

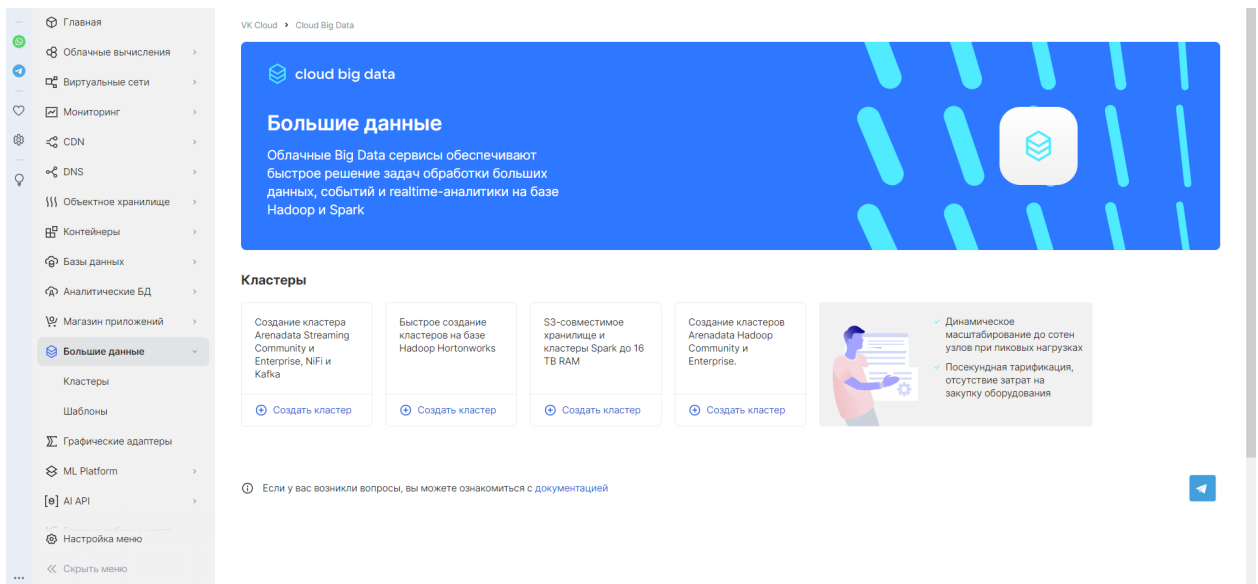
## Agenda

Часть 1. Создание кластера Arenadata Hadoop. Работа с HDFS	2
Создание кластера Arenadata Hadoop Community Test	2
Подключение к кластеру через ssh	4
Создание домашнего каталога в HDFS	5
Загрузка данных в домашний каталог HDFS	5
Вывод содержимого файлов в HDFS	6
Распределение блоков по узлам кластера	6
Расположение блоков на узлах в локальной ОС	7
Часть 2. Запуск задач Map Reduce	8
Запустите MR-задачу WordCount для загруженных данных	8
Выведите результат работы алгоритма	9
Отобразите информацию по выполненной задаче через консоль YARN.	9
Часть 3. Сжатие файлов в Hadoop. Запуск MR-задач для сжатых данных.	10
Произведите сжатие загруженных данных кодами: GzipCodec и BZip2Codec	10
Запустите MR-задачу WordCount для каждого из сжатых файлов	11

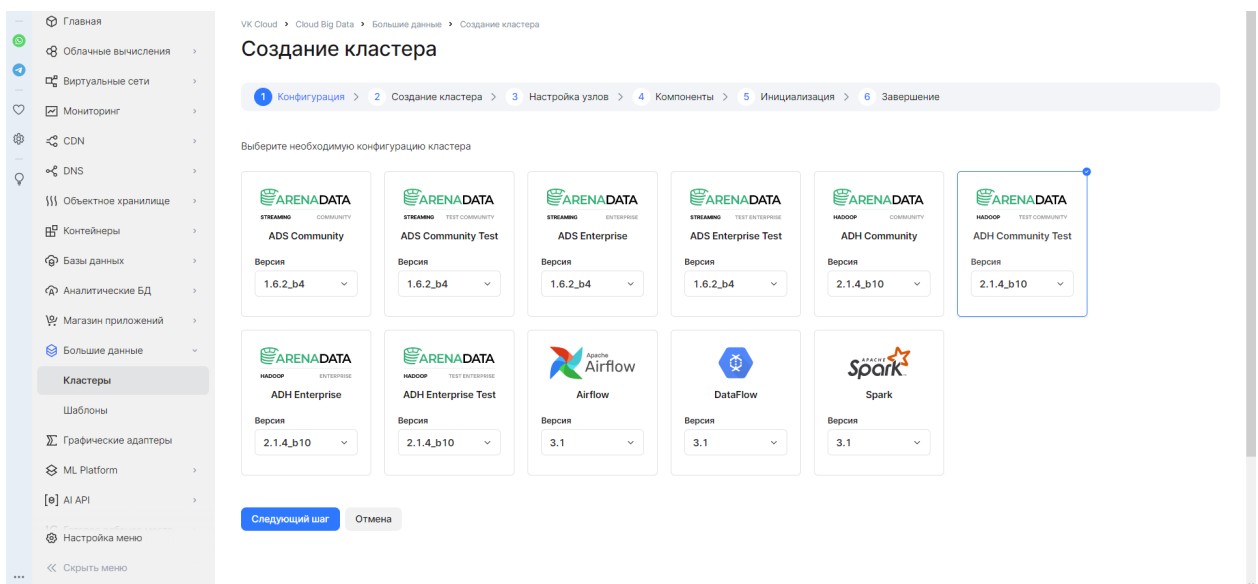
# Часть 1. Создание кластера Arenadata Hadoop. Работа с HDFS

## Создание кластера Arenadata Hadoop Community Test

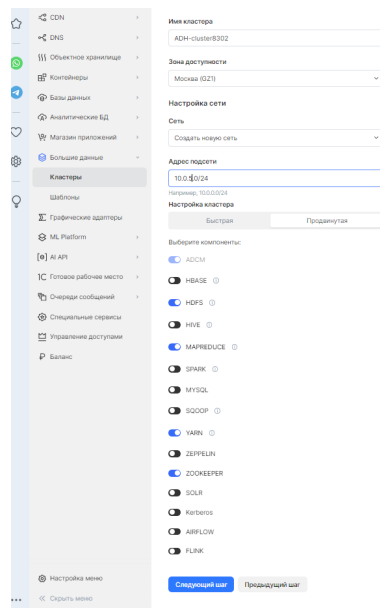
1. Перейдите в личный кабинет VK Cloud. Откройте раздел «**Большие Данные**» и выберите «**Создание кластеров Arenadata Hadoop Community и Enterprise**».



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):

**Master1**

Тип инстанса  
Standard-4-16 4 CPU 16 GB RAM

Количество узлов  
- 1 шт +

Количество дисков на один узел  
- 1 шт +

Размер диска  
- 100 ГБ +

Тип диска ⓘ  
SSD High-IOPS SSD

**Master2**

Тип инстанса  
Standard-4-16 4 CPU 16 GB RAM

Количество узлов  
- 1 шт +

Количество дисков на один узел  
- 1 шт +

Размер диска  
- 100 ГБ +

Тип диска ⓘ  
SSD High-IOPS SSD

**Workers**

Тип инстанса  
Standard-4-16 4 CPU 16 GB RAM

Количество узлов  
- 3 шт +

Количество дисков на один узел  
- 1 шт +

Размер диска  
- 100 ГБ +

Тип диска ⓘ  
SSD High-IOPS SSD

☒ Подключить Edge-узел ⓘ

**Edge**

Тип инстанса  
Basic-1-4 1 CPU 4 GB RAM

Количество узлов  
- 1 шт +

Количество дисков на один узел  
- 1 шт +

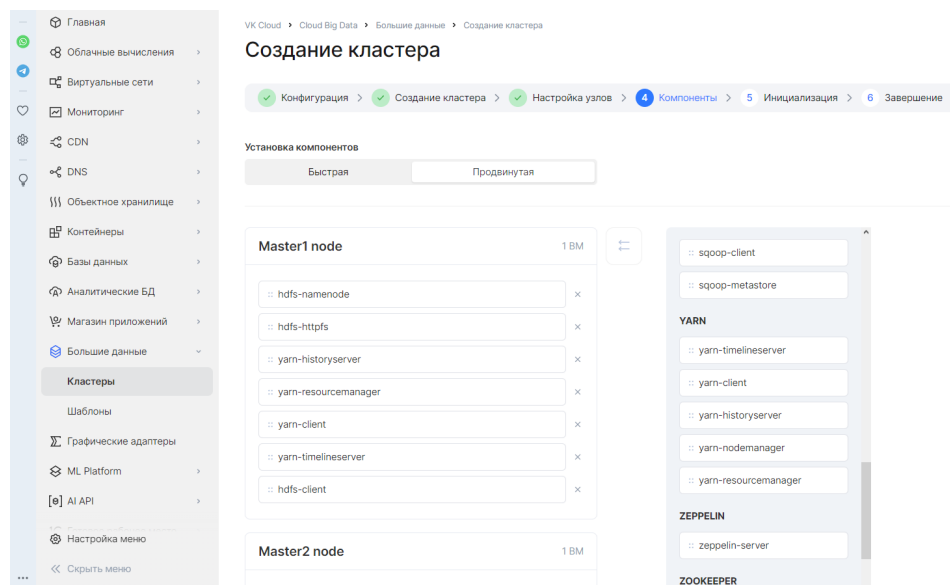
Размер диска  
- 10 ГБ +

Тип диска ⓘ  
SSD High-IOPS SSD

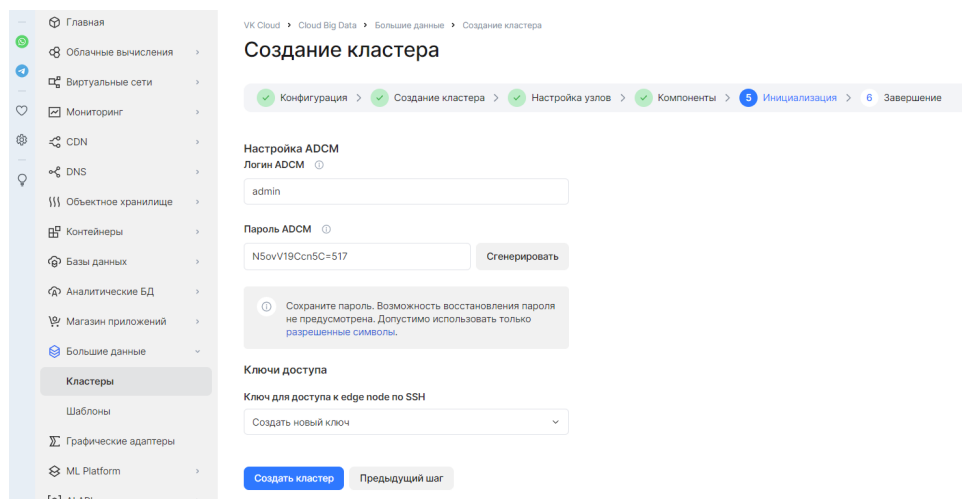
Назначьте внешний IP для инстанса ADCM

- ☐ Подключить Monitoring-узел ⓘ
- ☐ HA Cluster ⓘ
- ☒ Назначить внешний IP ⓘ
- ☐ Автомасштабирование дисков ⓘ

5. Компоненты оставьте без изменений

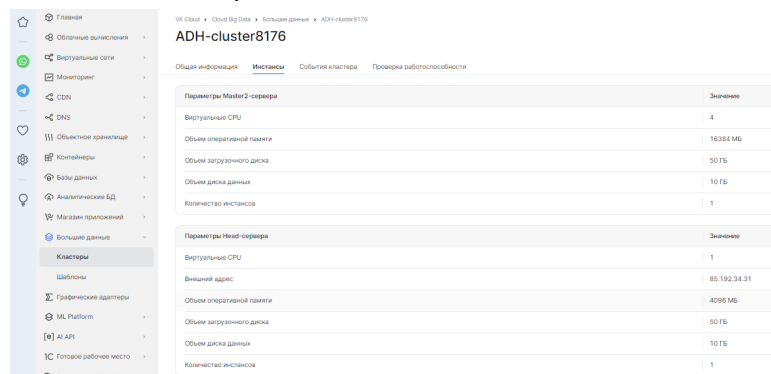


- Укажите пароль для ADCM (Arenadata Cluster Manager).  
Создайте SSH-ключ в формате RSA для подключения к узлу. Запустите создание кластера (длительность операции 20-30 минут).



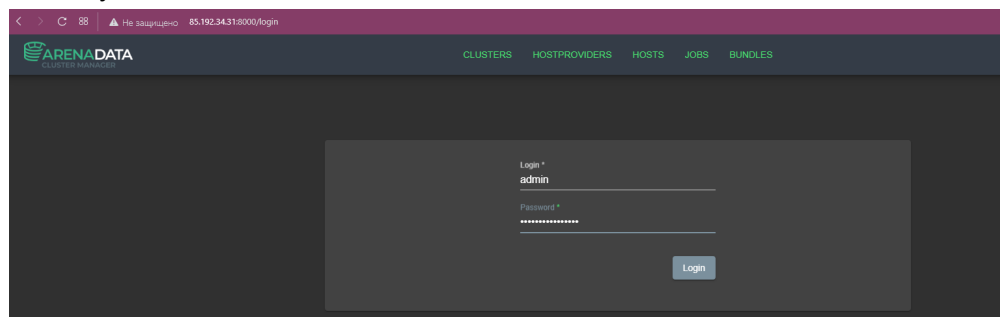
## Подключение к кластеру через ssh

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



- 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.

	Mar 25, 2023, 3:03:31 PM	Mar 25, 2023, 3:24:10 PM	
13 Install ^	ADH		✓
Services dependency check	Mar 25, 2023, 3:03:31 PM	Mar 25, 2023, 3:03:47 PM	✓
Pre-install check	Mar 25, 2023, 3:03:48 PM	Mar 25, 2023, 3:04:05 PM	✓
Pre-configure check	Mar 25, 2023, 3:04:05 PM	Mar 25, 2023, 3:04:20 PM	✓
Host-service os_family compatibility check	Mar 25, 2023, 3:04:20 PM	Mar 25, 2023, 3:04:42 PM	✓

4. Запустите команду для проверки работоспособности кластера:
- ```
hadoop fs -ls /
```

```
admin@ADH-cluster8176c60d9016-Edge-0 ~]$ hadoop fs -ls /
Found 4 items
drwxr-xr-x - yarn hadoop 0 2023-03-25 08:22 /logs
drwxr-xr-x - hdfs hadoop 0 2023-03-25 08:20 /system
drwxr-xr-x - hdfs hadoop 0 2023-03-25 08:25 /tmp
drwxr-xr-x - hdfs hadoop 0 2023-03-25 08:20 /user
admin@ADH-cluster8176c60d9016-Edge-0 ~]$
```

## 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
```

```
admin@ADH-cluster8176c60d9016-Edge-0 ~]$ hadoop fs -ls /user
Found 5 items
drwxr-xr-x - admin admin 0 2023-03-25 09:18 /user/admin
drwxr-xr-x - hdfs hadoop 0 2023-03-25 08:17 /user/hdfs
drwxr-xr-x - mapred hadoop 0 2023-03-25 08:21 /user/history
drwxr-xr-x - mapred mapred 0 2023-03-25 08:20 /user/mapred
drwxr-xr-x - yarn yarn 0 2023-03-25 08:26 /user/yarn
admin@ADH-cluster8176c60d9016-Edge-0 ~]$
```

## 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/PoXJ2YsKh2HZlw>  
<https://disk.yandex.ru/d/QbLv4KUFubO7A>
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
```

```
admin@ADH-cluster8176c60d9016-Edge-0-~$
[admin@ADH-cluster8176c60d9016-Edge-0 ~]$ hadoop fs -tail staging/input/stocks/nseCompSmall.csv
5,61.35,61.25,61.35,525
S803300,CHAMBLFERT,20150827,09:36:00,61.35,61.35,61.25,61.25,226
S803301,CHAMBLFERT,20150827,09:37:00,61.25,61.25,61.2,61.2,1475
S803302,CHAMBLFERT,20150827,09:42:00,61.2,61.2,61.2,61.2,5
S803303,CHAMBLFERT,20150827,09:44:00,61.2,61.2,61.2,61.2,150
S803304,CHAMBLFERT,20150827,09:45:00,61.35,61.35,61.2,61.2,200
S803305,CHAMBLFERT,20150827,09:48:00,61.4,61.4,61.4,61.4,250
S803306,CHAMBLFERT,20150827,09:49:00,61.2,61.4,61.2,61.4,513
S803307,CHAMBLFERT,20150827,09:50:00,61.4,61.4,61.4,61.4,1
S803308,CHAMBLFERT,20150827,09:53:00,61.4,61.5,61.3,61.3,1680
S803309,CHAMBLFERT,20150827,09:55:00,61.25,61.4,61.25,61.25,771
S803310,CHAMBLFERT,20150827,09:56:00,61.25,61.25,61.2,61.2,1245
S803311,CHAMBLFERT,20150827,09:58:00,61.2,61.2,61.2,61.2,1200
S803312,CHAMBLFERT,20150827,09:59:00,61.25,61.25,61.25,61.25,2
S803313,CHAMBLFERT,20150827,10:01:00,61.25,61.25,61.25,61.25,1
S803314,CHAMBLFERT,20150827,10:02:00,61.25,61.25,61.25,61.25,115
S803315,CHAMBLFERT,20150827,10:03:00,61.25,61.25,61.2,61.25,232
[admin@ADH-cluster8176c60d9016-Edge-0 ~]$
```

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

```
admin@ADH-cluster8176c60d9016-Edge-0-~$
[admin@ADH-cluster8176c60d9016-Edge-0 ~]$ hadoop fs -tail staging/input/stocks/nseCompBig.csv
L,20150428,15:13:00,96.8,96.8,96.65,96.8,1371
29016558,TWL,20150428,15:14:00,96.75,96.8,96.7,96.7,2840
29016559,TWL,20150428,15:15:00,96.7,96.8,96.5,96.8,1674
29016560,TWL,20150428,15:16:00,96.8,96.8,96.6,96.75,6162
29016561,TWL,20150428,15:17:00,96.55,96.7,96.4,96.5,2977
29016562,TWL,20150428,15:18:00,96.3,96.5,96.2,96.2,4196
29016563,TWL,20150428,15:19:00,96.35,96.9,96.1,96.9,15881
29016564,TWL,20150428,15:20:00,96.3,96.5,96.15,96.35,1532
29016565,TWL,20150428,15:21:00,96.35,96.35,96.96,96.25,9005
29016566,TWL,20150428,15:22:00,96.25,96.65,96.25,96.65,2450
29016567,TWL,20150428,15:23:00,96.7,96.75,96.4,96.4,3774
29016568,TWL,20150428,15:24:00,96.5,96.75,96.35,96.35,5197
29016569,TWL,20150428,15:25:00,96.3,96.4,96.2,96.25,3996
29016570,TWL,20150428,15:26:00,96.4,96.7,96.4,96.7,6459
29016571,TWL,20150428,15:27:00,96.75,96.85,96.65,96.8,3717
29016572,TWL,20150428,15:28:00,96.65,97.05,96.65,97.10686
29016573,TWL,20150428,15:29:00,97.05,97.3,97.25,97.25,9242
29016574,TWL,20150428,15:30:00,97.35,97.4,97.15,97.35,7245
[admin@ADH-cluster8176c60d9016-Edge-0 ~]$
```

## Распределение блоков по узлам кластера

Выведите распределение блоков по узлам кластера для файлов 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 - ssh hdfs fsck staging/input/stocks/nseCompSmall.csv -files -blocks -locations
Connecting to namednode via http://ADH-cluster83258ca307e7-Master1:0.mcs.local:1970/fsck?ugi=admin&files=1&blocks=1&locations=1&path=2&Fuser=2&Fadmin=2&Fstaging=2&Finput=2&Fstocks=2&FmseCompSmall.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_1073741889_1064 len=134217728 Live_rep1=3 [DatanodeInfoWithStorage[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], 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_rep1=3 [DatanodeInfoWithStorage[10.0.0.20:9866,DS-17d4f212-b9c1-46df-bc19-4c1496783666,DISK], DatanodeInfoWithStorage[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_rep1=3 [DatanodeInfoWithStorage[10.0.0.32:9866,DS-363d9040-21d4-49c2-8d67-f0f0fe7058d1,DISK], DatanodeInfoWithStorage[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-cluster8176c0d9016-Edge-0 - ssh hdfs fsck staging/input/stocks/nseCompBig.csv -files -blocks -locations
Connecting to namednode via http://ADH-cluster8176c0d9016-Master1:0.mcs.local:1970/fsck?ugi=admin&files=1&blocks=1&locations=1&path=2&Fuser=2&Fadmin=2&Fstaging=2&Finput=2&Fstocks=2&FmseCompBig.csv
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 1922975955 bytes, replicated: replication=3, 14 block(s): OK
0. BP-714958517-10.0.0.6-1679732120535:blk_1073741909_1075 len=134217728 Live_rep1=3 [DatanodeInfoWithStorage[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], 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_rep1=3 [DatanodeInfoWithStorage[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], 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_rep1=3 [DatanodeInfoWithStorage[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], 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_rep1=3 [DatanodeInfoWithStorage[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], DatanodeInfoWithStorage[10.0.0.31:9866,DS-ecf7e6b4-2572-446e-b5b4-45b4b56e338a,DISK]]
4. BP-714958517-10.0.0.6-1679732120535:blk_1073741903_1079 len=134217728 Live_rep1=3 [DatanodeInfoWithStorage[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], DatanodeInfoWithStorage[10.0.0.31:9866,DS-ecf7e6b4-2572-446e-b5b4-45b4b56e338a,DISK]]
5. BP-714958517-10.0.0.6-1679732120535:blk_1073741904_1080 len=134217728 Live_rep1=3 [DatanodeInfoWithStorage[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], 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_rep1=3 [DatanodeInfoWithStorage[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], 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_rep1=3 [DatanodeInfoWithStorage[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], DatanodeInfoWithStorage[10.0.0.31:9866,DS-ecf7e6b4-2572-446e-b5b4-45b4b56e338a,DISK]]
8. BP-714958517-10.0.0.6-1679732120535:blk_1073741907_1083 len=134217728 Live_rep1=3 [DatanodeInfoWithStorage[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], DatanodeInfoWithStorage[10.0.0.31:9866,DS-ecf7e6b4-2572-446e-b5b4-45b4b56e338a,DISK]]
9. BP-714958517-10.0.0.6-1679732120535:blk_1073741908_1084 len=134217728 Live_rep1=3 [DatanodeInfoWithStorage[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], DatanodeInfoWithStorage[10.0.0.31:9866,DS-ecf7e6b4-2572-446e-b5b4-45b4b56e338a,DISK]]
10. BP-714958517-10.0.0.6-1679732120535:blk_1073741909_1085 len=134217728 Live_rep1=3 [DatanodeInfoWithStorage[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], DatanodeInfoWithStorage[10.0.0.12:9866,DS-58d88aa8-38dd-4b11-abac-d267a7fdc7ae,DISK]]
11. BP-714958517-10.0.0.6-1679732120535:blk_1073741910_1086 len=134217728 Live_rep1=3 [DatanodeInfoWithStorage[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], DatanodeInfoWithStorage[10.0.0.31:9866,DS-ecf7e6b4-2572-446e-b5b4-45b4b56e338a,DISK]]
12. BP-714958517-10.0.0.6-1679732120535:blk_1073741911_1087 len=134217728 Live_rep1=3 [DatanodeInfoWithStorage[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], 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_rep1=3 [DatanodeInfoWithStorage[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], 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 130212569 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 %)

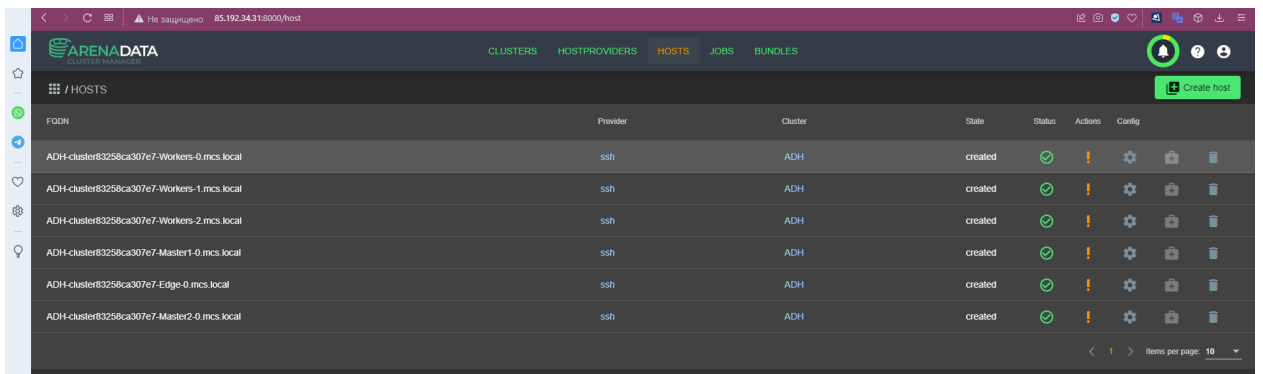
Erasure Coded Block Groups:
Total size: 0 B
Total files: 0
```

## Расположение блоков на узлах в локальной ОС

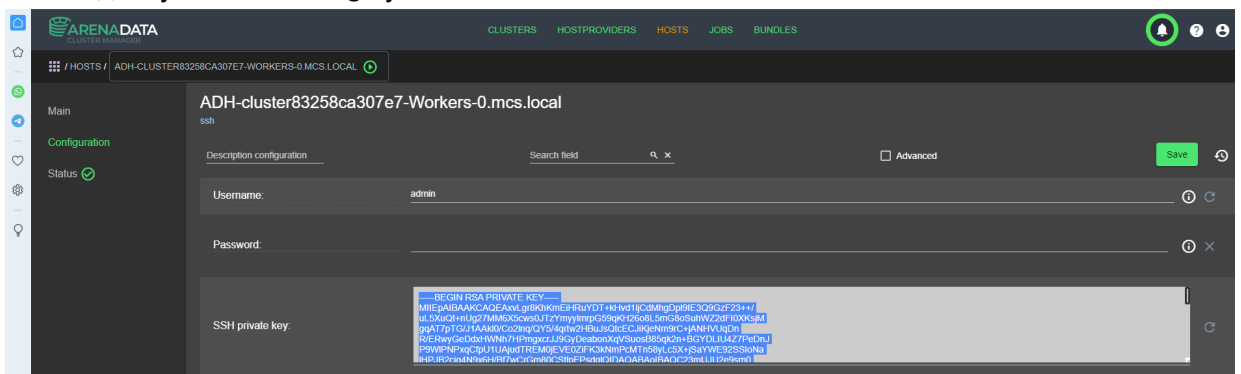
1. Найдите имя блока файла в выводе команды hdfs fsck

```
/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_1073741889_1064 len=134217728 Live_rep1=3 [DatanodeInfoWithStorage[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], 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_rep1=3 [DatanodeInfoWithStorage[10.0.0.20:9866,DS-17d4f212-b9c1-46df-bc19-4c1496783666,DISK], DatanodeInfoWithStorage[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_rep1=3 [DatanodeInfoWithStorage[10.0.0.32:9866,DS-363d9040-21d4-49c2-8d67-f0f0fe7058d1,DISK], DatanodeInfoWithStorage[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*
```

```
admin@ADH-cluster83258ca307e7-Workers-0-
[admin@ADH-cluster83258ca307e7-Workers-0 ~]$ sudo find /srv/hadoop-hdfs -name *blk_1073741888*
/srv/hadoop-hdfs/data/current/BP-1262089974-10.0.0.7-1679757265470/current/finalized/subdir0/subdir0/blk_1073741888_1064.meta
[admin@ADH-cluster83258ca307e7-Workers-0 ~]$
```

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/current/finalized/subdir0/subdir0/blk_1073741888
```

```
admin@ADH-cluster83258ca307e7-Workers-0-
[admin@ADH-cluster83258ca307e7-Workers-0 ~]$ sudo head /srv/hadoop-hdfs/data/current/BP-1262089974-10.0.0.7-1679757265470/current/finalized/subdir0/subdir0/blk_1073741888
1,31INFOTECH,20150703,09:16:00,4.55,4.55,4.55,4.55,835
2,31INFOTECH,20150703,09:17:00,4.55,4.55,4.55,4.55,390
3,31INFOTECH,20150703,09:18:00,4.55,4.55,4.55,4.55,1000
4,31INFOTECH,20150703,09:19:00,4.55,4.55,4.55,4.55,1150
5,31INFOTECH,20150703,09:20:00,4.55,4.55,4.55,4.55,1100
6,31INFOTECH,20150703,09:21:00,4.55,4.55,4.55,4.55,4500
7,31INFOTECH,20150703,09:22:00,4.55,4.55,4.55,4.55,1500
8,31INFOTECH,20150703,09:23:00,4.55,4.55,4.55,4.55,5540
9,31INFOTECH,20150703,09:24:00,4.55,4.55,4.55,4.55,1500
10,31INFOTECH,20150703,09:26:00,4.55,4.55,4.55,4.55,200
[admin@ADH-cluster83258ca307e7-Workers-0 ~]$
```

## 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
```



```

admin@ADI-cluster83258ca307e7-Edge-0 ~$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples-3.1.2.jar
An example program must be given as the first argument.

Valid program names are:
aggregatewordcount: An Aggregate based map/reduce program that counts the words in the input files.
aggregatewordhist: An Aggregate based map/reduce program that computes the histogram of the words in the input files.
bbp: A map/reduce program that uses Bailey-Borwein-Plouffe to compute exact digits of Pi.
dbcount: An example job that counts the pageview counts from a database.
distbp: A map/reduce program that uses a BPZ-type format to compute exact bits of Pi.
gperf: A map/reduce program that counts the matches of a regexp in the input.
join: A job that effects a join over sorted, equally partitioned datasets
multifilev: A job that counts words from several files.
pentomino: A map/reduce job laying programs to find solutions to pentomino problems.
pi: A map/reduce program that estimates Pi using a quasi-Monte Carlo method.
randomtextwriter: A map/reduce program that writes 10GB of random textual data per node.
randomwriter: A map/reduce program that writes 10GB of random data per node.
secondarysort: An example defining a secondary sort to the reducer.
sort: A map/reduce program that sorts the data written by the random writer.
sudoku: A sudoku solver.
terasort: Generate data for the terasort
terasort: Run the terasort
teravalidate: Checking results of terasort
wordcount: A map/reduce program that counts the words in the input files.
wordmean: A map/reduce program that counts the average length of the words in the input files.
wordmedian: A map/reduce program that counts the median length of the words in the input files.
wordstandarddeviation: A map/reduce program that counts the standard deviation of the length of the words in the input files.

admin@ADI-cluster83258ca307e7-Edge-0 ~$

```

- 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
```

```
[admin@ADH-cluster83258ca307e7-Edge-0] ~$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples-3.1.2.jar wordcount staging/output/stocks/nasCompSmall.csv output/mapreduce/wordcount/nasCompSmall
2023-03-25 16:06:43,418 INFO client.HMRProxy: Connecting to ResourceManager at ADH-cluster83258ca307e7-Master1-0.mcs.local/10.0.0.7:8032
2023-03-25 16:06:43,602 INFO client.HMRProxy: Connecting to Application History server at ADH-cluster83258ca307e7-Master1-0.mcs.local/10.0.0.7:10200
2023-03-25 16:06:43,914 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /user/admin/.staging/job_1679757608310_0003
2023-03-25 16:06:44,116 INFO InputFileFormat: Total input files to process : 1
2023-03-25 16:06:44,257 INFO mapreduce.JobSubmitter: number of splits:3
2023-03-25 16:06:44,524 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1679757608310_0003
2023-03-25 16:06:44,566 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-03-25 16:06:44,762 INFO conf.Configuration: resource-types.xml not found
2023-03-25 16:06:44,769 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-03-25 16:06:44,828 INFO impl.YarnClientImpl: Submitted application application_1679757608310_0003
2023-03-25 16:06:44,862 INFO mapreduce.Job: The url to track the job: http://ADH-cluster83258ca307e7-Master1-0.mcs.local:8080/proxy/application_1679757608310_0003/
2023-03-25 16:06:44,869 INFO mapreduce.Job: Running job: job_1679757608310_0003
2023-03-25 16:06:45,958 INFO mapreduce.Job: Job job_1679757608310_0003 running in uber mode : false
2023-03-25 16:06:46,359 INFO mapreduce.Job: map 0% reduce 0%
2023-03-25 16:07:01,052 INFO mapreduce.Job: map 33% reduce 0%
2023-03-25 16:07:03,065 INFO mapreduce.Job: map 67% reduce 0%
2023-03-25 16:07:05,056 INFO mapreduce.Job: map 100% reduce 0%
2023-03-25 16:07:15,175 INFO mapreduce.Job: map 100% reduce 100%
2023-03-25 16:07:15,185 INFO mapreduce.Job: Job job_1679757608310_0003 completed successfully
2023-03-25 16:07:15,268 INFO mapreduce.Job: Counter: 54
File System Counters
FILE: Number of bytes read=802353382
FILE: Number of bytes written=1204424515
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=366619401
HDFS: Number of bytes written=377963386
HDFS: Number of read operations=14
HDFS: Number of large read operations=0
HDFS: Number of write operations=0
Job Counters
Killed map tasks=1
Launched map tasks=3
Launched reduce tasks=1
Data-local map tasks=3
Total time spent by all maps in occupied slots (ms)=30106
Total time spent by all reduces in occupied slots (ms)=11003
Total time spent by all map tasks (ms)=30106
Total time spent by all reduce tasks (ms)=11003
Total vcores-milliseconds taken by all map tasks=30106
Total vcores-milliseconds taken by all reduce tasks=11003
Total megabyte-milliseconds taken by all map tasks=4248216
Total megabyte-milliseconds taken by all reduce tasks=1690608
Map-Reduce Framework
Map input records=5803318
Map output records=5803318
Map output bytes=389570022
Map output materialized bytes=401176676
Input split bytes=507
Combine input records=10047506
Combine output records=10047506
Reduce input groups=5803318
Reduce shuffle bytes=401176676
Reduce input records=5803318
Reduce output records=5803318
Spilled Records=740954
Shuffled Maps =3
Failed Shuffles=0
Wrote Map outputs=3
```

## 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
```

[illegible]

## 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
```

```
admin@ADH-cluster83258ca307e7-Edge-0: ~$ yarn application -appStates FINISHED -list
2023-03-25 16:21:03,680 INFO client.RMProxy: Connecting to ResourceManager at ADH-cluster83258ca307e7-Master1-0.mcs.local/10.0.0.7:8032
2023-03-25 16:21:03,904 INFO client.AHSProxy: Connecting to Application History server at ADH-cluster83258ca307e7-Master1-0.mcs.local/10.0.0.7:10200
Total number of applications (application-types: [], states: [FINISHED] and tags: [])=3
ApplicationId      ApplicationName      ApplicationType      User      Queue      default      State      Final-State      Progress      Tracking-URL
application_1679757608310_0001  QuasiMonteCarlo      MAPREDUCE      yarn      default      FINISHED      SUCCEEDED      100%      http://ADH-cluster83258ca307e7-Mast
eri-0.mcs.local:19888/jobhistory/job/job_1679757608310_0001
application_1679757608310_0002  QuasiMonteCarlo      MAPREDUCE      yarn      default      FINISHED      SUCCEEDED      100%      http://ADH-cluster83258ca307e7-Mast
eri-0.mcs.local:19888/jobhistory/job/job_1679757608310_0002
application_1679757608310_0003  word count      MAPREDUCE      admin      default      FINISHED      SUCCEEDED      100%      http://ADH-cluster83258ca307e7-Mast
eri-0.mcs.local:19888/jobhistory/job/job_1679757608310_0003
[admin@ADH-cluster83258ca307e7-Edge-0 ~]$
```

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

### Пояснение для преподавателя:

```
yarn logs -applicationId application_1679757608310_0003
```

```
admin@ADH-cluster83258ca307e7-Edge-0: ~$ yarn logs -applicationId application_1679757608310_0003
2023-03-25 16:06:55,353 INFO [main] org.apache.hadoop.mapred.MapTask: Spilling map output
2023-03-25 16:06:55,353 INFO [main] org.apache.hadoop.mapred.MapTask: bufstart = 0; bufend = 67934942; bufvoid = 104857600
2023-03-25 16:06:55,353 INFO [main] org.apache.hadoop.mapred.MapTask: kvstart = 26214396(104857584); kvend = 22226612(88906448); length = 3987785/6553600
2023-03-25 16:06:57,142 INFO [main] org.apache.hadoop.mapred.MapTask: (EQUATOR) 71929502 kv 17982368 (71929472)
2023-03-25 16:06:57,142 INFO [SpillThread] org.apache.hadoop.mapred.MapTask: Finished spill 0
2023-03-25 16:06:57,142 INFO [main] org.apache.hadoop.mapred.MapTask: (RESET) equator 71929502 kv 17982368 (71929472) kvi 16986012(67944048)
2023-03-25 16:06:57,795 INFO [main] org.apache.hadoop.mapred.MapTask: Spilling map output
2023-03-25 16:06:57,800 INFO [main] org.apache.hadoop.mapred.MapTask: bufstart = 71929502; bufend = 34953072; bufvoid = 104857571
2023-03-25 16:06:57,800 INFO [main] org.apache.hadoop.mapred.MapTask: kvstart = 17982368 (71929472); kvend = 13981140 (55924560); length = 4001229/6553600
2023-03-25 16:06:57,800 INFO [main] org.apache.hadoop.mapred.MapTask: (EQUATOR) 38947632 kvi 5736904 (38947616)
2023-03-25 16:06:57,894 INFO [main] org.apache.hadoop.mapred.MapTask: Starting flush of map output
2023-03-25 16:06:59,063 INFO [SpillThread] org.apache.hadoop.mapred.MapTask: Finished spill 1
2023-03-25 16:06:59,063 INFO [main] org.apache.hadoop.mapred.MapTask: (RESET) equator 38947632 kv 9736904 (38947616) kvi 9359900 (37439600)
2023-03-25 16:06:59,063 INFO [main] org.apache.hadoop.mapred.MapTask: Spilling map output
2023-03-25 16:06:59,063 INFO [main] org.apache.hadoop.mapred.MapTask: bufstart = 38947632; bufend = 45715300; bufvoid = 104857600
2023-03-25 16:06:59,063 INFO [main] org.apache.hadoop.mapred.MapTask: kvstart = 9736904 (38947616); kvend = 9359904 (37439616); length = 377001/6553600
2023-03-25 16:06:59,170 INFO [main] org.apache.hadoop.mapred.MapTask: Finished spill 2
2023-03-25 16:06:59,177 INFO [main] org.apache.hadoop.mapred.Mapper: Merging 3 sorted segments
2023-03-25 16:06:59,184 INFO [main] org.apache.hadoop.mapred.Mapper: Down to the last merge-pass, with 3 segments left of total size: 146766578 bytes
2023-03-25 16:07:01,030 INFO [main] org.apache.hadoop.mapred.Task: Task attempt_1679757608310_0003_m_000001_0 is done. And is in the process of committing
2023-03-25 16:07:01,115 INFO [main] org.apache.hadoop.mapred.Task: Task attempt_1679757608310_0003_m_000001_0 done.
2023-03-25 16:07:01,126 INFO [main] org.apache.hadoop.mapred.Task: Final Counters for attempt_1679757608310_0003_m_000001_0: Counters: 28
File System Counters
FILE: Number of bytes read=146766781
FILE: Number of bytes written=293757178
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=134348969
HDFS: Number of bytes written=0
HDFS: Number of read operations=3
HDFS: Number of large read operations=0
HDFS: Number of write operations=0
Map-Reduce Framework
Map input records=2091506
Map output records=2091506
Map output bytes=142839751
Map output materialized bytes=146766769
Input split bytes=169
Combine input records=4183012
Combine output records=4183012
Spilled Records=4183012
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=119
CPU time spent (ms)=9060
Physical memory (bytes) snapshot=506101760
Virtual memory (bytes) snapshot=2810118144
Total committed heap usage (bytes)=48824416
Peak Map Physical memory (bytes)=506101760
Peak Map Virtual memory (bytes)=2810118144
File Input Format Counters
Bytes Read=134348900
2023-03-25 16:07:01,226 INFO [main] org.apache.hadoop.metrics2.impl.MetricsSystemImpl: Stopping MapTask metrics system...
2023-03-25 16:07:01,227 INFO [main] org.apache.hadoop.metrics2.impl.MetricsSystemImpl: MapTask metrics system stopped.
2023-03-25 16:07:01,227 INFO [main] org.apache.hadoop.metrics2.impl.MetricsSystemImpl: MapTask metrics system shutdown complete.
End of LogType:syslog
*****
```

## Часть 13. Сжатие файлов в Hadoop. Запуск MR-задач для сжатых данных.

Произведите сжатие загруженных данных кодами: GzipCodec и BZip2Codec

Сжатие исходных данных можно выполнить с помощью MR-задачи:

- Для кода 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 кода 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
```

```
admin@ADH-cluster83258ca307e7-Edge-0-
2023-03-25 17:26:01,112 INFO mapreduce.Job: Job job_1679757608310_0006 completed successfully
2023-03-25 17:26:01,202 INFO mapreduce.Job: Counters: 53
File System Counters
  FILE: Number of bytes read=5502066127
  FILE: Number of bytes written=7495988760
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=368727090
  HDFS: Number of bytes written=1881009112
  HDFS: Number of read operations=8
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2
Job Counters
  Launched map tasks=1
  Launched reduce tasks=1
  Data-local map tasks=1
  Total time spent by all maps in occupied slots (ms)=106410
  Total time spent by all reduces in occupied slots (ms)=56579
  Total time spent by all map tasks (ms)=106410
  Total time spent by all reduce tasks (ms)=56579
  Total vcore-milliseconds taken by all map tasks=106410
  Total vcore-milliseconds taken by all reduce tasks=56579
  Total megabyte-milliseconds taken by all map tasks=163445760
  Total megabyte-milliseconds taken by all reduce tasks=36503344
Map-Reduce Framework
  Map input records=29016578
  Map output records=29016578
  Map output bytes=1939042268
  Map output materialized bytes=1997075430
  Input split bytes=171
  Combine input records=58033156
  Combine output records=58033156
  Reduce input groups=29016578
  Reduce shuffle bytes=1997075430
  Reduce input records=29016578
  Reduce output records=29016578
  Spilled Records=19063290
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=554
  CPU time spent (ms)=161740
  Physical memory (bytes) snapshot=850100224
  Virtual memory (bytes) snapshot=5615607808
  Total committed heap usage (bytes)=502267904
  Peak Map Physical memory (bytes)=521988216
  Peak Map Virtual memory (bytes)=2805694464
  Peak Reduce Physical memory (bytes)=462004224
  Peak Reduce Virtual memory (bytes)=2841116672
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_READER=0
File Input Format Counters
  Bytes Read=368726919
File Output Format Counters
  Bytes Written=1881009112
[admin@ADH-cluster83258ca307e7-Edge-0 ~]$
```

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
```

```
admin@ADM-cluster03258ca207e7-Edge0:
2022-03-25 17:51:46,257 INFO mapreduce.Job: Job job_1679757608310_0007 completed successfully
2023-03-25 17:55:46,340 INFO mapreduce.Job: Counters: 53
File System Counters
  FILE: Number of bytes read=4734863275
  FILE: Number of bytes written=6732609564
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=268196655
  HDFS: Number of bytes written=1881009112
  HDFS: Number of read operations=11
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2
Job Counters
  Launched map tasks=2
  Launched reduce tasks=1
Data-Local map tasks=2
  Total time spent by all maps in occupied slots (ms)=142431
  Total time spent by all reduces in occupied slots (ms)=40165
  Total time spent by all map tasks (ms)=142431
  Total time spent by all reduce tasks (ms)=40165
  Total vcore-milliseconds taken by all map tasks=142431
  Total vcore-milliseconds taken by all reduce tasks=40165
  Total megabyte-milliseconds taken by all map tasks=218774016
  Total megabyte-milliseconds taken by all reduce tasks=61693440
Map-Reduce Framework
  Map input records=29016575
  Map output records=29016578
  Map output bytes=1939042268
  Map output materialized bytes=1997075436
  Input split bytes=354
  Combine input records=58033156
  Combine output records=58033156
  Reduce input groups=29016578
  Reduce shuffle bytes=1997075436
  Reduce input records=29016578
  Reduce output records=29016578
  Spilled Records=97953017
  Shuffled Maps=2
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=607
  CPU time spent (ms)=208060
  Physical memory (bytes) snapshot=1324818432
  Virtual memory (bytes) snapshot=8429191168
  Total committed heap usage (bytes)=1942808832
  Peak Map Physical memory (bytes)=525430784
  Peak Map Virtual memory (bytes)=2808897536
  Peak Reduce Physical memory (bytes)=482662688
  Peak Reduce Virtual memory (bytes)=2816745472
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=268196655
File Output Format Counters
  Bytes Written=1881009112
admin@ADM-cluster03258ca207e7-Edge0 ~19
```

3. Объясните количество выделенных Mapper's MR-задачи для файлов:
- part-000000.gz – Launched map tasks=1,
  - part-000000.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. Кодок Bzip2 создает сплитуемый сжатый файл, что даёт возможность запустить для каждого сплита свой Mapper.

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