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

Башкатова Анна Денисовна

ОТЧЕТ

по дисциплине «Инструменты для хранения и обработки больших данных» Тема: «5.1. Установка и настройка Hadoop»

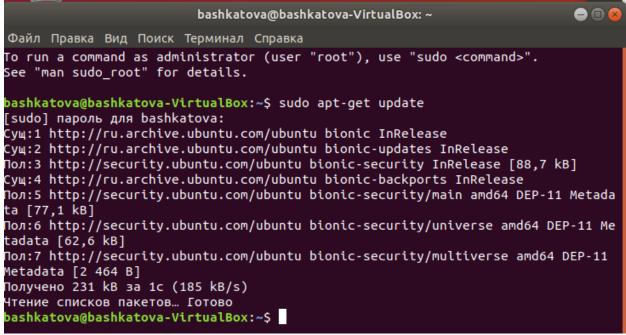
Направление подготовки (специальность) 38.03.05 – бизнес-информатика Направленность (профиль) образовательной программы «Аналитика данных и эффективное управление»

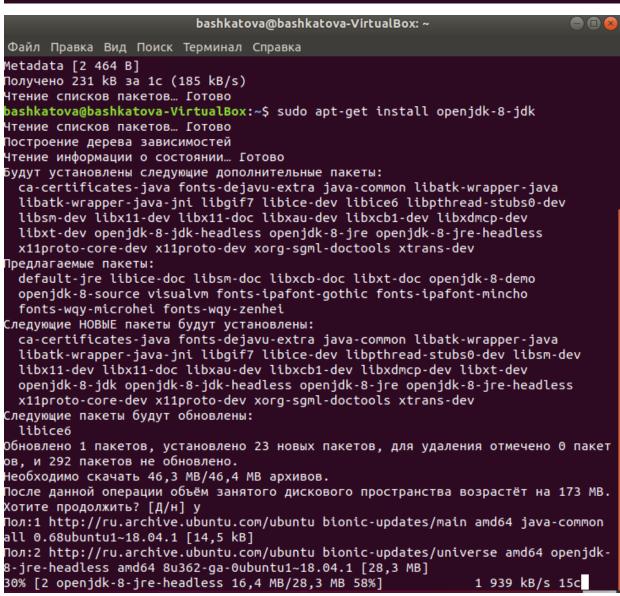
Курс обучения: 3

Форма обучения: очная

Руководитель: Босенко Т. М.

Установка Hadoop.





bashkatova@bashkatova-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) bashkatova@bashkatova-VirtualBox:~\$ openjuk o4-bit Server VM (Dulla 25.302-D09, Mixea Mode) bashkatova@bashkatova-VirtualBox:~\$ sudo nano /etc/environment bashkatova@bashkatova-VirtualBox:~S bashkatova@bashkatova-VirtualBox: ~ Файл Правка Вид Поиск Терминал Справка GNU nano 2.9.3 Изменён /etc/environment PATH="/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:/usr/game\$ JAVA_HOME="/usr/lib/jvm/java-8-openjdk-amd64" JRE_HOME="/usr/lib/jvm/java-8-openjdk-amd64/jre" Помощь Записать Поиск Вырезать ^Ј Выровнять

Замена

Отмен. вырез^Т

Словарь

ЧитФайл

Выход

```
bashkatova@bashkatova-VirtualBox:~$ sudo nano /etc/environment
bashkatova@bashkatova-VirtualBox:~$ sudo adduser hadoop
Добавляется пользователь «hadoop» ...
Добавляется новая группа «hadoop» (1001) ...
Добавляется новый пользователь «hadoop» (1001) в группу «hadoop» ...
Создаётся домашний каталог «/home/hadoop» ...
Копирование файлов из «/etc/skel» ...
Введите новый пароль UNIX:
Повторите ввод нового пароля UNIX:
passwd: пароль успешно обновлён
Изменение информации о пользователе hadoop
Введите новое значение или нажмите ENTER для выбора значения по умолчанию
        Полное имя []:
        Номер комнаты []:
        Рабочий телефон []:
        Домашний телефон []:
        Другое []:
Данная информация корректна? [Y/n]
bashkatova@bashkatova-VirtualBox:~$ sudo su hadoop
hadoop@bashkatova-VirtualBox:/home/bashkatova$
```

bashkatova@bashkatova-VirtualBox:~\$ sudo su hadoop hadoop@bashkatova-VirtualBox:/home/bashkatova\$ exit exit bashkatova@bashkatova-VirtualBox:~\$

```
bashkatova@bashkatova-VirtualBox: ~
Файл Правка Вид Поиск Терминал Справка
bashkatova@bashkatova-VirtualBox:~$ sudo apt-get install ssh pdsh
Чтение списков пакетов… Готово
Построение дерева зависимостей
Чтение информации о состоянии... Готово
Будут установлены следующие дополнительные пакеты:
  genders libgenders0 ncurses-term openssh-client openssh-server
 openssh-sftp-server ssh-import-id
Предлагаемые пакеты:
  rdist keychain libpam-ssh monkeysphere ssh-askpass molly-guard rssh
Следующие НОВЫЕ пакеты будут установлены:
  genders libgenders0 ncurses-term openssh-server openssh-sftp-server pdsh
  ssh ssh-import-id
Следующие пакеты будут обновлены:
 openssh-client
Обновлено 1 пакетов, установлено 8 новых пакетов, для удаления отмечено 0 пак
етов, и 291 пакетов не обновлено.
Необходимо скачать 812 kB/1 422 kB архивов.
После данной операции объём занятого дискового пространства возрастёт на 5 90
6 kB.
Хотите продолжить? [Д/н] у
Пол:1 http://ru.archive.ubuntu.com/ubuntu bionic-updates/main amd64 openssh-s
ftp-server amd64 1:7.6p1-4ubuntu0.7 [45,5 kB]
Пол:2 http://ru.archive.ubuntu.com/ubuntu bionic-updates/main amd64 openssh-s
erver amd64 1:7.6p1-4ubuntu0.7 [332 kB]
Пол:3 http://ru.archive.ubuntu.com/ubuntu bionic-updates/main amd64 ssh all 1
:7.6p1-4ubuntu0.7 [5 192 B]
Пол:4 http://ru.archive.ubuntu.com/ubuntu bionic/universe amd64 libgenders0 a
md64 1.21-1build5 [31,3 kB]
Пол:5 http://ru.archive.ubuntu.com/ubuntu bionic/universe amd64 genders amd64
1.21-1build5 [30,9 kB]
Пол:6 http://ru.archive.ubuntu.com/ubuntu bionic-updates/main amd64 ncurses-t
erm all 6.1-1ubuntu1.18.04 [248 kB]
Пол:7 http://ru.archive.ubuntu.com/ubuntu bionic/universe amd64 pdsh amd64 2.
31-3build2 [108 kB]
```

```
bashkatova@bashkatova-VirtualBox:~$ sudo su hadoop
hadoop@bashkatova-VirtualBox:/home/bashkatova$ cd
hadoop@bashkatova-VirtualBox:~$ ssh-keygen -t rsa -N "" -f /home/hadoop/.ssh/
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:j0VfHbBJ74fyC8Z2MwjAdI169M24zvRK5iy40/vWerQ hadoop@bashkatova-VirtualB
The key's randomart image is:
+---[RSA 2048]---
         . .00
        0 .0..=
         +0 .0+0
        =... 0000
          ..*00=
         o oBEo.+
         +Bo++..
+----[SHA256]----+
hadoop@bashkatova-VirtualBox:~$
```

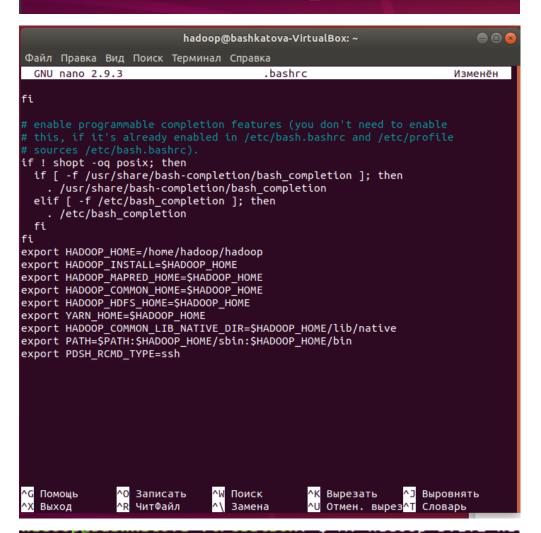
hadoop@bashkatova-VirtualBox:~\$ cat /home/hadoop/.ssh/id_rsa.pub >> /home/hadoop/.ssh/authorized_keys hadoop@bashkatova-VirtualBox:~\$ chmod 0600 /home/hadoop/.ssh/authorized_keys hadoop@bashkatova-VirtualBox:~\$

```
adoop@bashkatova-VirtualBox:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established. ECDSA key fingerprint is SHA256:jpvovkk842dcDhTqjk2Hrx0P0PgctoctPEf8B88vUp4.
Are you sure you want to continue connecting (yes/no)? yes Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
Welcome to Ubuntu 18.04.6 LTS (GNU/Linux 5.4.0-84-generic x86_64)
 * Documentation: https://help.ubuntu.com
                     https://landscape.canonical.com
 * Management:
 * Support:
                     https://ubuntu.com/advantage
297 updates can be applied immediately.
267 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 2023.
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@bashkatova-VirtualBox:~$
```

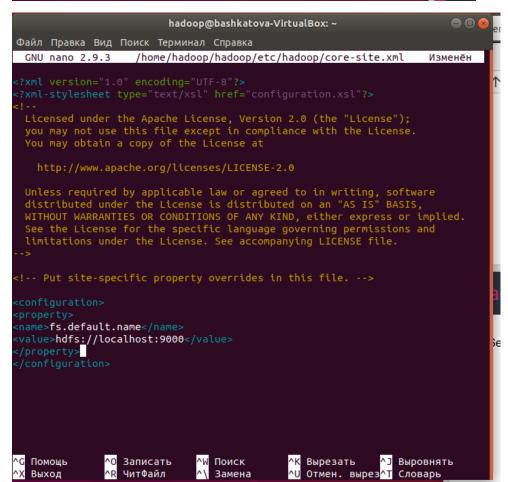
```
hadoop@bashkatova-VirtualBox:~$ wget https://archive.apache.org/dist/hadoop/common/hadoop-3.1.2/hadoop-3.1.2.tar.gz
--2023-04-22 15:31:39-- https://archive.apache.org/dist/hadoop/common/hadoop
-3.1.2/hadoop-3.1.2.tar.gz
Pacno3HaëTcя archive.apache.org (archive.apache.org)... 138.201.131.134, 2a01:4
f8:172:2ec5::2
Подключение к archive.apache.org (archive.apache.org)|138.201.131.134|:443...
соединение установлено.
HTTP-запрос отправлен. Ожидание ответа... 200 ОК
Длина: 332433589 (317M) [application/x-gzip]
Сохранение в: «hadoop-3.1.2.tar.gz»
hadoop-3.1.2.tar.gz 9%[> ] 29,62M 1,77MB/s ост 1m 50s
```

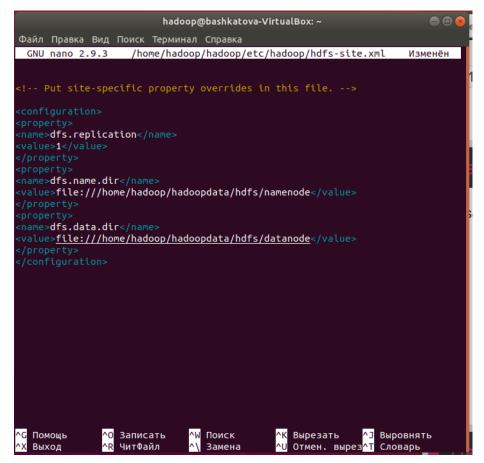
```
hadoop@bashkatova-VirtualBox:~$ wget https://archive.apache.org/dist/hadoop/common/hadoop-3.1.2/hadoop-3.1.2.tar.gz
--2023-04-22 15:31:39-- https://archive.apache.org/dist/hadoop/common/hadoop
-3.1.2/hadoop-3.1.2.tar.gz
Pacno3Haëtcя archive.apache.org (archive.apache.org)... 138.201.131.134, 2a01:4
f8:172:2ec5::2
Подключение к archive.apache.org (archive.apache.org)|138.201.131.134|:443...
соединение установлено.
HTTP-запрос отправлен. Ожидание ответа... 200 ОК
Длина: 332433589 (317M) [application/x-gzip]
Сохранение в: «hadoop-3.1.2.tar.gz»
hadoop-3.1.2.tar.gz 100%[============]] 317,03M 2,20MB/s за 2m 37s
2023-04-22 15:34:16 (2,02 MB/s) - «hadoop-3.1.2.tar.gz» сохранён [332433589/3 32433589]
hadoop@bashkatova-VirtualBox:~$
```

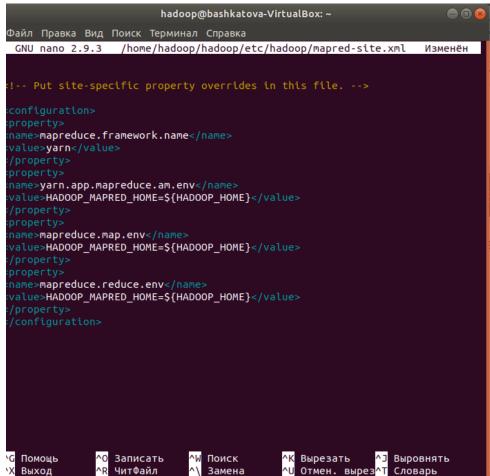
hadoop@bashkatova-VirtualBox:~\$ mv hadoop-3.1.2 hadoop hadoop@bashkatova-VirtualBox:~\$ [



hadoop@bashkatova-VirtualBox:~\$ nano .bashrc hadoop@bashkatova-VirtualBox:~\$ source .bashrc hadoop@bashkatova-VirtualBox:~\$



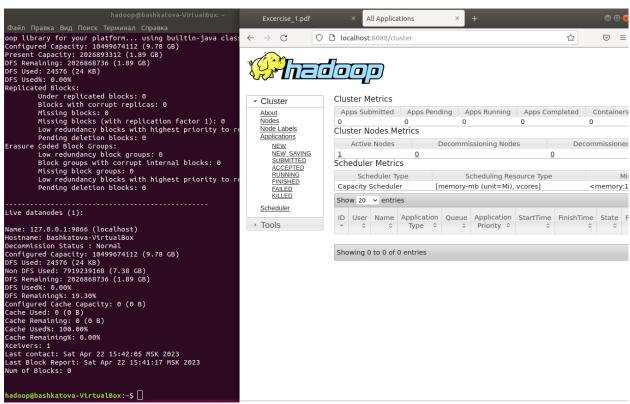


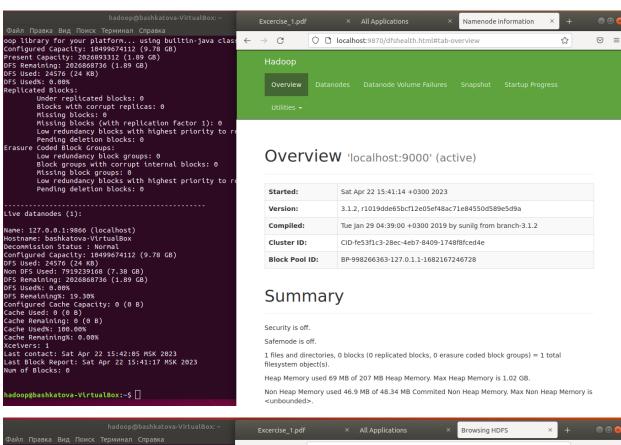


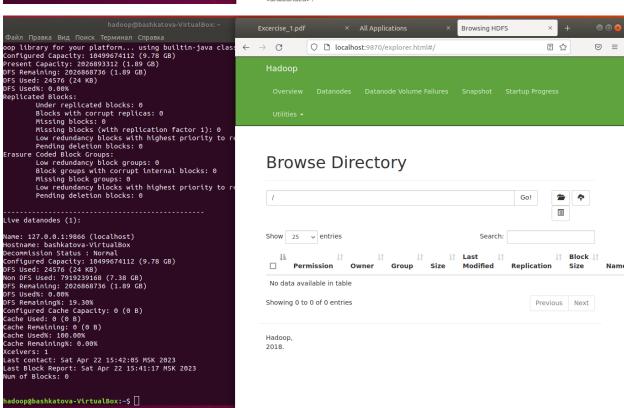
```
hadoop@bashkatova-VirtualBox: ~
                                                                                          Файл Правка Вид Поиск Терминал Справка
  GNU nano 2.9.3
                         /home/hadoop/hadoop/etc/hadoop/yarn-site.xml
                                                                                     Изменён
  Licensed under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at
    http://www.apache.org/licenses/LICENSE-2.0
  Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, 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.
 !-- Site specific YARN configuration properties -->
 name>yarn.nodemanager.aux-services</name>
 value>mapreduce_shuffle</value>
 name>yarn.nodemanager.resource.memory-mb</name>
 value>16384</value>
                                                                           ^] Выровнять
^G Помощь
                                      ^₩ Поиск
                      Записать
                                                            Вырезать
^Х Выход
                      ЧитФайл
                                         Замена
                                                            Отмен. вырез^Т
                                                                               Словарь
 hadoop@bashkatova-VirtualBox:~$ source .bashrc
hadoop@bashkatova-VirtualBox:~$ nano /home/hadoop/hadoop/etc/hadoop/hadoop-en
v.sh
hadoop@bashkatova-VirtualBox:~$ nano /home/hadoop/hadoop/etc/hadoop/core-site
.xml
hadoop@bashkatova-VirtualBox:~$ nano /home/hadoop/hadoop/etc/hadoop/hdfs-site
 .xml
hadoop@bashkatova-VirtualBox:~$ nano /home/hadoop/hadoop/etc/hadoop/mapred-si
te.xml
hadoop@bashkatova-VirtualBox:~$ nano /home/hadoop/hadoop/etc/hadoop/yarn-site
hadoop@bashkatova-VirtualBox:~$
hadoop@bashkatova-VirtualBox:~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [bashkatova-VirtualBox]
bashkatova-VirtualBox: Warning: Permanently added 'bashkatova-virtualbox' (EC
DSA) to the list of known hosts.
2023-04-22 15:41:22,041 WARN util.NativeCodeLoader: Unable to load native-had
oop library for your platform... using builtin-java classes where applicable
```

hadoop@bashkatova-VirtualBox:~\$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers

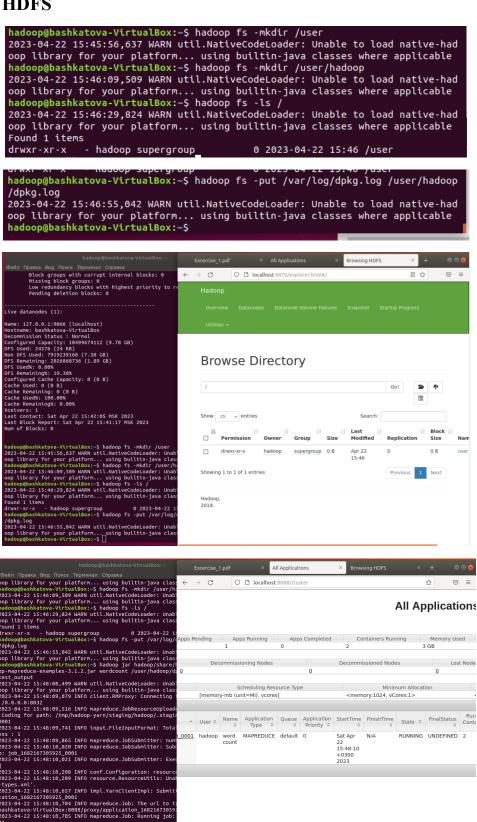
```
10000pg0aSinkatova-virtuatbox:~; indis disadmin -report
2023-04-22 15:42:06,396 WARN util.NativeCodeLoader: Unable to load native-had
oop library for your platform... using builtin-java classes where applicable
Configured Capacity: 10499674112 (9.78 GB)
Present Capacity: 2026893312 (1.89 GB)
DES Pemaining: 2026898736 (1.89 GB)
DFS Remaining: 2026868736 (1.89 GB)
DFS Used: 24576 (24 KB)
DFS Used%: 0.00%
Replicated Blocks:
            Under replicated blocks: 0
            Blocks with corrupt replicas: 0
           Missing blocks: 0
Missing blocks (with replication factor 1): 0
Low redundancy blocks with highest priority to recover: 0
Pending deletion blocks: 0
Erasure Coded Block Groups:
            Low redundancy block groups: 0
Block groups with corrupt internal blocks: 0
            Missing block groups: 0
            Low redundancy blocks with highest priority to recover: 0
            Pending deletion blocks: 0
Live datanodes (1):
Name: 127.0.0.1:9866 (localhost)
Hostname: bashkatova-VirtualBox
Decommission Status : Normal
Configured Capacity: 10499674112 (9.78 GB)
DFS Used: 24576 (24 KB)
Non DFS Used: 7919239168 (7.38 GB)
DFS Remaining: 2026868736 (1.89 GB)
DFS Used%: 0.00%
DFS Remaining%: 19.30%
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 15:42:05 MSK 2023
Last Block Report: Sat Apr 22 15:41:17 MSK 2023
Num of Blocks: 0
```



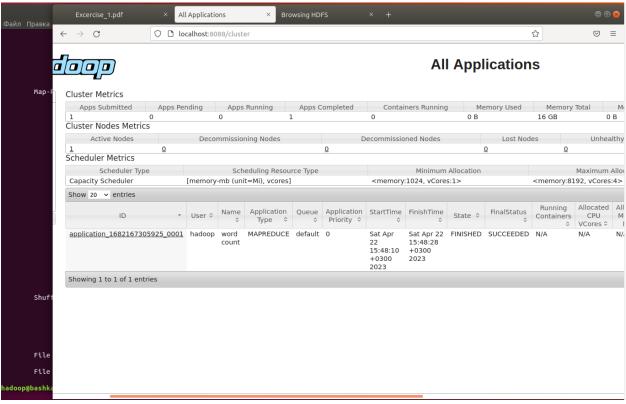




Работа в HDFS



23-04-22 15:48:18,880 INFO mapreduce.Job: Job job_168: In uber mode : false 23-04-22 15:48:18,882 INFO mapreduce.Job: map 0% redu 23-04-22 15:48:24,952 INFO mapreduce.Job: map 100% re



```
hadoop@bashkatova-VirtualBox:~$ hadoop fs -cat /user/hadoop/test_output/part-
r-00000
2023-04-22 15:49:17,008 WARN util.NativeCodeLoader: Unable to load native-had
oop library for your platform... using builtin-java classes where applicable
0+svn585-1build1
0.0.12 8
0.0.20060226-9ubuntu1
0.0.4-1 8
0.0.4-6ubuntu2 8
0.0.7+17.10.20170922-0ubuntu1
                               10
0.0.9-10
            8
0.01-5 8
0.032-1 8
0.04-7.1
              8
0.044-2 8
0.06-2.1ubuntu1 8
0.06-7.1
0.07-1 8
0.0~r131-2ubuntu3
0.0~r131-2ubuntu3.1
0.1-5 8
0.1-6
      8
0.1.0+qit20170815-2
                       16
0.1.0-2ubuntu1 8
0.1.10-2.1
0.1.2-1ubuntu1 8
0.1.2-2 8
0.1.26-2
               8
0.1.27-2
              16
0.1.3-1 8
0.1.32 8
0.1.33build1
               8
```

Выполнение упражнений

Клонирование репозитория git

```
hadoop@bashkatova-VirtualBox:~$ git clone https://github.com/BosenkoTM/Big-Data-Storage-a
nd-Processing-Tools.git Bosenko
Клонирование в «Bosenko»..
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
Получение объектов: 100% (123/123), 19.72 MiB | 1.77 MiB/s, готово.
Определение изменений: 100% (37/37), готово.
hadoop@bashkatova-VirtualBox:~$ ls
 Big-Data-Storage-and-Processing-Tools
                                            hadoopdata
                                                          Музыка
 Bosenko
                                                          Общедоступные
                                            Видео
 examples.desktop
                                            Документы
                                                         'Рабочий стол'
 hadoop
                                                          Шаблоны
                                            Загрузки
                                            Изображения
hadoop@bashkatova-VirtualBox:~$ cd Big-Data-Storage-and-Processing-Tools
hadoop@bashkatova-VirtualBox:~/Big-Data-Storage-and-Processing-Tools$ ls
kp lectures marketing_analytics_students.ipynb practice README.md
hadoop@bashkatova-VirtualBox:~/Big-Data-Storage-and-Processing-Tools$ cd
hadoop@bashkatova-VirtualBox:~$ cd Big-Data-Storage-and-Processing-Tools
hadoop@bashkatova-VirtualBox:~/Big-Data-Storage-and-Processing-Tools$ cd practice
hadoop@bashkatova-VirtualBox:~/Big-Data-Storage-and-Processing-Tools/practice$ ls
hadoop@bashkatova-VirtualBox:~/Big-Data-Storage-and-Processing-Tools/practice$ cd
hadoop@bashkatova-VirtualBox:~$ hadoop fs -put Big-Data-Storage-and-Processing-Tools/prac
tice/Faust_1.txt /user/hadoop/Faust_1.txt
2023-04-27 19:58:06,282 WARN util.NativeCodeLoader: Unable to load native-hadoop library
for your platform... using builtin-java classes where applicable hadoop@bashkatova-VirtualBox:~$
```

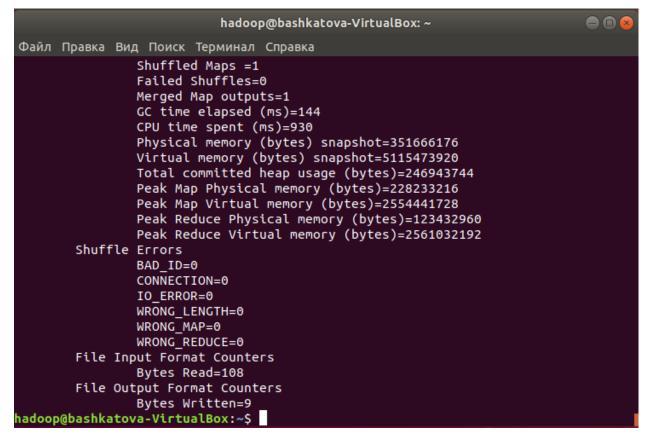
```
hadoop@bashkatova-VirtualBox:-$ hadoop jar hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.1.2.jar
wordcount /user/hadoop/Faust_1.txt /user/hadoop/Faust__Output
2023-04-27 20:57:52,168 WARN util.Nativcodecoader: Unable to load native-hadoop library for your platform...
using builtin-java classes where applicable
2023-04-27 20:57:53,204 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2023-04-27 20:57:53,812 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-ya
rn/staging/hadoop/.staging/job_1682168633199_0003
2023-04-27 20:57:54,221 INFO hapt-fileInputFormat: Total input files to process : 1
2023-04-27 20:57:54,221 INFO hapreduce.JobSubmitter: number of splits:1
2023-04-27 20:57:54,643 INFO mapreduce.JobSubmitter: submitting tokens for job: job_1682168633199_0003
2023-04-27 20:57:54,644 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-04-27 20:57:54,644 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-04-27 20:57:54,912 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-04-27 20:57:55,145 INFO mapreduce.Job: The url to track the job: http://bashkatova-VirtualBox:8088/proxy
/application_1682168633199_0003/
2023-04-27 20:57:55,145 INFO mapreduce.Job: The url to track the job: http://bashkatova-VirtualBox:8088/proxy
/application_1682168633199_0003/
2023-04-27 20:58:04,458 INFO mapreduce.Job: Dob job_1682168633199_0003 running in uber mode : false
2023-04-27 20:58:04,458 INFO mapreduce.Job: map 0% reduce 0%
2023-04-27 20:58:04,450 INFO mapreduce.Job: map 100% reduce 0%
2023-04-27 20:58:17,703 INFO mapreduce.Job: map 100% reduce 0%
2023-04-27 20:58:17,875 INFO mapreduce.Job: map 100% reduce 0%
2023-04-27 20:58:17,875 INFO mapreduce.Job: counters: 53
FILE: Number of by
```

hadoop@bashkatova-VirtualBox:~\$ hadoop fs -get /user/hadoop/Faust_1_Output/part-r-00000 Faust_1_Output.csv 2023-04-27 20:59:45,852 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable hadoop@bashkatova-VirtualBox:~\$

```
hadoop@bashkatova-VirtualBox:~$ head -10 Faust 1 Output.csv
"Allein,
"Alles
"Als
         1
         1
"Der
"Die
         2
"Ег
         2
         4
"Ich
"Im
         1
"Mein
         1
"Nur
         1
```

```
hadoop@bashkatova-VirtualBox:~$ hadoop fs -put Bosenko/practice/Faust_1.txt /use
r/hadoop/Faust_2.txt
2023-05-03 15:22:53,591 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
hadoop@bashkatova-VirtualBox:~$
```

```
hadoop@bashkatova-VirtualBox:~$ hadoop jar hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.1.2.jar grep /user/hadoop/Faust_2.txt /user/hadoop/Faust_1_Count_Output 'Faust'
2023-05-03 15:24:14,148 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable 2023-05-03 15:24:15,224 INFO client.RMProxy: Connecting to ResourceManager at /0 .0.0.88032
2023-05-03 15:24:15,910 INFO mapreduce.JobResourceUploader: Disabling Erasure Co ding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1683114645438_0003 2023-05-03 15:24:16,326 INFO input.FileInputFormat: Total input files to process: 1 2023-05-03 15:24:16,433 WARN hdfs.DataStreamer: Caught exception java.lang.InterruptedException at java.lang.Object.wait(Native Method)
```



Show 20 v entries									
ID *	User \$	Name \$	Application Type \$	Queue \$	Application Priority \$	StartTime \$	FinishTime	State \$	FinalStat
application_1683114645438_0004	hadoop	grep- sort	MAPREDUCE	default	0	Wed May 3 15:24:43 +0300 2023	Wed May 3 15:25:10 +0300 2023	FINISHED	SUCCEEL
application_1683114645438_0003	hadoop	grep- search	MAPREDUCE	default	0	Wed May 3 15:24:17 +0300 2023	Wed May 3 15:24:41 +0300 2023	FINISHED	SUCCEEL
application_1683114645438_0002	hadoop	word count	MAPREDUCE	default	0	Wed May 3 15:06:04 +0300 2023	Wed May 3 15:06:30 +0300 2023	FINISHED	SUCCEEL
application_1683114645438_0001	hadoop	word count	MAPREDUCE	default	0	Wed May 3 14:54:06 +0300 2023	Wed May 3 14:54:39 +0300 2023	FINISHED	SUCCEEL

hadoop@bashkatova-VirtualBox:~\$ hadoop fs -get /user/hadoop/Faust_1_Count_Output /part-r-00000 Faust_1_Count_Output.csv 2023-05-03 15:27:11,942 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable hadoop@bashkatova-VirtualBox:~\$

hadoop@bashkatova-VirtualBox:~\$ cat Faust_1_Count_Output.csv
50 Faust