



Министерство науки и высшего образования Российской Федерации  
Калужский филиал  
федерального государственного бюджетного  
образовательного учреждения высшего образования  
«Московский государственный технический университет имени Н.Э. Баумана  
(национальный исследовательский университет)»  
(КФ МГТУ им. Н.Э. Баумана)

**ФАКУЛЬТЕТ ИУК «Информатика и управление»**

**КАФЕДРА ИУК4 «Программное обеспечение ЭВМ, информационные технологии»**

## **ЛАБОРАТОРНАЯ РАБОТА №1**

**«ОСНОВЫ HADOOP. УСТАНОВКА HADOOP. ОСНОВНЫЕ КОМАНДЫ  
ФАЙЛОВОЙ СИСТЕМЫ HDFS»**

**ДИСЦИПЛИНА:** «Технологии обработки больших данных»

Выполнил: студент гр. ИУК4 -72Б \_\_\_\_\_ ( Калашников А.С. )  
(Подпись) (Ф.И.О.)

Проверил: \_\_\_\_\_ ( Голубева С.Е. )  
(Подпись) (Ф.И.О.)

Дата сдачи (защиты):

Результаты сдачи (защиты):

- Балльная оценка:

- Оценка:

Калуга, 2023

**Целью** выполнения лабораторной работы является формирование практических навыков по установке и настройке кластера Hadoop и работе с файловой системой HDFS.

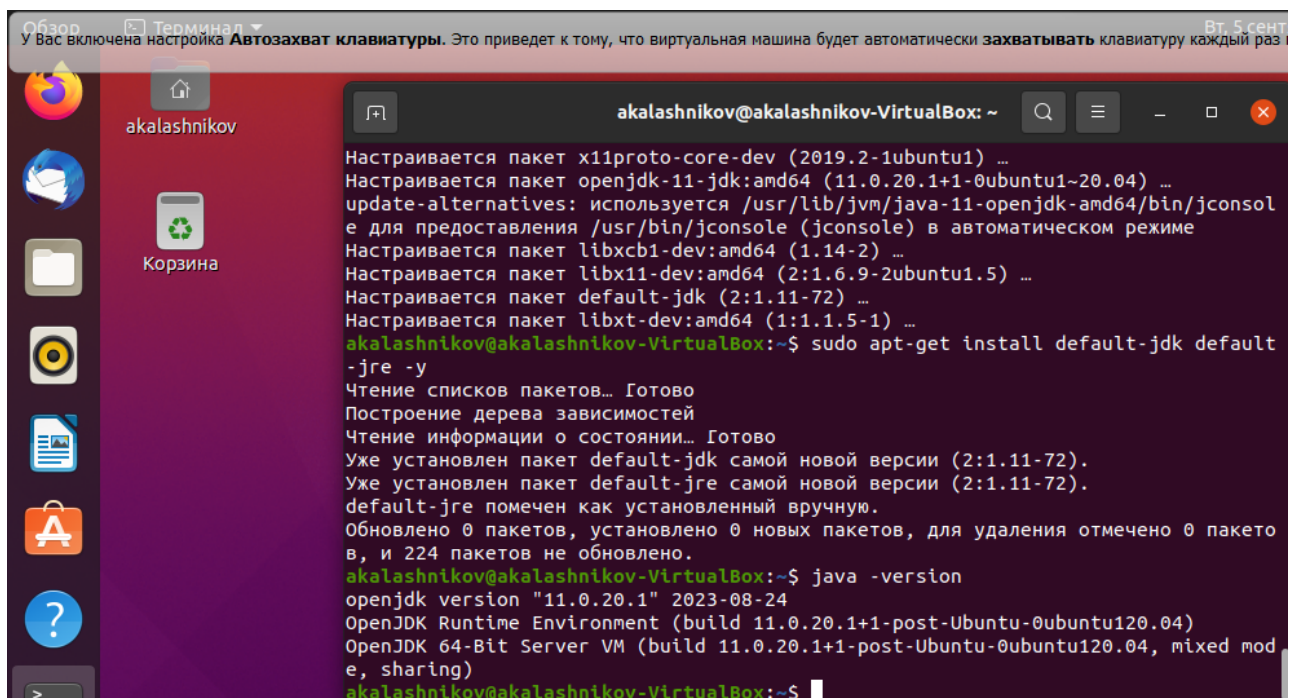
**Основными задачами** выполнения лабораторной работы являются:

1. Изучить основы Hadoop.
2. Научиться устанавливать и конфигурировать Hadoop
3. Изучить основные команды для работы с файловой системой HDFS.
4. Получить навыки написания программ для работы с HDFS

Напишите программу, которая будет рекурсивно выводить на экран список подкаталогов и файлов для заданного каталога в HDFS.

### Ход выполнения работы

Для работы Hadoop можно использовать Java версии 6 и выше. Установить можно как версию от Oracle, так и OpenJDK, для этого нужно выполнить следующую команду `sudo apt-get install default-jdk`



```
akalashnikov@akalashnikov-VirtualBox: ~  
Настраивается пакет x11proto-core-dev (2019.2-1ubuntu1) ...  
Настраивается пакет openjdk-11-jdk:amd64 (11.0.20.1+1-0ubuntu1~20.04) ...  
update-alternatives: используется /usr/lib/jvm/java-11-openjdk-amd64/bin/jconsole  
e для предоставления /usr/bin/jconsole (jconsole) в автоматическом режиме  
Настраивается пакет libxcb1-dev:amd64 (1.14-2) ...  
Настраивается пакет libx11-dev:amd64 (2:1.6.9-2ubuntu1.5) ...  
Настраивается пакет default-jdk (2:1.11-72) ...  
Настраивается пакет libxt-dev:amd64 (1:1.1.5-1) ...  
akalashnikov@akalashnikov-VirtualBox:~$ sudo apt-get install default-jdk default-jre -y  
Чтение списков пакетов... Готово  
Построение дерева зависимостей  
Чтение информации о состоянии... Готово  
Уже установлен пакет default-jdk самой новой версии (2:1.11-72).  
Уже установлен пакет default-jre самой новой версии (2:1.11-72).  
default-jre помечен как установленный вручную.  
Обновлено 0 пакетов, установлено 0 новых пакетов, для удаления отмечено 0 пакетов,  
и 224 пакетов не обновлено.  
akalashnikov@akalashnikov-VirtualBox:~$ java -version  
openjdk version "11.0.20.1" 2023-08-24  
OpenJDK Runtime Environment (build 11.0.20.1+1-post-Ubuntu-0ubuntu120.04)  
OpenJDK 64-Bit Server VM (build 11.0.20.1+1-post-Ubuntu-0ubuntu120.04, mixed mode,  
sharing)  
akalashnikov@akalashnikov-VirtualBox:~$
```

Рис. 1. Установка jdk

Сначала необходимо установить ssh и rsync, для этого нужно выполнить в терминале следующие команды:

```
sudo apt-get install  
ssh sudo apt-get install rsync
```

Hadoop требует доступ SSH для управления узлами. Необходимо настроить SSH доступ к каждому из узлов кластера для пользователя, команды для генерации нового ssh ключа и добавления созданного ключа в список

авторизованных:

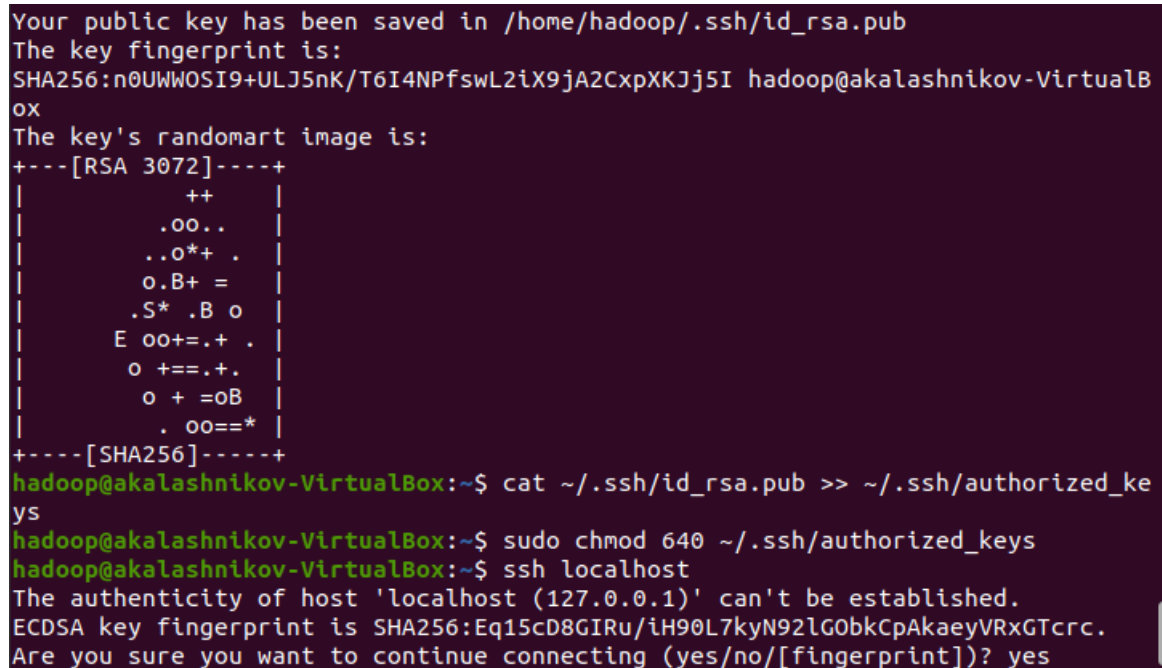
```
ssh-keygen -t rsa -P ""
```

```
cat $HOME/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys
```

Для проверки подключения к localhost нужно выполнить команду:

```
ssh localhost
```

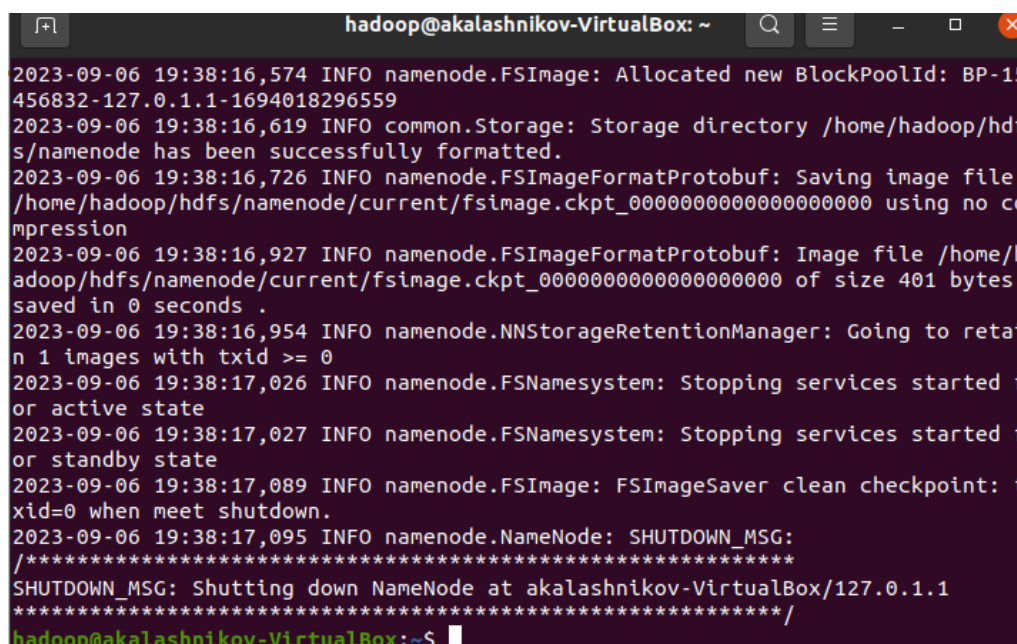
Результат выполнения команды представлен ниже (Рисунок 1).



```
Your public key has been saved in /home/hadoop/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:n0UWW0SI9+ULJ5nK/T6I4NPfswL2iX9jA2CxpXKJj5I  hadoop@akalashnikov-VirtualBox
The key's randomart image is:
+----[RSA 3072]-----+
|          ++          |
|        .oo..         |
|       ..o*+  .        |
|      o.B+  =         |
|     .S*  .B o        |
|    E oo+=.+  .       |
|   o  +=.+         |
|  o  +  =oB         |
| . oo==*          |
+----[SHA256]-----+
hadoop@akalashnikov-VirtualBox:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
hadoop@akalashnikov-VirtualBox:~$ sudo chmod 640 ~/.ssh/authorized_keys
hadoop@akalashnikov-VirtualBox:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ECDSA key fingerprint is SHA256:Eq15cD8GIRu/iH90L7kyN92LG0bkCpAkaeyVRxGTcrC.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
```

Рис. 2. Проверка ssh

Далее следует распаковать Hadoop. Затем настроить переменные окружения. В файл \$HOME/.bashrc добавить следующие переменные окружения:



```
hadoop@akalashnikov-VirtualBox: ~
2023-09-06 19:38:16,574 INFO namenode.FSImage: Allocated new BlockPoolId: BP-15456832-127.0.1.1-1694018296559
2023-09-06 19:38:16,619 INFO common.Storage: Storage directory /home/hadoop/hdfs/namenode has been successfully formatted.
2023-09-06 19:38:16,726 INFO namenode.FSImageFormatProtobuf: Saving image file /home/hadoop/hdfs/namenode/current/fsimage.ckpt_000000000000000000 using no compression
2023-09-06 19:38:16,927 INFO namenode.FSImageFormatProtobuf: Image file /home/hadoop/hdfs/namenode/current/fsimage.ckpt_000000000000000000 of size 401 bytes saved in 0 seconds .
2023-09-06 19:38:16,954 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2023-09-06 19:38:17,026 INFO namenode.FSNamesystem: Stopping services started for active state
2023-09-06 19:38:17,027 INFO namenode.FSNamesystem: Stopping services started for standby state
2023-09-06 19:38:17,089 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2023-09-06 19:38:17,095 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at akalashnikov-VirtualBox/127.0.1.1
*****/
hadoop@akalashnikov-VirtualBox:~$
```

Рис. 3. Hadoop

## #Hadoop variables

```
export JAVA_HOME= /usr/lib/jvm/java-8-openjdk-amd64/jre/
export HADOOP_INSTALL=/usr/local/hadoop
export PATH=$PATH:$HADOOP_INSTALL/bin
export PATH=$PATH:$HADOOP_INSTALL/sbin
export HADOOP_MAPRED_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_HOME=$HADOOP_INSTALL
export HADOOP_HDFS_HOME=$HADOOP_INSTALL
export YARN_HOME=$HADOOP_INSTALL
```

Настройка Hadoop Финальным шагом является конфигурирования работы кластера, для этого необходимо задать в файлах конфигурации значения соответствующих параметров. В файле \$HADOOP\_INSTALL/etc/hadoop/hadoop-env.sh необходимо задать переменную

```
JAVA_HOME: export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64
```

Основные настройки Hadoop выполняются в файле \$HADOOP\_INSTALL/etc/hadoop/core-site.xml, в котором указывается имя файловой системы (в одном кластере может физически быть несколько файловых систем, однако настроить взаимодействие между ними стандартными средствами не представляется возможным), а также порт, по которому можно к ней обратиться.

Настройки HDFS для каждого узла хранятся в файле \$HADOOP\_INSTALL/etc/hadoop/hdfs-site.xml. Параметр dfs.replication задает количество реплик, которые будут храниться на файловой системе. Также в этом файле прописываются все узлы файловой системы, присутствующие на данной машине (все dataNode и nameNode).

Настройки MapReduce прописываются в файле \$HADOOP\_INSTALL/etc/hadoop/mapred-site.xml.

Настройка фреймворка управления ресурсами кластера YARN производится в файле \$HADOOP\_INSTALL/etc/hadoop/yarn-site.xml.

Для стабильной работы Hadoop, необходимо отключить IPv6 в файле \$HADOOP\_INSTALL/etc/hadoop/hadoop-env.sh:

```
export HADOOP_OPTS=-Djava.net.preferIPv4Stack=true
```

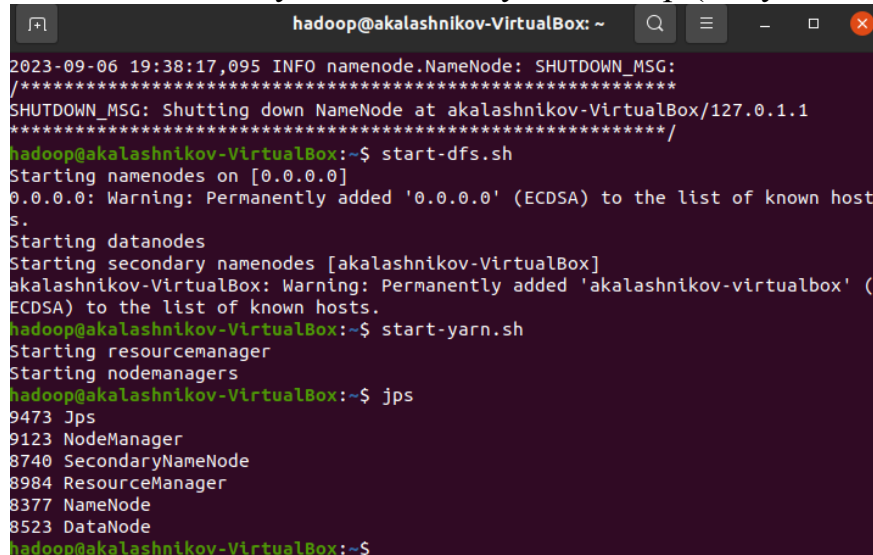
После завершения конфигурирования, необходимо отформатировать файловую систему HDFS. Для этого на NameNode необходимо выполнить команду: \$HADOOP\_INSTALL/bin/hadoop namenode -format Для запуска

Hadoop необходимо запустить следующие службы на master-узле (на всех дочерних узлах необходимые демоны запускаются автоматически, используя сконфигурированные подключения по ssh):

```
$HADOOP_INSTALL/sbin/start-dfs.sh
```

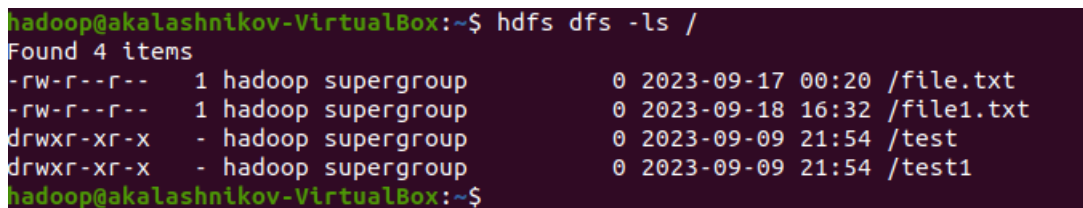
```
$HADOOP_INSTALL/sbin/start-yarn.sh
```

После этого был выполнен успешный запуск Hadoop (Рисунок 2, 3).



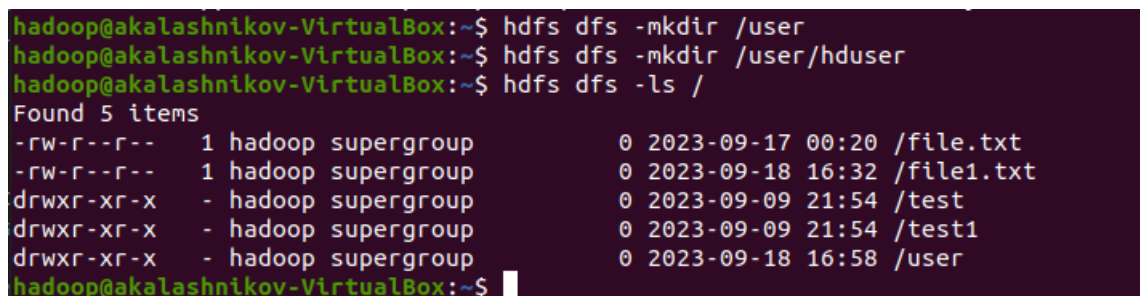
```
hadoop@akalashnikov-VirtualBox: ~
2023-09-06 19:38:17,095 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at akalashnikov-VirtualBox/127.0.1.1
*****/
hadoop@akalashnikov-VirtualBox:~$ start-dfs.sh
Starting namenodes on [0.0.0.0]
0.0.0.0: Warning: Permanently added '0.0.0.0' (ECDSA) to the list of known hosts.
Starting datanodes
Starting secondary namenodes [akalashnikov-VirtualBox]
akalashnikov-VirtualBox: Warning: Permanently added 'akalashnikov-virtualbox' (
ECDSA) to the list of known hosts.
hadoop@akalashnikov-VirtualBox:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hadoop@akalashnikov-VirtualBox:~$ jps
9473 Jps
9123 NodeManager
8740 SecondaryNameNode
8984 ResourceManager
8377 NameNode
8523 DataNode
hadoop@akalashnikov-VirtualBox:~$
```

Рис. 4. Проверка работоспособности



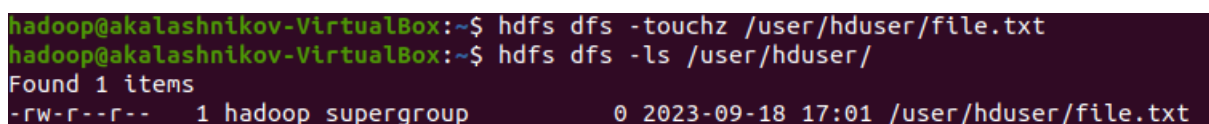
```
hadoop@akalashnikov-VirtualBox:~$ hdfs dfs -ls /
Found 4 items
-rw-r--r-- 1 hadoop supergroup 0 2023-09-17 00:20 /file.txt
-rw-r--r-- 1 hadoop supergroup 0 2023-09-18 16:32 /file1.txt
drwxr-xr-x - hadoop supergroup 0 2023-09-09 21:54 /test
drwxr-xr-x - hadoop supergroup 0 2023-09-09 21:54 /test1
hadoop@akalashnikov-VirtualBox:~$
```

Рис. 5. Содержимое корневой директории



```
hadoop@akalashnikov-VirtualBox:~$ hdfs dfs -mkdir /user
hadoop@akalashnikov-VirtualBox:~$ hdfs dfs -mkdir /user/hduser
hadoop@akalashnikov-VirtualBox:~$ hdfs dfs -ls /
Found 5 items
-rw-r--r-- 1 hadoop supergroup 0 2023-09-17 00:20 /file.txt
-rw-r--r-- 1 hadoop supergroup 0 2023-09-18 16:32 /file1.txt
drwxr-xr-x - hadoop supergroup 0 2023-09-09 21:54 /test
drwxr-xr-x - hadoop supergroup 0 2023-09-09 21:54 /test1
drwxr-xr-x - hadoop supergroup 0 2023-09-18 16:58 /user
hadoop@akalashnikov-VirtualBox:~$
```

Рис. 6. Создание директории /user/hduser



```
hadoop@akalashnikov-VirtualBox:~$ hdfs dfs -touchz /user/hduser/file.txt
hadoop@akalashnikov-VirtualBox:~$ hdfs dfs -ls /user/hduser/
Found 1 items
-rw-r--r-- 1 hadoop supergroup 0 2023-09-18 17:01 /user/hduser/file.txt
```

Рис. 7. Создание текстового файла

```
hadoop@akalashnikov-VirtualBox:~$ hdfs dfs -appendToFile - /user/hduser/file.txt
asdasd
zdfsfsdf
sdf
^Chadoop@akalashnikov-VirtualBox:~$ hdfs dfs -cat /user/hduser/file.txt
asdasd
zdfsfsdf
sdf
hadoop@akalashnikov-VirtualBox:~$
```

Рис. 8. Заполнение файла из консоли

	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	-rw-r--r--	hadoop	supergroup	0 B	Sep 17 00:20	1	128 MB	file.txt	
<input type="checkbox"/>	-rw-r--r--	hadoop	supergroup	0 B	Sep 18 16:32	1	128 MB	file1.txt	
<input type="checkbox"/>	drwxr-xr-x	hadoop	supergroup	0 B	Sep 09 21:54	0	0 B	test	
<input type="checkbox"/>	drwxr-xr-x	hadoop	supergroup	0 B	Sep 09 21:54	0	0 B	test1	
<input type="checkbox"/>	drwxr-xr-x	hadoop	supergroup	0 B	Sep 18 16:58	0	0 B	user	

Рис. 9. Создание текстового файла

```
hadoop@akalashnikov-VirtualBox:~$ hdfs dfs -lsr /
lsr: DEPRECATED: Please use 'ls -R' instead.
-rw-r--r-- 1 hadoop supergroup 0 2023-09-17 00:20 /file.txt
-rw-r--r-- 1 hadoop supergroup 0 2023-09-18 16:32 /file1.txt
drwxr-xr-x - hadoop supergroup 0 2023-09-09 21:54 /test
drwxr-xr-x - hadoop supergroup 0 2023-09-09 21:54 /test1
drwxr-xr-x - hadoop supergroup 0 2023-09-18 16:58 /user
drwxr-xr-x - hadoop supergroup 0 2023-09-18 17:05 /user/hduser
-rw-r--r-- 1 hadoop supergroup 17 2023-09-18 17:07 /user/hduser/file.txt
hadoop@akalashnikov-VirtualBox:~$
```

Рис. 10. Просмотр прав

```
hadoop@akalashnikov-VirtualBox:~$ hdfs dfs -chmod 770 /file.txt
hadoop@akalashnikov-VirtualBox:~$ hdfs dfs -lsr /
lsr: DEPRECATED: Please use 'ls -R' instead.
-rwxrwx--- 1 hadoop supergroup 0 2023-09-17 00:20 /file.txt
-rw-r--r-- 1 hadoop supergroup 0 2023-09-18 16:32 /file1.txt
drwxr-xr-x - hadoop supergroup 0 2023-09-09 21:54 /test
drwxr-xr-x - hadoop supergroup 0 2023-09-09 21:54 /test1
drwxr-xr-x - hadoop supergroup 0 2023-09-18 16:58 /user
drwxr-xr-x - hadoop supergroup 0 2023-09-18 17:05 /user/hduser
-rw-r--r-- 1 hadoop supergroup 17 2023-09-18 17:07 /user/hduser/file.txt
hadoop@akalashnikov-VirtualBox:~$
```

Рис. 11. Изменение прав

In operation

DataNode State

All

Show

25

entries

Search:

Node	Http Address	Last contact	Last Block Report	Used	Non DFS Used	Capacity	Blocks	Block pool used	Version	
<div> <div>✓</div> <div>default-rack/akalashnikov-VirtualBox:9866 (127.0.0.1:9866)</div> </div>	http://akalashnikov-VirtualBox:9864	0s	1m	32 KB	13.77 GB	40.63 GB	<div> <div></div> <div></div> </div>	0	32 KB (0%)	3.3.6

Рис. 12. Кластер из двух серверов

### Листинг программы

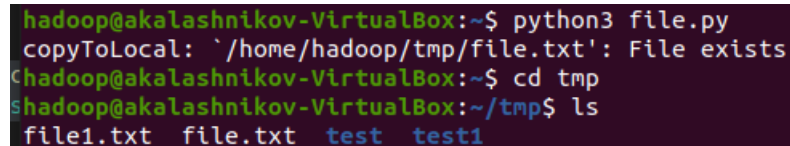
```
import subprocess
import os

def copy_hdfs_to_local(hdfs_dir, local_dir):
    subprocess.run(["hadoop", "fs", "-get", hdfs_dir, local_dir])

# Пример вызова функции
hdfs_dir = "/"
local_dir = "/home/akalashnikov/tmp"

os.system('hdfs dfs -copyToLocal / /home/hadoop/tmp')
#copy_hdfs_to_local(hdfs_dir, local_dir)
```

### Результат программы



```
hadoop@akalashnikov-VirtualBox:~$ python3 file.py
copyToLocal: `/home/hadoop/tmp/file.txt': File exists
hadoop@akalashnikov-VirtualBox:~$ cd tmp
hadoop@akalashnikov-VirtualBox:~/tmp$ ls
file1.txt file.txt test test1
```

Рис.13. Результат

**Выводы:** в результате выполнения лабораторной работы были сформированы практические навыки по установке и настройке кластера Hadoop и работе с файловой системой HDFS.