

**Федеральное агентство связи
Ордена трудового Красного Знамени
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский технический университет связи и информатики»**



**Практическая работа № 1
По дисциплине
Введение в большие данные**

Группа: МБД2431
ФИО: Киреев Артём Александрович

Москва, 2025

Цель работы: получить навыки работы с файловой системой HDFS.

Pipeline IPA_VPN

Для этой сборки необходимы следующие параметры:

Login

Обязательно к заполнению! Введите логин в виде ivanov_bvt-21-1 *Обязательно к заполнению!*

Givenname

Обязательно к заполнению! Введите Своё имя (Иван) *Обязательно к заполнению!*

Surname

Обязательно к заполнению! Введите свою фамилию (Иванов) *Обязательно к заполнению!*

Password

Обязательно к заполнению! Придумайте транспортный пароль. Вы смените его при первом подключении к кластеру

 **Собрать**

Cancel

Рисунок 1 - Создание сборки Pipeline

Скачиваем клиент для создания OpenVPN-подключения и загружаем *.ovpn-файл с почты и установите VPN-соединение.

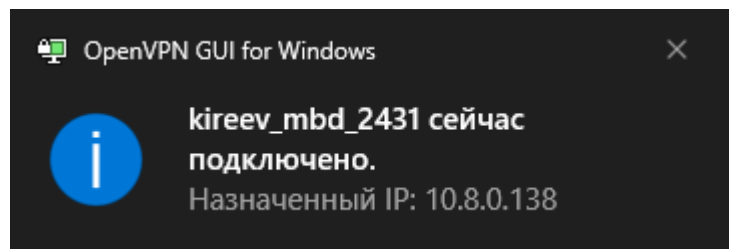


Рисунок 2 - Подключение по OpenVPN

Залогиньтесь на хост по протоколу ssh

```
ssh kireev_mbd_2431@192.168.0.5
```

```
PS C:\Users\User> ssh kireev_mbd_2431@192.168.0.5
(kireev_mbd_2431@192.168.0.5) Password:
Last login: Sun Apr  6 19:58:17 2025 from 192.168.0.3
-sh-4.2$
```

Рисунок 3 - Подключение по ssh

Опробуем консольные утилиты для работы с кластером через HDFS Shell

Напишите команды, с помощью которых выполняли действия и скриншоты с результатами.

1. Создать локально на сервере тестовый файл

```
echo "test text" >> test
```

2. Увеличить размер файла, чтобы он превышал размер одного блока HDFS

```
fallocate -l 333M test
```

```
Last login: Sun Apr  6 19:58:17 2025 from 192.168.0.3
-sh-4.2$ echo "test text" >> test
-sh-4.2$ fallocate -l 333M test
-sh-4.2$ ls -lh test
-rw-rw-r--. 1 kireev_mbd_2431 kireev_mbd_2431 333M Apr  8 20:56 test
-sh-4.2$
```

Рисунок 4 – Создать и увеличить размер файла

3. Создать новую директорию в hdfs по пути /data/<ваш_логин>

```
hdfs dfs -mkdir /data/kireev_mbd_2431
```

4. Создать новую директорию в hdfs по пути /data/<ваш_логин>/test_dir

```
hdfs dfs -mkdir /data/kireev_mbd_2431/test_dir
```

```
hdfs dfs -ls /data/kireev_mbd_2431
```

```
-sh: syntax error near unexpected token `newline'
-sh-4.2$ hdfs dfs -mkdir /data/kireev_mbd_2431
-sh-4.2$ hdfs dfs -mkdir /data/kireev_mbd_2431/test_dir
-sh-4.2$ hdfs dfs -ls /data/kireev_mbd_2431
Found 1 items
drwxr-xr-x  - kireev_mbd_2431 hdfs          0 2025-04-08 20:58 /data/kireev_mbd_2431/test_dir
-sh-4.2$
```

Рисунок 5 - Создание директории в hdfs

5. Положить в HDFS файл test по пути /data/<ваш_логин>/test_dir/test

```
hdfs dfs -put test /data/kireev_mbd_2431/test_dir/test
```

```
hdfs dfs -ls /data/kireev_mbd_2431/test_dir
```

```
-sh-4.2$ hdfs dfs -put test /data/kireev_mbd_2431/test_dir/test
-sh-4.2$ hdfs dfs -ls /data/kireev_mbd_2431/test_dir
Found 1 items
-rw-r--r-- 3 kireev_mbd_2431 hdfs 349175808 2025-04-06 10:35 /data/kireev_mbd_2431/test_dir/test
-sh-4.2$
```

Рисунок 6 - Перекладывание файла

6. Скопировать директорию /data/<ваш_логин>/test_dir в /data/<ваш_логин>/test_dir_1

```
hdfs dfs -cp /data/kireev_mbd_2431/test_dir /data/kireev_mbd_2431/test_dir_1
hdfs dfs -ls /data/kireev_mbd_2431
```

```
-sh-4.2$ hdfs dfs -cp /data/kireev_mbd_2431/test_dir /data/kireev_mbd_2431/test_dir_1
-sh-4.2$ hdfs dfs -ls /data/kireev_mbd_2431
Found 2 items
drwxr-xr-x - kireev_mbd_2431 hdfs 0 2025-04-06 10:35 /data/kireev_mbd_2431/test_dir
drwxr-xr-x - kireev_mbd_2431 hdfs 0 2025-04-06 10:35 /data/kireev_mbd_2431/test_dir_1
-sh-4.2$
```

Рисунок 7 - Скопировать директорию

7. Удалить файл test из директории test_dir_1 без сохранения файла в корзине.

```
hdfs dfs -rm -skipTrash /data/kireev_mbd_2431/test_dir_1/test
hdfs dfs -ls /data/kireev_mbd_2431/test_dir_1
```

```
drwxr-xr-x - kireev_mbd_2431 hdfs 0 2025-04-06 10:35 /data/kireev_mbd_2431/test_dir_1
-sh-4.2$ hdfs dfs -rm -skipTrash /data/kireev_mbd_2431/test_dir_1/test
Deleted /data/kireev_mbd_2431/test_dir_1/test
-sh-4.2$ hdfs dfs -ls /data/kireev_mbd_2431/test_dir_1
-sh-4.2$
```

Рисунок 8 - Удалить файл test из директории

8. Просмотреть размер любой директории

```
hdfs dfs -du -h /data/kireev_mbd_2431
hdfs dfs -du -h /data/kireev_mbd_2431/test_dir
```

```
-sh-4.2$ hdfs dfs -du -h /data/kireev_mbd_2431
333 M 999 M /data/kireev_mbd_2431/test_dir
0 0 /data/kireev_mbd_2431/test_dir_1
-sh-4.2$
```

```
-sh-4.2$ hdfs dfs -du -h /data/kireev_mbd_2431/test_dir
333 M 999 M /data/kireev_mbd_2431/test_dir/test
-sh-4.2$
```

Рисунок 9 - Просмотреть размер любой директории

9. Посмотреть, как файл /data/<ваш_логин>/test_dir/test хранится на файловой системе (см. команду hdfs fsck).

```
hdfs fsck /data/kireev_mbd_2431/test_dir/test -files -blocks -locations
```

```

-sh-4.2$ hdfs fsck /data/kireev_mbd_2431/test_dir/test -files -blocks -locations
Connecting to namenode via http://node1.mtuci.cloud.ru:50070/fsck?ugi=kireev_mbd_24
FSCK started by kireev_mbd_2431 (auth:SIMPLE) from /192.168.0.5 for path /data/kire
/data/kireev_mbd_2431/test_dir/test 349175808 bytes, replicated: replication=3, 3 b
0. BP-2089730104-192.168.0.3-1694299343161:blk_1074610886_870493 len=134217728 Live
DatanodeInfoWithStorage[192.168.0.4:50010,DS-49c0449b-0964-43e2-8cc4-200fa7acf471,D
1. BP-2089730104-192.168.0.3-1694299343161:blk_1074610887_870494 len=134217728 Live
DatanodeInfoWithStorage[192.168.0.8:50010,DS-6efcc178-736f-4d4d-9ae3-ae298b12f96e,D
2. BP-2089730104-192.168.0.3-1694299343161:blk_1074610888_870495 len=80740352 Live
atanodeInfoWithStorage[192.168.0.8:50010,DS-6efcc178-736f-4d4d-9ae3-ae298b12f96e,D

Status: HEALTHY
Number of data-nodes: 5
Number of racks: 1
Total dirs: 0
Total symlinks: 0

Replicated Blocks:
Total size: 349175808 B
Total files: 1
Total blocks (validated): 3 (avg. block size 116391936 B)
Minimally replicated blocks: 3 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 3
Average block replication: 3.0
Missing blocks: 0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)

Erasure Coded Block Groups:
Total size: 0 B
Total files: 0
Total block groups (validated): 0
Minimally erasure-coded block groups: 0
Over-erasure-coded block groups: 0
Under-erasure-coded block groups: 0
Unsatisfactory placement block groups: 0
Average block group size: 0.0
Missing block groups: 0
Corrupt block groups: 0
Missing internal blocks: 0
FSCK ended at Sun Apr 06 10:38:49 MSK 2025 in 2 milliseconds

The filesystem under path '/data/kireev_mbd_2431/test_dir/test' is HEALTHY
-sh-4.2$

```

Рисунок 10 - Хранение файла на файловой системе

1. Какой фактор репликации установлен на кластере?

Default replication factor: 3 в выводе команды `hdfs fsck`.

2. Сколько блоков составляют файл?

Total blocks (validated): 3 в выводе команды `hdfs fsck`.

Когда выполняется команду `hdfs fsck /data/<ваш_логин>/test_dir/test -blocks -files -locations`, вы получаете подробную информацию о каждом блоке, из которого состоит файл

0. BP-2089730104-192.168.0.3-1694299343161:blk_1074610886_870493 len=134217728 Live_repl=3 [DatanodeInfoWithStorage[192.168.0.8:50010,DS-6efcc178-736f-4d4d-9ae3-ae298b12f96e,DISK],

DatanodeInfoWithStorage[192.168.0.4:50010,DS-49c0449b-0964-43e2-8cc4-200fa7acf471,DISK], DatanodeInfoWithStorage[192.168.0.5:50010,DS-78f1e807-efc1-4cc0-875d-44182c85393b,DISK]]

3. Заполните таблицу для данных первого блока Вашего тестового файла

Таблица 1. Для данных первого блока

0.	Номер блока по порядку
BP-2089730104-192.168.0.3-1694299343161	Идентификатор block pull
blk_1074610886	Идентификатор блока
870493	Generation stamp.
134217728	Объем блока
3	Количество живых реплик блока
192.168.0.8:50010	IP-адрес и порт, по которому доступен блок
DS-6efcc178-736f-4d4d-9ae3-ae298b12f96e	Data Storage ID идентификатор ноды (Если у ноды изменится hostname или IP-адрес, нода всё равно будет идентифицироваться внутри HDFS)
DISK	Способ хранения блока (Может также храниться в S3-хранилище)

4. Скопируйте результат работы команды для любого из блоков, составляющих ваш тестовый файл. Какие данные мы получили?

```
-sh-4.2$ hdfs fsck -blockId blk_1074610886
Connecting to namenode via http://node1.mtuci.cloud.ru:50070/fsck?ugi=kireev_mbd_2431&blockId=1074610886&path=%2F
FSCK started by kireev_mbd_2431 (auth:SIMPLE) from /192.168.0.5 at Sun Apr 06 11:11:10 MSK 2020

Block Id: blk_1074610886
Block belongs to: /data/kireev_mbd_2431/test_dir/test
No. of Expected Replica: 3
No. of live Replica: 3
No. of excess Replica: 0
No. of stale Replica: 0
No. of decommissioned Replica: 0
No. of decommissioning Replica: 0
No. of corrupted Replica: 0
Block replica on datanode/rack: node3.mtuci.cloud.ru/default-rack is HEALTHY
Block replica on datanode/rack: node2.mtuci.cloud.ru/default-rack is HEALTHY
Block replica on datanode/rack: node6.mtuci.cloud.ru/default-rack is HEALTHY
[1]+  Done                  curl -i -L --negotiate -u : http://node1.mtuci.cloud.ru:50070/w
v1/data/kireev_mbd_2431/test_dir/test?op=OPEN
-sh-4.2$
```

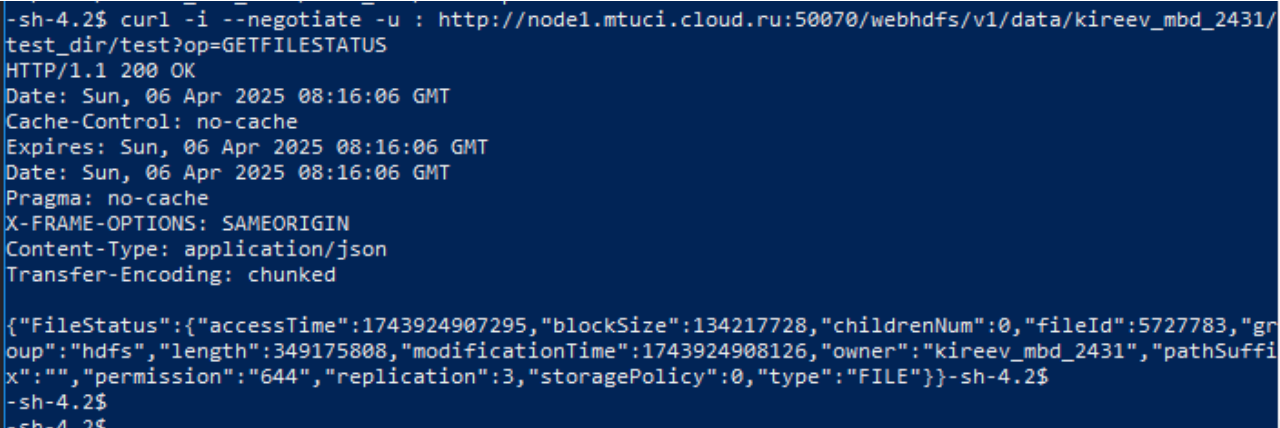
Рисунок 11 - Результат работы команды для любого из блоков

Блок blk_1074610886 находится в хорошем состоянии. Все реплики блока доступны, не повреждены и находятся на разных DataNodes.

Фактор репликации для этого блока установлен на 3, и все реплики в порядке. Реплики блока распределены по трем разным DataNodes (node3, node2, node6), что обеспечивает высокую доступность и надежность данных.

5. Выполним вывод информации о файле /data/<ваш_логин>/test_dir/test:

```
curl -i --negotiate -u :  
http://node1.mtuci.cloud.ru:50070/webhdfs/v1/data/kireev_mbd_2431/test_dir/test?  
op=GETFILESTATUS
```



```
-sh-4.2$ curl -i --negotiate -u : http://node1.mtuci.cloud.ru:50070/webhdfs/v1/data/kireev_mbd_2431/  
test_dir/test?op=GETFILESTATUS  
HTTP/1.1 200 OK  
Date: Sun, 06 Apr 2025 08:16:06 GMT  
Cache-Control: no-cache  
Expires: Sun, 06 Apr 2025 08:16:06 GMT  
Date: Sun, 06 Apr 2025 08:16:06 GMT  
Pragma: no-cache  
X-FRAME-OPTIONS: SAMEORIGIN  
Content-Type: application/json  
Transfer-Encoding: chunked  
  
{\"FileStatus\":{\"accessTime\":1743924907295,\"blockSize\":134217728,\"childrenNum\":0,\"fileId\":5727783,\"gr  
oup\":\"hdfs\",\"length\":349175808,\"modificationTime\":1743924908126,\"owner\":\"kireev_mbd_2431\",\"pathSuffi  
x\":\"\",\"permission\":\"644\",\"replication\":3,\"storagePolicy\":0,\"type\":\"FILE\"}}-sh-4.2$  
-sh-4.2$  
-sh-4.2$
```

Рисунок 12 - Вывод информации о файле

6. Какая информация выводится в результате работы команды?

Таблица 2. Информация о состоянии и характеристиках файла в HDFS.

Параметр	Значение	Описание
accessTime	1743924907295	Время последнего доступа к файлу в миллисекундах с 1 января 1970 года
blockSize	134217728	Размер блока в байтах (128 MB)
childrenNum	0	Количество дочерних элементов (для файла всегда 0)
fileId	5727783	Уникальный идентификатор файла
group	hdfs	Группа, которой принадлежит файл
length	349175808	Размер файла в байтах (около 333 MB)
modificationTime	1743924908126	Время последней модификации файла в миллисекундах с 1 января 1970 года
owner	kireev_mbd_2431	Владелец файла
pathSuffix	""	Суффикс пути (пустой в этом случае)
permission	644	Разрешения файла (читать и записывать для владельца, читать для группы и других)
replication	3	Фактор репликации файла
storagePolicy	0	Политика хранения (0 - по умолчанию)
type	FILE	Тип элемента (файл)

7. Теперь попробуем прочитать первые 10 символов тестового файла, используя тот же синтаксис команды:

```
curl -i --negotiate -u :  
http://node1.mtuci.cloud.ru:50070/webhdfs/v1/data/kireev_mbd_2431/test_dir/test?  
op=OPEN&length=10
```

```

-sh-4.2$ curl -i --negotiate -u : http://node1.mtuci.cloud.ru:50070/webhdfs/v1/data/kireev_mbd_2431/
test_dir/test?op=OPEN&length=10
[1] 10563
-sh-4.2$ HTTP/1.1 307 Temporary Redirect
Date: Sun, 06 Apr 2025 08:29:20 GMT
Cache-Control: no-cache
Expires: Sun, 06 Apr 2025 08:29:20 GMT
Date: Sun, 06 Apr 2025 08:29:20 GMT
Pragma: no-cache
X-FRAME-OPTIONS: SAMEORIGIN
Location: http://node3.mtuci.cloud.ru:50075/webhdfs/v1/data/kireev_mbd_2431/test_dir/test?op=OPEN&na
menoderpcaddress=node1.mtuci.cloud.ru:8020&offset=0
Content-Type: application/octet-stream
Content-Length: 0

[1]+  Done                  curl -i --negotiate -u : http://node1.mtuci.cloud.ru:50070/webhdfs/v1/
data/kireev_mbd_2431/test_dir/test?op=OPEN
-sh-4.2$

```

Рисунок 13 - Первые 10 символов тестового файла

Параметр	Значение	Описание
sber-node	node1.mtuci.cloud.ru	Адрес ноды, на которой установлен клиент HDFS
Порт подключения к REST API	50070	Порт, используемый для подключения к REST API
--negotiate	--negotiate	Включает SPNEGO в curl для аутентификации Kerberos
/webhdfs/v1	/webhdfs/v1	Адрес Web API
/data/<ваш_логин>/test_dir/test	/data/kireev_mbd_2431/test_dir/test	Путь к нужному файлу в HDFS
op=	OPEN	Производимое действие (открыть файл)
length=	10	Желаемая длина считывания символов (10 символов)

8. Почему мы не получили требуемых данных?

NameNode перенаправляет к DataNode, на котором хранится запрашиваемый блок данных

9. Проанализируйте ссылку раздела location из ответа сервера

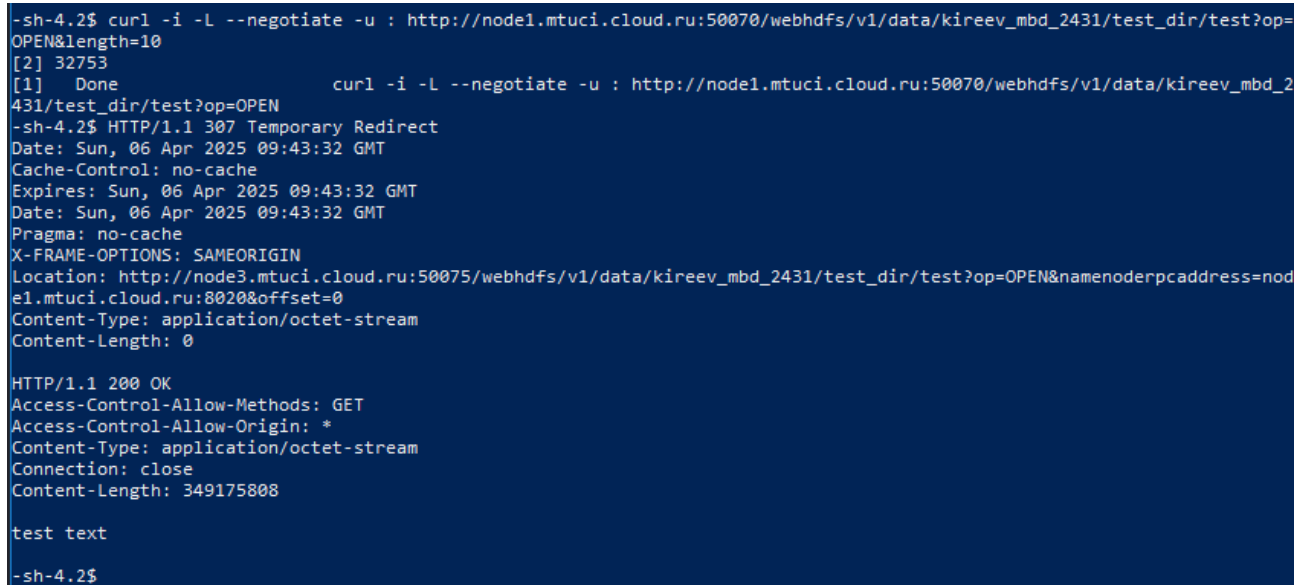
Перенаправление: NameNode перенаправляет вас к DataNode, на котором хранится запрашиваемый блок данных.

http://node3.mtuci.cloud.ru:50075/webhdfs/v1/data/kireev_mbd_2431/test_dir/test?op=OPEN&namenoderpcaddress=node1.mtuci.cloud.ru:8020&offset=0

Это позволяет оптимизировать доступ к данным, минуя NameNode после первоначального запроса.

10. Для того, чтобы получить желаемые данные мы можем добавить в curl-запрос флаг -L (location). Эта опция заставит Curl повторить запрос для нового адреса, который мы получили ранее в Location.

```
curl -i -L --negotiate -u : http://node1.mtuci.cloud.ru:50070/webhdfs/v1/data/kireev_mbd_2431/test_dir/test?op=OPEN&length=10
```



```
-sh-4.2$ curl -i -L --negotiate -u : http://node1.mtuci.cloud.ru:50070/webhdfs/v1/data/kireev_mbd_2431/test_dir/test?op=OPEN&length=10
[2] 32753
[1] Done curl -i -L --negotiate -u : http://node1.mtuci.cloud.ru:50070/webhdfs/v1/data/kireev_mbd_2431/test_dir/test?op=OPEN
-sh-4.2$ HTTP/1.1 307 Temporary Redirect
Date: Sun, 06 Apr 2025 09:43:32 GMT
Cache-Control: no-cache
Expires: Sun, 06 Apr 2025 09:43:32 GMT
Date: Sun, 06 Apr 2025 09:43:32 GMT
Pragma: no-cache
X-FRAME-OPTIONS: SAMEORIGIN
Location: http://node3.mtuci.cloud.ru:50075/webhdfs/v1/data/kireev_mbd_2431/test_dir/test?op=OPEN&namenoderpcaddress=node1.mtuci.cloud.ru:8020&offset=0
Content-Type: application/octet-stream
Content-Length: 0

HTTP/1.1 200 OK
Access-Control-Allow-Methods: GET
Access-Control-Allow-Origin: *
Content-Type: application/octet-stream
Connection: close
Content-Length: 349175808

test text
-sh-4.2$
```

Рисунок 14 - Вывод верной информации

11. Скопируйте файл /test из /data/<ваш_логин>/test_dir в test_dir_1 и удалите его /data/<ваш_логин>/test_dir_1/test с помощью curl-команды. Приложите скриншоты.

```
curl -i --negotiate -u : -X PUT \
"http://node1.mtuci.cloud.ru:50070/webhdfs/v1/data/kireev_mbd_2431/test_dir_1/test?op=CREATE&overwrite=true"
```

```
curl -i --negotiate -u : -X DELETE
"http://192.168.0.3:50070/webhdfs/v1/data/kireev_mbd_2431/test_dir_1/test?op=DELETE&recursive=true"
```

```
hdfs dfs -ls /data/kireev_mbd_2431/test_dir_1/
```

```

.50070/webhdfs/v1/data/kireev_mbd_2431/test_dir_1/test?op=OPEN
-sh-4.2$ curl -i --negotiate -u : -X PUT "http://node1.mtuci.cloud.ru:50070/webhdfs/v1/data/kireev_mbd_2431/test_dir_1/test?op=CREATE&overwrite=true" -T test
HTTP/1.1 100 Continue

HTTP/1.1 307 Temporary Redirect
Date: Tue, 08 Apr 2025 18:30:21 GMT
Cache-Control: no-cache
Expires: Tue, 08 Apr 2025 18:30:21 GMT
Date: Tue, 08 Apr 2025 18:30:21 GMT
Pragma: no-cache
X-FRAME-OPTIONS: SAMEORIGIN
Location: http://node3.mtuci.cloud.ru:50075/webhdfs/v1/data/kireev_mbd_2431/test_dir_1/test?op=CREATE&namenoderpcaddress=node1.mtuci.cloud.ru:8020&createflag=&createparent=true&overwrite=true
Content-Type: application/octet-stream
Content-Length: 0

```

Рисунок 15 - Копирование файла

```

-sh-4.2$ curl -i --negotiate -u : -X DELETE "http://node1.mtuci.cloud.ru:50070/webhdfs/v1/data/kireev_mbd_2431/test_dir_1/test?op=DELETE&recursive=true"
HTTP/1.1 403 Forbidden
Date: Tue, 08 Apr 2025 18:30:41 GMT
Cache-Control: no-cache
Expires: Tue, 08 Apr 2025 18:30:41 GMT
Date: Tue, 08 Apr 2025 18:30:41 GMT
Pragma: no-cache
X-FRAME-OPTIONS: SAMEORIGIN
Content-Type: application/json
Transfer-Encoding: chunked

{"RemoteException":{"exception":"AccessControlException","javaClassName":"org.apache.hadoop.security.AccessControlException","message":"Permission denied: user=dr.w
ho, access=WRITE, inode=\"/data/kireev_mbd_2431/test_dir_1\":kireev_mbd_2431:hdfs:drwxr-xr-x"}}-sh-4.2$

```

Рисунок 16 - Удаление файла

```

-sh-4.2$ curl -i --negotiate -u : http://node1.mtuci.cloud.ru:50070/webhdfs/v1/data/kireev_mbd_2431/test_dir/test?op=GETFILESTATUS
HTTP/1.1 200 OK
Date: Tue, 08 Apr 2025 18:34:22 GMT
Cache-Control: no-cache
Expires: Tue, 08 Apr 2025 18:34:22 GMT
Date: Tue, 08 Apr 2025 18:34:22 GMT
Pragma: no-cache
X-FRAME-OPTIONS: SAMEORIGIN
Content-Type: application/json
Transfer-Encoding: chunked

{"FileStatus":{"accessTime":1744135366843,"blockSize":134217728,"childrenNum":0,"fileId":5737153,"group":"hdfs","length":349175808,"modificationTime":1744135367608,"owner":"kireev_mbd_2431","pathSuffix":"","permission":"644","replication":3,"storagePolicy":0,"type":"FILE"}}-sh-4.2$

```

Рисунок 17 - Проверка удаления

Работа с UI

12. Залогиньтесь в Hadoop Manager Ambari <http://192.168.0.3:8080/>

Логин и пароль для входа – monitor/monitor

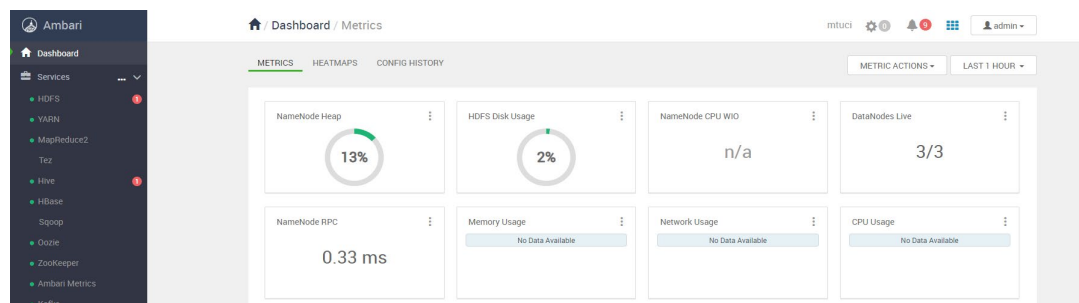


Рисунок 18 – Интерфейс Hadoop

Через Ambari осуществляется настройка и мониторинг кластера Hadoop.

13. Войдите в NameNode UI <http://192.168.0.3:50070/>

(Адрес можно увидеть в разделе Services->HDFS->QuickLinks)

В интерфейсе собрана наиболее важная информация о состоянии HDFS

14. Зайдите в Utilities->Browse FileSystem.

15. Найдите созданный вами файл в директории /data.

16. Нажмите на название файла и внимательно изучите информацию о блоках данных.

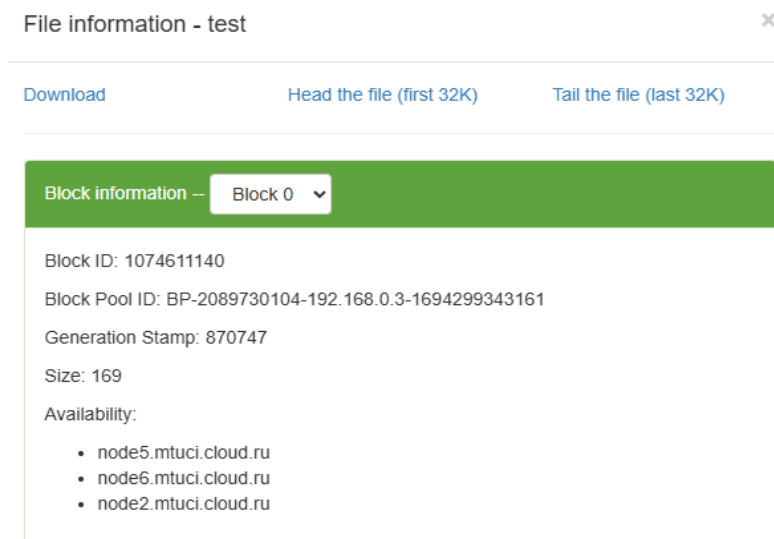


Рисунок 20 - Информация о блоках данных

17. Ответьте на вопрос: какой объём стандартного блока HDFS исходя из показателя Size?

Ответ: 169 байт

18. Есть ли блок, который отличается по размеру от остальных?

Ответ: Да, последний блок файла может быть меньше стандартного размера блока, если размер файла не кратен стандартному размеру блока.

19. Сколько пространства занимает последний блок файла в локальной файловой системе?

Ответ: 169 байт в локальной файловой системе

20. Подумайте, при каких обстоятельствах количество реплик блоков файла может превышать стандартный фактор репликации?

Ответ: если администратор кластера изменил стандартный фактор репликации. Если файл был помечен для хранения с более высоким уровнем репликации (например, для критически важных данных). В случае ошибок или неисправностей в кластере, HDFS может автоматически увеличить количество реплик для обеспечения надежности данных.