



Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет
имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ

КАФЕДРА КОМПЬЮТЕРНЫЕ СИСТЕМЫ И СЕТИ (ИУ6)

НАПРАВЛЕНИЕ ПОДГОТОВКИ 09.04.01 Информатика и вычислительная техника

МАГИСТЕРСКАЯ ПРОГРАММА 09.04.01/07 Интеллектуальные системы анализа,
обработки и интерпретации больших данных

О Т Ч Е Т

по лабораторной работе № 7-8

Название: Работа с Hadoop Spark

Дисциплина: Технология параллельных систем баз данных

Студент

ИУ6-12М

(Группа)

(Подпись, дата)

Д.С. Каткова

(И.О. Фамилия)

Преподаватель

(Подпись, дата)

А.Д. Пономарев

(И.О. Фамилия)

Москва, 2023

1 Цель лабораторной работы

Цель работы – приобретение навыков инсталляции продуктов Apache Hadoop и Apache Spark, поддерживающих технологию MapReduce, которые используются для обработки больших данных (Big Data).

Ход работы

Установка Hadoop

Перепишем 3 файла из яндекс-папки в свою ВМ. В этом каталоге хранятся 3 файла: архив Hadoop, архив Spark и текстовый файл для анализа:

- 1) hadoop-2.9.2.tar.gz;
- 2) spark-2.4.6-bin-hadoop2.7.tgz;
- 3) gamlet_en.txt.

Создадим пользователя hduser в группе Hadoop. Теперь пользователю hduser добавим права для выполнения команд типа от рута (sudo). Вставим строку hduser ALL=(ALL:ALL) ALL

```
# Host alias specification

# User alias specification

# Cmnd alias specification

# User privilege specification
root    ALL=(ALL:ALL) ALL
hduser  ALL=(ALL:ALL) ALL
# Members of the admin group may gain root privileges
%admin  ALL=(ALL) ALL

# Allow members of group sudo to execute any command
%sudo   ALL=(ALL:ALL) ALL

# See sudoers(5) for more information on "#include" directives:

#includedir /etc/sudoers.d
```

Далее устанавливаем и настраиваем SSH. SSH-сервер не установлен на ВМ, установим его командой `sudo apt-get install openssh-server`.

Сгенерируем SSH-ключ под hduser.

```

hduser@daria-VirtualBox:~$ ssh-keygen -t rsa -P ""
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hduser/.ssh/id_rsa):
Created directory '/home/hduser/.ssh'.
Your identification has been saved in /home/hduser/.ssh/id_rsa
Your public key has been saved in /home/hduser/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:aSSipLqk9p3gyhR6n0klxRwp78g4uKFNgxbLisUpWKY hduser@daria-VirtualBox
The key's randomart image is:
+---[RSA 3072]-----+
|      ..          |
|      .o..        |
|      . .o= .     |
|    +o. o.o .     |
|  +O++ .o. S      |
| E+X oo..         |
| BX.+            |
| X+oooo.          |
| +ooo+o          |
+----[SHA256]-----+
hduser@daria-VirtualBox:~$ cat $HOME/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys
hduser@daria-VirtualBox:~$

```

Далее установим HADOOP. Распакуем архив, после чего переместим распакованный каталог в рабочий каталог /usr/local/hadoop.

Затем создадим рабочие каталоги для HDFS (NameNode и DataNode) и назначим владельца каталога hadoop (пользователь hduser).

```

hduser@daria-VirtualBox:~$ cd /usr/local
hduser@daria-VirtualBox:/usr/local$ ls -l
total 36
drwxr-xr-x 2 root root 4096 апр 2 2020 bin
drwxr-xr-x 2 root root 4096 апр 2 2020 etc
drwxr-xr-x 2 root root 4096 апр 2 2020 games
drwxr-xr-x 4 hduser hadoop 4096 дек 19 21:28 hadoop
drwxr-xr-x 2 root root 4096 апр 2 2020 include
drwxr-xr-x 3 root root 4096 апр 2 2020 lib
lrwxrwxrwx 1 root root 9 дек 19 20:43 man -> share/man
drwxr-xr-x 2 root root 4096 апр 2 2020 sbin
drwxr-xr-x 5 root root 4096 апр 2 2020 share
drwxr-xr-x 2 root root 4096 апр 2 2020 src
hduser@daria-VirtualBox:/usr/local$

```

Проверим наличие /usr/lib/jvm/java-1.8.0-openjdk-amd64 :

```

hduser@daria-VirtualBox:~$ update-java-alternatives -l
java-1.8.0-openjdk-amd64 1081 /usr/lib/jvm/java-1.8.0-openjdk-amd64
hduser@daria-VirtualBox:~$

```

Настроим hadoop-env.sh

```

/usr/local/hadoop/hadoop-2.9.2/etc/hadoop/hadoop-env.sh Modified
# These options will be appended to the options specified as HADOOP_OPTS
# and therefore may override any similar flags set in HADOOP_OPTS
#
# export HADOOP_DFSROUTER_OPTS=""
###
###
# Advanced Users Only!
###
# The directory where pid files are stored. /tmp by default.
# NOTE: this should be set to a directory that can only be written to by
#       the user that will run the hadoop daemons. Otherwise there is the
#       potential for a symlink attack.
export HADOOP_PID_DIR=${HADOOP_PID_DIR}
export HADOOP_SECURE_DN_PID_DIR=${HADOOP_PID_DIR}
# A string representing this instance of hadoop. $USER by default.
export HADOOP_IDENT_STRING=$USER
export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64

```

Настроим файла core-site.xml

```

/usr/local/hadoop/hadoop-2.9.2/etc/hadoop/core-site.xml Modified
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.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>fs.default.name</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>

```


Настроим файла hdfs-site.xml

```
/usr/local/hadoop/hadoop-2.9.2/etc/hadoop/hdfs-site.xml Modified
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.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.namenode.name.dir</name>
<value>file:/usr/local/hadoop/hadoop_tmp/hdfs/namenode</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>file:/usr/local/hadoop/hadoop_tmp/hdfs/datanode</value>
</property>
```

Настроим yarn-site.xml.

```
/usr/local/hadoop/hadoop-2.9.2/etc/hadoop/yarn-site.xml Modified
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.
-->
<configuration>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
<property>
<name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
<!-- Site specific YARN configuration properties -->

</configuration>
```

```

/usr/local/hadoop/hadoop-2.9.2/etc/hadoop/mapred-site.xml Modified
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.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
</configuration>

```

Успешность запуска HADOOP проверим с помощью команды `jps`.

```
hduser@daria-VirtualBox:~$ jps
2531 SecondaryNameNode
2948 Jps
2790 NodeManager
2296 DataNode
2171 NameNode
2653 ResourceManager
hduser@daria-VirtualBox:~$
```

После выполнения всех команд, приведенных методических указания, наем следующий вывод в консоль.

6

Запустим Hadoop, а затем создадим каталог в HDFS и посмотрим список в корневой папке.

```
hduser@daria-VirtualBox:~$ hdfs dfs -mkdir /chapter5
hduser@daria-VirtualBox:~$ hdfs dfs -ls /
Found 1 items
drwxr-xr-x  - hduser supergroup          0 2023-12-19 21:49 /chapter5
hduser@daria-VirtualBox:~$
```

Запустим оболочку для работы с Python в Spark и введем операторы для проверки записи/чтения файла из HDFS.

+-----+-----+	
	_c0 _c1
+-----+-----+	
	First 1
	Second 2
	Third 3
	Fourth 4
	Fifth 5
+-----+-----+	

4 Подсчет числа слов в тексте

Перепишем файл с пьесой «Гамлет» из локальной файловой системы в файловую систему Hadoop

```
hduser@daria-VirtualBox:~$ hdfs dfs -ls /chapter5
Found 2 items
drwxr-xr-x  - hduser supergroup          0 2023-12-19 21:50 /chapter5/example.csv
-rw-r--r--  1 hduser supergroup    188041 2023-12-19 21:55 /chapter5/gamlet_en.txt
hduser@daria-VirtualBox:~$
```

Выполним следующий код.

```
from pyspark import SparkContext
from datetime import datetime
f = sc.textFile("hdfs://localhost:9000/chapter5/gamlet_en.txt")
counts = f.flatMap(lambda line:line.split(" ")) \
    .map(lambda word: (word , 1)) \
    .reduceByKey(lambda a, b: a + b) \
    .sortBy(lambda a: a[1], ascending=False)
counts.saveAsTextFile("/home/hduser/res/")
sm = counts.map(lambda x: x[1]).sum()
print ("Summa %d" % (sm))
quit()
```

Результат работы программы представлен ниже.

```
GNU nano 4.8
(u'the', 995)
(u'and', 701)
(u'of', 641)
(u'to', 606)
(u'I', 511)
(u'a', 449)
(u'my', 444)
(u'in', 385)
(u'you', 363)
(u'Ham.', 358)
(u'is', 297)
(u'his', 281)
(u'it', 268)
(u'not', 255)
(u'And', 250)
(u']', 244)
(u'that', 225)
(u'your', 224)
(u'with', 222)
(u'this', 203)
```

Вывод: в ходе выполнения лабораторной работы был изучен процесс инсталляции и настройки продуктов Apache Hadoop и Apache Spark. Корректность установки была проверена на примере подсчета количества слов в тексте пьесы Шекспира «Гамлет».