Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования

«Национальный исследовательский ядерный университет «МИФИ»

ИНСТИТУТ ИНТЕЛЛЕКТУАЛЬНЫХ КИБЕРНЕТИЧЕСКИХ СИСТЕМ

КАФЕДРА КОМПЬЮТЕРНЫХ СИСТЕМ И ТЕХНОЛОГИЙ

**Лабораторная работа №1**

**«Hadoop»**

**Вариант a (3), b (1), c (3), d (5)**

**Выполнил: Куренков Д.А.**

**Группа: М19-502**

**Москва, 2020**

**Постановка задачи:**

Реализовать программу, которая считает количество пользователей IE, Mozilla или других браузеров.

Дополнительное требование: счетчики используются для сбора статистики о неправильных строках

Отчётный материал должен содержать:

1.ZIP-архив с реализацией бизнес-логики;

2.Скриншоты успешного выполнения тестов;

3.Скриншоты загруженного файла в HDFS;

4.Скриншоты успешного выполнения задания и результата;

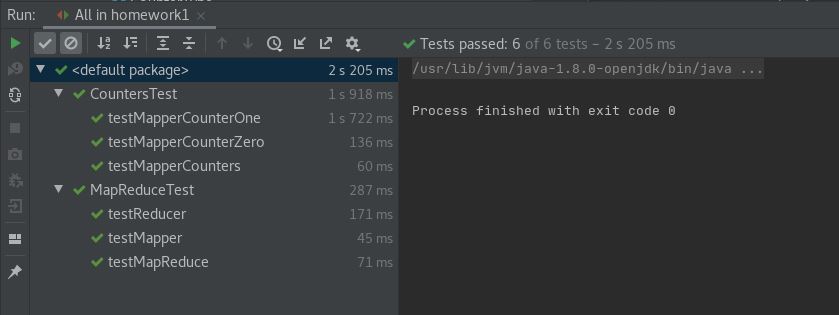
5.Быстрое руководство по сборке и развертыванию;

6.Скриншот чтения данных из csv файла.

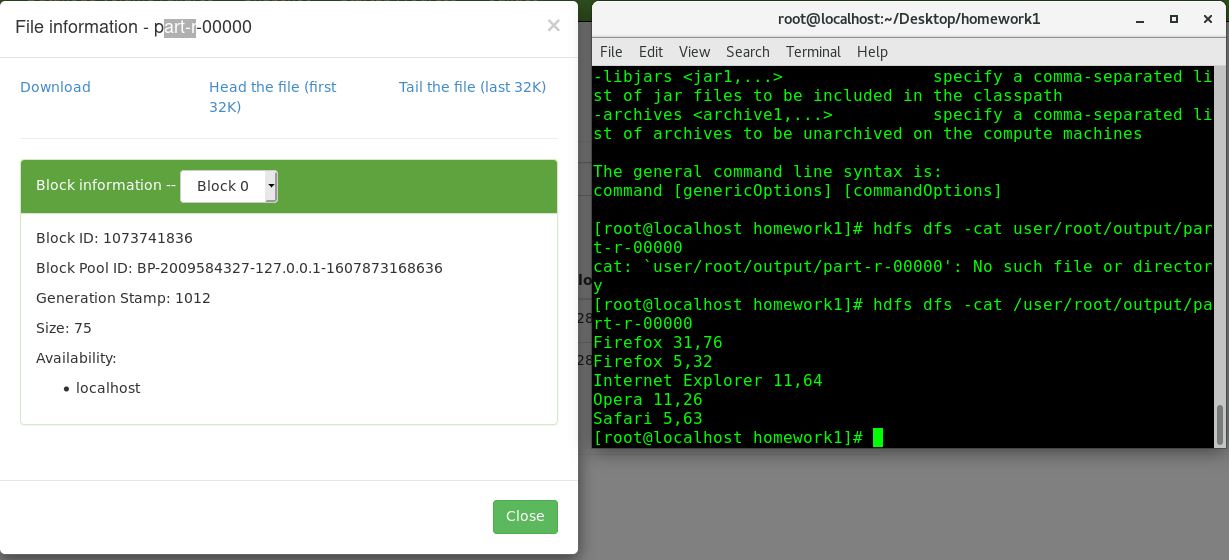
1.ZIP-архив с реализацией бизнес-логики

<https://github.com/zpbkda/lab1/blob/main/lab1.zip>

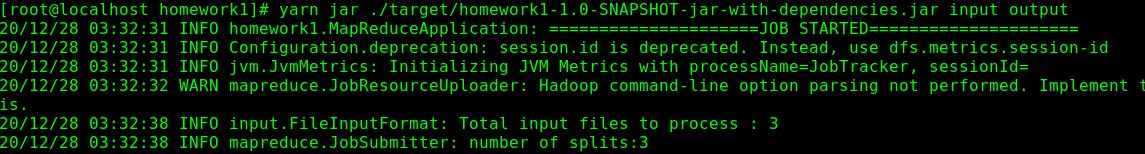
2.Скриншоты успешного выполнения тестов;



3.Скриншоты загруженного файла в HDFS;

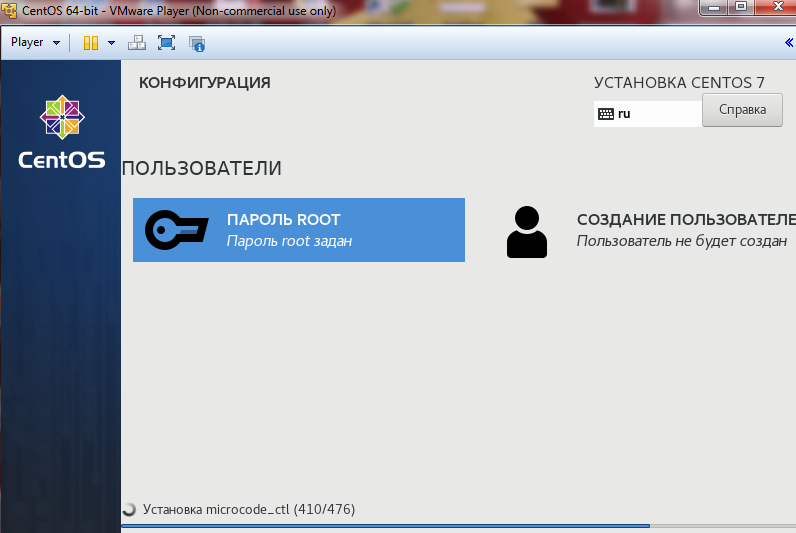


4.Скриншоты успешного выполнения задания;



5. Руководство по установке CentOS 7, по сборке и развертыванию.

Установить дистрибутив на свежесозданную виртуалку, в качестве варианта установки выбрать вариант по умолчанию Minimal Install.



При установке задаем пароль для root (1234567890). Система потребует перезагрузить для дальнейшей работы. После перезагрузки вводим логин и пароль в соответствие с тем, как мы установили.

Для того, чтобы начать устанавливать необходимо подключить интернет-соединение:  
$ONBOOT=no  
$dhclient  
$yum update

Далее необходимо установить оболочку GNOME на сервер:  
$sudo yum groups install “GNOME Desktop”

$sudo yum groups install “Server with GUI”

После успешной установки перезагружаемся, авторизируемся под root и в консоли пишем:  
$sudo systemctl set-default graphical.target  
Ждем загрузки графической оболочки CentOS 7.  
Переходим к установки Hadoop и IDE.

1.Скачиваем IntelliJ IDEA Comunity с веб-адреса:  
 [https://www.jetbrains.com/ru-ru/idea/download/download thanks.html?platform=linux&code=IIC](https://www.jetbrains.com/ru-ru/idea/download/download%20thanks.html?platform=linux&code=IIC)

Распаковываем файл idea-2020.1.tar.gz при помощи команды :

$tar –xzf idea-2020.1.tar.gz

Запускаем idea.sh из подкаталога bin, находясь в директории выгруженного архива.

2.Загрузить дистрибутив hadoop можно через веб-адрес или через консоль:

$wget <https://apache-mirror.rbc.ru/pub/apache/hadoop/common/hadoop-2.10.1/hadoop-2.10.1.tar.gz>

Распаковываем архив в /opt с помощью команды:

$tar -zxvf hadoop-2.10.0.tar.gz -C /opt

4.Отредактировать файл также можно двумя способами через терминал или графическую оболочку.

$sudo vi /opt/hadoop-2.10.0/etc/hadoop/hadoop-env.shзаменив export JAVA\_HOME=$на​ JAVA\_HOME=$(readlink -f /usr/bin/java | sed"s:bin/java::")

добавить исполняемые файлы hadoop в PATH

$echo "export PATH=/opt/hadoop-2.10.0/bin:$PATH" | sudotee -a /etc/profilesource /etc/profile

5.Проверить работоспособность hadoop

mkdir in cp /opt/hadoop-2.10.0/LICENSE.txt in

6.Установить jdk c помощью командной строки(необязательно, т.к openjdk устанавливается с пакетами GNOME):

$yum install java-1.8.0-openjdk-devel.x86\_64

7.Перевести hadoop в псевдо-распределенный режим

$sudo apt-get install ssh sudo apt-get install pdsh

Внести изменения в файл etc/hadoop/core-site.xml(через терминал или графический интерфейс):

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

Внести изменения в файл ​etc/hadoop/hdfs-site.xml(аналогично):

<configuration> <property> <name>dfs.replication</name> <value>1</value> </property></configuration>

Проверяем подключение SSH без passphrase:

$ssh localhost

Если система запросила пароль, выполняем следующие команды:

$ssh-keygen -t rsa -P '' -f ~/.ssh/id\_rsa

$cat ~/.ssh/id\_rsa.pub >> ~/.ssh/authorized\_keyschmod 0600 ~/.ssh/authorized\_keys

8.Форматируем файловую систему HDFS:

bin/hdfs namenode -format

9.Стартуем узел:

sbin/start-dfs.sh

а. HADOOP-демон ведёт запись в логи по пути ​$HADOOP\_LOG\_DIR​ (default:$HADOOP\_HOME/logs​)б. Запущенный узел доступен по адресу ​<http://localhost:9870/>

10.Создаем HDFS-директории для MapReduce job’ов:

$bin/hdfs dfs -mkdir /userbin/hdfs dfs -mkdir /user/<username>

11.Копируем файлы в HDFS

$bin/hdfs dfs -mkdir inputbin/hdfs dfs -put etc/hadoop/\*.xml input

12.Теперь можно запустить пример:

$bin/hadoop jarshare/hadoop/mapreduce/hadoop-mapreduce-examples-3.2.1.jar grep input output 'dfs[a-z.]+'

13.Проверяем выходные файлы: копируем из HDFS файлы из папки output себе в ФС.

$bin/hdfs dfs -get output outputcat output/\*или просматриваем их внутри HDFS:bin/hdfs dfs -cat output/\*

14.Отключение демона​ при необходимости​ выполняется следующей командой:

$sbin/stop-dfs.sh

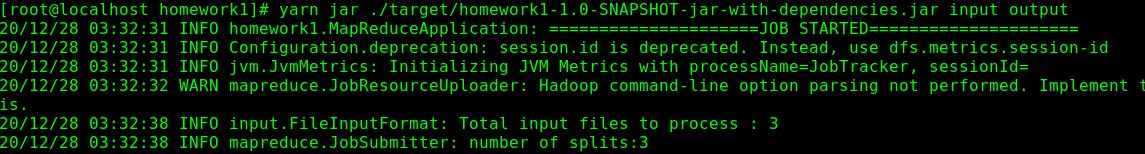
Запущенные сервисы можно просмотреть командой ​jps​.

После запуска HDFS и YARN WebUI будет доступен по ссылкам​ localhost:50070и ​localhost:8088.

Для запуска homework1-1.0-SNAPSHOT-jar-with-depencies.jar необходимо:  
1) $bin/hdfs dfs –rm –r output(удалить файлы, если имеются)

2) $bin/hdfs dfs –rm –r input(удалить файлы, если имеются)

3) $bin/hdfs dfs –put input input(необходимо быть в директории cd Desktop/homework1)



6. График зависимости времени выполнения от объема данных.

Для определения зависимости был в качестве примера использован подготовленный файл log.1, который содержал 200строк. С каждой итерацией количество строк уменьшалось вдвое.

