"anything that makes you sweat"
0000014,skanky,9,48,485e4db7,"Anything of or pertaining to a $10,000 hooker."
0000015,ho-bag,26,27,b37fba05,"a term of endearment, used affectionately for your roommate. First used in the schools' park
ing lot after an incident with the hall moniters."
0000016,massive,36,45,b9dcf126,"really really good. excellently good."
0000017,wtf,183,99,a6c97ba3,"what the fuck? ;; use it in place of expletives. a more polite alternative."
cat: Unable to write to output stream.
[student3_10@bigdataanalytics2-head-shdpt-v31-1-0 ~]$
```

Выведения конца файла: hdfs dfs -tail **urbandict.csv**

Изменяем файл репликации 2: hdfs dfs -setrep 2 urbandict.csv

Удаляем файл **urbandict.csv** из HDFS без помещения в корзину: hdfs dfs -rm -skipTrash **urbandict.csv**

Захожу на Ambari

Захожу на YARN

Подключаюсь к Namenode

просмотрим корневую директорию: hdfs dfs -ls /

Создаем директорию в HDFS: hdfs dfs -mkdir /data