|  |  |
| --- | --- |
| **Gerb-BMSTU_01** | **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

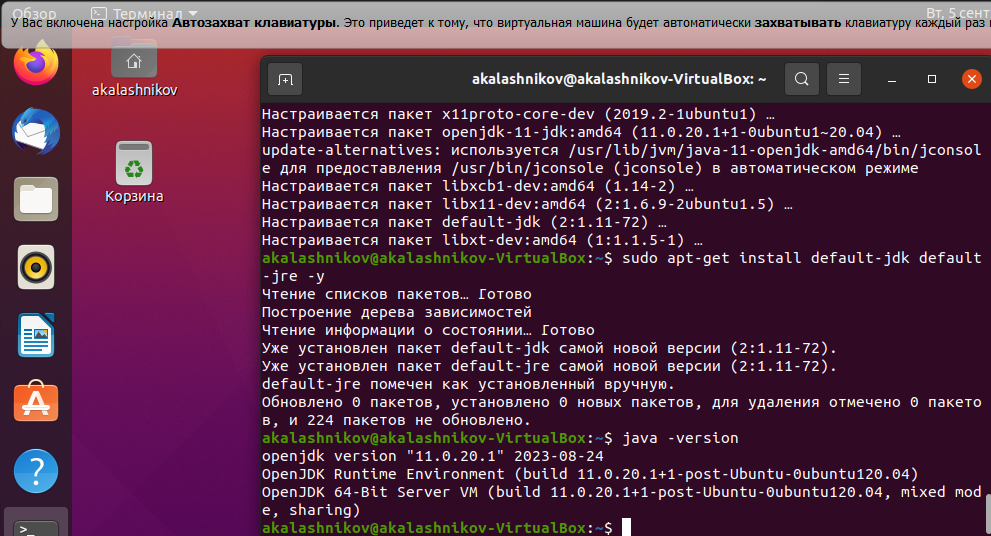


Рис. 1. Установка jdk

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).



Рис. 2. Проверка ssh

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

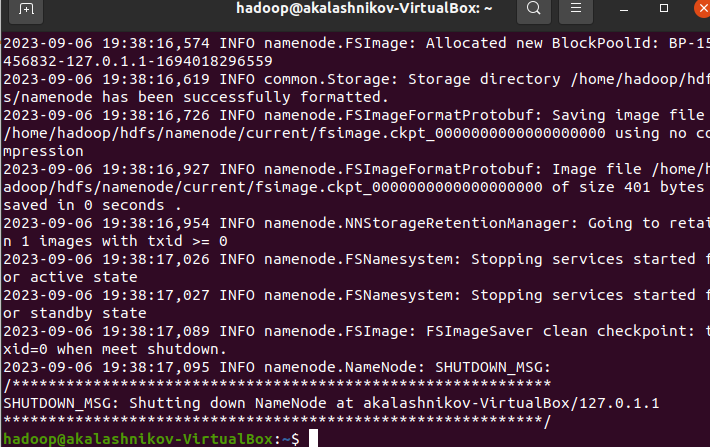


Рис. 3. Hadoop

#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/cor-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 хранится на файловой системе. Также в этом файле прописываются все узлы файловый 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 запустятся 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).

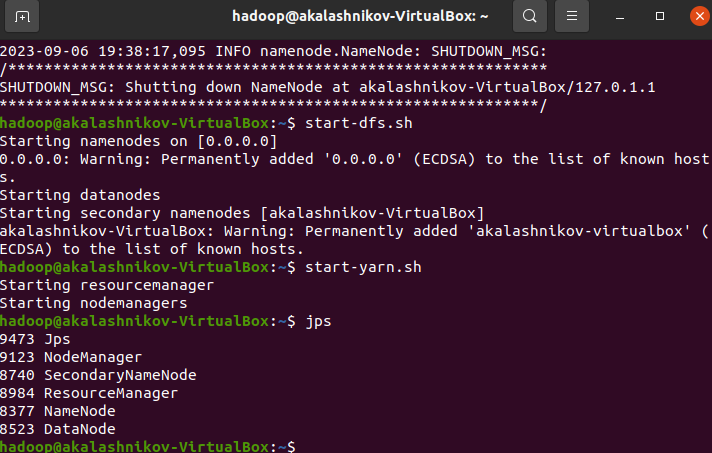


Рис. 4. Проверка работоспособности

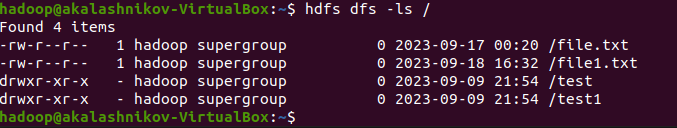


Рис. 5. Содержимое корневой директории

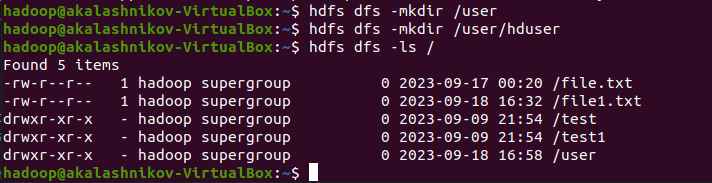


Рис. 6. Создание директории /user/hduser

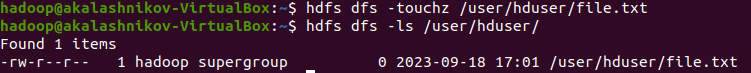


Рис. 7. Создание текстового файла

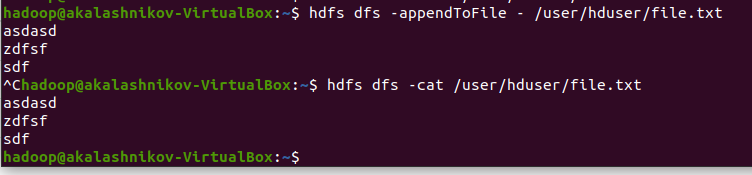


Рис. 8. Заполнение файла из консоли

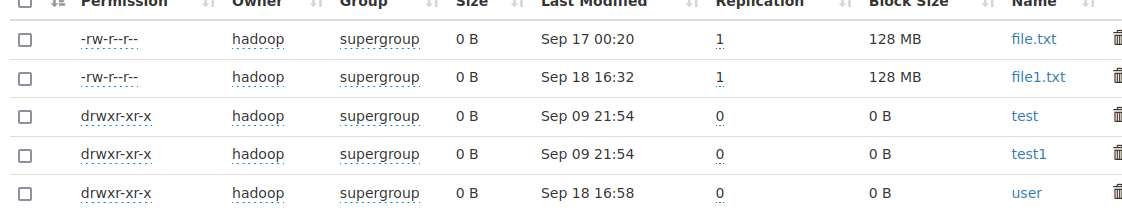


Рис. 9. Создание текстового файла

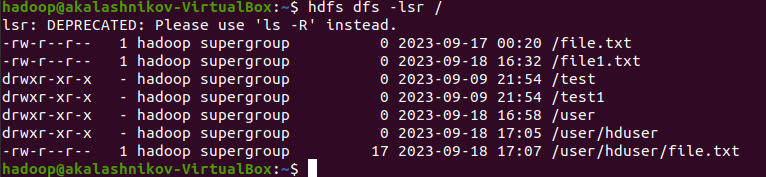


Рис. 10. Просмотр прав

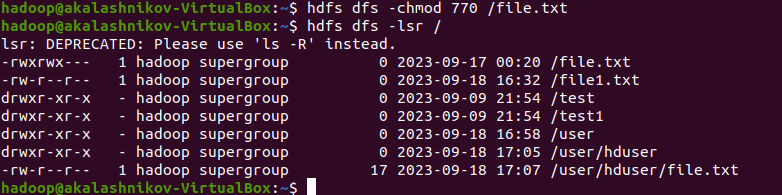


Рис. 11. Изменение прав

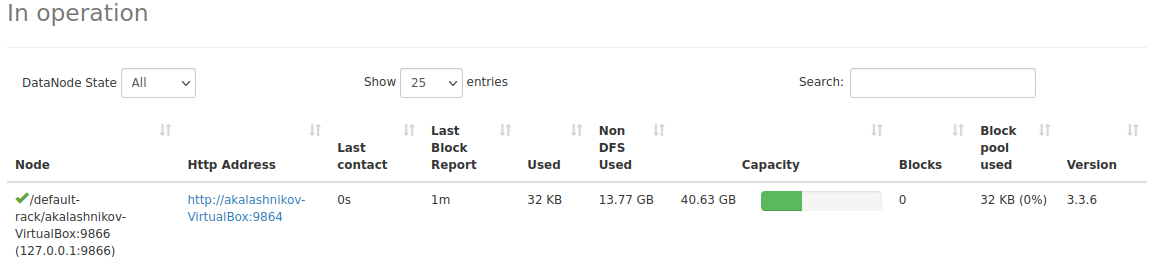


Рис. 12. Кластер из двух серверов

**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**

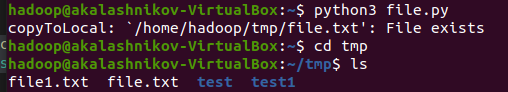
****

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.