|  |  |
| --- | --- |
|  | **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

**ДОМАШНЯЯ РАБОТА 2**

**«»**

**по дисциплине: «Технологии обработки больших данных»**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Выполнил: студент группы ИУК4-72Б | |  |  | Губин Е.В. | |
|  | | (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 реализации pig-скриптов для обработки

больших данных.

**Задачи:**

1. Получить навыки обработки больших данных, использую Pig Latin.

2. Изучить принцип работы Pig Latin.

3. Изучить синтаксис Pig Latin.

4. Уметь писать запросы, комбинируя несколько источников данных.

**Вариант №9**

**Задание 1:**

Выполнить задание из лабораторной работы №2, используя язык Pig Latin:

Построить индекс файла. Для каждого слова в файле результат должен

содержать номера всех строк, в которых появляется данное слово. Индекс

должен быть регистро-независимым. Результат должен быть сохранен в

файле в виде:

((word1 (1 42 58)), (word2 (34 55 776 3456), …).

**Задание 2:**

База данных твитов состоит из двух файлов:

Файл tweets.csv имеет формат: tweet\_id, tweet, login

Файл users.csv имеет формат: login, user\_name, state

Необходимо создать эти файлы и выполнить задание по варианту, используя Pig Latin:

Найти всех пользователей, написавших менее 3 твитов. Подсчитать общее количество твитов, написанных этими пользователями. Вычислить долю, которую составляют эти

твиты от общего количества твитов в базе.

**Ход выполнения работы:**

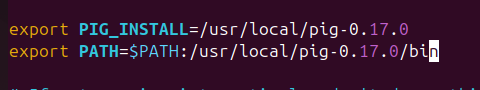


Рисунок Переменные стреды для Pig Latin

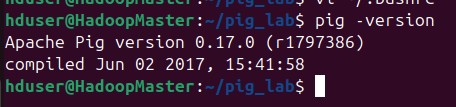


Рисунок Проверка выерсии pig'

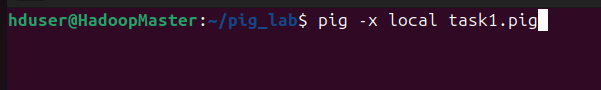


Рисунок Пример запуска скрипта



Рисунок Результат выполнения задания 1

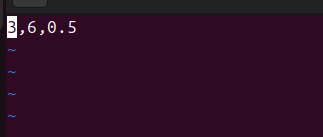


Рисунок Результат выполнения задания 2 (статистические результаты)

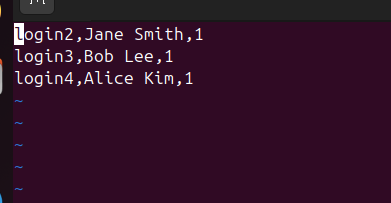


Рисунок Результат выполнения задания 2 (кто сделал меньше 3 твитов)

**Листинги программ:**

*task1.pig:*

STOPWORDS = LOAD 'input/stopwords.csv' USING PigStorage(',') AS (word:chararray);

BOOK = LOAD 'input/book.txt' USING PigStorage('\n') AS (line:chararray);

BOOK\_WITH\_NUM = RANK BOOK;

BOOK\_LOWER = FOREACH BOOK\_WITH\_NUM GENERATE

rank\_BOOK AS line\_num,

LOWER(line) AS line;

BOOK\_WORDS = FOREACH BOOK\_LOWER GENERATE

line\_num,

FLATTEN(TOKENIZE(line)) AS word;

CLEAN\_WORDS = FOREACH BOOK\_WORDS GENERATE

line\_num,

REPLACE(LOWER(word), '[^a-z]', '') AS word;

FILTERED\_WORDS = FILTER CLEAN\_WORDS BY word IS NOT NULL AND word != '';

STOPWORDS\_LOWER = FOREACH STOPWORDS GENERATE LOWER(word) AS word;

JOINED = JOIN FILTERED\_WORDS BY word, STOPWORDS\_LOWER BY word USING 'replicated';

GROUPED = GROUP JOINED BY FILTERED\_WORDS::word;

RESULT = FOREACH GROUPED GENERATE

group AS word,

JOINED.FILTERED\_WORDS::line\_num AS lines\_bag;

STORE RESULT INTO 'output/task1\_2025-10-20' USING PigStorage('\t');

*task2.pig:*

TWEETS = LOAD 'input/tweets.csv' USING PigStorage(',') AS (tweet\_id:int, tweet:chararray, login:chararray);

USERS = LOAD 'input/users.csv' USING PigStorage(',') AS (login:chararray, user\_name:chararray, state:chararray);

TWEET\_COUNT = GROUP TWEETS BY login;

USER\_TWEETS = FOREACH TWEET\_COUNT GENERATE

group AS login,

COUNT(TWEETS) AS tweet\_count;

FEW\_TWEETS = FILTER USER\_TWEETS BY tweet\_count < 3;

FEW\_USERS = JOIN FEW\_TWEETS BY login, USERS BY login;

FEW\_USERS\_FINAL = FOREACH FEW\_USERS GENERATE

FEW\_TWEETS::login AS login,

USERS::user\_name AS user\_name,

FEW\_TWEETS::tweet\_count AS tweet\_count;

FEW\_TWEETS\_SUM = FOREACH (GROUP FEW\_USERS\_FINAL ALL) GENERATE SUM(FEW\_USERS\_FINAL.tweet\_count) AS few\_total;

TOTAL\_TWEETS = FOREACH (GROUP TWEETS ALL) GENERATE COUNT(TWEETS) AS total;

JOINED = CROSS FEW\_TWEETS\_SUM, TOTAL\_TWEETS;

RESULT\_STATS = FOREACH JOINED GENERATE

few\_total,

total,

(double)few\_total / (double)total AS fraction;

STORE RESULT\_STATS INTO 'output/task2\_stats\_2025-10-20' USING PigStorage(',');

STORE FEW\_USERS\_FINAL INTO 'output/task2\_users\_2025-10-20' USING PigStorage(',');

**Вывод:** в ходе выполнения лабораторной работы были получены практические навыки по написанию программ на языке Pig Latin.