Министерство науки и высшего образования Российской Федерации
Калужский филиал федерального государственного автономного 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

«ОСНОВЫ HADOOP. УСТАНОВКА HADOOP. ОСНОВНЫЕ КОМАНДЫ ФАЙЛОВОЙ СИСТЕМЫ HDFS»

по дисциплине: «Технологии обработки больших данных»

Выполнил: студент группы ИУК4-72Б _____ Моряков В.Ю.
(Подпись) (И.О. Фамилия)

Проверил: _____ Голубева С.Е.
(Подпись) (И.О. Фамилия)

Дата сдачи (защиты):

Результаты сдачи (защиты):

- Балльная оценка:
- Оценка:

Калуга, 2025

Цель: формирование практических навыков по установке и настройке кластера Hadoop и работе с файловой системой HDFS.

Задачи:

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.

Формулировка задания (17 вариант):

Напишите программу, которая будет сравнивать содержимое двух текстовых файлов в HDFS.

Ход выполнения:

```
WARN[0000] /home/hronoz/BMSTU_FINISH_LINE/docker-hadoop/lab1/docker-compose.yaml: the attribute `version` is obsolete, it will be ignored, please remove it to avoid
potential confusion
WARN[0000] The "HADOOP_HOME" variable is not set. Defaulting to a blank string.
WARN[0000] The "HADOOP_HOME" variable is not set. Defaulting to a blank string.
WARN[0000] The "HADOOP_HOME" variable is not set. Defaulting to a blank string.
WARN[0000] The "HADOOP_HOME" variable is not set. Defaulting to a blank string.
WARN[0000] The "HADOOP_HOME" variable is not set. Defaulting to a blank string.
WARN[0000] The "HADOOP_HOME" variable is not set. Defaulting to a blank string.
WARN[0000] The "HADOOP_HOME" variable is not set. Defaulting to a blank string.
WARN[0000] The "HADOOP_HOME" variable is not set. Defaulting to a blank string.
WARN[0000] The "HADOOP_HOME" variable is not set. Defaulting to a blank string.
WARN[0000] The "HADOOP_HOME" variable is not set. Defaulting to a blank string.
WARN[0000] The "HADOOP_HOME" variable is not set. Defaulting to a blank string.
WARN[0000] The "HADOOP_HOME" variable is not set. Defaulting to a blank string.
[+] Running 4/4
 ✓ Container lab1-nodemanager-1      Started                                                                              0.4s
 ✓ Container lab1-namenode-1         Started                                                                              0.5s
 ✓ Container lab1-datanode-1         Started                                                                              0.4s
 ✓ Container lab1-resourcemanager-1  Started                                                                              0.4s
CONTAINER ID   IMAGE             COMMAND                  CREATED                  STATUS                  PORTS                                        NAMES
a7d3c991d9c9   apache/hadoop:3   "/usr/local/bin/dumb…"   Less than a second ago   Up Less than a second   0.0.0.0:9870->9870/tcp, [::]:9870->9870/tcp   lab1-namenod
e-1
7bcc9fdf97db   apache/hadoop:3   "/usr/local/bin/dumb…"   Less than a second ago   Up Less than a second   0.0.0.0:8088->8088/tcp, [::]:8088->8088/tcp   lab1-resourc
emanager-1
a8f6792d7e53   apache/hadoop:3   "/usr/local/bin/dumb…"   Less than a second ago   Up Less than a second                                                lab1-datanod
e-1
3dac64c1d8b3   apache/hadoop:3   "/usr/local/bin/dumb…"   Less than a second ago   Up Less than a second                                                lab1-nodeman
ager-1
bash-4.2$ █
```

Рисунок 1 Установка hadoop

```
ager-1
bash-4.2$ bash /run_it_docker.sh
Creating HDFS directories...
```

Рисунок 2 Запуск скрипта

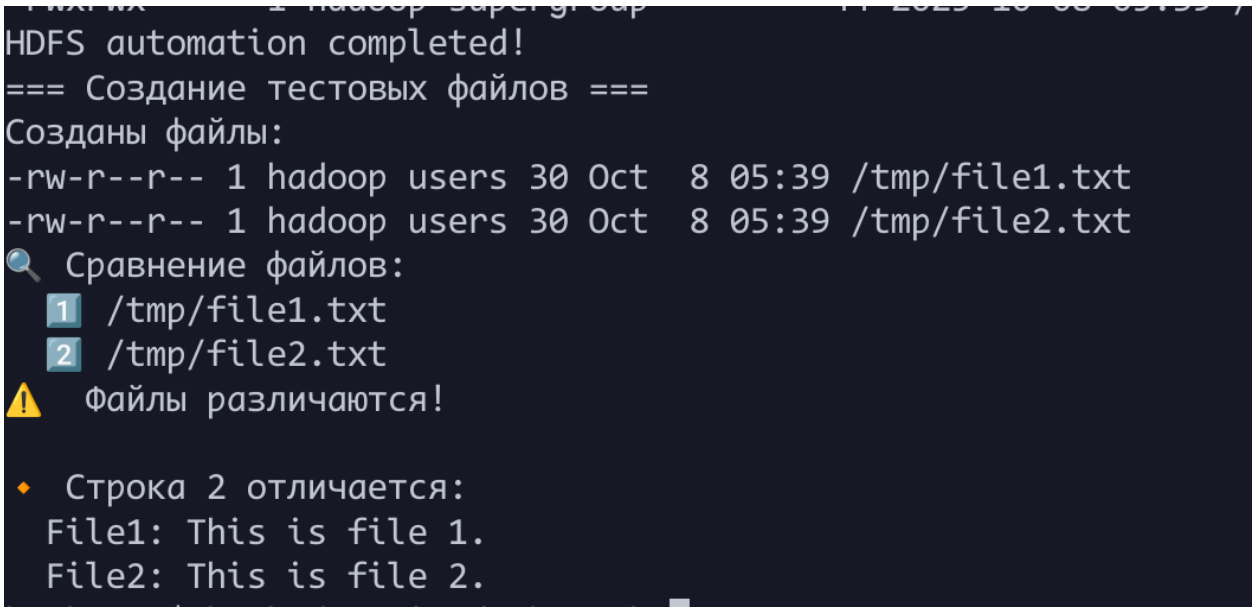

Рисунок 3 Результаты работы программы

Листинги программ:

compare_hdfs_files.py

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-

import subprocess
import sys

def read_hdfs_file(path):
    """Возвращает содержимое HDFS файла как список строк"""
    try:
        output = subprocess.check_output(
            ["hdfs", "dfs", "-cat", path],
            stderr=subprocess.STDOUT
        )
        # В Python 2 output - bytes, декодируем в utf-8
        if isinstance(output, bytes):
            output = output.decode('utf-8')
        return output.splitlines()
    except subprocess.CalledProcessError:
        print "❌ Ошибка: невозможно прочитать {}".format(path)
        sys.exit(1)

def main():
    if len(sys.argv) != 3:
        print "Использование: python compare_hdfs_files.py <HDFS_file1>
<HDFS_file2>"
        sys.exit(1)

    file1, file2 = sys.argv[1], sys.argv[2]
```

```
    print "🔍 Сравнение файлов:\n  1️⃣ {}\n  2️⃣ {}".format(file1,
file2)

    lines1 = read_hdfs_file(file1)
    lines2 = read_hdfs_file(file2)

    if lines1 == lines2:
        print "✅ Файлы идентичны"
    else:
        print "⚠️  Файлы различаются!\n"
        max_len = max(len(lines1), len(lines2))
        for i in range(max_len):
            line1 = lines1[i] if i < len(lines1) else "<no line>"
            line2 = lines2[i] if i < len(lines2) else "<no line>"
            if line1 != line2:
                print " 🔸 Строка {} отличается:\n  File1: {}\n  File2:
{}".format(i+1, line1, line2)

if __name__ == "__main__":
    main()
```

Вывод: в ходе лабораторной работы были получены практические навыки по работе с hadoop и python.