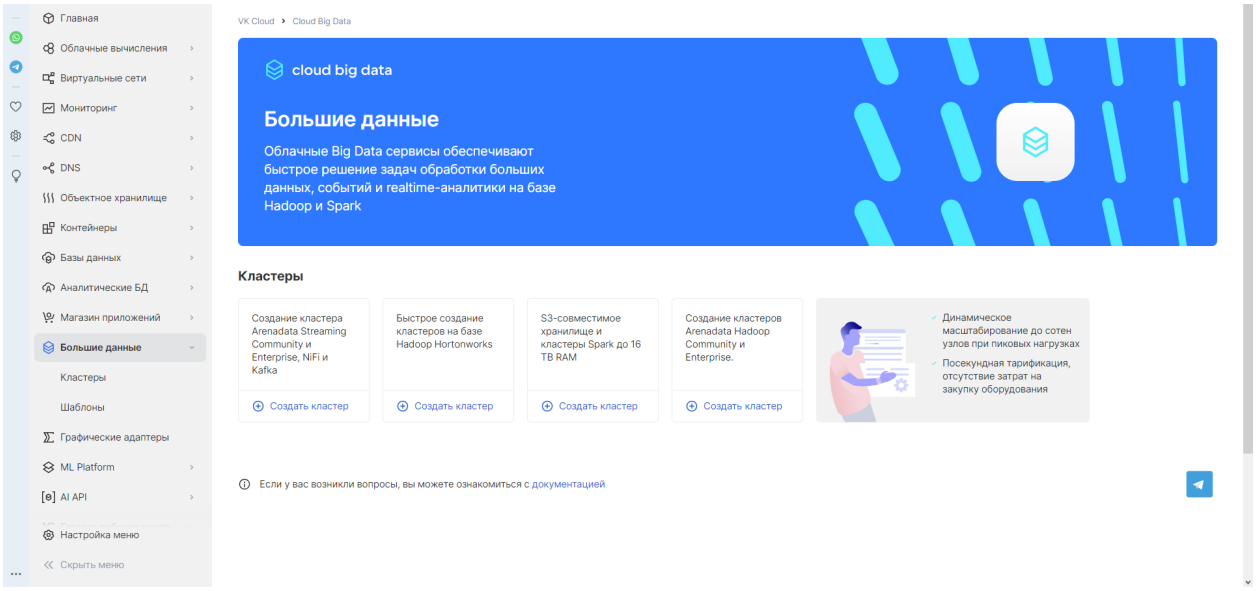
Agenda

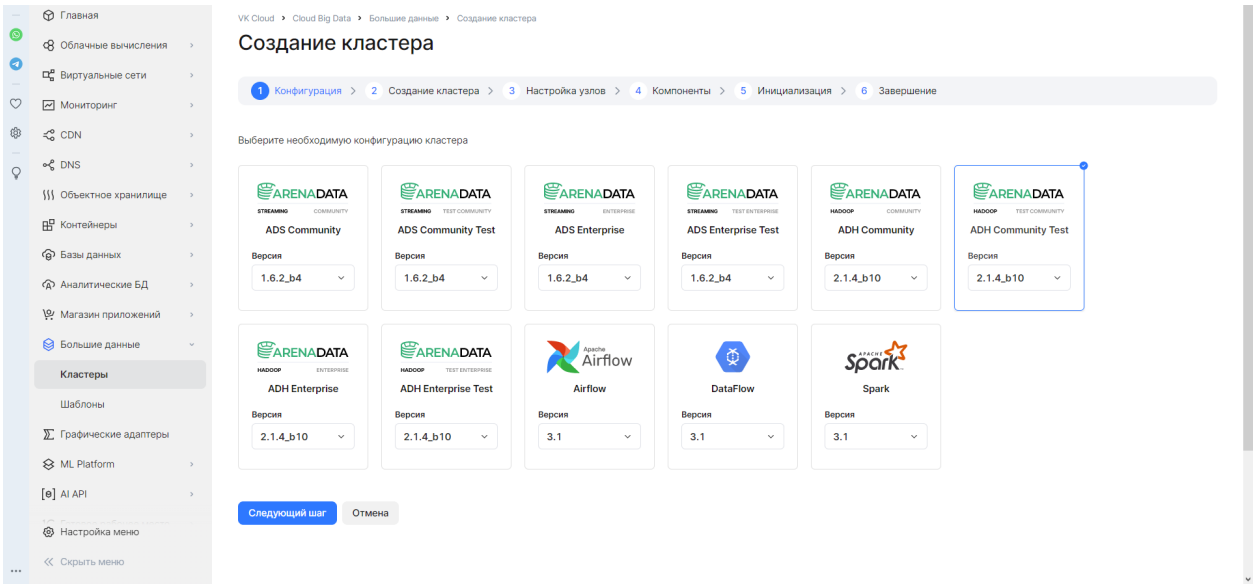
Часть 1. Создание кластера Arenadata Hadoop. Работа с HDFS

Создание кластера Arenadata Hadoop Community Test

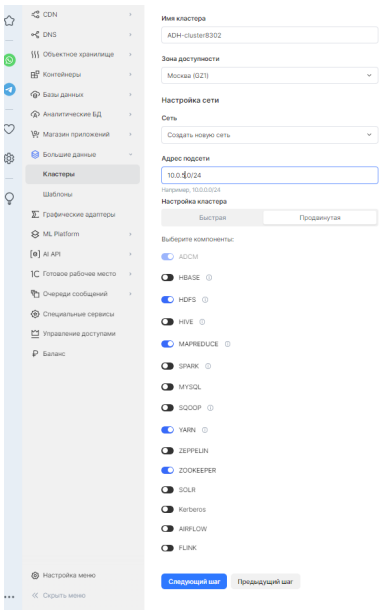
1. Перейдите в личный кабинет VK Cloud. Откройте раздел «**Большие Данные**» и выберите «**Создание кластеров Arenadata Hadoop Community и Enterprise**».



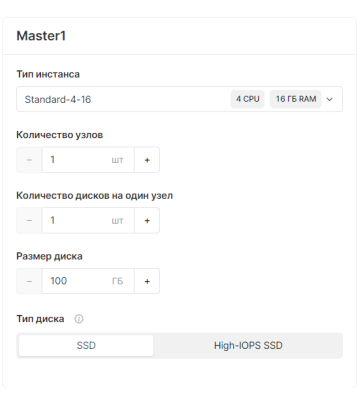
2. Выберите конфигурацию «**ADH Community Test v. 2.1.4_b10**»

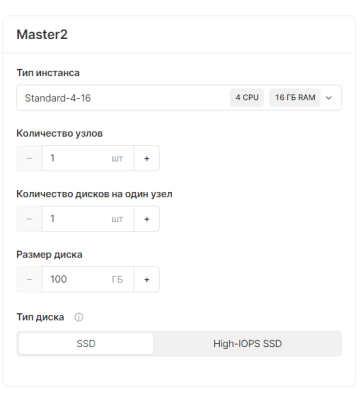


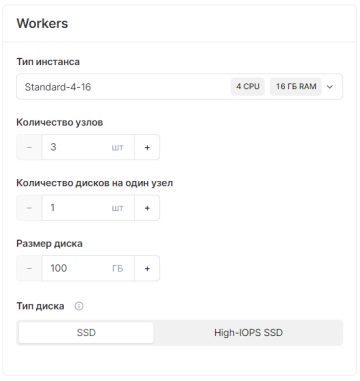
3. Настройте сеть для вашего кластера и добавьте следующие компоненты: **ADCM, HDFS, MAPREDUCE, YARN, ZOOKEEPER**

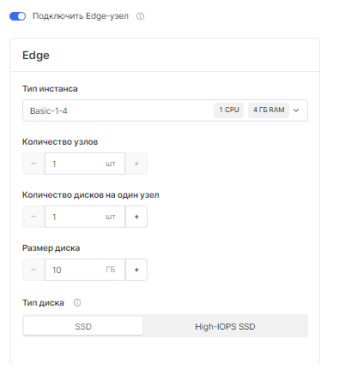


4. Укажите конфигурацию узлов Master1, Master2, Workers и Edge (установите соответствующий переключатель):

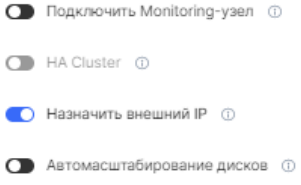




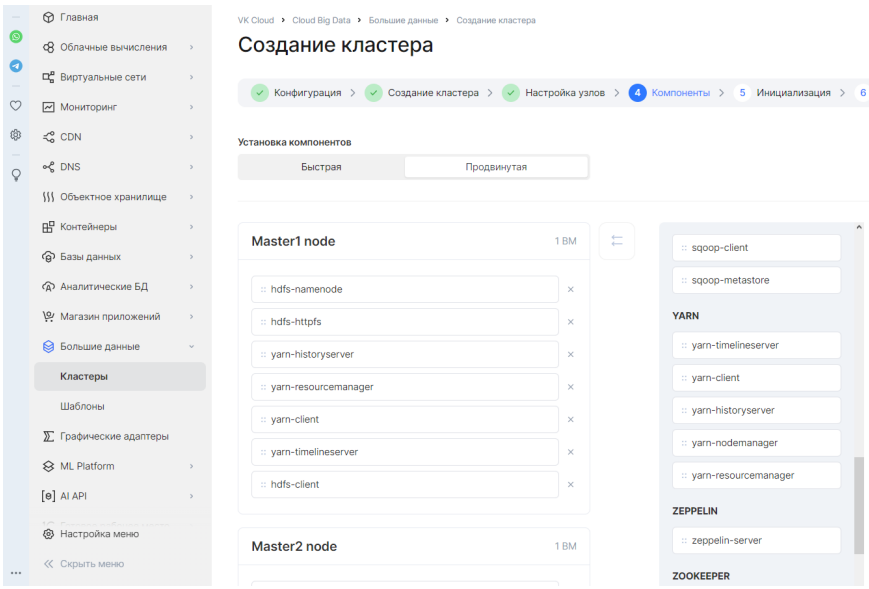




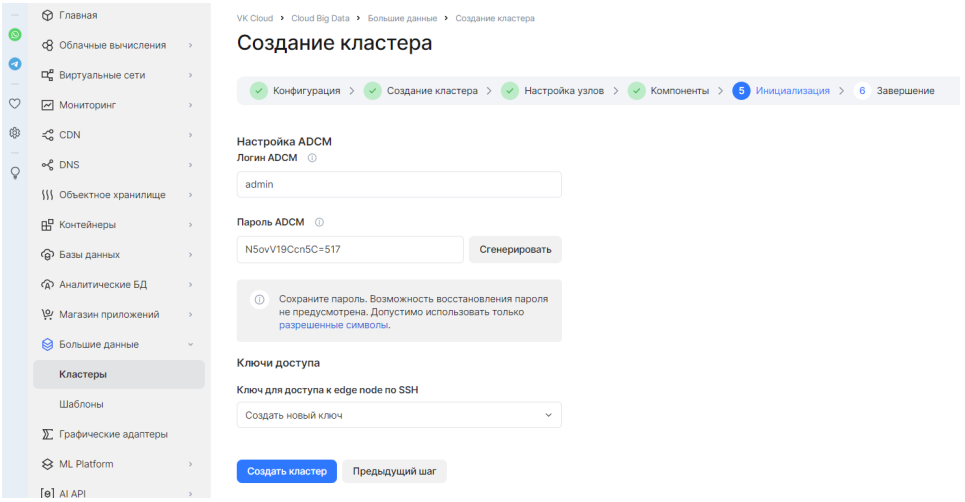
Назначьте внешний IP для инстанса ADCM



5. Компоненты оставьте без изменений

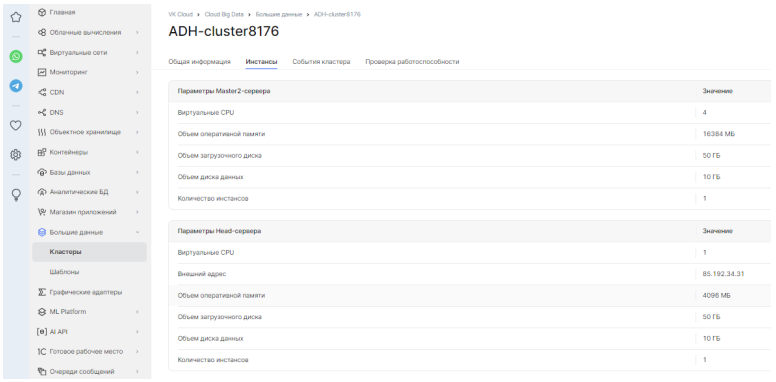


6. Укажите пароль для ADCM (Arenadata Cluster Manager).
   Создайте SSH-ключ в формате RSA для подключения к узлу. Запустите создание кластера (длительность операции 20-30 минут).



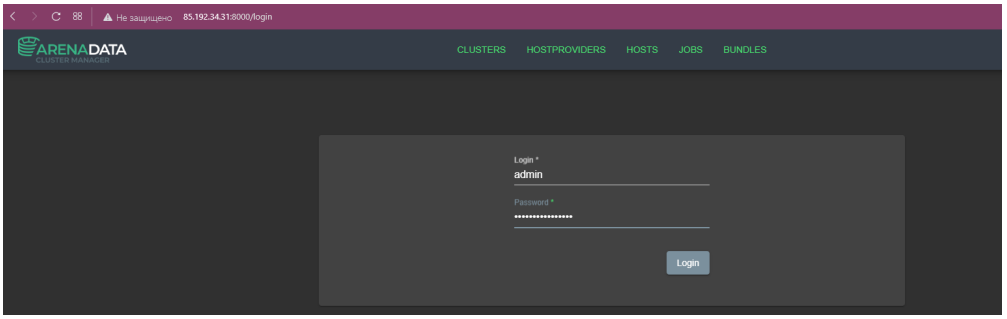
Подключение к кластеру через ssh

1. Установите или используйте стандартный ssh-клиент. Для настройки подключения воспользуйтесь инструкцией – Подключение к Linux ВМ.
2. Используйте внешний IP адрес для ssh-клиента (пользователь **admin**):

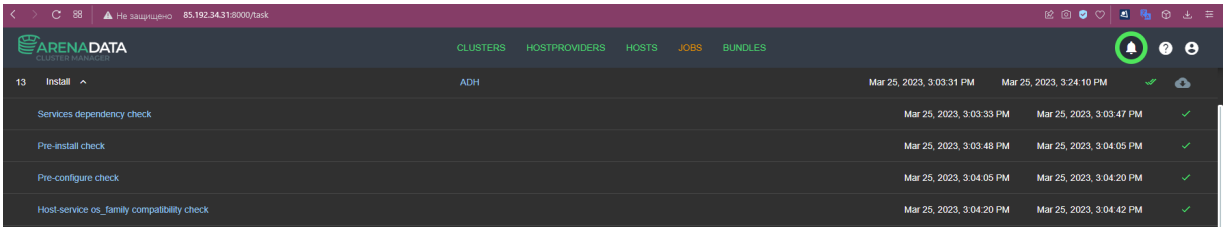


3. Дождитесь установки кластера.

Для проверки статуса установки кластера перейдите по ссылке «http://<Внешний IP адрес>:8000» (например, http://85.192.34.31:8000).
Введите учетные данные для ADCM:



Проверьте статус выполнения задачи Install/ADH.



4. Запустите команду для проверки работоспособности кластера:

```
hadoop fs -ls /
```

Создание домашнего каталога в HDFS

1. Создайте домашнюю директорию для текущего пользователя Linux в HDFS

Пояснение для преподавателя:

```
sudo -u hdfs hadoop fs -mkdir /user/admin
```

2. Смените владельца и группу созданного каталога на admin

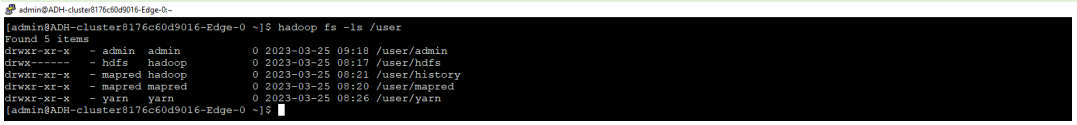
Пояснение для преподавателя:

```
sudo -u hdfs hadoop fs -chown admin:admin /user/admin
```

3. Проверьте владельца каталога в HDFS

Пояснение для преподавателя:

```
hadoop fs -ls /user
```



Загрузка данных в домашний каталог HDFS

1. Загрузите файлы nseCompBig.zip и nseCompSmall.zip на узел кластера (например, используя curl или wget). Скачать (временно) файлы можно по ссылкам, соответственно:
   https://disk.yandex.ru/d/PoXJ2YsKh2HZIw
   https://disk.yandex.ru/d/QbLvf4KUFubO7A
2. Разархивируйте загруженные файлы на узле в домашнюю директорию

Пояснение для преподавателя:

```
unzip /tmp/nseCompSmall.zip
unzip /tmp/nseCompBig.zip
```

3. Создайте директорию staging/input/stocks/ в HDFS

Пояснение для преподавателя:

```
hadoop fs -mkdir -p staging/input/stocks/
```

4. Загрузите файлы nseCompBig.csv и nseCompSmall.csv в HDFS каталог staging/input/stocks/

Пояснение для преподавателя:

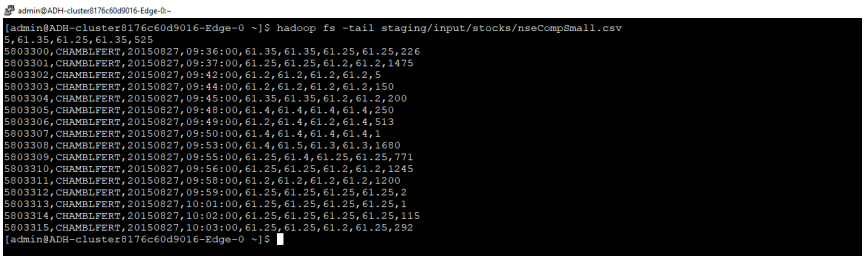
```
hadoop fs -put nseCompSmall.csv staging/input/stocks/nseCompSmall.csv
hadoop fs -put nseCompBig.csv staging/input/stocks/nseCompBig.csv
```

Вывод содержимого файлов в HDFS

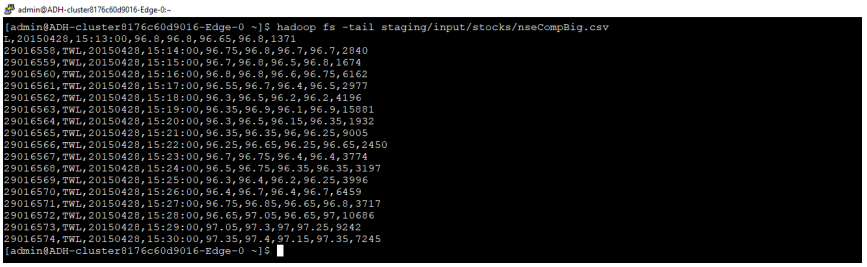
Выведите несколько строк файлов nseCompBig.csv и nseCompSmall.csv в HDFS (hadoop fs [-tail|cat|head])

Пояснение для преподавателя:

```
hadoop fs -tail staging/input/stocks/nseCompSmall.csv
```




```
hadoop fs -tail staging/input/stocks/nseCompBig.csv
```



Распределение блоков по узлам кластера

Выведите распределение блоков по узлам кластера для фалов nseCompBig.csv и nseCompSmall.csv в HDFS (hdfs fsck)

Пояснение для преподавателя:

```
hdfs fsck staging/input/stocks/nseCompSmall.csv -files -blocks
-locations
```

```
admin@ADH-cluster83258ca307e7-Edge-0:~
[admin@ADH-cluster83258ca307e7-Edge-0 ~]$ hdfs fsck staging/input/stocks/nseCompSmall.csv -files -blocks -locations
Connecting to namenode via http://ADH-cluster83258ca307e7-Master1-0.mcs.local:9870/fsck?ugi=admin&files=1&blocks=1&locations=1&path=%2Fuser%2Fadmin%2Fstaging%2Finput%2Fstocks%2FnseCompSmall.
csv
FSCK started by admin (auth:SIMPLE) from /10.0.0.17 for path /user/admin/staging/input/stocks/nseCompSmall.csv at Sat Mar 25 15:35:36 UTC 2023
/user/admin/staging/input/stocks/nseCompSmall.csv 366356750 bytes, replicated: replication=3, 3 block(s):  OK
0. BP-1262089974-10.0.0.7-1679757265470:blk_1073741888_1064 len=134217728 Live_repl=3  [DatanodeInfoWithStorage[10.0.0.18:9866,DS-9e48a95c-7da9-414c-b669-b755be3fa943,DISK], DatanodeInfoWith
Storage[10.0.0.20:9866,DS-17d4f212-b9c1-46df-bc19-4c1496783666,DISK], DatanodeInfoWithStorage[10.0.0.32:9866,DS-363d9040-21d4-49c2-8d67-f0f0fe7058d1,DISK]]
1. BP-1262089974-10.0.0.7-1679757265470:blk_1073741889_1065 len=134217728 Live_repl=3  [DatanodeInfoWithStorage[10.0.0.20:9866,DS-17d4f212-b9c1-46df-bc19-4c1496783666,DISK], DatanodeInfoWith
Storage[10.0.0.18:9866,DS-9e48a95c-7da9-414c-b669-b755be3fa943,DISK], DatanodeInfoWithStorage[10.0.0.32:9866,DS-363d9040-21d4-49c2-8d67-f0f0fe7058d1,DISK]]
2. BP-1262089974-10.0.0.7-1679757265470:blk_1073741890_1066 len=97921294 Live_repl=3  [DatanodeInfoWithStorage[10.0.0.32:9866,DS-363d9040-21d4-49c2-8d67-f0f0fe7058d1,DISK], DatanodeInfoWithS
torage[10.0.0.18:9866,DS-9e48a95c-7da9-414c-b669-b755be3fa943,DISK], DatanodeInfoWithStorage[10.0.0.20:9866,DS-17d4f212-b9c1-46df-bc19-4c1496783666,DISK]]


Status: HEALTHY
 Number of data-nodes:  3
 Number of racks:               1
 Total dirs:                    0
 Total symlinks:                0

Replicated Blocks:
 Total size:    366356750 B
 Total files:   1
 Total blocks (validated):      3 (avg. block size 122118916 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 Sat Mar 25 15:35:36 UTC 2023 in 1 milliseconds


The filesystem under path '/user/admin/staging/input/stocks/nseCompSmall.csv' is HEALTHY
[admin@ADH-cluster83258ca307e7-Edge-0 ~]$
```

```
hdfs fsck staging/input/stocks/nseCompBig.csv -files -blocks -locations
```

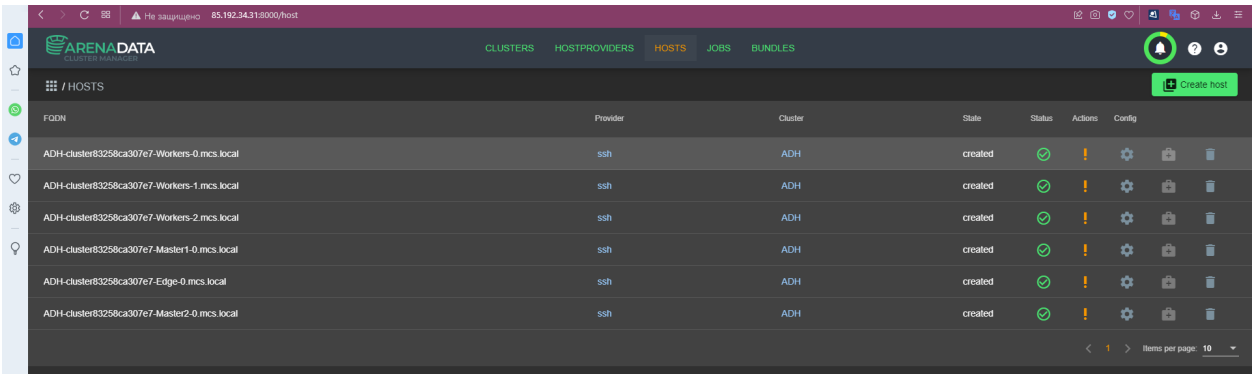
```
admin@ADH-cluster8176c60d9016-Edge-0:~
[admin@ADH-cluster8176c60d9016-Edge-0 ~]$ hdfs fsck staging/input/stocks/nseCompBig.csv  -files -blocks -locations
Connecting to namenode via http://ADH-cluster8176c60d9016-Master1-0.mcs.local:9870/fsck?ugi=admin&files=1&blocks=1&locations=1&path=%2Fuser%2Fadmin%2Fstaging%2Finput%2Fstocks%2FnseCompBig.cs
v
FSCK started by admin (auth:SIMPLE) from /10.0.0.8 for path /user/admin/staging/input/stocks/nseCompBig.csv at Sat Mar 25 10:15:00 UTC 2023
/user/admin/staging/input/stocks/nseCompBig.csv 1822975955 bytes, replicated: replication=3, 14 block(s):  OK
0. BP-714958517-10.0.0.6-1679732120535:blk_1073741899_1075 len=134217728 Live_repl=3  [DatanodeInfoWithStorage[10.0.0.16:9866,DS-9305a2f3-c692-43d1-ad7b-9ce4d859db63,DISK], DatanodeInfoWithS
torage[10.0.0.31:9866,DS-ecf7e6b4-2572-446e-b5b4-45b4b56e338a,DISK], DatanodeInfoWithStorage[10.0.0.12:9866,DS-58d88aa8-38dd-4b11-abac-d267a7fdc7ae,DISK]]
1. BP-714958517-10.0.0.6-1679732120535:blk_1073741900_1076 len=134217728 Live_repl=3  [DatanodeInfoWithStorage[10.0.0.31:9866,DS-ecf7e6b4-2572-446e-b5b4-45b4b56e338a,DISK], DatanodeInfoWithS
torage[10.0.0.16:9866,DS-9305a2f3-c692-43d1-ad7b-9ce4d859db63,DISK], DatanodeInfoWithStorage[10.0.0.12:9866,DS-58d88aa8-38dd-4b11-abac-d267a7fdc7ae,DISK]]
2. BP-714958517-10.0.0.6-1679732120535:blk_1073741901_1077 len=134217728 Live_repl=3  [DatanodeInfoWithStorage[10.0.0.16:9866,DS-9305a2f3-c692-43d1-ad7b-9ce4d859db63,DISK], DatanodeInfoWithS
torage[10.0.0.12:9866,DS-58d88aa8-38dd-4b11-abac-d267a7fdc7ae,DISK], DatanodeInfoWithStorage[10.0.0.31:9866,DS-ecf7e6b4-2572-446e-b5b4-45b4b56e338a,DISK]]
3. BP-714958517-10.0.0.6-1679732120535:blk_1073741902_1078 len=134217728 Live_repl=3  [DatanodeInfoWithStorage[10.0.0.16:9866,DS-9305a2f3-c692-43d1-ad7b-9ce4d859db63,DISK], DatanodeInfoWithS
torage[10.0.0.31:9866,DS-ecf7e6b4-2572-446e-b5b4-45b4b56e338a,DISK], DatanodeInfoWithStorage[10.0.0.12:9866,DS-58d88aa8-38dd-4b11-abac-d267a7fdc7ae,DISK]]
4. BP-714958517-10.0.0.6-1679732120535:blk_1073741903_1079 len=134217728 Live_repl=3  [DatanodeInfoWithStorage[10.0.0.12:9866,DS-58d88aa8-38dd-4b11-abac-d267a7fdc7ae,DISK], DatanodeInfoWithS
torage[10.0.0.31:9866,DS-ecf7e6b4-2572-446e-b5b4-45b4b56e338a,DISK], DatanodeInfoWithStorage[10.0.0.16:9866,DS-9305a2f3-c692-43d1-ad7b-9ce4d859db63,DISK]]
5. BP-714958517-10.0.0.6-1679732120535:blk_1073741904_1080 len=134217728 Live_repl=3  [DatanodeInfoWithStorage[10.0.0.12:9866,DS-58d88aa8-38dd-4b11-abac-d267a7fdc7ae,DISK], DatanodeInfoWithS
torage[10.0.0.16:9866,DS-9305a2f3-c692-43d1-ad7b-9ce4d859db63,DISK], DatanodeInfoWithStorage[10.0.0.31:9866,DS-ecf7e6b4-2572-446e-b5b4-45b4b56e338a,DISK]]
6. BP-714958517-10.0.0.6-1679732120535:blk_1073741905_1081 len=134217728 Live_repl=3  [DatanodeInfoWithStorage[10.0.0.31:9866,DS-ecf7e6b4-2572-446e-b5b4-45b4b56e338a,DISK], DatanodeInfoWithS
torage[10.0.0.16:9866,DS-9305a2f3-c692-43d1-ad7b-9ce4d859db63,DISK], DatanodeInfoWithStorage[10.0.0.12:9866,DS-58d88aa8-38dd-4b11-abac-d267a7fdc7ae,DISK]]
7. BP-714958517-10.0.0.6-1679732120535:blk_1073741906_1082 len=134217728 Live_repl=3  [DatanodeInfoWithStorage[10.0.0.31:9866,DS-ecf7e6b4-2572-446e-b5b4-45b4b56e338a,DISK], DatanodeInfoWithS
torage[10.0.0.16:9866,DS-9305a2f3-c692-43d1-ad7b-9ce4d859db63,DISK], DatanodeInfoWithStorage[10.0.0.12:9866,DS-58d88aa8-38dd-4b11-abac-d267a7fdc7ae,DISK]]
8. BP-714958517-10.0.0.6-1679732120535:blk_1073741907_1083 len=134217728 Live_repl=3  [DatanodeInfoWithStorage[10.0.0.12:9866,DS-58d88aa8-38dd-4b11-abac-d267a7fdc7ae,DISK], DatanodeInfoWithS
torage[10.0.0.31:9866,DS-ecf7e6b4-2572-446e-b5b4-45b4b56e338a,DISK], DatanodeInfoWithStorage[10.0.0.16:9866,DS-9305a2f3-c692-43d1-ad7b-9ce4d859db63,DISK]]
9. BP-714958517-10.0.0.6-1679732120535:blk_1073741908_1084 len=134217728 Live_repl=3  [DatanodeInfoWithStorage[10.0.0.12:9866,DS-58d88aa8-38dd-4b11-abac-d267a7fdc7ae,DISK], DatanodeInfoWithS
torage[10.0.0.31:9866,DS-ecf7e6b4-2572-446e-b5b4-45b4b56e338a,DISK], DatanodeInfoWithStorage[10.0.0.16:9866,DS-9305a2f3-c692-43d1-ad7b-9ce4d859db63,DISK]]
10. BP-714958517-10.0.0.6-1679732120535:blk_1073741909_1085 len=134217728 Live_repl=3  [DatanodeInfoWithStorage[10.0.0.31:9866,DS-ecf7e6b4-2572-446e-b5b4-45b4b56e338a,DISK], DatanodeInfoWith
Storage[10.0.0.12:9866,DS-58d88aa8-38dd-4b11-abac-d267a7fdc7ae,DISK], DatanodeInfoWithStorage[10.0.0.16:9866,DS-9305a2f3-c692-43d1-ad7b-9ce4d859db63,DISK]]
11. BP-714958517-10.0.0.6-1679732120535:blk_1073741910_1086 len=134217728 Live_repl=3  [DatanodeInfoWithStorage[10.0.0.12:9866,DS-58d88aa8-38dd-4b11-abac-d267a7fdc7ae,DISK], DatanodeInfoWith
Storage[10.0.0.31:9866,DS-ecf7e6b4-2572-446e-b5b4-45b4b56e338a,DISK], DatanodeInfoWithStorage[10.0.0.16:9866,DS-9305a2f3-c692-43d1-ad7b-9ce4d859db63,DISK]]
12. BP-714958517-10.0.0.6-1679732120535:blk_1073741911_1087 len=134217728 Live_repl=3  [DatanodeInfoWithStorage[10.0.0.31:9866,DS-ecf7e6b4-2572-446e-b5b4-45b4b56e338a,DISK], DatanodeInfoWith
Storage[10.0.0.16:9866,DS-9305a2f3-c692-43d1-ad7b-9ce4d859db63,DISK], DatanodeInfoWithStorage[10.0.0.12:9866,DS-58d88aa8-38dd-4b11-abac-d267a7fdc7ae,DISK]]
13. BP-714958517-10.0.0.6-1679732120535:blk_1073741912_1088 len=78145491 Live_repl=3  [DatanodeInfoWithStorage[10.0.0.16:9866,DS-9305a2f3-c692-43d1-ad7b-9ce4d859db63,DISK], DatanodeInfoWithS
torage[10.0.0.31:9866,DS-ecf7e6b4-2572-446e-b5b4-45b4b56e338a,DISK], DatanodeInfoWithStorage[10.0.0.12:9866,DS-58d88aa8-38dd-4b11-abac-d267a7fdc7ae,DISK]]


Status: HEALTHY
 Number of data-nodes:  3
 Number of racks:               1
 Total dirs:                    0
 Total symlinks:                0

Replicated Blocks:
 Total size:    1822975955 B
 Total files:   1
 Total blocks (validated):      14 (avg. block size 130212568 B)
 Minimally replicated blocks:   14 (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
```

Расположение блоков на узлах в локальной ОС

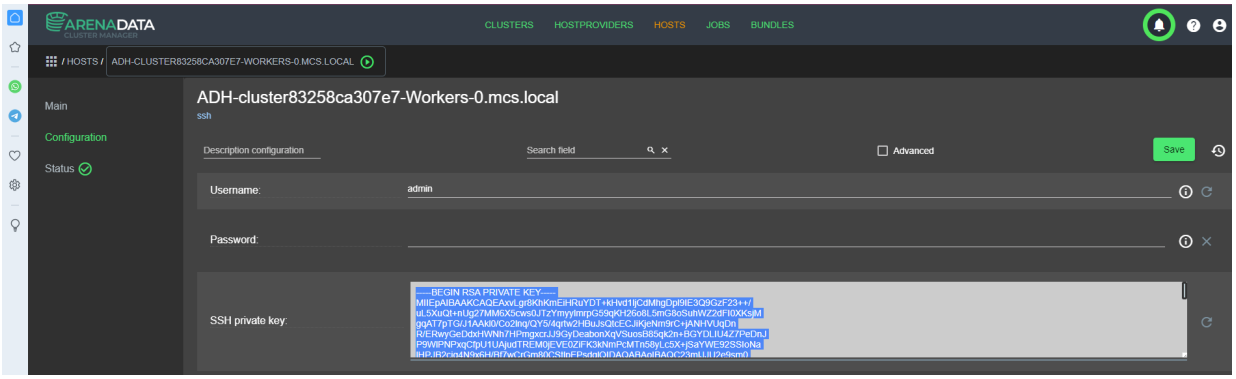
1. Найдите имя блока файла в выводе команды hdfs fsck

```
/user/admin/staging/input/stocks/nseCompSmall.csv 366356750 bytes, replicated: replication=3, 3 block(s):  OK
0. BP-1262089974-10.0.0.7-1679757265470:blk_1073741888_1064 len=134217728 Live_repl=3  [DatanodeInfoWithStorage[10.0.0.18:9866,DS-9e48a95c-7da9-414c-b669-b755be3fa943,DISK], DatanodeInfoWith
Storage[10.0.0.20:9866,DS-17d4f212-b9c1-46df-bc19-4c1496783666,DISK], DatanodeInfoWithStorage[10.0.0.32:9866,DS-363d9040-21d4-49c2-8d67-f0f0fe7058d1,DISK]]
1. BP-1262089974-10.0.0.7-1679757265470:blk_1073741889_1065 len=134217728 Live_repl=3  [DatanodeInfoWithStorage[10.0.0.20:9866,DS-17d4f212-b9c1-46df-bc19-4c1496783666,DISK], DatanodeInfoWith
Storage[10.0.0.18:9866,DS-9e48a95c-7da9-414c-b669-b755be3fa943,DISK], DatanodeInfoWithStorage[10.0.0.32:9866,DS-363d9040-21d4-49c2-8d67-f0f0fe7058d1,DISK]]
2. BP-1262089974-10.0.0.7-1679757265470:blk_1073741890_1066 len=97921294 Live_repl=3  [DatanodeInfoWithStorage[10.0.0.32:9866,DS-363d9040-21d4-49c2-8d67-f0f0fe7058d1,DISK], DatanodeInfoWithS
torage[10.0.0.18:9866,DS-9e48a95c-7da9-414c-b669-b755be3fa943,DISK], DatanodeInfoWithStorage[10.0.0.20:9866,DS-17d4f212-b9c1-46df-bc19-4c1496783666,DISK]]
```

2. Подключитесь через SSH к любой hdfs-datanode.
   Выберите любой узел Workers во вкладке **ADCM –> HOSTS** (например, ADH-CLUSTER83258CA307E7-WORKERS-0.MCS.LOCAL)



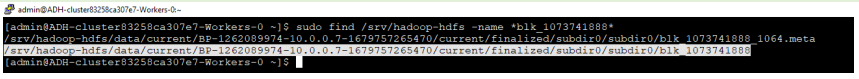
Откройте настройки узла (**Configuration**) и скопируйте поле «**SSH private key**» для доступа по ssh с Edge-узла.



3. Выполните поиск в локальной ОС в директории /srv/hadoop-hdfs -name (find)

Пояснение для преподавателя:

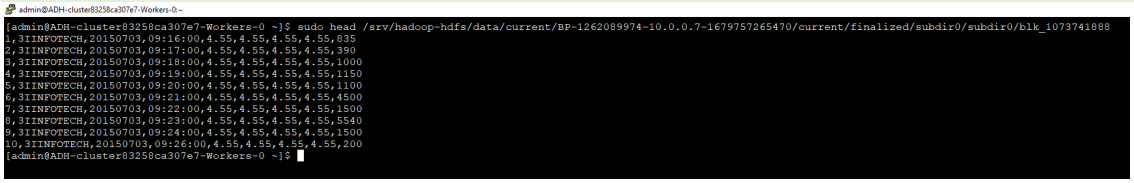
```
sudo find /srv/hadoop-hdfs -name *blk_1073741888*
```



4. Выведите содержимое блока в локальной ОС (cat, head, tail)

Пояснение для преподавателя:

```
sudo head
/srv/hadoop-hdfs/data/current/BP-1262089974-10.0.0.7-1679757265470/curr
ent/finalized/subdir0/subdir0/blk_1073741888
```



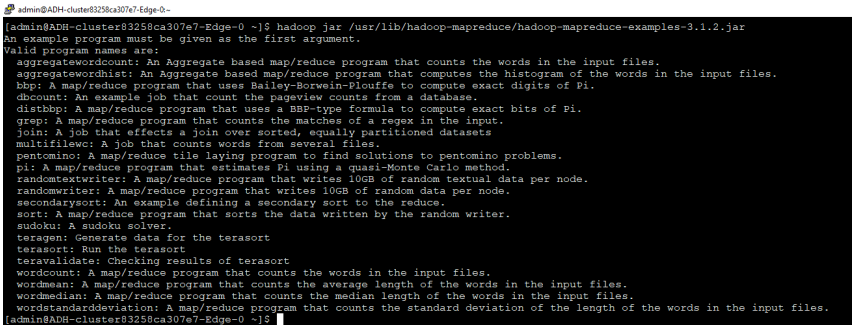
Часть 2. Запуск задач Map Reduce

Запустите MR-задачу WordCount для загруженных данных

1. Ознакомьтесь с возможностями стандартных примеров запуска MR-задач Hadoop (hadoop-mapreduce-examples-3.1.2.jar)

Пояснение для преподавателя:

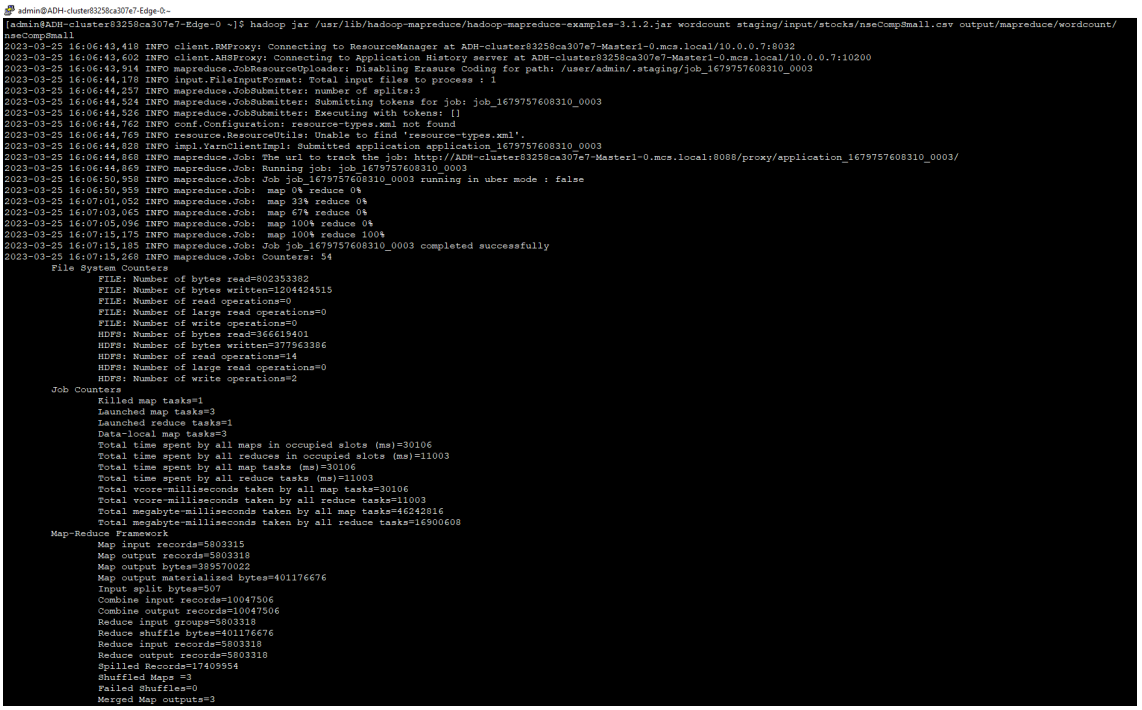
```
hadoop jar
/usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples-3.1.2.jar
```



2. Запустите MR-задачу wordcount для файла nseCompSmall.csv. Сохраните результат работы алгоритма в HDFS каталог output/mapreduce/wordcount/nseCompSmall

Пояснение для преподавателя:

```
hadoop jar
/usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples-3.1.2.jar wordcount
staging/input/stocks/nseCompSmall.csv
output/mapreduce/wordcount/nseCompSmall
```

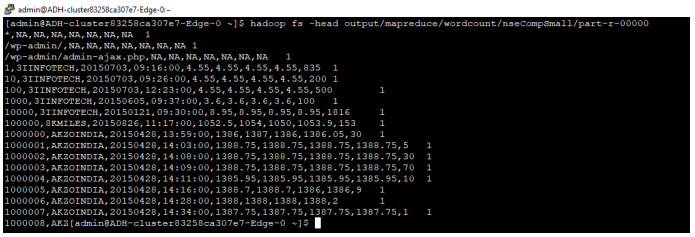


Выведите результат работы алгоритма

Выведите содержимое файла результата вычислений (используйте команду hadoop fs -head, т.к. вывод команды hadoop fs -cat может быть объемным)

Пояснение для преподавателя:

```
hadoop fs -head output/mapreduce/wordcount/nseCompSmall/part-r-00000
```

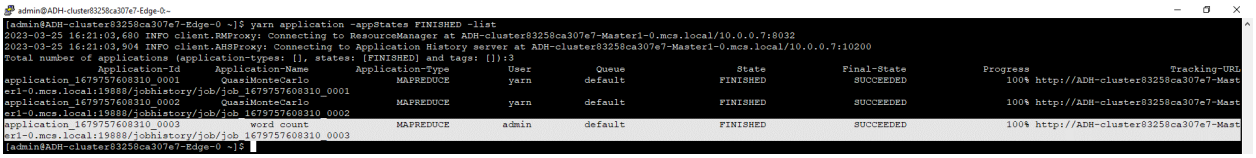


Отобразите информацию по выполненной задаче через консоль YARN.

1. Найдите имя MR-задачи с помощью вывода списка выполненных (FINISHED) задач (yarn application -list).

Пояснение для преподавателя:

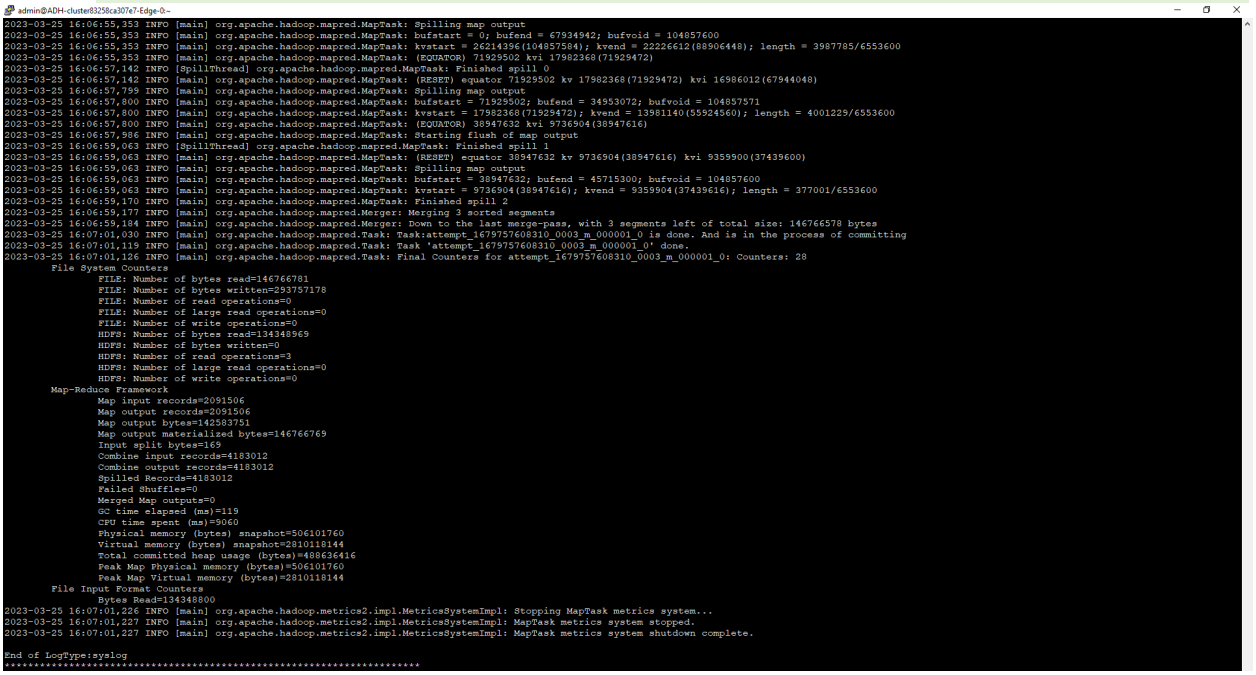
```
yarn application -appStates FINISHED -list
```



2. Изучите вывод лога MR-задачи (yarn logs -applicationId)

Пояснение для преподавателя:

```
yarn logs -applicationId application_1679757608310_0003
```



Часть 3. Сжатие файлов в Hadoop. Запуск MR-задач для сжатых данных.

Произведите сжатие загруженных данных кодаками: GzipCodec и BZip2Codec

Сжатие исходных данных можно выполнить с помощь MR-задачи:

- Для кодака Gzip:

```
hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming-*.jar -D mapreduce.job.reduces=1 -D mapred.output.compress=true -D mapred.output.compression.codec=org.apache.hadoop.io.compress.GzipCodec -D mapreduce.output.fileoutputformat.compress.type=RECORD -mapper /bin/cat -reducer /bin/cat -input staging/input/stocks/nseCompBig.csv -output staging/input/stocks/gzip_big
```

- Для кодака Bzip2:

```
hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming-*.jar -D mapreduce.job.reduces=1 -D mapred.output.compress=true -D
```

mapred.output.compression.codec=org.apache.hadoop.io.compress.BZip2Codec -D mapreduce.output.fileoutputformat.compress.type=RECORD -mapper /bin/cat -reducer /bin/cat -input staging/input/stocks/nseCompBig.csv -output staging/input/stocks/bzip2_big

Результатами выполнения указанных задач будут сжатые файлы:

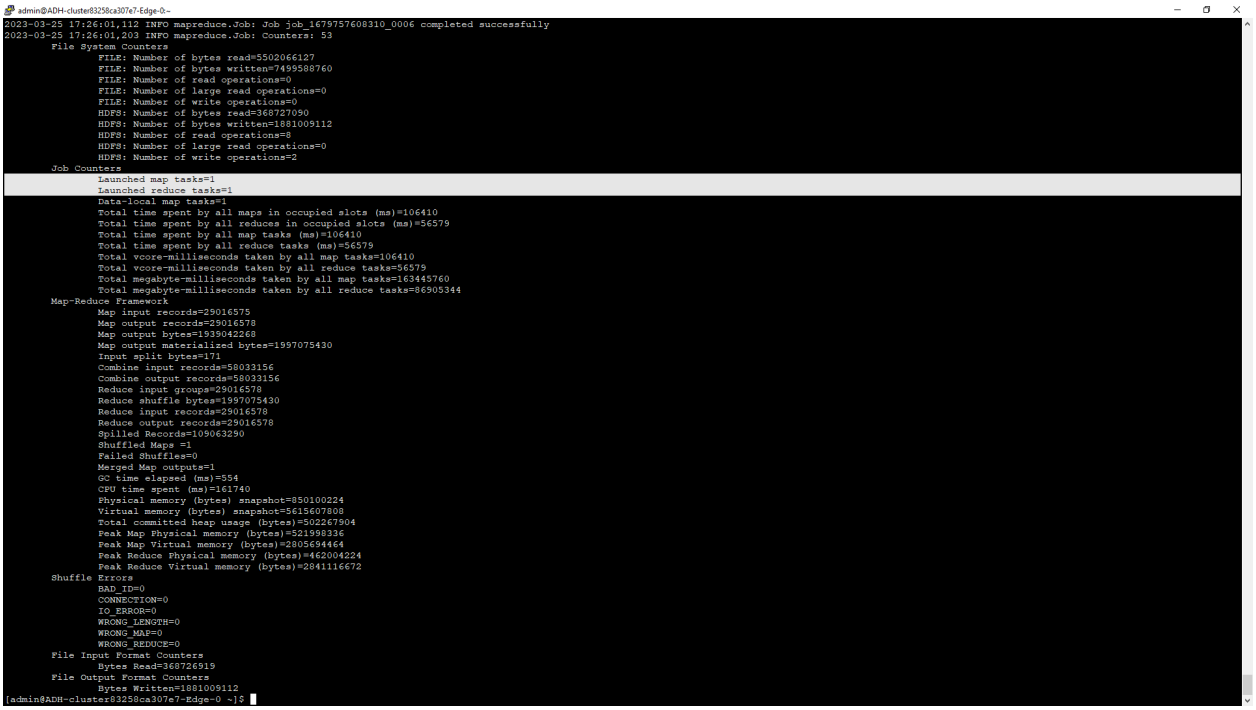
```
-rw-r--r--   3 admin admin  267980233 staging/input/stocks/bzip2_big/part-00000.bz2

-rw-r--r--   3 admin admin  368726919 staging/input/stocks/gzip_big/part-00000.gz
```

Запустите MR-задачу WordCount для каждого из сжатых файлов

1. Запустите MR-задачу WordCount для файла staging/input/stocks/gzip_big/part-00000.gz и каталогом вывода результатов output/mapreduce/wordcount/nseCompGzip.

Пояснение для преподавателя:

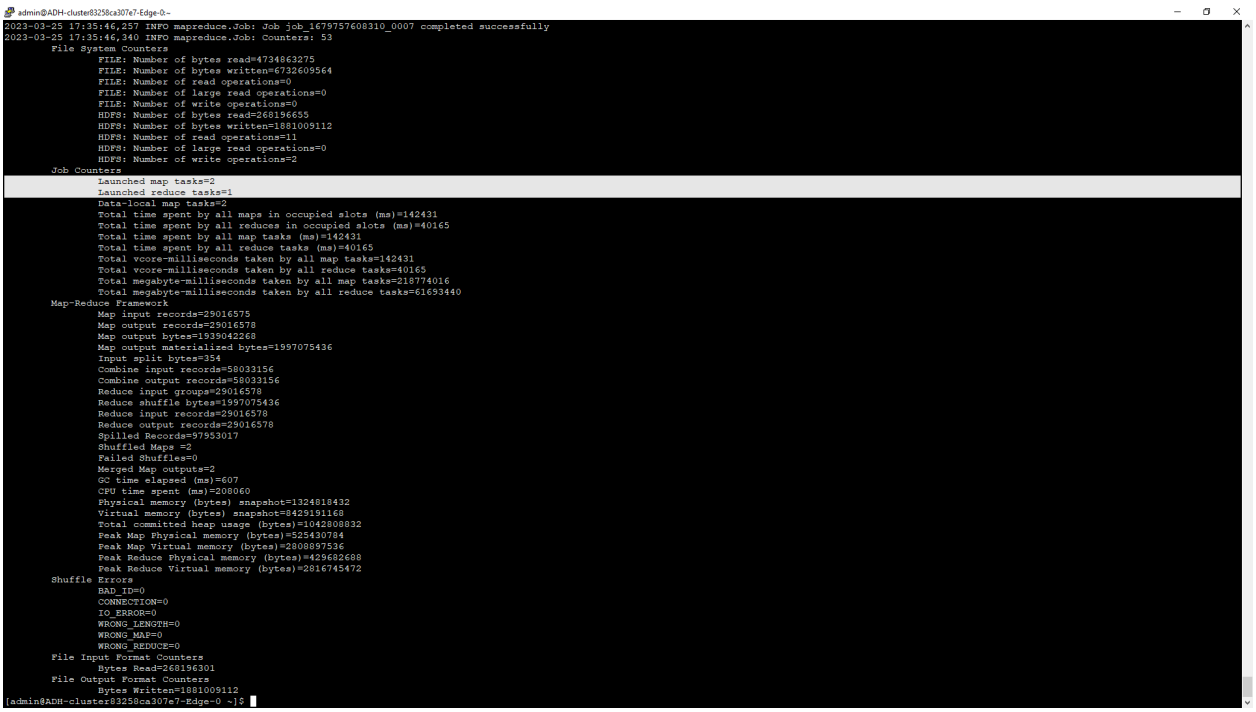
```
hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples-3.1.2.jar wordcount staging/input/stocks/gzip_big/part-00000.gz output/mapreduce/wordcount/nseCompGzip
```



2. Запустите MR-задачу WordCount для файла staging/input/stocks/bzip2_big/part-00000.bz2 и каталогом вывода результатов output/mapreduce/wordcount/nseCompBzip2.

Пояснение для преподавателя:

```
hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples-3.1.2.jar wordcount staging/input/stocks/bzip2_big/part-00000.bz2 output/mapreduce/wordcount/nseCompBzip2
```



3. Объясните количество выделенных Mapper’s MR-задачи для файлов:

- `part-00000.gz` – Launched map tasks=1,
- part-00000.bz2 – Launched map tasks=2?

Пояснение для преподавателя:

При вычислении MR-задачи для сжатых файлов Hadoop выполняет декомпрессию локально на узлах, где будут запускаться Mapper’s. Количество блоков в HDFS для сжатых данных не будет влиять на количество выделяемых Mapper’s для MR-задачи.

Количество выделенных Mapper’s объясняется способом сжатия. Кодак Bzip2 создает сплитируемый сжатый файл, что даёт возможность запустить для каждого сплита свой Mapper.

Кодак Gzip создает сжатый файл без сплитов, т.е. Mapper будет один!