Департамент образования и науки города Москвы
Государственное автономное образовательное учреждение
высшего образования города Москвы
«Московский городской педагогический университет»
Институт цифрового образования
Департамент информатики, управления и технологий

Практическая работа № 6

Тема: «HADOOP администрирование»

Выполнил студент ТП-191: Эльмукова О. Г.
Руководитель: Босенко Т. М.

Москва

2022

1. Установка и настройка Java
   - Install OpenJDK (JDK 8):

```
lemp001@u20-17:~$ sudo apt-get update
[sudo] password for lemp001:
Sorry, try again.
[sudo] password for lemp001:
Hit:1 http://ru.archive.ubuntu.com/ubuntu focal InRelease
Get:2 http://ru.archive.ubuntu.com/ubuntu focal-updates InRelease [114 kB]
Get:3 http://ru.archive.ubuntu.com/ubuntu focal-backports InRelease [108 kB]
```

```
lemp001@u20-17:~$ sudo apt-get install openjdk-8-jdk
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following packages were automatically installed and are no longer required:
  chromium-codecs-ffmpeg-extra gstreamer1.0-vaapi libfwupdplugin1
  libgstreamer-plugins-bad1.0-0 libva-wayland2
Use 'sudo apt autoremove' to remove them.
The following additional packages will be installed:
```

   - Verify installation:

```
lemp001@u20-17:~$ java -version
openjdk version "1.8.0_342"
OpenJDK Runtime Environment (build 1.8.0_342-8u342-b07-0ubuntu1~20.04-b07)
OpenJDK 64-Bit Server VM (build 25.342-b07, mixed mode)
```

   - SET JAVA_HOME and JRE_HOME:

```
lemp001@u20-17:~$ sudo nano /etc/environment
```

2. Настройка пользователя Hadoop

```
lemp001@u20-17:~$ sudo adduser hadoop
Adding user `hadoop' ...
Adding new group `hadoop' (1001) ...
Adding new user `hadoop' (1001) with group `hadoop'
Creating home directory `/home/hadoop' ...
Copying files from `/etc/skel' ...
New password:
Retype new password:
passwd: password updated successfully
Changing the user information for hadoop
Enter the new value, or press ENTER for the default
        Full Name []:
        Room Number []:
        Work Phone []:
        Home Phone []:
        Other []:
Is the information correct? [Y/n] y
lemp001@u20-17:~$ sudo passwd hadoop
New password:
Retype new password:
passwd: password updated successfully
```

```
lemp001@u20-17:~$ sudo su hadoop
hadoop@u20-17:/home/lemp001$ exit
exit
```

3. Настройка SSH (требуется для компонентов Hadoop)
   - Install SSH and PDSH:

```
lemp001@u20-17:~$ sudo apt-get install ssh pdsh
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following packages were automatically installed and are no longer required:
  chromium-codecs-ffmpeg-extra gstreamer1.0-vaapi libfwupdplugin1 libgstreamer-plugins-bad
  libva-wayland2
```

- Create Private/Public Keypair for hadoop user (without passphrase):

```
lemp001@u20-17:~$ sudo su hadoop
hadoop@u20-17:/home/lemp001$ cd
hadoop@u20-17:~$ ssh-keygen -t rsa -N "" -f /home/hadoop/.ssh/id_rsa
Generating public/private rsa key pair.
Created directory '/home/hadoop/.ssh'.
Your identification has been saved in /home/hadoop/.ssh/id_rsa
Your public key has been saved in /home/hadoop/.ssh/id_rsa.pub
```

- Add Public Key To Authorized Keys file (to enable passwordless ssh login):

cat /home/hadoop/.ssh/id_rsa.pub >>/home/hadoop/.ssh/authorized_keys chmod 0600 /home/hadoop/.ssh/authorized_keys

4. Настройка SSH (требуется для компонентов Hadoop)

```
hadoop@u20-17:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ECDSA key fingerprint is SHA256:h5WebNqc8BlmaFv/IyFBqqFeSk3CyvnPrpDJpJCcl3o.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
Welcome to Ubuntu 20.04.4 LTS (GNU/Linux 5.13.0-51-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

210 updates can be applied immediately.
161 of these updates are standard security updates.
To see these additional updates run: apt list --upgradable

Your Hardware Enablement Stack (HWE) is supported until April 2025.

The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

hadoop@u20-17:~$ exit
```

5. Установка Hadoop
   - Download Hadoop (v3.1.1):

wget https://archive.apache.org/dist/hadoop/common/hadoop-3.1.2/hadoop-3.1.2.tar.gz

   - Extract Binaries:

tar -xvzf hadoop-3.1.2.tar.gz

   - Move Binaries:

```
hadoop@u20-17:~$ mv hadoop-3.1.2 hadoop
```

6. Настройка Hadoop

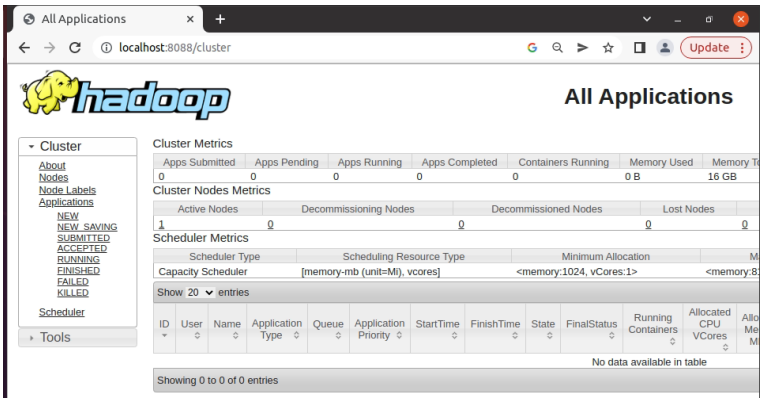
```
hadoop@u20-17:~$ mv hadoop-3.1.2 hadoop
hadoop@u20-17:~$ nano .bashrc
hadoop@u20-17:~$ source .bashrc
hadoop@u20-17:~$ nano /home/hadoop/hadoop/etc/hadoop/hadoop-env.sh
hadoop@u20-17:~$ nano /home/hadoop/hadoop/etc/hadoop/core-site.xml
hadoop@u20-17:~$ nano /home/hadoop/hadoop/etc/hadoop/hdfs-site.xml
hadoop@u20-17:~$ nano /home/hadoop/hadoop/etc/hadoop/mapred-site.xml
hadoop@u20-17:~$ nano /home/hadoop/hadoop/etc/hadoop/yarn-site.xml
hadoop@u20-17:~$ hdfs namenode -format
WARNING: /home/hadoop/hadoop/logs does not exist. Creating.
2022-11-01 16:57:53,179 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
```

```
hadoop@u20-17:~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [u20-17]
u20-17: Warning: Permanently added 'u20-17,172.26.35.137' (ECDSA) to the list o
f known hosts.
2022-11-01 16:58:25,101 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
hadoop@u20-17:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
```

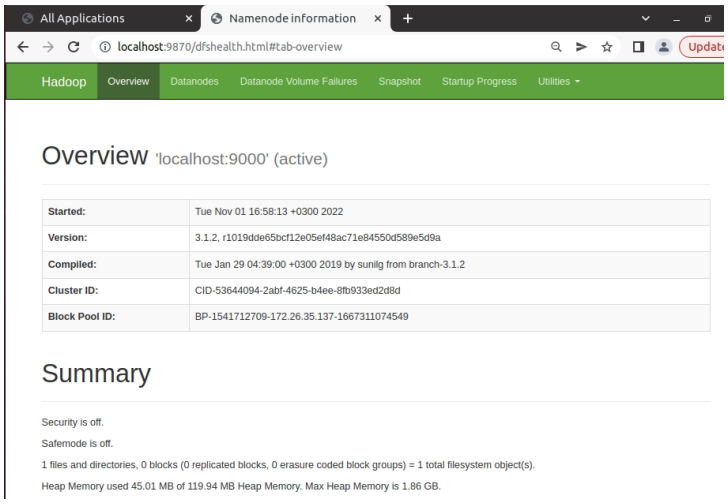
7. Проверка Hadoop/HDFS

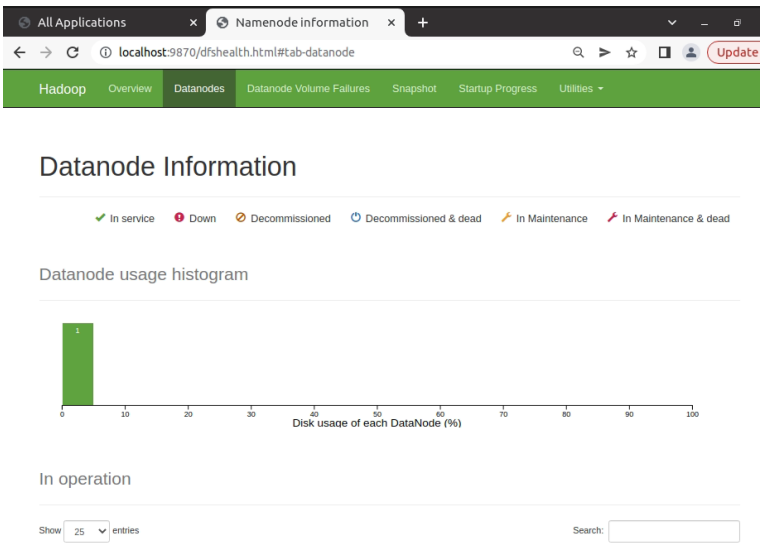
```
hadoop@u20-17:~$ hdfs dfsadmin -report
2022-11-01 16:59:06,612 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
Configured Capacity: 52044496896 (48.47 GB)
```

- Check Ressource Manager Landing Page

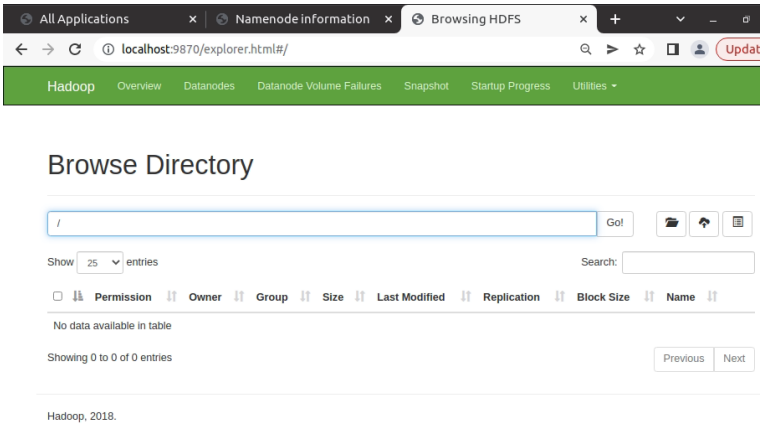


- Check NameNode Landing and Status Page





- Check HDFS File Browser



8. Работа в HDFS

- Create User Directory (on HDFS):

```
hadoop@u20-17:~$ hadoop fs -mkdir /user
2022-11-03 19:34:45,348 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
mkdir: `/user': File exists
hadoop@u20-17:~$ hadoop fs -mkdir /user/hadoop
2022-11-03 19:36:09,207 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
mkdir: `/user/hadoop': File exists
```

- List Directories (on HDFS):

```
hadoop@u20-17:~$ hadoop fs -ls /
2022-11-03 19:36:17,881 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
Found 1 items
drwxr-xr-x   - hadoop supergroup          0 2022-11-03 17:07 /user
```

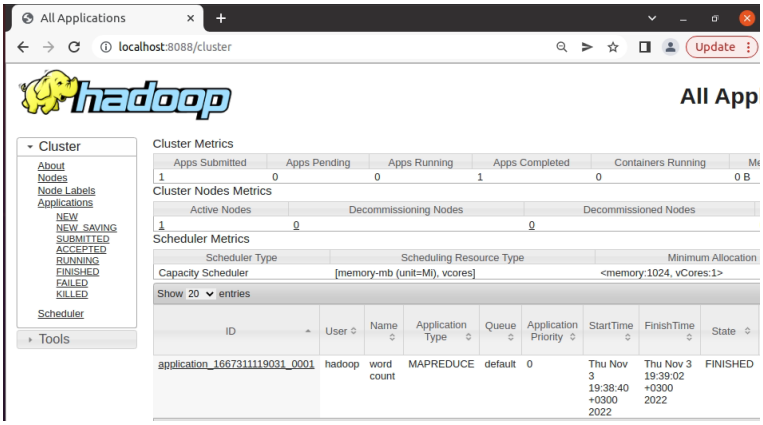
- Copy File (just a random log file) from local directory to HDFS:

```
hadoop@u20-17:~$ hadoop fs -put /var/log/dpkg.log /user/hadoop/dpkg.log
2022-11-03 19:37:05,588 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
```

9. Запуск примера MapReduce Job
   - Использование MapReduce WordCount Jar, предоставляемого Hadoop, для подсчета слов в файле

```
hadoop@u20-17:~$ hadoop jar hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examp
les-3.1.2.jar wordcount /user/hadoop/dpkg.log /user/hadoop/test_output
2022-11-03 19:38:37,936 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
2022-11-03 19:38:38,792 INFO client.RMProxy: Connecting to ResourceManager at /0
.0.0.0:8032
2022-11-03 19:38:39,368 INFO mapreduce.JobResourceUploader: Disabling Erasure Co
ding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1667311119031_0001
2022-11-03 19:38:39,653 INFO input.FileInputFormat: Total input files to process
 : 1
```

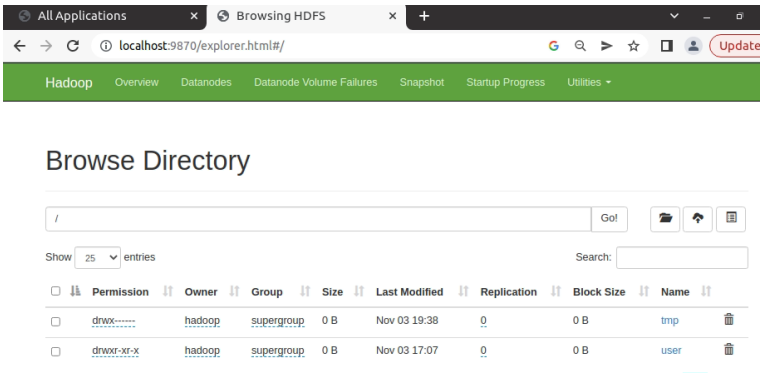
   - Просмотр запущенного MapReduce Job:



10. Запуск примера MapReduce Job
   - Проверить результат на Output/Result (via Bash):

```
hadoop@u20-17:~$  hadoop fs -cat /user/hadoop/test_output/part-r-00000
2022-11-03 20:08:19,837 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
0.0.2-5ubuntu1  7
0.136ubuntu6.7  4
0.17-2  24
0.24-1ubuntu3   32
0.36-6ubuntu1   4
0.37.1-1        14
0.4-1   7
0.72    7
0.86.1-0ubuntu1 16
0.86.1-0ubuntu1.1       24
0.9.12-1        7
06:31:07        5
06:31:08        6
```

11. Запуск Примера MapReduce Job
   - Проверить результат на Output/Result (via Web HDFS File Browser)



Упражнение 1 (вычисление количества слов в текстовом файле)

1. Клонирование репозитория git

```
hadoop@u20-17:~$ git clone https://github.com/BosenkoTM/ds_practice.git
Cloning into 'ds_practice'...
remote: Enumerating objects: 76, done.
remote: Total 76 (delta 0), reused 0 (delta 0), pack-reused 76
Unpacking objects: 100% (76/76), 6.24 MiB | 9.61 MiB/s, done.
```

2. Копирование образца файла из репозитория GIT в HDFS каталог пользователя

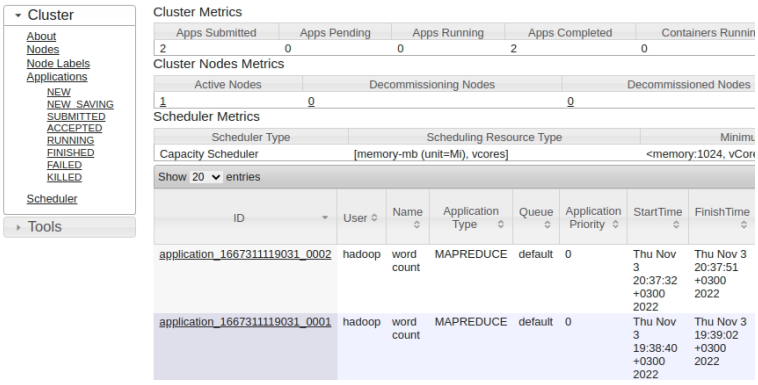
```
hadoop@u20-17:~$ hadoop fs -put ds_practice/exercises/winter_semester_2021-2022/
05_hadoop/sample_data/ Faust_1.txt
2022-11-03 20:31:07,396 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
```

3. Запуск MapReduce Jar по умолчанию (hadoop-mapreduceexa
4. mples-3.1.2.jar) для вычисления количества слов для текстового файла „Faust_1.txt “

```
hadoop@u20-17:~$ hadoop jar hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examp
les-3.1.2.jar wordcount /user/hadoop/Faust_1.txt /user/hadoop/Faust_1_Output
2022-11-03 20:37:30,742 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
2022-11-03 20:37:31,554 INFO client.RMProxy: Connecting to ResourceManager at /0
.0.0.0:8032
```

```
                Virtual memory (bytes) snapshot=4994744320
                Total committed heap usage (bytes)=356388864
                Peak Map Physical memory (bytes)=263286784
                Peak Map Virtual memory (bytes)=2495430656
                Peak Reduce Physical memory (bytes)=146731008
                Peak Reduce Virtual memory (bytes)=2499313664
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
```

5. Проверить в Диспетчере ресурсов выполнение задания



6. Скопировать полученный файл MapReduce обратно в локальную файловую систему ubuntu (using bash):

```
hadoop@u20-17:~$ hadoop fs -get /user/hadoop/Faust_1_Output/part-r-00000 Faust_1
_Output.csv
2022-11-03 21:01:26,678 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
hadoop@u20-17:~$ head -10 Faust_1_Output.csv
"Allein,        1
"Alles  1
"Als    1
"Der    1
"Die    2
"Er     2
"Ich    4
"Im     1
"Mein   1
"Nur    1
```

Упражнение 2 (подсчитать количество повторений данного слова)

1. Скопировать образец файла из репозитория GIT в каталог пользователя HDFS
2. Запустить MapReduce Jar по умолчанию (hadoop-mapreduceexamples-3.1.2.jar) для поиска в grep строки „Faust“ в текстовом файле „Faust_1.txt “ и подсчитайте количество повторений данного слова

```
hadoop@u20-17:~$ hadoop jar hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examp
les-3.1.2.jar grep /user/hadoop/Faust_1.txt /user/hadoop/Faust_1_Count_Output 'F
aust'
2022-11-03 21:18:09,927 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
2022-11-03 21:18:10,869 INFO client.RMProxy: Connecting to ResourceManager at /0
.0.0.0:8032
```

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Runnir
5	0	0	5	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes
1	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimu
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCor

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime
application_1667311119031_0005	hadoop	grep-search	MAPREDUCE	default	0	Thu Nov 3 21:25:10 +0300 2022	Thu Nov 3 21:25:28 +0300 2022
application_1667311119031_0004	hadoop	grep-sort	MAPREDUCE	default	0	Thu Nov 3 21:18:32 +0300 2022	Thu Nov 3 21:18:55 +0300 2022
application_1667311119031_0003	hadoop	grep-search	MAPREDUCE	default	0	Thu Nov 3 21:18:12 +0300 2022	Thu Nov 3 21:18:30 +0300 2022

```
hadoop@u20-17:~$ cat Faust_1_Count_Output.csv
50      Faust
```