

Практическая работа 3-1. Корпоративные платформы. Cloudera (Гадильшина В.Е., БД-231м)

Задание к теме: Работа в HDFS в экосистеме cloudera

3.1.1. Развернуть виртуальное окружение.

+

3.1.2. Вывести с помощью команды help описание основных команды shell-клиента.

```
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ hdfs dfs -help
Usage: hadoop fs [generic options]
    [-appendToFile <localsrc> ... <dst>]
    [-cat [-ignoreCrc] <src> ...]
    [-checksum <src> ...]
    [-chgrp [-R] GROUP PATH...]
    [-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]
    [-chown [-R] [OWNER][:[GROUP]] PATH...]
    [-copyFromLocal [-f] [-p] [-l] <localsrc> ... <dst>]
    [-copyToLocal [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
```

3.1.3. Просмотреть корневую директорию HDFS.

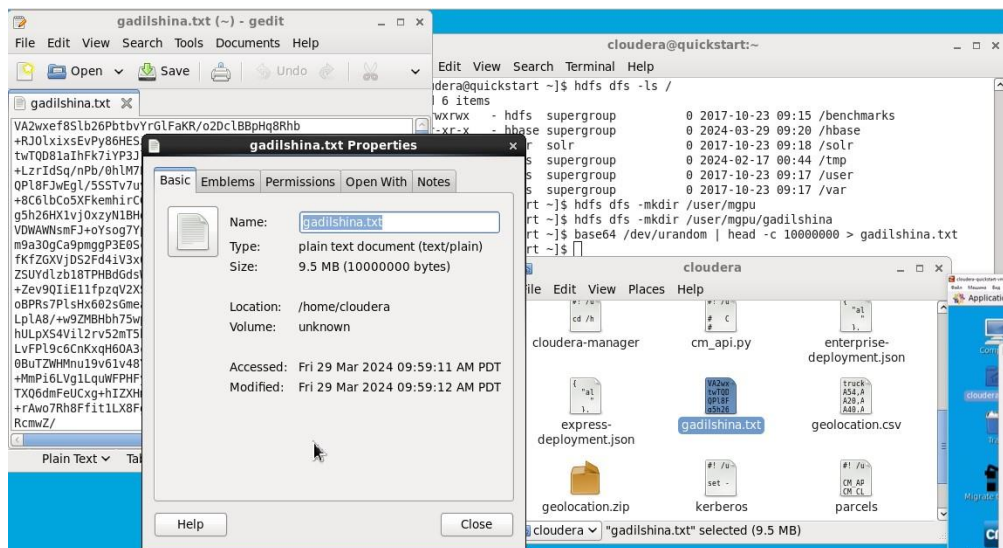
```
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 6 items
drwxrwxrwx - hdfs supergroup          0 2017-10-23 09:15 /benchmarks
drwxr-xr-x - hbase supergroup          0 2024-03-29 09:20 /hbase
drwxr-xr-x - solr solr                  0 2017-10-23 09:18 /solr
drwxrwxrwt - hdfs supergroup           0 2024-02-17 00:44 /tmp
drwxr-xr-x - hdfs supergroup           0 2017-10-23 09:17 /user
drwxr-xr-x - hdfs supergroup           0 2017-10-23 09:17 /var
```

3.1.4. Создать в HDFS в директории /user/mgpu поддиректорию ваше_фио.

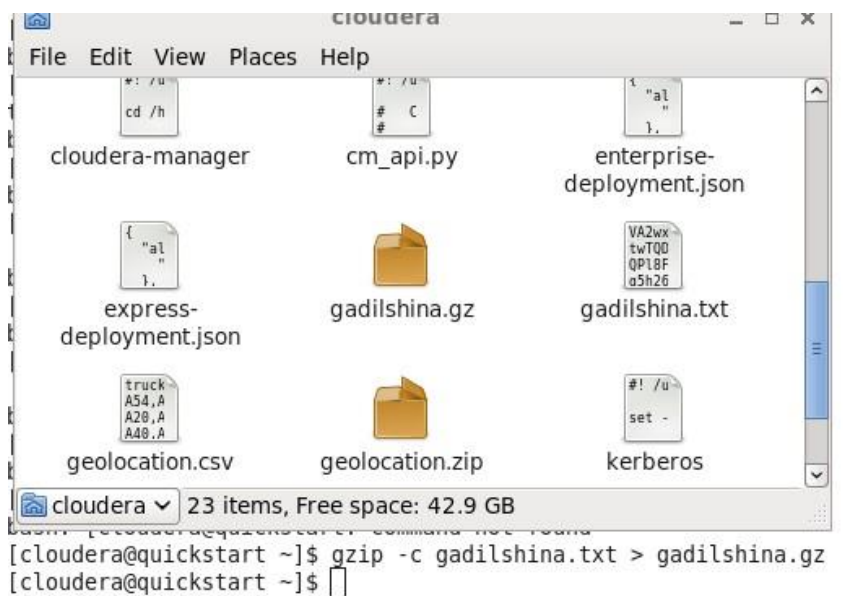
```
[cloudera@quickstart ~]$ hdfs dfs -mkdir /user/mgpu
[cloudera@quickstart ~]$ hdfs dfs -mkdir /user/mgpu/gadilshina
[cloudera@quickstart ~]$
```

3.1.5. Создать в локальной файловой системе случайный текстовый файл размером 10 Mb с именем, образованным вашими инициалами base64 /dev/urandom | head -c 10000000 > file.txt

```
[cloudera@quickstart ~]$ base64 /dev/urandom | head -c 10000000 > gadilshina.txt
[cloudera@quickstart ~]$ ls
cloudera-manager  eclipse                geolocation.zip  parcels          Videos
cm_api.py         enterprise-deployment.json  kerberos         Pictures         workspace
Desktop          express-deployment.json   lib              Public
Documents        gadilshina.txt           __MACOSX        Templates
Downloads        geolocation.csv          Music            trucks.csv
```



3.1.6. Заархивировать созданный текстовый файл `gzip -c file.txt > file.gz`.



3.1.7. Скопировать текстовый файл и архив в директорию `/user/mgpu/fio` HDFS виртуальной машины.

Home / user / mgpu / gadilshina

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	↑		cloudera	supergroup	drwxr-xr-x	March 29, 2024 09:56 AM
<input type="checkbox"/>	.		cloudera	supergroup	drwxr-xr-x	March 29, 2024 10:10 AM
<input type="checkbox"/>	gadilshina.gz	7.2 MB	cloudera	supergroup	-rw-r--r--	March 29, 2024 10:10 AM
<input type="checkbox"/>	gadilshina.txt	9.5 MB	cloudera	supergroup	-rw-r--r--	March 29, 2024 10:09 AM

Show 45 of 2 items Page 1 of 1

3.1.8. Просмотреть файл и архив с помощью утилит cat, text в комбинации с каналами и утилитами head, tail -- привести не менее 3 вариантов команд и просмотра файла.

```
[cloudera@quickstart ~]$ cat gadilshina.txt
```

```
KprashoTCqIBzzdxZ0XN7rffXYBu4k6UNmljKvYJCyckwmMk80WgrDuEu+EjbnvS6PxIACtDo0iS
uNezk2Lkq3r2DraC99XmhY/UdL00S0szdgUCeFn7wofHZPga878C4sJ2ErMeQh00CP26oF0L/psx
TyQ5zu7GWUyawxKHGLi9CwNgdyZnWweJsF17kHFcTw0k0xBs1CTviLkmB0MxyVSUDtNwfs0jNA31
99yH0Lur/M5xBcwZxGCWF6HXXN00FZwIjIVqP0xJwLYvR7B7Q7AdqfSig/INHnToGE7jXUefvhi
vdJ2po6lIISud89sRw4nNibAIWlH+o9SWD01mH6Pso0QmTDJmZAUbjH8c4cPKUQPvFKIx0GXckk7
eM8Y6elWFmgBVQf043pnrzkd8pMQ1EtPrI2Bilz4y0Qxg45ahiGrRdWRUVSkbmEyK71/kj5QfKYh
KFa05UW4sW39fu4gdA+ilFiwx4U/gV0Dk8opP/100pW5r6Hjje2yyFUEIuWNfSwqF8ZtEwREnrRn
t4CisqwhsYqYslpTq6txGx0peeJ60Vs9G9a8qRxa70aZF00d6Hks1HJpVxrdRPQzGcQ96jBhj
21cTx0srG84tG9JX20zhC0niW1CE2iGz+jivK35sssVYzIkGEXkwc6AFw5qzre+uPQwNZv14bgjg
etCHn2lQNXTyqA7MrwvvpJpqrW8dttZhIUxWZCLPj+Je15FgifD97xXLQS8P8d0cx77PakE0D8qB
yl0SwLggQeVYB7auCT6CvUmq+TlwYmD8nYNFWbjqG91lyrXXerI1ZAQy/uSlW0eLiVrzHpq7vpwV
w9iCTBQfa399PrVbTx4YjcknCh7TeXNvZThdWI+vm/wNCFIjNA+BXl/vI8RNBfvvWP3/YvJ82U1B
jJdgmGIGYVJ6sjkgzIaVkbNSF0Uicf4zn10JxtKN0+KaY9Igf5gA0rJRrdexEfpvd0uhU6WYl0oN
k8fSMs90Unvd7GyZm4zqrP/oBf5NQpa3Uu4RD+gQlTrFRq9jtx/MjId2BJo0fR25QZ1pJubP/2Pk
HwrQGH4imTLQa0r+LfpNdLpThkTlCs1oJ0zHiQXiWlqFCPwV6qj35/Wc/bF0P9pMVHLUNbCh/r
W+CMK/cWIf4bzcVcQ+uNP0x9d7UYi/09zKbvFQuNR9Fw5rq37e0MaPqgHDXMPiTNvPIjymipR3Zy
QksVY8YRe+hd/bynRk093/elkaGApr7DVSe4D7rVyb1hlCJT5b3Vhxs8LnlFK0dI4WNfM8sXZT1Z
OThQC60W9j4s3h9WgkVQzg6ovHBkhaXp3V55ZbbdFKIG3xT6oQ7GJVvWB3PHlMYkCuzE3zyVaLzu
BpuNvFa6WVRPE5GHstJxzzV0G8q6i6kLw99Y0TJbuN/8J2X0vAUP32XCRQjycrwxISFrVwRK+Jp6
D3iM19PXB178AHikW/7WNsepgCeyoSDcl8XuRlyHDctWRTzKBzmniFa7l0M8THP2NDxWaFrJeqjy
Au3Bf4XJXV2Rb/lhKla89C58UeLI+01QpEaY7hYEJEH8Gs1082GKQVdrSjkoIQYhRaPutLMk7vDZ
m84wwkJYzP0DK7gaJlaJI7Z0SrDniVDMFpsjes4vMXo+ltgFXzFwR82w30xfBlusMMoApgybsGMe
BqSgDnL40TnAdDnBDZPhnprPbw7t7TJzCq10Xlwh9e00hz4XnopMP1LlaP2UdLtraejhEwtTYA3f
dT3//m0IYAPEGKkiGiO/Uu+e47xkBgR0C2AuQLXKasCPSk8JyLXmqS2io2o1e8PhUY4qKiH5F1c
```

```
File Edit View Search Terminal Help
```

```
[cloudera@quickstart ~]$ cat -n gadilshina.txt
```

```
File Edit View Search Terminal Help
```

```
129360 KprashoTCqIBzzdxZ0XN7rffXYBu4k6UNmljKvYJCyckwmMk80WgrDuEu+EjbnvS6PxIACtDo0iS
129361 uNezk2Lkq3r2DraC99XmhY/UdL00S0szdgUCeFn7wofHZPga878C4sJ2ErMeQh00CP26oF0L/psx
129362 TyQ5zu7GWUyawxKHGLi9CwNgdyZnWweJsF17kHFcTw0k0xBs1CTviLkmB0MxyVSUDtNwfs0jNA31
129363 99yH0Lur/M5xBcwZxGCWF6HXXN00FZwIjIVqP0xJwLYvR7B7Q7AdqfSig/INHnToGE7jXUefvhi
129364 vdJ2po6lIISud89sRw4nNibAIWlH+o9SWD01mH6Pso0QmTDJmZAUbjH8c4cPKUQPvFKIx0GXckk7
129365 eM8Y6elWFmgBVQf043pnrzkd8pMQ1EtPrI2Bilz4y0Qxg45ahiGrRdWRUVSkbmEyK71/kj5QfKYh
129366 KFa05UW4sW39fu4gdA+ilFiwx4U/gV0Dk8opP/100pW5r6Hjje2yyFUEIuWNfSwqF8ZtEwREnrRn
129367 t4CisqwhsYqYslpTq6txGx0peeJ60Vs9G9a8qRxa70aZF00d6Hks1HJpVxrdRPQzGcQ96jBhj
129368 21cTx0srG84tG9JX20zhC0niW1CE2iGz+jivK35sssVYzIkGEXkwc6AFw5qzre+uPQwNZv14bgjg
129369 etCHn2lQNXTyqA7MrwvvpJpqrW8dttZhIUxWZCLPj+Je15FgifD97xXLQS8P8d0cx77PakE0D8qB
129370 yl0SwLggQeVYB7auCT6CvUmq+TlwYmD8nYNFWbjqG91lyrXXerI1ZAQy/uSlW0eLiVrzHpq7vpwV
129371 w9iCTBQfa399PrVbTx4YjcknCh7TeXNvZThdWI+vm/wNCFIjNA+BXl/vI8RNBfvvWP3/YvJ82U1B
129372 jJdgmGIGYVJ6sjkgzIaVkbNSF0Uicf4zn10JxtKN0+KaY9Igf5gA0rJRrdexEfpvd0uhU6WYl0oN
```

```
[cloudera@quickstart ~]$ cat -b gadilshina.txt
```



```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
129360 KprashoTCqIBzzdxZ0XN7rffXYBu4k6UNmLjKvYJCyckwmMk80WgrDuEu+EjbnvS6PxIAcTDo0iS  
129361 uNezk2Lkq3r2DraC99XmhY/UdL00S0szdgUCeFn7wofHZPga878C4sJ2ErMeQh00CP26oFOL/psx  
129362 TyQ5zu7GWUyaxXKHGLi9CwNgdyZnWweJsF17KHFcTw0k0xBs1CTviLkmB0MxyVSUDtNwfs0jNA31  
129363 99yH0LUr/M5xBcwWZxGCWF6HXXN00FZwIJIvQp0xJwLYvR7B7Q7AdqfSig/INHnToGE7jXUefvhi  
129364 vdJ2po61IISUd89sRw4nNibAIWlH+o9SWD01mH6Pso0QmTDJmZAUBJH8c4cPKUQPvFKIx0GXckk7  
129365 eM8Y6e1WFmgBVQf043pnrzkd8pMQ1EtPrI2Bi1z4y0Qxg45ahiGrRdWRUVskbmEyk71/kj5QfKYh  
129366 KFa05UW4sW39fu4gdA+i1Fiwx4U/gV0Dk8opP/100pW5r6Hjje2yyFueIuWNfSwqF8ZtEwEnrRn  
129367 t4GicqwhsYqYclnTg6tvGv0nqa160Vc9G9a8cRvuAur70a7E00d6Hke1H1nVvrdR0n7Gc006iRhi
```

```
[cloudera@quickstart ~]$ head gadilshina.txt
```

```
[cloudera@quickstart ~]$ head gadilshina.txt  
VA2wxef8S1b26PbtbvYrGlFaKR/o2DclBBpHq8Rhb+RJ0lxixsEvPy86HESJLFztpJPcTIDTmuuu  
twTQD81aIhFk7iYP3JLIa7ftQBJ+AADlcAI7rLgc39AtTnL6xl+LzrIdSq/nPb/0hLM7PL8hVklL  
QPl8FJwEgl/5SSTv7uya4W6bA0Q0+Imm2aHmajdIyX+8C6lbCo5XfkemhirCGl0AI0pzFsXfgR2i  
g5h26HX1vj0xzyN1BHdf0V9NPAGRg4LuIO/2zKxi1IOlqLU7BMZQL/VDWAWNsmFJ+oYsog7YpzLI  
m9a30gCa9pmggP3E0Sc+d47QLViUioviEFXVKW0DVLV/ByvQo/fKfZGXVjDS2Fd4iV3x0W9hzXmz  
ZSUYdlzb18TPHBdGdsWuQcC2xCbXlp+g3q+Zev9QIiE1lfpzqV2XSNUUCN7YzhNxmNRBx8nnJymC  
oBPRs7PlsHx602sGmea5VrV1FKv0CN4vBbjZLA4ZYWbHYCVumSahkG3Fw85RQaqTP99bcxCiU0HG  
LplA8/+w9ZMBHbh75wp7ihF/zg/hULpXS4Vil2rv52mT5hr5Bi5BaUbtxEkMLHk1Sg6USENGYH5U  
LvFPL9c6CnKxqH60A3c7Ueb6fn2hYW7ypoI/6r8qm4N6cyc4GHYfMD9KUdwbXWHHskdDMIQdt8RQ  
0BuTZWHMnu19v61v48YnluEE3Ds+MmPi6LVq1LquWFPHFynQDdih1HidlK+RyRam134muk6B8qNY
```

```
[cloudera@quickstart ~]$ head -n 5 gadilshina.txt
```

VA2wx

```
VA2wxef[cloudera@quickstart ~]$ head -n 5 gadilshina.txt  
VA2wxef8S1b26PbtbvYrGlFaKR/o2DclBBpHq8Rhb+RJ0lxixsEvPy86HESJLFztpJPcTIDTmuuu  
twTQD81aIhFk7iYP3JLIa7ftQBJ+AADlcAI7rLgc39AtTnL6xl+LzrIdSq/nPb/0hLM7PL8hVklL  
QPl8FJwEgl/5SSTv7uya4W6bA0Q0+Imm2aHmajdIyX+8C6lbCo5XfkemhirCGl0AI0pzFsXfgR2i  
g5h26HX1vj0xzyN1BHdf0V9NPAGRg4LuIO/2zKxi1IOlqLU7BMZQL/VDWAWNsmFJ+oYsog7YpzLI  
m9a30gCa9pmggP3E0Sc+d47QLViUioviEFXVKW0DVLV/ByvQo/fKfZGXVjDS2Fd4iV3x0W9hzXmz
```

```
[cloudera@quickstart ~]$ tail gadilshina.txt  
XlPUQHq+h/QG0i86N4Q8NchXJScp/XQGNnZbz5VTh2U6lg1CiNhD3nkdLY23XR54xfeN135VfAY6  
QFaU+QDg3b0hRcLRK2X8HtBpIb7ETKQy60ncxSUGtKvbT67v8e+lRFwpWlrMn0YiHK8Pf6dwzYB6  
0kTYaLHNw1J9gLYEmY64VTDEPZ5E0j1ZmpQDAul3i25C45EH8FBQE0sEFy+kJY0cqxdrrSXG/eVq  
X7GgpzekT/N4HgikeeDuS/Kzlf8q5VP1+EtUlCdnsEpPE5/GNyXDhREfzDfjxmT2LlQAmoUkZxtZ  
79AkSvk8S5M0AZwAdbBueb0mnvr2i7CcJRcl9JkkWRpZUSIYqS3UaK+bojCrsa0t8/Y9YS0MMVHo  
lhivALEvPEcCHM10YXQMp8davlModQdLVtZ6mSsle00Yk2GPVSgjKJIbZElmw0EhLetuB2dER6vH  
5C+7r6NgTE0lleEBCVe5B1lJQnyb8iA0bNxUxi64+0nt/0QrCAzY6XAlj5sLkDThHaytZs3JQsb8  
4j2H1Yyq9f0zh4HVWasnU/dgGsRINMQiSwf4cWk5lWvN5b0w64Mt2J3iDe+uD9MYm9guU76RJKgk  
mD/ltjPQCzJXjYjokjOWHuL2D9MQgMgBxXgzRe+geRubhWvsGhdCuW6yrJdWk+Lwt3ukjYUVhE9z  
jURg/L8hr+[cloudera@quickstart ~]$
```

```
File Edit View Search Terminal Help  
[cloudera@quickstart ~]$ tail -c 5 gadilshina.txt  
L8hr+[cloudera@quickstart ~]$
```

```
L8hr+[cloudera@quickstart ~]$ tail -n 5 gadilshina.txt
\hivALevPEcCHM10YXQMp8davlModQdLVtZ6mSsle00Yk2GpVsgjKJIbzElmw0EhLetuB2dER6vH
5C+7r6NgtE0lleEBCVe5B1lJQnyb8iA0bNxUxi64+0Nt/0QrCAzY6XAlj5sLkDTHAyZs3JQsb8
4j2H1Yyq9f0zh4HVWasnU/dgGsRINMQiSwf4cWk5lwvN5b0w64Mt2J3iDe+uD9MYm9guU76RJKgk
mD/1tjPQCzJXjYjokjOWHuL2D9MQgMgBxXgzRe+geRubhWvsGhdCuW6yrJdWk+Lwt3ukjYUVhE9z
jURg/L8hr+[cloudera@quickstart ~]$
```

3.1.9. Создать копию файла file.txt вида date_file.txt, где в начале имени файла-копии указана текущая дата. Вывести листинг.

```
[cloudera@quickstart ~]$ cp gadilshina.txt 29.03.2024_gadilshina.txt
[cloudera@quickstart ~]$ ls
29.03.2024_gadilshina.txt  Downloads          gadilshina.txt      _MACOSX  Templates
cloudera-manager         eclipse            geolocation.csv     Music     trucks.csv
cm_api.py                enterprise-deployment.json geolocation.zip     parcels   Videos
Desktop                  express-deployment.json  kerberos            Pictures  workspace
Documents                gadilshina.gz        lib                 Public
[cloudera@quickstart ~]$
```

```
[cloudera@quickstart ~]$ hdfs dfs -cp /user/mgpu/gadilshina/gadilshina.txt /user/mgpu/gadilshina/29.03.2024_gadilshina.txt
[cloudera@quickstart ~]$ hdfs dfs -ls /user/mgpu/gadilshina/
Found 3 items
-rw-r--r-- 1 cloudera supergroup 10000000 2024-03-29 10:58 /user/mgpu/gadilshina/29.03.2024_gadilshina.txt
-rw-r--r-- 1 cloudera supergroup 7599428 2024-03-29 10:10 /user/mgpu/gadilshina/gadilshina.gz
-rw-r--r-- 1 cloudera supergroup 10000000 2024-03-29 10:09 /user/mgpu/gadilshina/gadilshina.txt
[cloudera@quickstart ~]$
```

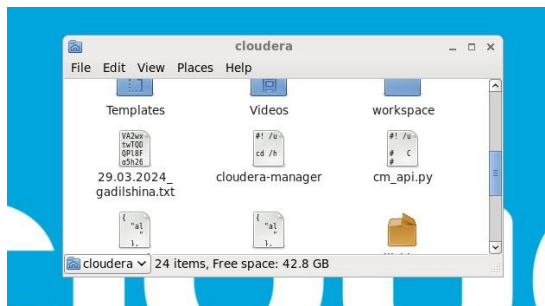
3.1.10. Вывести статистику по директории /user/mgpu/fo виртуальной машины.

```
[cloudera@quickstart ~]$ hdfs dfs -cp /user/mgpu/gadilshina/gadilshina.txt /user/mgpu/gadilshina/29.03.2024_gadilshina.txt
[cloudera@quickstart ~]$ hdfs dfs -ls /user/mgpu/gadilshina/
Found 3 items
-rw-r--r-- 1 cloudera supergroup 10000000 2024-03-29 10:58 /user/mgpu/gadilshina/29.03.2024_gadilshina.txt
-rw-r--r-- 1 cloudera supergroup 7599428 2024-03-29 10:10 /user/mgpu/gadilshina/gadilshina.gz
-rw-r--r-- 1 cloudera supergroup 10000000 2024-03-29 10:09 /user/mgpu/gadilshina/gadilshina.txt
[cloudera@quickstart ~]$
```

Home / user / mgpu / gadilshina

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>			cloudera	supergroup	drwxr-xr-x	March 29, 2024 09:56 AM
<input type="checkbox"/>			cloudera	supergroup	drwxr-xr-x	March 29, 2024 10:58 AM
<input type="checkbox"/>	 29.03.2024_gadilshina.txt	9.5 MB	cloudera	supergroup	-rw-r--r--	March 29, 2024 10:58 AM
<input type="checkbox"/>	 gadilshina.gz	7.2 MB	cloudera	supergroup	-rw-r--r--	March 29, 2024 10:10 AM
<input type="checkbox"/>	 gadilshina.txt	9.5 MB	cloudera	supergroup	-rw-r--r--	March 29, 2024 10:09 AM

Show 45 of 3 items Page 1 of 1



3.1.11. Удалить поддиректорию /fo со всем содержимым.


```
-rw-r--r-- 1 cloudera supergroup 10000000 2024-03-29 10:09 /u:
[cloudera@quickstart ~]$ hdfs dfs -rm -R /user/mgpu/gadilshina
Deleted /user/mgpu/gadilshina
[cloudera@quickstart ~]$
```

Home / user / mgpu

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	↓		hdfs	supergroup	drwxr-xr-x	March 29, 2024 09:55 AM
<input type="checkbox"/>	↓		cloudera	supergroup	drwxr-xr-x	March 29, 2024 11:04 AM

Show 45 of 0 items Page 1 of 1

3.1.12. Подсчитать количество слов в файле внутри HDFS с помощью методологии Map Reduce (размер файла не менее 128 Мб).

```
[cloudera@quickstart ~]$ yarn jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar wordcount /user/mgpu/gadilshina/3.1.12_gadilshina_VE.txt /user/mgpu/gadilshina/output
24/03/29 11:15:32 INFO client.RMProxy: Connecting to ResourceManager at 0.0.0.0:8032
24/03/29 11:15:33 INFO mapreduce.JobSubmitter: Cleaning up the staging area /tmp/hadoop-yarn/staging/cloudera/.staging/job_1711729182730_0001
24/03/29 11:15:33 WARN security.UserGroupInformation: PrivilegedActionException as:cloudera (auth:SIMPLE) cause:org.apache.hadoop.mapreduce.lib.input.InvalidInputException: Input path does not exist: hdfs://quickstart.cloudera:8020/user/mgpu/gadilshina/3.1.12_gadilshina_VE.txt
org.apache.hadoop.mapreduce.lib.input.InvalidInputException: Input path does not exist: hdfs://quickstart.cloudera:8020/user/mgpu/gadilshina/3.1.12_gadilshina_VE.txt
    at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.singleThreadedListStatus(FileInputFormat.java:323)
    at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.listStatus(FileInputFormat.java:265)
    at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.getSplits(FileInputFormat.java:367)
    at org.apache.hadoop.mapreduce.JobSubmitter.writeNewSplits(JobSubmitter.java:365)
    at org.apache.hadoop.mapreduce.JobSubmitter.writeSplits(JobSubmitter.java:322)
    at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:208)
    at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1307)
    at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1304)
    at java.security.AccessController.doPrivileged(Native Method)
    at javax.security.auth.Subject.doAs(Subject.java:415)
    at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1917)
    at org.apache.hadoop.mapreduce.Job.submit(Job.java:1304)
    at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1325)
    at org.apache.hadoop.examples.WordCount.main(WordCount.java:87)
    at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
    at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
    at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:606)
    at org.apache.hadoop.util.ProgramDriver$ProgramDescription.invoke(ProgramDriver.java:71)
    at org.apache.hadoop.util.ProgramDriver.run(ProgramDriver.java:144)
    at org.apache.hadoop.examples.ExampleDriver.main(ExampleDriver.java:74)
    at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
    at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
    at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:606)
    at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
    at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
```

3.2. Создание таблицы в Hive

1. Скачать [датасет](#) или [тут](#)

```
[cloudera@quickstart ~]$ wget https://github.com/BosenkoTM/cloudera-quickstart/blob/main/data/athlete.snappy.parquet
--2024-03-29 11:45:30-- https://github.com/BosenkoTM/cloudera-quickstart/blob/main/data/athlete.snappy.parquet
Resolving github.com... 140.82.121.3
Connecting to github.com|140.82.121.3|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [text/html]
Saving to: "athlete.snappy.parquet"

[ <=> ] 147,480 --.-K/s in 0.1s

2024-03-29 11:45:31 (1.26 MB/s) - "athlete.snappy.parquet" saved [147480]
```

```
[cloudera@quickstart ~]$ ls
29.03.2024 gadilshina.txt cm_api.py enterprise-deployment.json geolocation.zip parcels Videos
3.1.12_gadilshina_SV.txt Desktop express-deployment.json kerberos Pictures workspace
3.1.12_gadilshina_VE.txt Documents gadilshina.gz lib Public Templates
athlete.snappy.parquet Downloads gadilshina.txt __MACOSX Music trucks.csv
cloudera-manager eclipse geolocation.csv
```

2. Через HUE загрузите файл в папку /user/cloudera/athlete.

+

3. В навигационном меню выберите Files.\

+

4. Создайте

Home / user / cloudera

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	↑		hdfs	supergroup	drwxr-xr-x	March 29, 2024 09:55 AM
<input type="checkbox"/>	.		cloudera	cloudera	drwxr-xr-x	March 29, 2024 11:54 AM
<input type="checkbox"/>	athlete		cloudera	cloudera	drwxr-xr-x	March 29, 2024 11:54 AM
<input type="checkbox"/>	data		cloudera	cloudera	drwxr-xr-x	March 29, 2024 09:47 AM
<input type="checkbox"/>	geolocation.csv	514.3 KB	cloudera	cloudera	-rw-r--r--	March 29, 2024 09:42 AM

5. Загрузите файл в HDFS, нажав Upload.

Home / user / cloudera / athlete

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	↑		cloudera	cloudera	drwxr-xr-x	March 29, 2024 11:54 AM
<input type="checkbox"/>	.		cloudera	cloudera	drwxr-xr-x	March 29, 2024 11:54 AM
<input type="checkbox"/>	athlete.snappy.parquet	144.0 KB	cloudera	cloudera	-rw-r--r--	March 29, 2024 11:52 AM

Show 45 of 1 itemsPage 1 of 1

6. Перейдите в “Editor > Hive” и выполните запрос:

Impala

0s default text

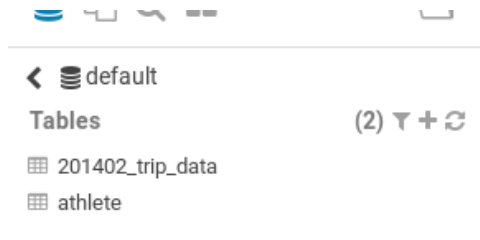
```
1 CREATE EXTERNAL TABLE athlete (  
2   ID INT,  
3   Name STRING,  
4   Sex STRING,  
5   Age INT,  
6   Height INT,  
7   Weight INT,  
8   Team STRING,  
9   NOC STRING,  
10  Games STRING,  
11  `Year` INT,  
12  Season STRING,  
13  City STRING,  
14  Sport STRING,  
15  Event STRING,  
16  Medal STRING  
17 )  
18 STORED AS PARQUET  
19 LOCATION '/user/cloudera/athlete'
```

Success.

Query History

a minute ago

CREATE EXTERNAL TABLE athlete (ID INT, Name STRING, Sex STRING, Age INT, Height INT, Weight INT, Team STRING, NOC STRING, Games STRING, `Year` INT, Season STRING, City STRING, Sport STRING, Event STRING, Medal STRING) STORED AS PARQUET LOCATION '/user/cloudera/athlete'



3.3 Проанализировать и визуализировать данные с помощью Impala (высокоскоростной механизм запросов SQL) или Hive.

- Загрузить и разархивировать babs_open_data_year_1.zip.

```
[cloudera@quickstart ~]$ wget https://community.cloudera.com/xgkfq28377/attachments/xgkfq28377/Questions/87306/1/babs_open_data_year_1.zip
--2024-03-29 12:38:13-- https://community.cloudera.com/xgkfq28377/attachments/xgkfq28377/Questions/87306/1/babs_open_data_year_1.zip
Resolving community.cloudera.com... 143.204.237.55, 143.204.237.25, 143.204.237.129, ...
Connecting to community.cloudera.com|143.204.237.55|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 74878 (73K) [application/zip]
Saving to: "babs_open_data_year_1.zip"

100%[=====>] 74,878 --K/s in 0.04s

2024-03-29 12:38:15 (1.74 MB/s) - "babs_open_data_year_1.zip" saved [74878/74878]
```

```
[cloudera@quickstart ~]$ ls
29.03.2024_gadilshina.txt  babs_open_data_year_1.zip  Downloads  gadilshina.txt  __MACOSX  Templates
3.1.12_gadilshina_SV.txt  cloudera-manager          eclipse     geolocation.csv  Music      trucks.csv
3.1.12_gadilshina_VE.txt  cm_api.py                 enterprise-deployment.json  geolocation.zip  parcels    Videos
athlete.snappy.parquet    Desktop                  express-deployment.json    kerberos         Pictures   workspace
babs_open_data_year_1     Documents                gadilshina.gz             lib              Public

[cloudera@quickstart ~]$ unzip babs_open_data_year_1.zip
Archive:  babs_open_data_year_1.zip
  creating: 201402_babs_open_data/
  inflating: 201402_babs_open_data/201402_station_data.csv
  inflating: 201402_babs_open_data/201402_status_data.csv
  inflating: 201402_babs_open_data/201402_trip_data.csv
  inflating: 201402_babs_open_data/201402_weather_data.csv
  inflating: 201402_babs_open_data/README.txt
  creating: 201408_babs_open_data/
  inflating: 201408_babs_open_data/201408_station_data.csv
  inflating: 201408_babs_open_data/201408_status_data.csv
  inflating: 201408_babs_open_data/201408_trip_data.csv
  inflating: 201408_babs_open_data/201408_weather_data.csv
  inflating: 201408_babs_open_data/README.txt
```

- Перенести данные 201402_trip_data.csv в HDFS.

Home / user / cloudera

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	↓		hdfs	supergroup	drwxr-xr-x	March 29, 2024 09:55 AM
<input type="checkbox"/>	.		cloudera	cloudera	drwxr-xr-x	March 29, 2024 12:53 PM
<input type="checkbox"/>	201402_trip_data.csv	16.4 MB	cloudera	cloudera	-rw-r--r--	March 29, 2024 12:53 PM
<input type="checkbox"/>	athlete		cloudera	cloudera	drwxr-xr-x	March 29, 2024 11:54 AM
<input type="checkbox"/>	data		cloudera	cloudera	drwxr-xr-x	March 29, 2024 09:47 AM
<input type="checkbox"/>	geolocation.csv	514.3 KB	cloudera	cloudera	-rw-r--r--	March 29, 2024 09:42 AM

Show 45 of 4 items Page 1 of 1

- Создать таблицу в Hive с привязкой к внешним данным 201402_trip_data.csv.

Os default text

```

1 CREATE EXTERNAL TABLE 201402_trip_data (
2   TripID INT,
3   Duration INT,
4   StartDate STRING,
5   startstation STRING,
6   StartTerminal INT,
7   EndDate STRING,
8   endstation STRING,
9   EndTerminal INT,
10  Bike INT,
11  SubscriptionType STRING,
12  ZipCode STRING
13 )
14 ROW FORMAT DELIMITED
15 FIELDS TERMINATED BY ','
16 LOCATION '/user/cloudera/trip_data'

```

Success.

Query History Saved Queries

a few seconds ago

```

CREATE EXTERNAL TABLE 201402_trip_data ( TripID INT, Duration INT, StartDate STRING, startstation STRING, StartTerminal INT, EndDate STRING, endstation STRING, EndTerminal INT, Bike INT, SubscriptionType STRING, ZipCode STRING ) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/cloudera/trip_data'

```

default

Tables (2)

- 201402_trip_data
- athlete

- **выполнить запрос**

```

select `startstation`, `endstation`, count(*) as trips
from `default`.`201402_trip_data`
group by `startstation`, `endstation`
order by trips desc;

```

```

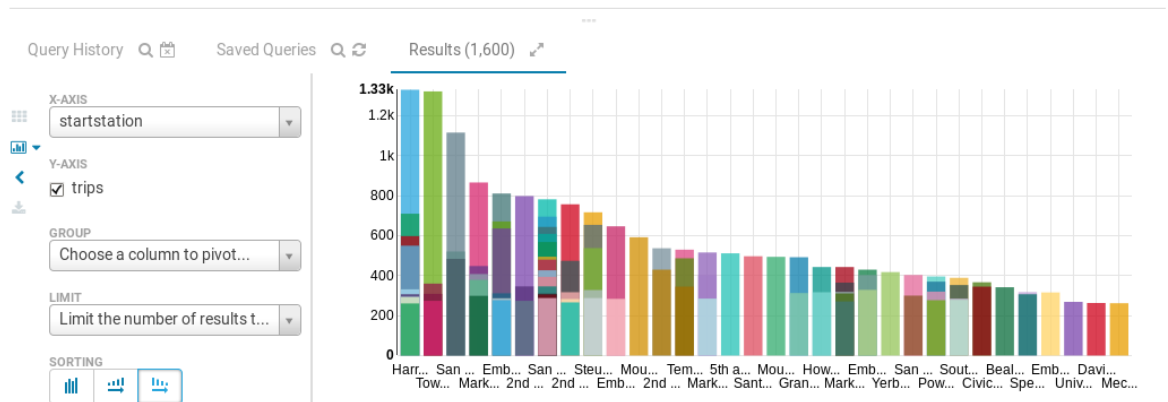
1 select `startstation`, `endstation`, count(*) as trips
2 from `default`.`201402_trip_data`
3 group by `startstation`, `endstation`
4 order by trips desc;

```

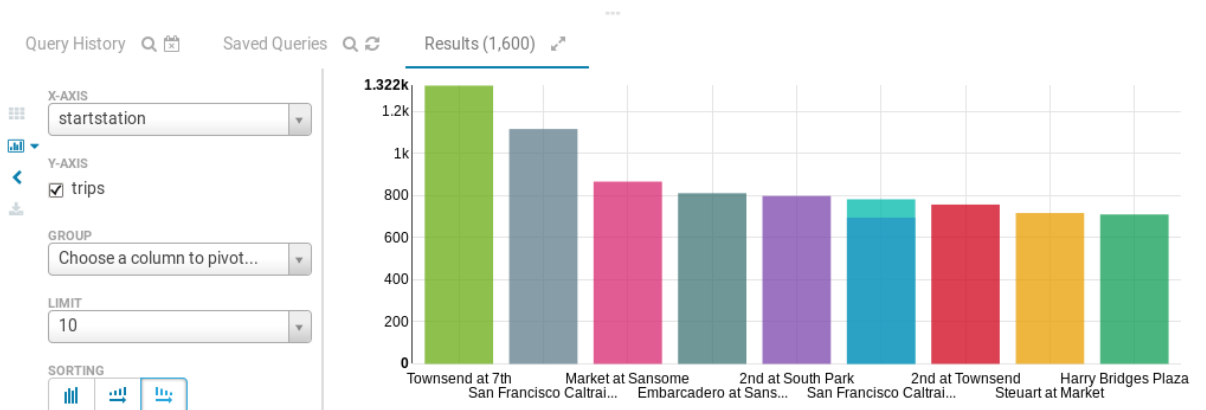
Query History Saved Queries Results (1,024+)

	startstation	endstation	trips
1	Harry Bridges Plaza (Ferry Building)	Embarcadero at Sansome	1330
2	Townsend at 7th	San Francisco Caltrain (Townsend at 4th)	1322
3	San Francisco Caltrain 2 (330 Townsend)	Townsend at 7th	1116
4	Market at Sansome	2nd at South Park	866
5	Embarcadero at Sansome	Steuart at Market	811
6	2nd at South Park	Market at Sansome	798
7	San Francisco Caltrain (Townsend at 4th)	Harry Bridges Plaza (Ferry Building)	782
8	2nd at Townsend	Harry Bridges Plaza (Ferry Building)	757
9	Steuart at Market	Embarcadero at Sansome	717
10	Harry Bridges Plaza (Ferry Building)	2nd at Townsend	710
11	San Francisco Caltrain (Townsend at 4th)	Embarcadero at Folsom	695
12	Embarcadero at Sansome	Harry Bridges Plaza (Ferry Building)	671
13	Steuart at Market	2nd at Townsend	654
14	Embarcadero at Folsom	San Francisco Caltrain (Townsend at 4th)	647

- Создать гистограмму, щелкнув значок «Hue Bar»:



- Установить ось X в качестве начальной станции, а ось Y — в качестве маршрута. Установить лимит 10.



- Выгрузить результаты, выбрав CSV или Excel.

