ДЗ 1

Кондраев Дмитрий

1 Разверните у себя hadoop кластер внутри docker контейнера (1 балл)

1. Установка Docker

```
mhq@mhq-envy:~$ curl -s https://get.docker.com/ | sudo sh
```

```
mhq@mhq-envy:~$ sudo docker version
[sudo] password for mhq:
Sorry, try again.
[sudo] password for mhq:
Client: Docker Engine - Community
 Version:           20.10.18
 API version:       1.41
 Go version:        go1.18.6
 Git commit:        b40c2f6
 Built:             Thu Sep  8 23:11:43 2022
 OS/Arch:           linux/amd64
 Context:           default
 Experimental:      true

Server: Docker Engine - Community
 Engine:
  Version:          20.10.18
  API version:      1.41 (minimum version 1.12)
  Go version:       go1.18.6
  Git commit:       e42327a
  Built:            Thu Sep  8 23:09:30 2022
  OS/Arch:          linux/amd64
  Experimental:     false
 containerd:
  Version:          1.6.8
  GitCommit:        9cd3357b7fd7218e4aec3eae239db1f68a5a6ec6
 runc:
  Version:          1.1.4
  GitCommit:        v1.1.4-0-g5fd4c4d
 docker-init:
  Version:          0.19.0
  GitCommit:        de40ad0
```

Проверка:

```
mhq@mhq-envy:~$ docker compose version
Docker Compose version v2.10.2
```

2. Развертывание Hadoop

Docker-образ Hadoop распакован в директорию `~/dev/img-hdp-hadoop/`. Сборка образа:

```
mhq@mhq-envy:~/dev/img-hdp-hadoop$ sudo docker build -t img-hdp-hadoop .
<OUTPUT TRIMMED FOR BREVITY>
Successfully built 573156a441d0
Successfully tagged img-hdp-hadoop:latest
```

Проверим список образов:

```
mhq@mhq-envy:~$ sudo docker images
REPOSITORY       TAG       IMAGE ID       CREATED          SIZE
img-hdp-hadoop   latest    573156a441d0   16 minutes ago   1.82GB
ubuntu           18.04     35b3f4f76a24   3 weeks ago      63.1MB
```

Запустим контейнер:

```
mhq@mhq-envy:~$ sudo docker run -it --name hdp \
-p 50090:50090 \
-p 50075:50075 \
-p 50070:50070 \
-p 8042:8042 \
-p 8088:8088 \
-p 8888:8888 \
-p 4040:4040 \
-p 4044:4044 \
--hostname localhost \
img-hdp-hadoop
```

Команда для второго и следующих запусков:

```
mhq@mhq-envy:~$ sudo docker start hdp -i
 * Starting OpenBSD Secure Shell server sshd [ OK ]
Starting namenodes on [localhost]
localhost: Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
localhost: starting namenode, logging to /home/hduser/hadoop/logs/hadoop-hduser-namenode-localhost.out
localhost: Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
localhost: starting datanode, logging to /home/hduser/hadoop/logs/hadoop-hduser-datanode-localhost.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: Warning: Permanently added '0.0.0.0' (ECDSA) to the list of known hosts.
0.0.0.0: starting secondarynamenode, logging to /home/hduser/hadoop/logs/hadoop-hduser-secondarynamenode-localhost.out
starting yarn daemons
starting resourcemanager, logging to /home/hduser/hadoop/logs/yarn--resourcemanager-localhost.out
localhost: Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
localhost: starting nodemanager, logging to /home/hduser/hadoop/logs/yarn-hduser-nodemanager-localhost.out
```

2 Проверьте работоспособность кластера, посмотрев на статус ресурс менеджера, нейм ноды и дата ноды (1 балл)

Кластер готов к работе (рис. 1–4).

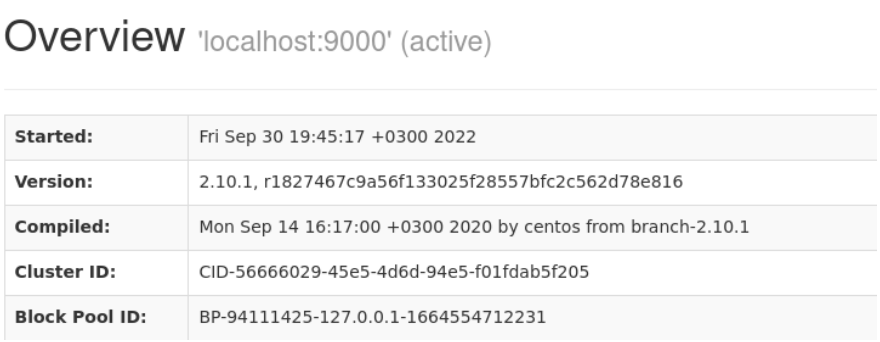

Рис. 1: NameNode information http://localhost:50070

Summary

Security is off.

Safemode is off.

7 files and directories, 0 blocks = 7 total filesystem object(s).

Heap Memory used 206.12 MB of 306.5 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 56.37 MB of 57.45 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	29.17 GB
DFS Used:	32 KB (0%)
Non DFS Used:	13.45 GB
DFS Remaining:	14.22 GB (48.74%)
Block Pool Used:	32 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion	0
Block Deletion Start Time	Fri Sep 30 19:45:17 +0300 2022
Last Checkpoint Time	Fri Sep 30 19:18:32 +0300 2022

Рис. 2: NameNode Summary

DataNode on localhost:50010

Cluster ID:	CID-56666029-45e5-4d6d-94e5-f01fdab5f205
Version:	2.10.1

Block Pools

Namenode Address	Block Pool ID	Actor State	Last Heartbeat	Last Block Report	Last Block Report Size (Max Size)
localhost:9000	BP-94111425-127.0.0.1-1664554712231	RUNNING	0s	17 minutes	0 B (64 MB)

Volume Information

Directory	StorageType	Capacity Used	Capacity Left	Capacity Reserved	Reserved Space for Replicas	Blocks
/tmp/hadoop-hduser/dfs/data/current	DISK	32 KB	14.22 GB	0 B	0 B	0

Рис. 3: DataNode http://localhost:50075/datanode.html

NodeManager information	
Total Vmem allocated for Containers	8.40 GB
Vmem enforcement enabled	true
Total Pmem allocated for Container	4 GB
Pmem enforcement enabled	true
Total VCores allocated for Containers	8
NodeHealthyStatus	true
LastNodeHealthTime	Fri Sep 30 17:19:36 GMT 2022
NodeHealthReport	
NodeManager started on	Fri Sep 30 16:45:33 GMT 2022
NodeManager Version:	2.10.1 from 1827467c9a56f133025f28557bfc2c562d78e816 by centos source checksum 2da9946ffe56799794b77621fbe0be1a on 2020-09-14T13:24Z
Hadoop Version:	2.10.1 from 1827467c9a56f133025f28557bfc2c562d78e816 by centos source checksum 3114edef868f1f3824e7d0f68be03650 on 2020-09-14T13:17Z

Рис. 4: Resource manager node http://localhost:8042/node

3 Поместите датасет ppkm_sentiment у себя в HDFS и дайте всем пользователям на них полные права (1 балл)

Копируем архив в ФС контейнера:

```
mhq@mhq-envy:~/Downloads$ sudo docker cp archive.zip hdp:/home/hduser/
```

Распаковываем:

```
hduser@localhost:~$ unzip archive.zip -d ppkm && rm archive.zip
Archive:  archive.zip
  inflating: ppkm/ppkm_dataset.csv
  inflating: ppkm/ppkm_test.csv
  inflating: ppkm/stopwordv1.txt
hduser@localhost:~$ ls
hadoop  ppkm
```

Копируем директорию `ppkm` в hdfs:

```
hduser@localhost:~$ hdfs dfs -put ppkm /user/hduser/
```

Даем полные права на файлы датасета всем пользователям:

```
hduser@localhost:~$ hdfs dfs -chmod -R a+rwx /user/hduser/ppkm
```

Проверяем:

```
hduser@localhost:~$ hdfs dfs -ls /user/hduser/ppkm
Found 3 items
-rwxrwxrwx 1 hduser supergroup 43320 2022-09-30 17:37 /user/hduser/ppkm/ppkm_dataset.csv
-rwxrwxrwx 1 hduser supergroup   476 2022-09-30 17:37 /user/hduser/ppkm/ppkm_test.csv
-rwxrwxrwx 1 hduser supergroup  4015 2022-09-30 17:37 /user/hduser/ppkm/stopwordv1.txt
```

4 Определите расположение блоков файла ppkm_dataset.csv в файловой системе (3 балла)

Открываем http://localhost:50070/explorer.html#/user/hduser/ppkm и смотрим подробности файла:

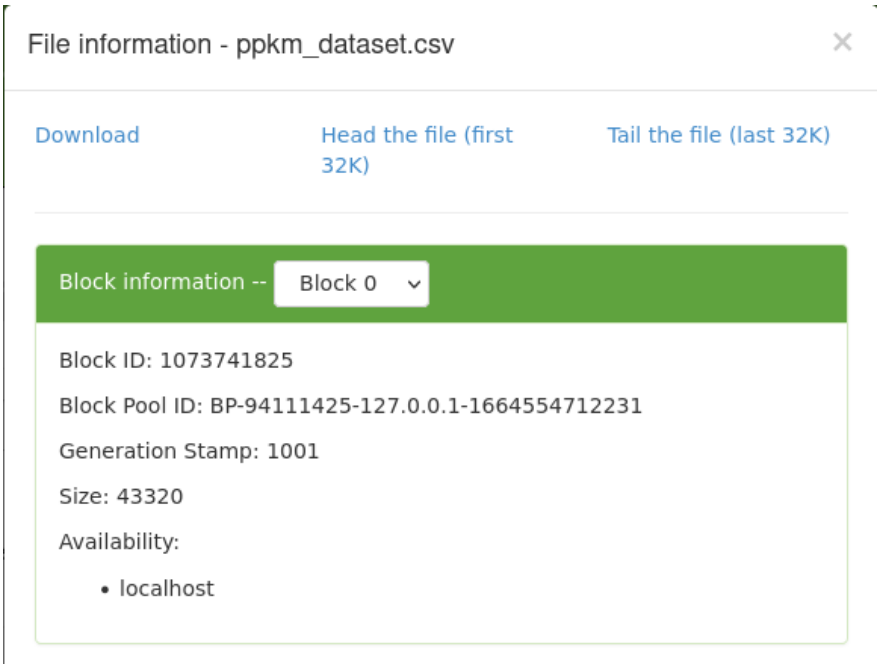

Рис. 5: File information - ppkm_dataset.csv

Запоминаем подробности:

```
Block ID: 1073741825
Block Pool ID: BP-94111425-127.0.0.1-1664554712231
Generation Stamp: 1001
Size: 43320
```

Находим на странице DataNode Information в разделе *Volume Information* в колонке *Directory* путь к блокам:

```
/tmp/hadoop-hduser/dfs/data/current
```

Командой ls находим директорию с блоками.

```
hduser@localhost:~$ ls /tmp/hadoop-hduser/dfs/data/current/\
BP-94111425-127.0.0.1-1664554712231/current/finalized/subdir0/subdir0/
blk_1073741825  blk_1073741825_1001.meta  blk_1073741826
blk_1073741826_1002.meta  blk_1073741827  blk_1073741827_1003.meta
```

Таким образом, блок файла `ppkm_dataset.csv` хранится по пути

```
/tmp/hadoop-hduser/dfs/data/current/BP-94111425-127.0.0.1-1664554712231/
current/finalized/subdir0/subdir0/blk_1073741825
```

5 У вас 20 файлов, каждый размером 130 Мб. Сколько блоков будет аллоцировано в NameNode, при условии, что размер блока по умолчанию у вас 128 Мб, а фактор репликации равен 3? (2 балла)

Один блок пренадлежит только одному файлу, файл, если он больше блока, занимает несколько. Replication Factor задает количество копий блока. Таким образом,

$$\left\lceil \frac{130}{128} \right\rceil \cdot 20 \cdot 3 = 120$$

блоков будет аллоцировано.

6 У вас 1 файл, размером 1.56 Тб. Сколько блоков будет аллоцировано в NameNode, при условии, что размер блока по умолчанию у вас 128 Мб, а фактор репликации равен 3? (2 балла)

$$1.56 \text{ Тб} = 1.56 \cdot 2^{20} \text{ Мб}, \quad 128 = 2^7$$

$$\left\lceil \frac{1.56 \cdot 2^{20}}{2^7} \right\rceil \cdot 3 = 38340$$

блоков будет аллоцировано.

P.S. Ответы на предыдущие 2 задачи будут верными, если количество DataNode $n \geqslant 3$ (фактора репликации). Иначе будет аллоцировано не $3\times$, а $n\times$ блоков, а оставшиеся $3 - n$ будут в `Missing replicas`, и будут аллоцированы, как только в кластере появятся новые DataNode.