**Практическое задание 5.**

**Реализовать MAPREDUCE JAR подсчета слов в файле и подсчета определенного, наперед заданного слова, в предложении. Задание реализовать в экосистеме Hadoop. Результаты представить в консоли и HDFS WEB BROWSER.**

**Место выполнения задания Виртуальная машина U20-04.**

**Материал для подготовки к практическому заданию**

[**https://github.com/BosenkoTM/ds\_practice/tree/main/exercises/winter\_semester\_2021-2022/05\_hadoop**](https://github.com/BosenkoTM/ds_practice/tree/main/exercises/winter_semester_2021-2022/05_hadoop)

**Установка Java**

**sudo apt-get update**

**sudo apt-get install openjdk-8-jdk**

проверка:

**java -version**

**sudo nano /etc/environment**

(может попросить пароль - нажимаем 1 и enter)

(открывается конфигурационный файл, там в самый конец вставляем):

*JAVA\_HOME="/usr/lib/jvm/java-8-openjdk-amd64"*

*JRE\_HOME="/usr/lib/jvm/java-8-openjdk-amd64/jre"*

(чтобы сохранить и выйти нажимаем ctrl X, затем Y, затем enter)

(аналогичным образом дальше заполняются остальные файлы)

**Установка HADOOP**

**sudo adduser hadoop**

**sudo passwd hadoop**

**sudo su hadoop**

**exit**

**sudo apt-get install ssh pdsh**

**sudo su hadoop**

**cd**

**ssh-keygen -t rsa -N "" -f /home/hadoop/.ssh/id\_rsa**

**cat /home/hadoop/.ssh/id\_rsa.pub >> /home/hadoop/.ssh/authorized\_keys**

**chmod 0600 /home/hadoop/.ssh/authorized\_keys**

**ssh localhost**

**exit**

**wget** [**https://archive.apache.org/dist/hadoop/common/hadoop-3.1.2/hadoop-3.1.2.tar.gz**](https://archive.apache.org/dist/hadoop/common/hadoop-3.1.2/hadoop-3.1.2.tar.gz)

**tar -xvzf hadoop-3.1.2.tar.gz**

**mv hadoop-3.1.2 hadoop**

**nano .bashrc**

(также вставляем в конец файла):

*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*

**source .bashrc**

**nano /home/hadoop/hadoop/etc/hadoop/hadoop-env.sh**

(также вставляем в конец файла):

*export JAVA\_HOME=/usr/lib/jvm/java-8-openjdk-amd64*

**nano /home/hadoop/hadoop/etc/hadoop/core-site.xml**

(листаем в конец файла и находим строчки <configuration></configuration> между ними вставляем):

*<property>*

*<name>fs.default.name</name>*

*<value>hdfs://localhost:9000</value>*

*</property>*

(аналогичным образом между делаются все остальные)

**nano /home/hadoop/hadoop/etc/hadoop/hdfs-site.xml**

*<property>*

*<name>dfs.replication</name>*

*<value>1</value>*

*</property>*

*<property>*

*<name>dfs.name.dir</name>*

*<value>file:///home/hadoop/hadoopdata/hdfs/namenode</value>*

*</property>*

*<property>*

*<name>dfs.data.dir</name>*

*<value>file:///home/hadoop/hadoopdata/hdfs/datanode</value>*

*</property>*

**nano /home/hadoop/hadoop/etc/hadoop/mapred-site.xml**

*<property>*

*<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>*

**nano /home/hadoop/hadoop/etc/hadoop/yarn-site.xml**

*<property>*

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

*<value>mapreduce\_shuffle</value>*

*</property>*

*<property>*

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

*<value>16384</value>*

*</property>*

**hdfs namenode -format**

**start-dfs.sh**

**start-yarn.sh**

**hdfs dfsadmin -report**

Проверяем localhost:8088 и localhost:9870, localhost:9870/explorer.html#/

1. Create User Directory (on HDFS):

**hadoop fs -mkdir /user**

**hadoop fs -mkdir /user/hadoop**

2. List Directories (on HDFS):

**hadoop fs -ls /**

3. Copy File (just a random log file) from local directory to HDFS:

**hadoop fs -put /var/log/dpkg.log /user/hadoop/dpkg.log**

**САМО ЗАДАНИЕ**

Подсчет слов в файле:

**git clone** [**https://github.com/BosenkoTM/ds\_practice.git**](https://github.com/BosenkoTM/ds_practice.git)

**hadoop fs -put ds\_practice/exercises/winter\_semester\_2021-2022/01\_hadoop/sample\_data/Faust\_1.txt /user/hadoop/Faust\_1.txt**

**hadoop jar hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.1.2.jar wordcount /user/hadoop/Faust\_1.txt /user/hadoop/Faust\_1\_Output**

Проверяем localhost:8088 и localhost:9870

**hadoop fs -get /user/hadoop/Faust\_1\_Output/part-r-00000 Faust\_1\_Output.csv**

Далее напишу код который выведет строки с подсчетом слов из нашего файла

**head -10 Faust\_1\_Output.csv**

Подсчет определенного слова в файле:

**hadoop fs -put ds\_practice/exercises/winter\_semester\_2021-2022/01\_hadoop/sample\_data/Faust\_1.txt /user/hadoop/Faust\_1.txt**

**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 'Faust'**

**hadoop fs -get /user/hadoop/Faust\_1\_Count\_Output/part-r-00000 Faust\_1\_Count\_Output.csv**

**cat Faust\_1\_Count\_Output.csv**