Федеральное государственное автономное

образовательное учреждение

высшего образования

«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

|  |
| --- |
| Институт космических и информационных технологий |
| институт |
| Программная инженерия |
| кафедра |

**ОТЧЕТ О ПРАКТИЧЕСКОЙ РАБОТЕ №3**

|  |
| --- |
| Инженерия машинного обучения |
| тема |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Преподаватель | |  |  |  | А.С. Михалев |
|  | |  | подпись, дата |  | инициалы, фамилия |
| Студент | КИ22-16/1б, 032210172 |  |  |  | М.В. Аверьянова |
|  | номер группы, зачётной книжки |  | подпись, дата |  | инициалы, фамилия |

Красноярск 2025

**ВВЕДЕНИЕ**

**Цель:**

Знакомство с инструментами для организации распределенной обработки данных, получение практических навыков по работе с Hadoop и подходом MapReduce.

**Задачи:**

Решение задачи представляет собой получение четырёх пользователей, выполнивших максимальное количество одинаковых запросов. Выполнение практической работы предполагает решение следующих задач.

1. Реализовать первый шаг обработки данных – подсчет количества одинаковых запросов от каждого пользователя. Для преобразования каждой строки реализовать mapper, для объединения результатов – reducer.

2. Реализовать второй шаг обработки данных – нахождение требуемых трех пользователей. Должно быть указано имя пользователя, тип запроса и количество. Для сортировки результатов предыдущего шага реализовать reducer.

3. Установить Apache Hadoop и запустить обработку файла в соответствии с вариантом задания.

4. В отчете привести результаты работы каждого шага.

1. **Ход работы**
   1. **Подготовка окружения и данных**
      1. Установка WSL2 на Windows для работы с Linux-окружением
      2. Установка Hadoop в WSL Ubuntu:

* Установка Java:

sudo apt install openjdk-8-jdk

* Скачивание и распаковка Hadoop:

Wget https://archive.apache.org/dist/hadoop/common/hadoop-3.3.1/hadoop-3.3.1.tar.gz

tar -xzf hadoop-3.3.1.tar.gz

mv hadoop-3.3.1 hadoop

* + 1. Настройка Hadoop (конфиги, SSH, переменные окружения)
    2. Запуск Hadoop-кластер:

hdfs namenode -format

start-dfs.sh

* + 1. Загрузка файла log\_3.txt в HDFS (Hadoop Distributed File System)

hdfs dfs -mkdir /input

hdfs dfs -put log\_3.txt /input/

* 1. **Реализация MapReduce**

Создаём python файлы. В папке Step1 Mapper читает данные и парсит каждую строку лог-файла для извлечения имени пользователя и типа запроса. Reduсer получает отсортированные данные от Mapper из папки Step1, суммирует количество одинаковых запросов для каждого пользователя, группирует данные по ключу (пользователь, запрос). В папке Step2 Mapper передаёт данные в Reducer папки Step2. Reducer из Step2 получает данные, сортирует все записи по количеству запросов по убыванию и выводит топ-4 записи с наибольшим количеством одинаковых запросов у пользователя. Код представлен на рисунках 1-4.

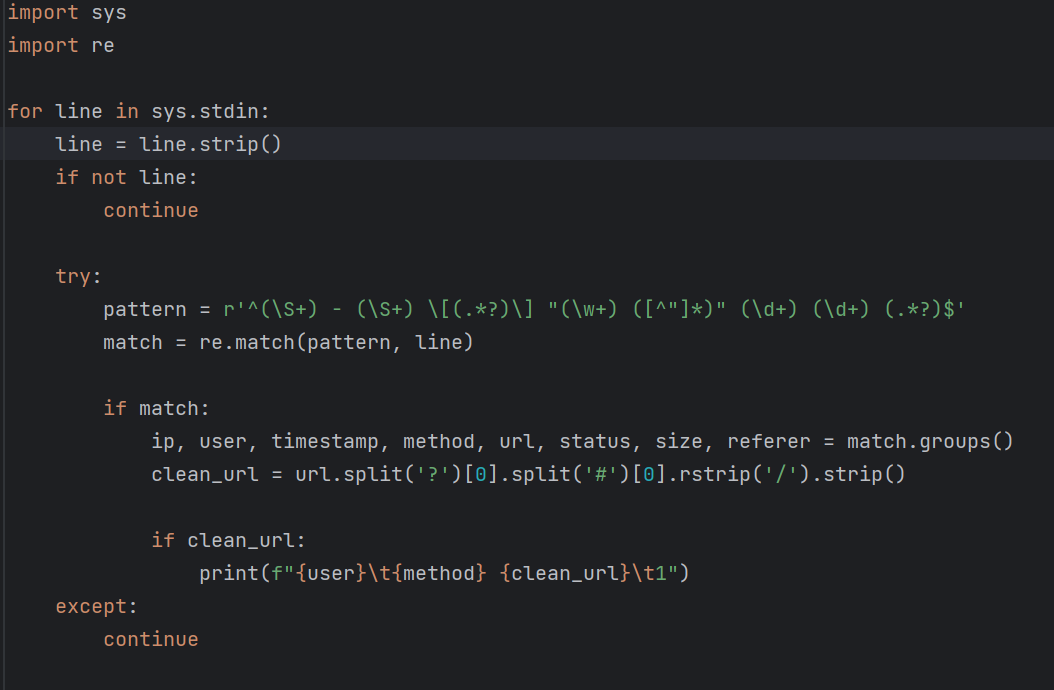


Рисунок 1 – Mapper Step1

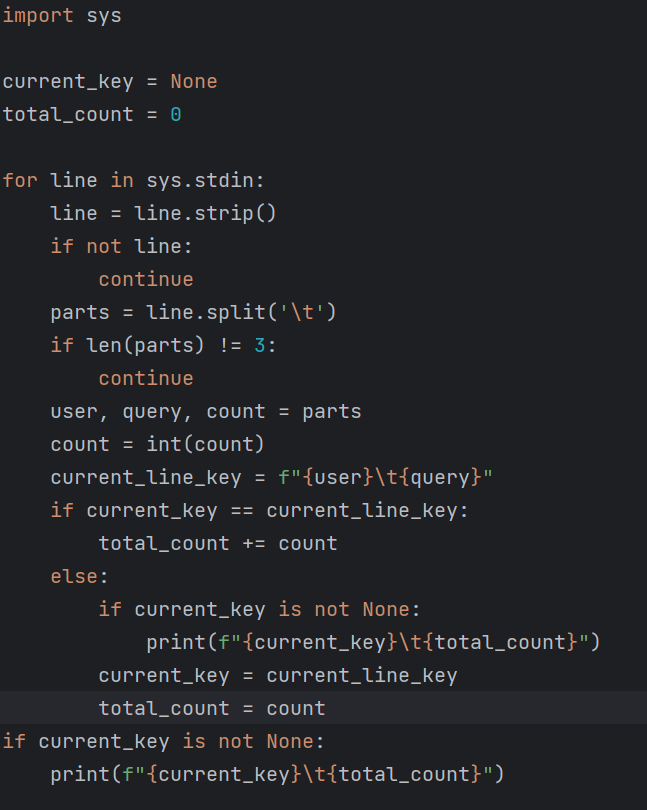


Рисунок 2 – Reducer Step1

Делаем скрипты исполняемыми для шага 1. Эта команда даёт разрешение запускать файлы как программы:

chmod +x ~/step1/mapper.py

chmod +x ~/step1/reducer.py

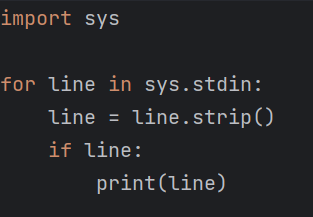


Рисунок 3 – Mapper Step2

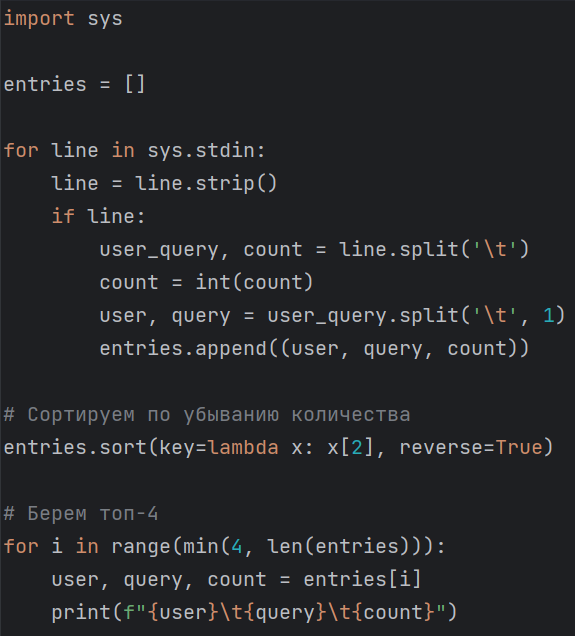


Рисунок 4 – Reducer Step2

Делаем скрипты исполняемыми для шага 2:

chmod +x ~/step2/mapper.py

chmod +x ~/step2/reducer.py

Запускаем первый и второй шаг. Команды, часть логов и результат выполнения первого шага представлены на рисунках 5-7.

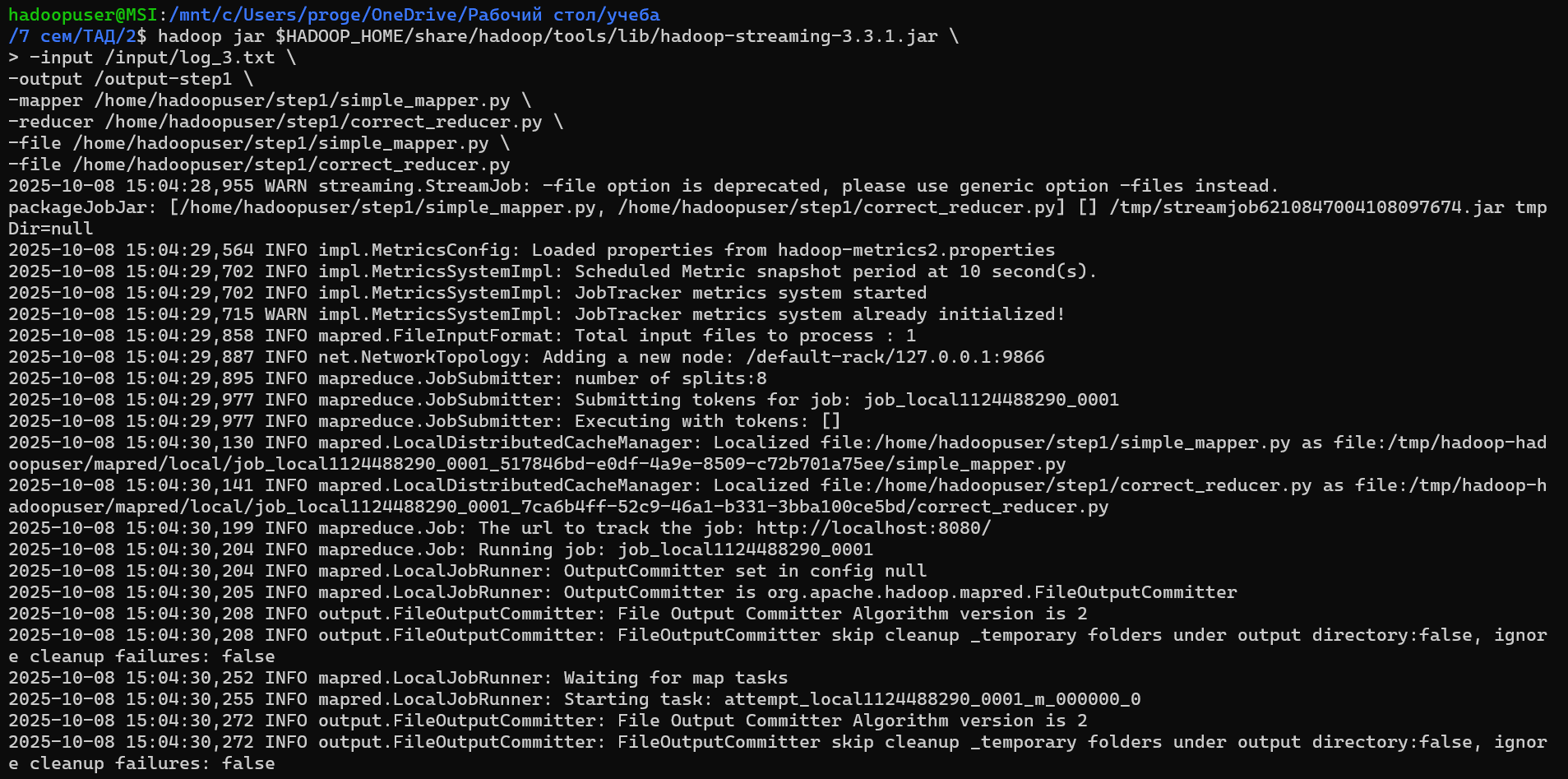


Рисунок 5 – Запуск шага 1

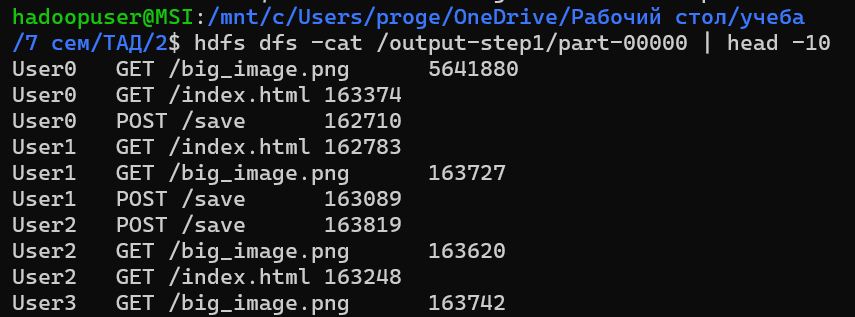


Рисунок 6 – Результат первого шага

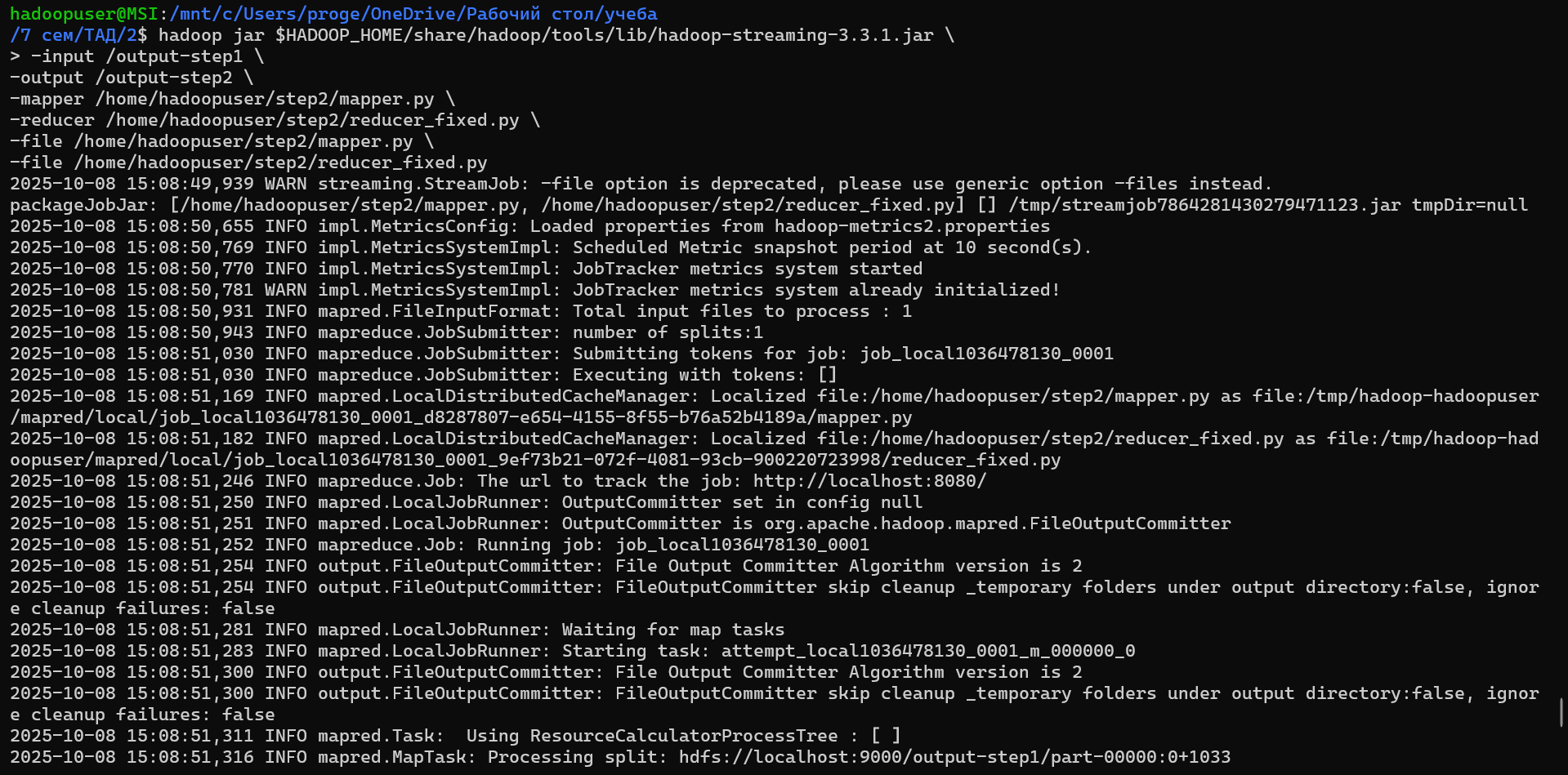


Рисунок 7 – Запуск шага 2

* 1. **Результат**

В результате проделанной работы, программа показала нам результат: четыре пользователя, выполнивших максимальное количество одинаковых запросов. Команда и результат представлены на рисунке 8.

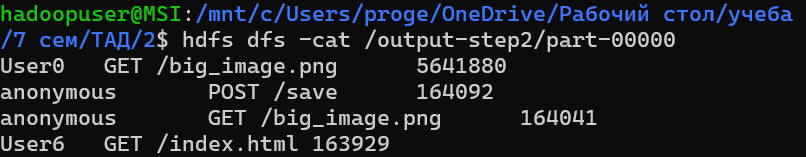


Рисунок 8 – Результат

То, что в результатах мы видим пользователей anonymous – это совершенно нормально. Такое может происходить если пользователь не авторизован, запросы от сервисов без пользовательской сессии и т.д. Их запросы также можно учитывать в статистике.

1. **Заключение:**

Были получены навыки работы с Hadoop и подходом MapReduse. Так же была написана программа, которая выдаёт четырёх пользователей, выполнивших максимальное количество одинаковых запросов.