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

ДИСЦИПЛИНА:

«Инструменты для хранения и обработки больших данных»

пр_05-01

Тема:

«Introduction to HDFS»

Выполнила: Николаева С.Г., АДЭУ-201

Преподаватель: Босенко Тимур Муртазович

Москва



```
⊕
                                lemp001@u20-04: ~
                                                                                ×
Setting up libxt-dev:amd64 (1:1.1.5-1) ...
lemp001@u20-04:~$ java -version
openjdk version "1.8.0 362"
OpenJDK Runtime Environment (build 1.8.0 362-8u362-ga-0ubuntu1~20.04.1-b09)
OpenJDK 64-Bit Server VM (build 25.362-b09, mixed mode)
lemp001@u20-04:~$ sudo nano /etc/environment
lemp001@u20-04:~$ 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]
lemp001@u20-04:~$
```

```
⊞
                                lemp001@u20-04: ~
                                                                     Q
                                                                          ≡
                                                                               ×
lemp001@u20-04:~$ sudo su hadoop
hadoop@u20-04:/home/lemp001$ exit
exit
lemp001@u20-04:~$ 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-bad1.0-0 libva-wayland2
Use 'sudo apt autoremove' to remove them.
The following additional packages will be installed:
  genders libgenders0 ncurses-term openssh-server openssh-sftp-server
  ssh-import-id
Suggested packages:
  rdist molly-guard monkeysphere ssh-askpass
The following NEW packages will be installed:
  genders libgenders0 ncurses-term openssh-server openssh-sftp-server pdsh ssh
  ssh-import-id
O upgraded, 8 newly installed, O to remove and 85 not upgraded.
Need to get 860 kB of archives.
After this operation, 6 649 kB of additional disk space will be used.
Do you want to continue? [Y/n] y
Get:1 http://ru.archive.ubuntu.com/ubuntu focal-updates/main amd64 openssh-sftp-
```

```
⊞
                                 hadoop@u20-04: ~
                                                                      Q
                                                                           ≡
                                                                                ×
Processing triggers for libc-bin (2.31-0ubuntu9.9) ...
Processing triggers for ufw (0.36-6ubuntul) ...
lemp001@u20-04:~$ sudo su hadoop
hadoop@u20-04:/home/lemp001$ cd
hadoop@u20-04:~$ ssh-kevgen -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
The key fingerprint is:
SHA256:TtV5FqL+PYRENvF80skbxrsHoIqvv6z203mqqeCK2uq hadoop@u20-04
The key's randomart image is:
+---[RSA 3072]----+
           .oB..
           . + Boo.|
           o = B*ol
          . 0 *.0+
         S . . +o
        + .
                0. |
       00.
                 . . |
       .+0 .
      .+*0*
IBE.
+----[SHA256]----+
hadoop@u20-04:~$
 ⊞
                                                                          \equiv
                                hadoop@u20-04: ~
                                                                               ×
|BE. .+*0*
+----[SHA256]----+
hadoop@u20-04:~$ cat /home/hadoop/.ssh/id rsa.pub >> /home/hadoop/.ssh/authorize
hadoop@u20-04:~$ chmod 0600 /home/hadoop/.ssh/authorized keys
hadoop@u20-04:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ECDSA key fingerprint is SHA256:9nmZ5vcfVQAFeKCmaDa8AsvQLbBRcTFQ1T+tjRnUrog.
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
86 updates can be applied immediately.
To see these additional updates run: apt list --upgradable
Your Hardware Enablement Stack (HWE) is supported until April 2025.
*** System restart required ***
The programs included with the Ubuntu system are free software;
```

the exact distribution terms for each program are described in the

hadoop@u20-04: ~





×

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@u20-04:~\$ exit

logout

Connection to localhost closed.

hadoop@u20-04:~\$ wget https://archive.apache.org/dist/hadoop/common/hadoop-3.1.2
/hadoop-3.1.2.tar.gz

--2023-04-21 13:51:50-- https://archive.apache.org/dist/hadoop/common/hadoop-3.1.2/hadoop-3.1.2.tar.gz

Resolving archive.apache.org (archive.apache.org)... 138.201.131.134, 2a01:4f8:1 72:2ec5::2

Connecting to archive.apache.org (archive.apache.org)|138.201.131.134|:443... co

HTTP request sent, awaiting response... 200 OK Length: 332433589 (317M) [application/x-gzip] Saving to: 'hadoop-3.1.2.tar.gz'

hadoop-3.1.2.tar.gz 19%[==>

63,16M 795KB/s

eta 4m 43s

```
Q
 ⊞
                               hadoop@u20-04: ~
                                                                           ×
  GNU nano 4.8
                                                                   Modified
                                     .bashrc
# enable programmable completion features (you don't need to enable
# this, if it's already enabled in /etc/bash.bashrc and /etc/profile
# sources /etc/bash.bashrc).
if ! shopt -oq posix; then
  if [ -f /usr/share/bash-completion/bash completion ]; then
    . /usr/share/bash-completion/bash completion
  elif [ -f /etc/bash completion ]; then
    . /etc/bash completion
  fi
fi
export HADOOP HOME=/home/hadoop/hadoop
export HADOOP INSTALL=$HADOOP HOME
export HADOOP MAPRED HOME=$HADOOP HOME
export HADOOP COMMON HOME=$HADOOP_HOME
export HADOOP HDFS HOME=$HADOOP HOME
export YARN HOME=$HADOOP HOME
export HADOOP_COMMON_LIB NATIVE DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
export PDSH RCMD TYPE=ssh
File Name to Write: .bashrc
                                                         M-B Backup File
`G Get Help
                   M-D DOS Format
                                      M-A Append
^C Cancel
                                                         ^T To Files
                   M-M Mac Format
                                      M-P Prepend
 ⊞
                                  hadoop@u20-04: ~
                                                                         Q
                                                                              ≡
hadoop-3.1.2/share/doc/hadoop/images/icon warning sml.gif
hadoop-3.1.2/share/doc/hadoop/images/newwindow.png
hadoop-3.1.2/share/doc/hadoop/images/maven-logo-2.gif
hadoop-3.1.2/share/doc/hadoop/images/banner.jpg
hadoop-3.1.2/share/doc/hadoop/images/logo maven.jpg
hadoop-3.1.2/share/doc/hadoop/images/collapsed.gif
hadoop-3.1.2/share/doc/hadoop/images/external.png
hadoop-3.1.2/share/doc/hadoop/images/h5.jpg
hadoop-3.1.2/share/doc/hadoop/images/logos/
hadoop-3.1.2/share/doc/hadoop/images/logos/build-by-maven-white.png
hadoop-3.1.2/share/doc/hadoop/images/logos/build-by-maven-black.png
hadoop-3.1.2/share/doc/hadoop/images/logos/maven-feather.png
hadoop-3.1.2/share/doc/hadoop/index.html
hadoop-3.1.2/share/doc/hadoop/project-reports.html
hadoop-3.1.2/include/
hadoop-3.1.2/include/hdfs.h
hadoop-3.1.2/include/Pipes.hh
hadoop-3.1.2/include/SerialUtils.hh
hadoop-3.1.2/include/StringUtils.hh
hadoop-3.1.2/include/TemplateFactory.hh
hadoop@u20-04:~$ mv hadoop-3.1.2 hadoop
hadoop@u20-04:~$ nano
                        .bashrc
```

hadoop@u20-04:~\$ source .bashrc

hadoop@u20-04:~\$

```
⊞
                                                                   Q
                               hadoop@u20-04: ~
GNU nano 4.8
                   /home/hadoop/hadoop/etc/hadoop/hadoop-env.sh
                                                                    Modified
# export HDFS DFSROUTER OPTS=""
###
###
# Advanced Users Only!
###
# When building Hadoop, one can add the class paths to the commands
# via this special env var:
# export HADOOP ENABLE BUILD PATHS="true"
# To prevent accidents, shell commands be (superficially) locked
# to only allow certain users to execute certain subcommands.
# It uses the format of (command) (subcommand) USER.
# For example, to limit who can execute the namenode command,
# export HDFS NAMENODE USER=hdfs
export JAVA HOME=/usr/lib/jvm/java-8-openjdk-amd64
            ^O Write Out ^W Where Is
                                      ^K Cut Text ^J Justify
^X Exit
            ^R Read File ^\ Replace
                                      ^U Paste Text^T To Spell
                                                                Go To Line
hadoop@u20-04:~$ mv hadoop-3.1.2 hadoop
 hadoop@u20-04:~$ nano .bashrc
hadoop@u20-04:~$ source .bashrc
hadoop@u20-04:~$ nano /home/hadoop/hadoop/etc/hadoop/hadoop-env.sh
hadoop@u20-04:~$
```

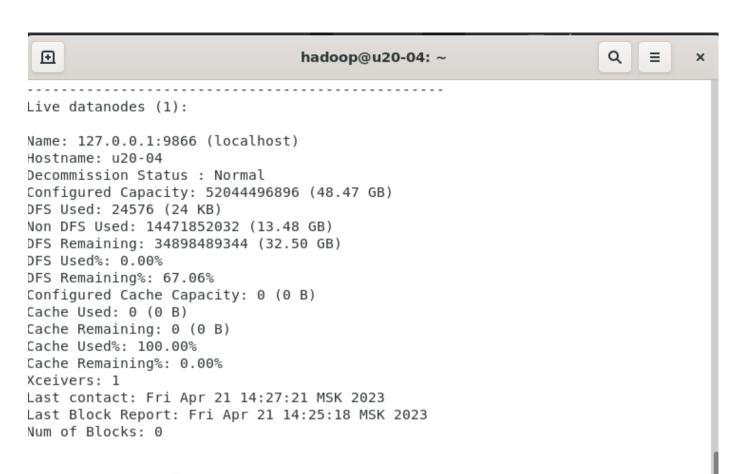


```
naaoopgaro ot. 4 Jource ibasiire
hadoop@u20-04:~$ nano /home/hadoop/hadoop/etc/hadoop/hadoop-env.sh
hadoop@u20-04:~$ nano /home/hadoop/hadoop/etc/hadoop/core-site.xml
hadoop@u20-04:~$ nano /home/hadoop/hadoop/etc/hadoop/hdfs-site.xml
hadoop@u20-04:~$ nano /home/hadoop/hadoop/etc/hadoop/mapred-site.xml
hadoop@u20-04:~$ nano /home/hadoop/hadoop/etc/hadoop/yarn-site.xml
 ⊞
                             hadoop@u20-04: ~
                                                             Q
 total heap and retry cache entry expiry time is 600000 millis
2023-04-21 14:24:20,842 INFO util.GSet: Computing capacity for map NameNodeRet
Cache
2023-04-21 14:24:20,842 INFO util.GSet: VM type
                                                = 64-bit
2023-04-21 14:24:20,842 INFO util.GSet: 0.029999999329447746% max memory 1.9 G
= 584.4 KB
2023-04-21 14:24:20,842 INFO util.GSet: capacity
                                                = 2^16 = 65536 entries
2023-04-21 14:24:20,894 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1
4279312-172.26.35.24-1682076260884
2023-04-21 14:24:20,927 INFO common.Storage: Storage directory /home/hadoop/ha
opdata/hdfs/namenode has been successfully formatted.
2023-04-21 14:24:20,947 INFO namenode.FSImageFormatProtobuf: Saving image file
ing no compression
2023-04-21 14:24:21,094 INFO namenode.FSImageFormatProtobuf: Image file /home/
doop/hadoopdata/hdfs/namenode/current/fsimage.ckpt 000000000000000000 of size
90 bytes saved in 0 seconds .
2023-04-21 14:24:21,116 INFO namenode.NNStorageRetentionManager: Going to reta
 1 images with txid >= 0
2023-04-21 14:24:21.120 INFO namenode.NameNode: SHUTDOWN MSG:
SHUTDOWN MSG: Shutting down NameNode at u20-04/172.26.35.24
hadoop@u20-04:~$
 hadoop@u20-04:~$ start-dfs.sh
 Starting namenodes on [localhost]
 Starting datanodes
 Starting secondary namenodes [u20-04]
 u20-04: Warning: Permanently added 'u20-04,172.26.35.24' (ECDSA) to the list of
```

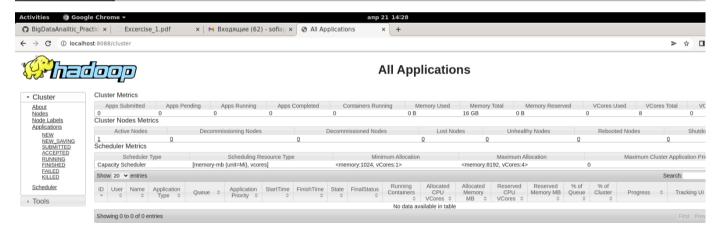
2023-04-21 14:25:22,422 WARN util.NativeCodeLoader: Unable to load native-hadoop

library for your platform... using builtin-java classes where applicable

hadoop@u20-04:~\$



hadoop@u20-04:~\$



hadoop@u20-04:~\$ hadoop fs -mkdir /user

2023-04-21 14:34:48,307 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
hadoop@u20-04:~\$ hadoop fs -mkdir /user/hadoop

2023-04-21 14:35:08,857 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
hadoop@u20-04:~\$

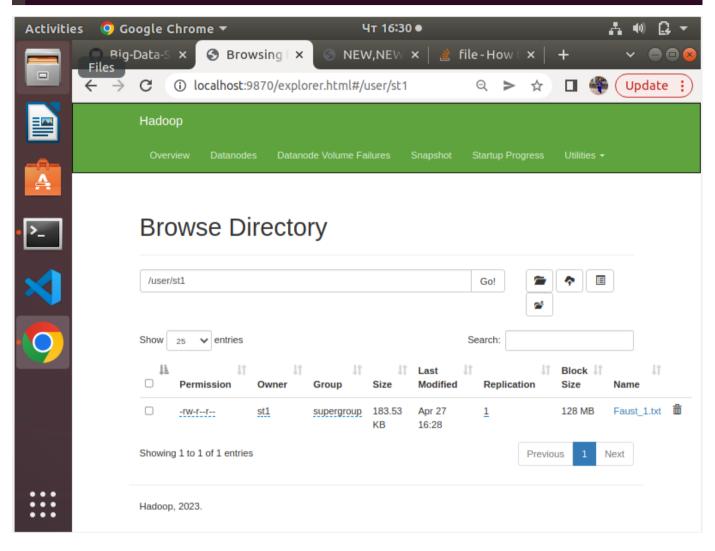
hadoop@u20-04:~\$ hadoop fs -put /var/log/dpkg.log /user/hadoop/dpkg.log 2023-04-21 14:37:15,591 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable hadoop@u20-04:~\$



Доп задания

```
st1@st1-vbox:~$ git clone https://github.com/BosenkoTM/Big-Data-Storage-and-Pro
cessing-Tools
Cloning into 'Big-Data-Storage-and-Processing-Tools'...
remote: Enumerating objects: 123, done.
remote: Counting objects: 100% (27/27), done.
remote: Compressing objects: 100% (27/27), done.
remote: Total 123 (delta 12), reused 0 (delta 0), pack-reused 96
Receiving objects: 100% (123/123), 19.72 MiB | 1.90 MiB/s, done.
Resolving deltas: 100% (37/37), done.
```

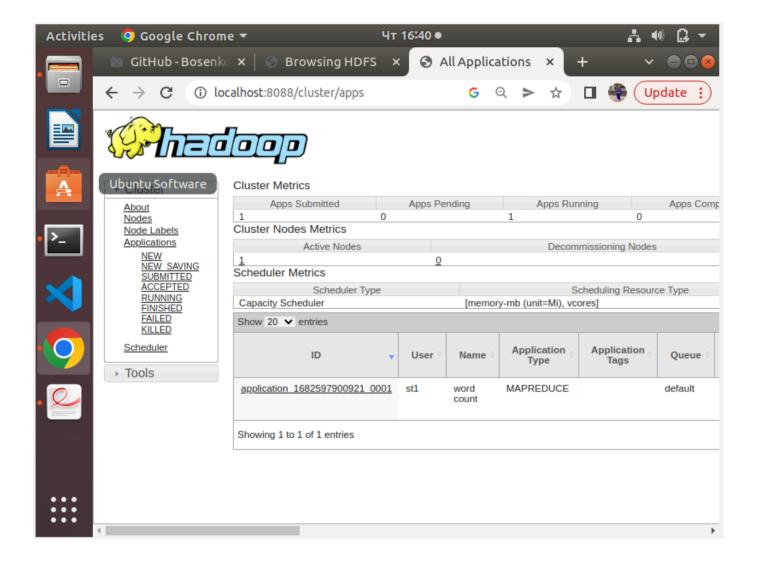
st1@st1-vbox:~\$ \$HADOOP_HOME/bin/hadoop fs -mkdir -p /user/st1/
st1@st1-vbox:~\$ \$HADOOP_HOME/bin/hadoop fs -put Big-Data-Storage-and-Processing
-Tools/practice/Faust_1.txt /user/st1
st1@st1-vbox:~\$

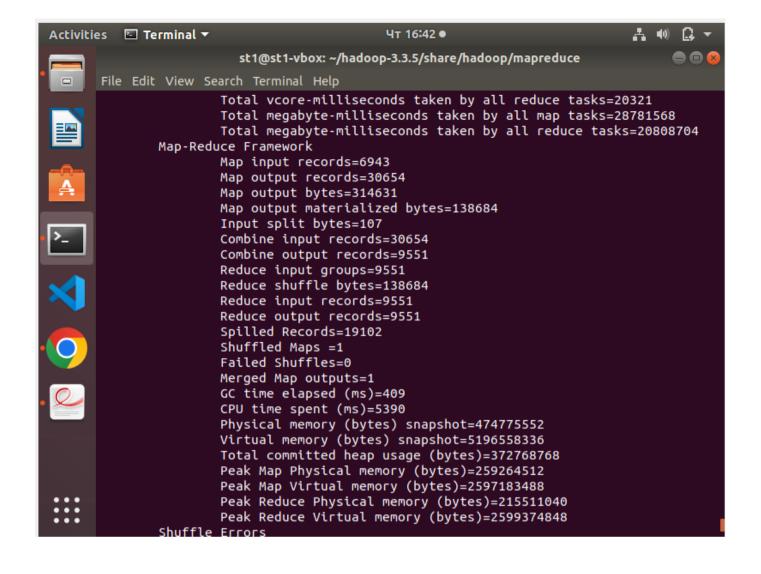


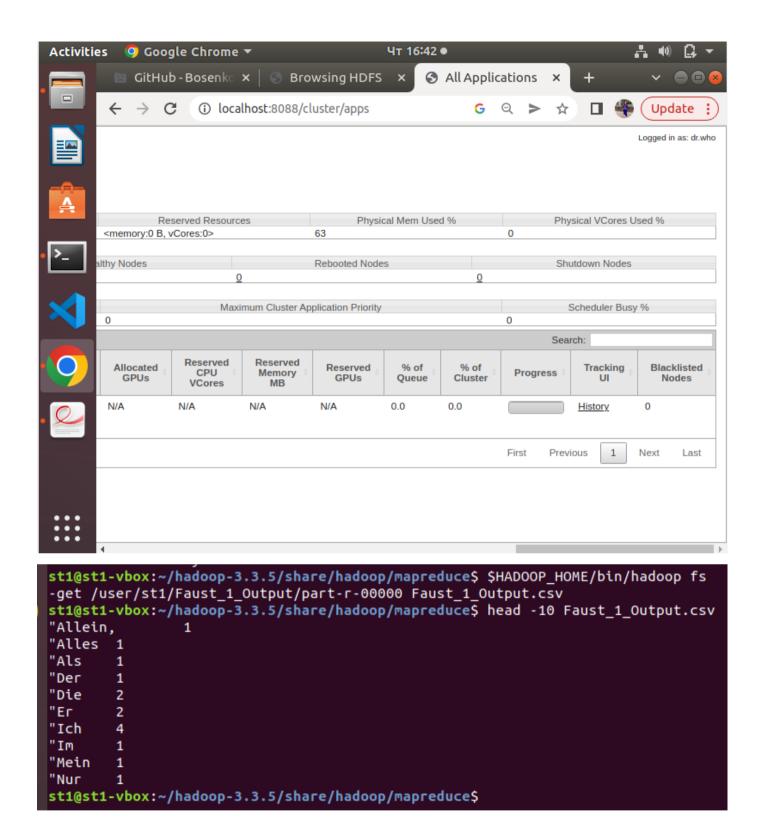
1-vbox:8088/proxy/application_1682597900921_0001/ 2023-04-27 16:38:52,696 INFO mapreduce.Job: Running job: job_1682597900921_0001 2023-04-27 16:39:42,326 INFO mapreduce.Job: Job job 1682597900921 0001 running

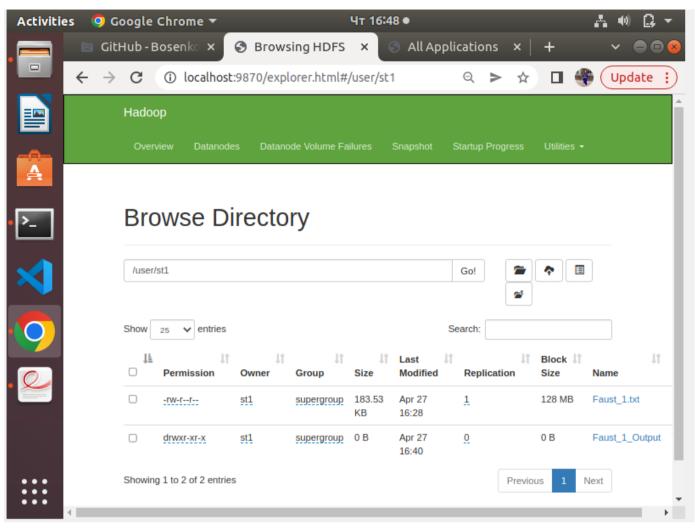
2023-04-27 16:39:42,334 INFO mapreduce.Job: map 0% reduce 0%

in uber mode : false



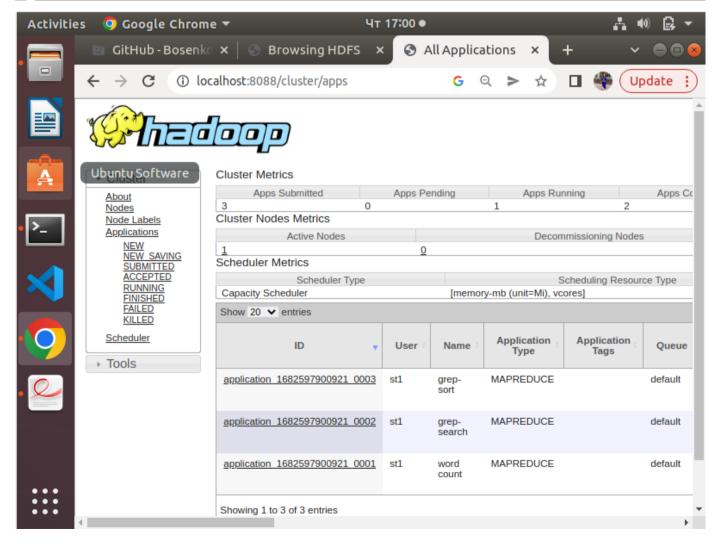


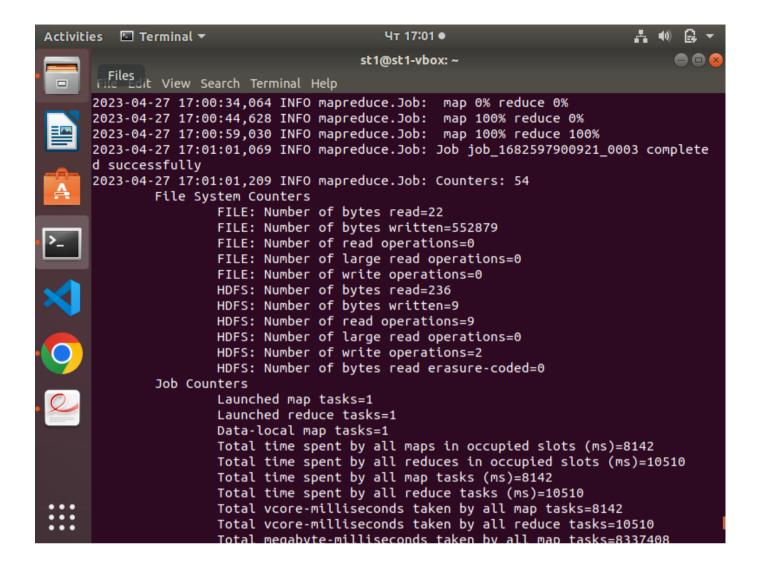


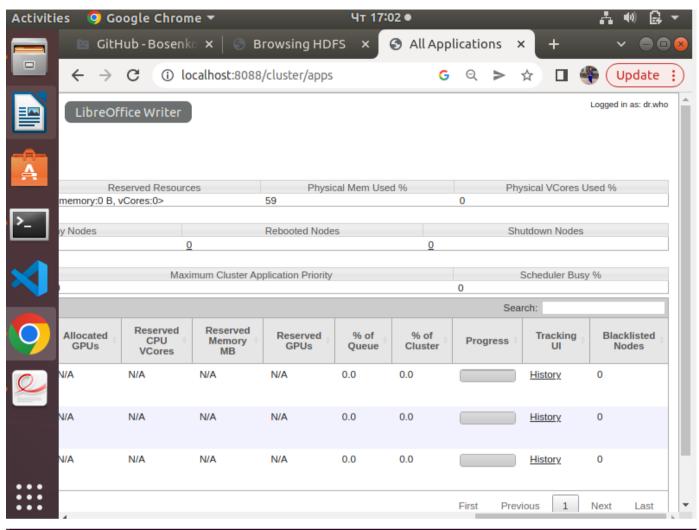


```
st1@st1-vbox:~$ $HADOOP_HOME/bin/hadoop fs -put Big-Data-Storage-and-Processing
-Tools/practice/Faust_1.txt /user/st1
put: `/user/st1/Faust_1.txt': File exists
st1@st1-vbox:~$ $HADOOP_HOME/bin/hadoop fs -mkdir -p /user/nik/
st1@st1-vbox:~$ $HADOOP_HOME/bin/hadoop fs -put Big-Data-Storage-and-Processing
-Tools/practice/Faust 1.txt /user/nik
```

```
st1@st1-vbox:~$ $HADOOP HOME/bin/hadoop jar $HADOOP HOME/share/hadoop/mapreduce
/hadoop-mapreduce-examples-3.3.5.jar grep /user/nik/Faust_1.txt /user/nik/Faust
_1_Count_Output 'Faust'
2023-04-27 16:59:14.070 INFO client.DefaultNoHARMFailoverProxyProvider: Connect
ing to ResourceManager at /0.0.0.0:8032
2023-04-27 16:59:15,128 INFO mapreduce.JobResourceUploader: Disabling Erasure C
oding for path: /tmp/hadoop-yarn/staging/st1/.staging/job_1682597900921_0002
2023-04-27 16:59:15,714 INFO input.FileInputFormat: Total input files to proces
s:1
2023-04-27 16:59:16,032 INFO mapreduce.JobSubmitter: number of splits:1
2023-04-27 16:59:16.570 INFO mapreduce.JobSubmitter: Submitting tokens for job:
 job 1682597900921 0002
2023-04-27 16:59:16.570 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-04-27 16:59:16,979 INFO conf.Configuration: resource-types.xml not found
2023-04-27 16:59:16,980 INFO resource.ResourceUtils: Unable to find 'resource-t
ypes.xml'.
2023-04-27 16:59:17,149 INFO impl. YarnClientImpl: Submitted application applica
tion 1682597900921 0002
2023-04-27 16:59:17,249 INFO mapreduce.Job: The url to track the job: http://st
1-vbox:8088/proxy/application 1682597900921 0002/
2023-04-27 16:59:17,251 INFO mapreduce.Job: Running job: job_1682597900921_0002
2023-04-27 16:59:31,126 INFO mapreduce.Job: Job job 1682597900921 0002 running
in uber mode : false
2023-04-27 16:59:31,137 INFO mapreduce.Job:
                                             map 0% reduce 0%
2023-04-27 16:59:52,841 INFO mapreduce.Job:
                                             map 100% reduce 0%
```







st1@st1-vbox:~\$ \$HADOOP_HOME/bin/hadoop fs -get /user/nik/Faust_1_Count_Output/
part-r-00000 Faust_1_Count_Output.csv
st1@st1-vbox:~\$ cat Faust_1_Count_Output.csv

50 Faust

