

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

Практическая работа 05_1

Тема:

Introduction to HDFS.

Выполнила: Соколова М. С., группа: АДЭУ-201

Преподаватель: Босенко Т. М.

Москва

2023

```
marina@marina-VirtualBox: ~  
File Edit View Search Terminal Help  
marina@marina-VirtualBox:~$ sudo apt-get install openjdk-8-jdk  
  
marina@marina-VirtualBox:~$ java -version  
openjdk version "1.8.0_362"  
OpenJDK Runtime Environment (build 1.8.0_362-8u362-ga-0ubuntu1~18.04.1-b09)  
OpenJDK 64-Bit Server VM (build 25.362-b09, mixed mode)  
  
marina@marina-VirtualBox:~$ sudo nano /etc/environment  
File Edit View Search Terminal Help  
GNU nano 2.9.3 /etc/environment  
PATH="/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:/usr/ga$  
JAVA_HOME="/usr/lib/jvm/java-8-openjdk-amd64"  
JRE_HOME="/usr/lib/jvm/java-8-openjdk-amd64/jre"  
  
marina@marina-VirtualBox:~$ 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' ...  
Enter new UNIX password:  
Retype new UNIX 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  
marina@marina-VirtualBox:~$ sudo su hadoop  
hadoop@marina-VirtualBox:/home/marina$ exit  
exit  
marina@marina-VirtualBox:~$  
marina@marina-VirtualBox:~$ sudo apt-get install ssh pdsh  
Reading package lists... Done  
Building dependency tree  
Reading state information... Done  
The following additional packages will be installed:  
  genders libgenders0 ncurses-term openssh-client openssh-server  
  openssh-sftp-server ssh-import-id  
Suggested packages:  
  rdist keychain libpam-ssh monkeysphere ssh-askpass molly-guard rssh  
The following NEW packages will be installed:  
  ssh pdsh  
0 upgraded, 2 newly installed, 0 to remove and 0 not installed.  
Need to get 1,104 kB of archives.  
After this operation, 4,096 kB of additional disk space will be used.  
Do you want to continue? [Y/n] y  
Get:1 http://ppa.launchpad.net/openssh/openssh-stable/ubuntu/ubuntu18.04/ubuntu18.04 amd64 openssh-server amd64 1:8.9p1-3ubuntu0.1 [1,040 kB]  
Get:2 http://ppa.launchpad.net/openssh/openssh-stable/ubuntu/ubuntu18.04/ubuntu18.04 amd64 pdsh amd64 2.16.1-1 [64.0 kB]  
Fetched 1,104 kB in 1s (1,104 kB/s)  
debconf: delaying package configuration, since apt-utils is not installed  
Setting up openssh-server (1:8.9p1-3ubuntu0.1) ...  
Setting up pdsh (2.16.1-1) ...  
Processing triggers for man-db (2.7.8-1) ...
```

```

marina@marina-VirtualBox:~$ sudo su hadoop
hadoop@marina-VirtualBox:/home/marina$ cd
hadoop@marina-VirtualBox:~$ 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.
The key fingerprint is:
SHA256:dJY/tN+D1WjHfTqnQg5bQfk+Jltm7qADRRJ6t7zjmY hadoop@marina-VirtualBox
The key's randomart image is:
+---[RSA 2048]-----+
|      .oo  .      |
|      .o  .o      |
|     ... +...     |
|    . .o o...oo   |
|    . So  +oo.*    |
|    . . o ++X+.    |
|    . o. *.X+oo    |
|      E+o.+...+    |
|    o..+. oo      |
+-----[SHA256]-----+

hadoop@sokolova-VirtualBox:~$ cat /home/hadoop/.ssh/id_rsa.pub >> /home/hadoop/.ssh/authorized_keys
hadoop@sokolova-VirtualBox:~$ chmod 0600 /home/hadoop/.ssh/authorized_keys

hadoop@marina-VirtualBox:~$ exit
ВЫХОД
Connection to localhost closed.
hadoop@marina-VirtualBox:~$ wget https://archive.apache.org/dist/hadoop/common/hadoop-3.1.2/hadoop-3.1.2.tar.gz
--2023-04-22 14:16:01-- 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, 2a01:4f8:172:2ec5::2
Connecting to archive.apache.org (archive.apache.org)|138.201.131.134|:443.. connected.
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.g 100%[=====>] 317,03M  1,37MB/s   in 3m 44s

2023-04-22 14:19:46 (1,41 MB/s) - 'hadoop-3.1.2.tar.gz' saved [332433589/332433589]

hadoop@marina-VirtualBox:~$

```

```
GNU nano 2.9.3 .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

GNU nano 2.9.3 /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

^G Get Help      ^O Write Out    ^W Where Is     ^K Cut Text     ^J Justify
^X Exit          ^R Read File    ^\ Replace      ^U Uncut Text   ^T To Linter

GNU nano 2.9.3 /home/hadoop/hadoop/etc/hadoop/core-site.xml

<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
<property>
<name>yarn.app.mapreduce.am.env</name>
<value>HADOOP_MAPRED_HOME=${HADOOP_HOME}</value>
</property>
<property>
<name>mapreduce.map.env</name>
<value>HADOOP_MAPRED_HOME=${HADOOP_HOME}</value>
</property>
<property>
<name>mapreduce.reduce.env</name>
<value>HADOOP_MAPRED_HOME=${HADOOP_HOME}</value>
</property>
</configuration>
```

```
GNU nano 2.9.3 /home/hadoop/hadoop/etc/hadoop/yarn-site.xml
```

```
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.  
See the License for the specific language governing permissions and  
limitations under the License. See accompanying LICENSE file.
```

```
-->
```

```
<configuration>
```

```
<!-- Site specific YARN configuration properties -->
```

```
<property>
```

```
<name>yarn.nodemanager.aux-services</name>
```

```
<value>mapreduce_shuffle</value>
```

```
</property>
```

```
<property>
```

```
<name>yarn.nodemanager.resource.memory-mb</name>
```

```
<value>16384</value>
```

```
</property>
```

```
</configuration>
```

```
*****/
```

```
hadoop@marina-VirtualBox:~$ start-dfs.sh
```

```
Starting namenodes on [localhost]
```

```
Starting datanodes
```

```
Starting secondary namenodes [marina-VirtualBox]
```

```
2023-04-22 14:42:49,573 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your  
platform... using builtin-java classes where applicable
```

```
hadoop@marina-VirtualBox:~$ start-yarn.sh
```

```
Starting resourcemanager
```

```
resourcemanager is running as process 12627. Stop it first.
```

```
Starting nodemanagers
```

```
localhost: nodemanager is running as process 12768. Stop it first.
```

```
pdsh@marina-VirtualBox: localhost: ssh exited with exit code 1
```

```
hadoop@marina-VirtualBox:~$ hdfs dfsadmin -report
```

```
2023-04-22 14:43:39,308 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your  
platform... using builtin-java classes where applicable
```

```
Configured Capacity: 10499674112 (9.78 GB)
```

```
Present Capacity: 2112827392 (1.97 GB)
```

```
DFS Remaining: 2112802816 (1.97 GB)
```

```
DFS Used: 24576 (24 KB)
```

```
DFS Used%: 0.00%
```

```
-----  
Live datanodes (1):
```

```
Name: 127.0.0.1:9866 (localhost)
```

```
Hostname: marina-VirtualBox
```

```
Decommission Status : Normal
```

```
Configured Capacity: 10499674112 (9.78 GB)
```

```
DFS Used: 24576 (24 KB)
```

```
Non DFS Used: 7833305088 (7.30 GB)
```

```
DFS Remaining: 2112802816 (1.97 GB)
```

```
DFS Used%: 0.00%
```

```
DFS Remaining%: 20.12%
```

```
Configured Cache Capacity: 0 (0 B)
```

```
Cache Used: 0 (0 B)
```

```
Cache Remaining: 0 (0 B)
```

```
Cache Used%: 100.00%
```

```
Cache Remaining%: 0.00%
```

```
Xceivers: 1
```

```
Last contact: Sat Apr 22 14:43:40 MSK 2023
```

```
Last Block Report: Sat Apr 22 14:42:46 MSK 2023
```


```
Num of Blocks: 0
```

```

hadoop@marina-VirtualBox:~$ hadoop fs -mkdir /user
2023-04-22 14:44:46,507 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
hadoop@marina-VirtualBox:~$ hadoop fs -mkdir /user/hadoop
2023-04-22 14:45:02,502 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
hadoop@marina-VirtualBox:~$ hadoop fs -ls /
2023-04-22 14:45:29,184 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 2023-04-22 14:45 /user
hadoop@marina-VirtualBox:~$ hadoop fs -put /var/log/dpkg.log /user/hadoop/dpkg.log
2023-04-22 14:45:48,958 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
hadoop@marina-VirtualBox:~$ hadoop jar hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.
1.2.jar wordcount /user/hadoop/dpkg.log /user/hadoop/test_output
2023-04-22 14:47:07,174 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
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=1340963
File Output Format Counters
    Bytes Written=63683

```

```
sokolova@sokolova-VirtualBox:~$ sudo apt update
[sudo] password for sokolova:
Hit:1 http://ru.archive.ubuntu.com/ubuntu bionic InRelease
Hit:2 http://ru.archive.ubuntu.com/ubuntu bionic-updates InRelease
Hit:3 http://ru.archive.ubuntu.com/ubuntu bionic-backports InRelease
Hit:4 http://security.ubuntu.com/ubuntu bionic-security InRelease
Reading package lists... Done
Building dependency tree
Reading state information... Done
291 packages can be upgraded. Run 'apt list --upgradable' to see them.
sokolova@sokolova-VirtualBox:~$ sudo apt install git
sokolova@sokolova-VirtualBox:~$ git --version
git version 2.17.1
hadoop@sokolova-VirtualBox:~$ git clone https://github.com/BosenkoTM/Big-Data-Storage-and-Processing-Tools.git
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 | 1.40 MiB/s, done.
Resolving deltas: 100% (37/37), done.
hadoop@sokolova-VirtualBox:~$ hadoop fs -put Big-Data-Storage-and-Processing-Tools/partice/Faust_1.txt /user/hadoop/Faust_1.txt
2023-04-27 22:15:52,107 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hadoop@sokolova-VirtualBox:~$ hadoop jar hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.1.2.jar wordcount /user/hadoop/Faust_1.txt /user/hadoop/Faust_1_Output
```



All Applications

Cluster

About

Nodes

Node Labels

Applications

NEW

NEW SAVING

SUBMITTED

ACCEPTED

RUNNING

FINISHED

FAILED

KILLED

Scheduler

Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used
2	0	0	2	0	0 B	16 GB	0 B	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
1	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maxin
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCoers	Allocated Memory MB	Reserved CPU VCoers	Reserved Memory MB	% of Queue
application_1682621421576_0002	hadoop	word count	MAPREDUCE	default	0	Thu Apr 27 22:19:33 +0300 2023	Thu Apr 27 22:20:31 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0
application_1682621421576_0001	hadoop	word count	MAPREDUCE	default	0	Thu Apr 27 21:58:54 +0300 2023	Thu Apr 27 21:59:35 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0

Showing 1 to 2 of 2 entries

Browse Directory

/user/hadoop/Go!

Show 25 entriesSearch:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	183.53 KB	Apr 27 22:15	1	128 MB	Faust_1.txt
drwxr-xr-x	hadoop	supergroup	0 B	Apr 27 22:20	0	0 B	Faust_1_Output
-rw-r--r--	hadoop	supergroup	1.28 MB	Apr 27 21:58	1	128 MB	dpkg.log
drwxr-xr-x	hadoop	supergroup	0 B	Apr 27 21:59	0	0 B	test_output

Showing 1 to 4 of 4 entries

Previous1Next


```

hadoop@sokolova-VirtualBox:~$ hadoop fs -get /user/hadoop/Faust_1_Output/part-
r-00000 Faust_1_Output.csv
2023-04-27 22:21:42,293 WARN util.NativeCodeLoader: Unable to load native-hado
op library for your platform... using builtin-java classes where applicable
hadoop@sokolova-VirtualBox:~$ 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
hadoop@sokolova-VirtualBox:~$

```

Open

```

"Allein,      1
"Alles  1
"Als  1
"Der  1
"Die  2
"Er  2
"Ich  4
"Im  1
"Mein  1
"Nur  1
"Rette  1
"So  1
"Uhu!  1
"Wenn  1
"Wie",  1

```

```

hadoop@sokolova-VirtualBox:~$ hadoop fs -put Big-Data-Storage-and-Processing-Tools/prac
tice/Faust_1.txt /user/hadoop/Faust_1.txt
2023-04-27 22:25:55,998 WARN util.NativeCodeLoader: Unable to load native-hadoop librar
y for your platform... using builtin-java classes where applicable
put: `/user/hadoop/Faust_1.txt': File exists
hadoop@sokolova-VirtualBox:~$ hadoop jar hadoop/share/hadoop/mapreduce/hadoop-mapreduce
-examples-3.1.2.jar grep /user/hadoop/Faust_1.txt /user/hadoop/Faust_1_Count_Output 'Fa
ust'

```

All Applications

Cluster

About
Nodes
Node Labels
Applications
NEW
NEW SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED
Scheduler

Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores
4	0	1	3	1	2 GB	16 GB	0 B	1

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Reboot
1	0	0	0	0	0

Scheduler Metrics





Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>

Show: 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCoers	Allocated Memory MB	Reserved CPU VCoers	Reserved Memory MB	% of Queue
application_1682621421576_0004	hadoop	grep-sort	MAPREDUCE	default	0	Thu Apr 27 22:27:26 +0300 2023	N/A	ACCEPTED	UNDEFINED	1	1	2048	0	0	12.5
application_1682621421576_0003	hadoop	grep-search	MAPREDUCE	default	0	Thu Apr 27 22:26:57 +0300 2023	Thu Apr 27 22:27:23 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0
application_1682621421576_0002	hadoop	word count	MAPREDUCE	default	0	Thu Apr 27 22:19:33 +0300 2023	Thu Apr 27 22:20:31 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0
application_1682621421576_0001	hadoop	word count	MAPREDUCE	default	0	Thu Apr 27 21:58:54 +0300 2023	Thu Apr 27 21:59:35 +0300 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0




Showing 1 to 4 of 4 entries


```
bytes written=9
hadoop@sokolova-VirtualBox:~$ hadoop fs -get /user/hadoop/Faust_1_Count_Output/part-r-00000 Faust_1_Count_Output.csv
2023-04-27 22:29:23,655 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hadoop@sokolova-VirtualBox:~$ cat Faust_1_Count_Output.csv
50      Faust
```







Open  part-r-00000 ~/Downloads Save   

50 Faust

Browse Directory

Show entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	-rw-r--r--	hadoop	supergroup	183.53 KB	Apr 27 22:15	1	128 MB	Faust_1.txt	
<input type="checkbox"/>	drwxr-xr-x	hadoop	supergroup	0 B	Apr 27 22:28	0	0 B	Faust_1_Count_Output	
<input type="checkbox"/>	drwxr-xr-x	hadoop	supergroup	0 B	Apr 27 22:20	0	0 B	Faust_1_Output	
<input type="checkbox"/>	-rw-r--r--	hadoop	supergroup	1.28 MB	Apr 27 21:58	1	128 MB	dpkg.log	
<input type="checkbox"/>	drwxr-xr-x	hadoop	supergroup	0 B	Apr 27 21:59	0	0 B	test_output	

Showing 1 to 5 of 5 entries Previous **1** Next