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

**Отчет**

**по лабораторной работе № 7-8**

**Название:** Работа с Hadoop Spark

**Дисциплина:** Технология параллельных систем баз данных

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Студент | ИУ6-12М |  |  | Д.С. Каткова |
|  | (Группа) |  | (Подпись, дата) | (И.О. Фамилия) |
|  |  |  |  |  |
| Преподаватель |  |  |  | А.Д. Пономарев |
|  |  |  | (Подпись, дата) | (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).

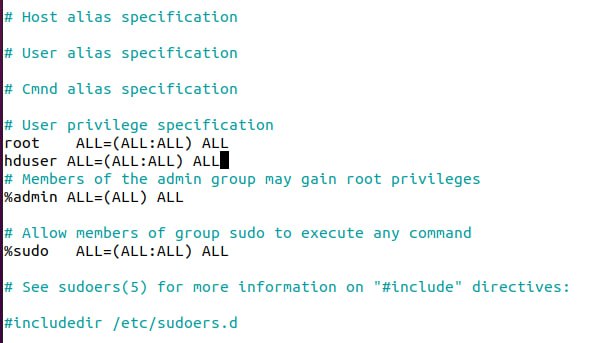
**Ход работы**

**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:

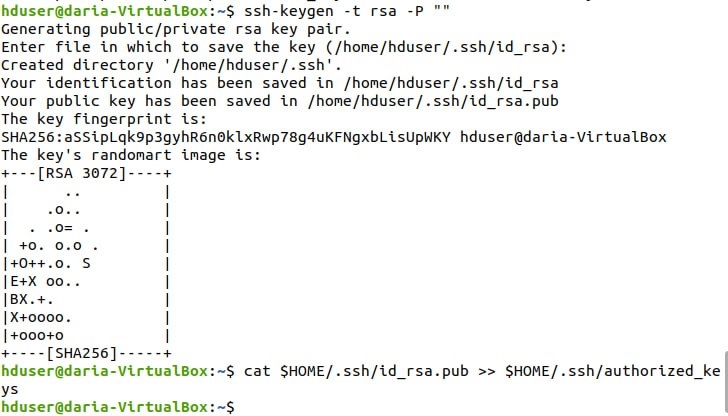
1. hadoop-2.9.2.tar.gz;
2. spark-2.4.6-bin-hadoop2.7.tgz;
3. gamlet\_en.txt.

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

****

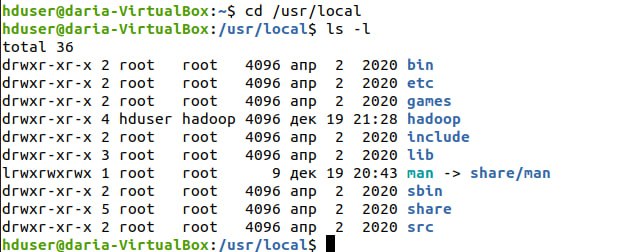
Далее устанавливаем и настраиваем SSH. SSH-сервер не установлен на ВМ, установим его командой sudo apt-get install openssh-server.

Сгенерируем SSH-ключ под hduser.



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

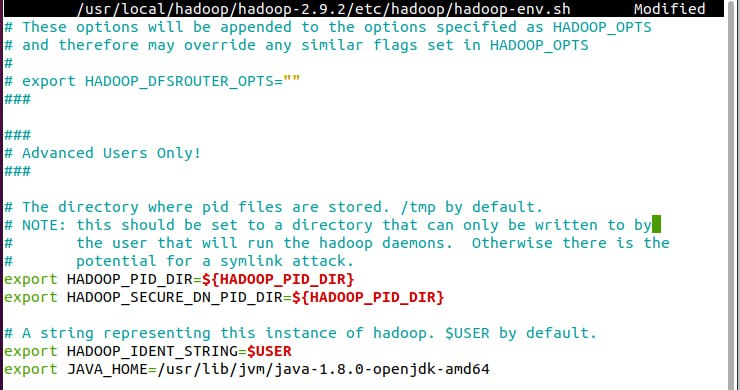
Затем создадим рабочие каталоги для HDFS (NameNode и DataNode) и назначим владельца каталога hadoop (пользователь hduser).



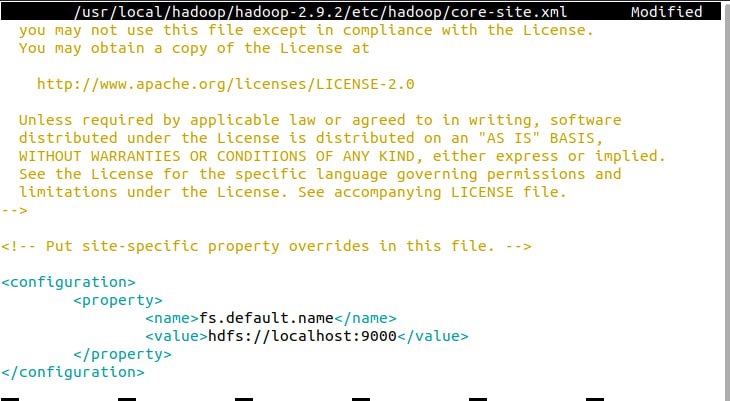
Проверим наличие /usr/lib/jvm/java-1.8.0-openjdk-amd64 :



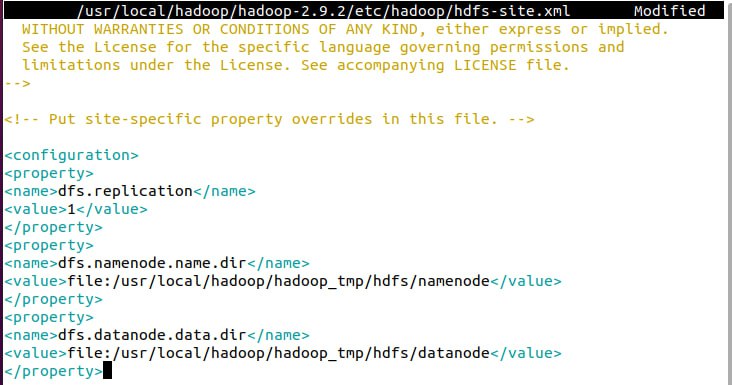
Настроим hadoop-env.sh



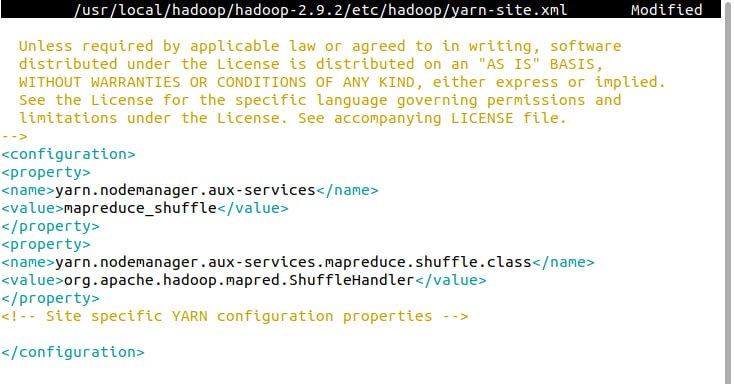
Настроим файла core-site.xml



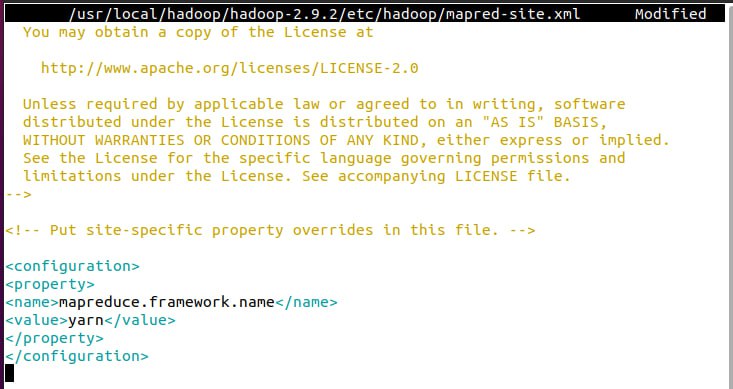
Настроим файла hdfs-site.xml



Настроим yarn-site.xml.

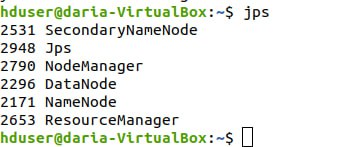


Настроим mapred-site.xml.



После перезагрузим систему с помощью команды sudo reboot 0.

Успешность запуска HADOOP проверим с помощью команды jps.

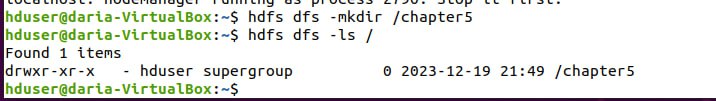
****

**3 Установка, настройка и проверка работы Spark**

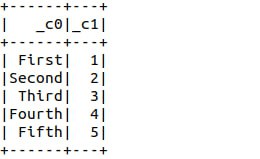
После выполнения всех команд, приведенных методических указания, получаем следующий вывод в консоль.



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

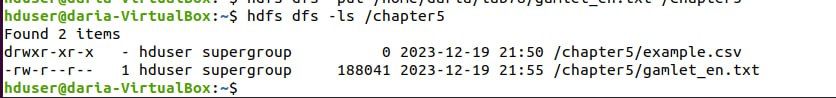


Запустим оболочку для работы с Python в Spark и введем операторы для проверки записи/чтения файла из HDFS.



**4 Подсчет числа слов в тексте**

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

****

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)

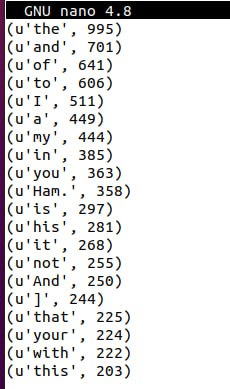
counts.saveAsTextFile("/home/hduser/res/")

sm = counts.map(lambda x: x[1]).sum()

print ("Summa %d" % (sm))

quit()

Результат работы программы представлен ниже.



**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».