Міністерство освіти і науки України

Національний технічний університет України

“Київський політехнічний інститут імені Ігоря Сікорського”

Кафедра автоматизації проектування енергетичних процесів і систем

Звіт до практичного заняття №1

Тема «ПРОГРАМНА ПЛАТФОРМА APACHE HADOOP. ВИКОРИСТАННЯ ШАБЛОНІВ ПРОЕКТУВАННЯ MAPREDUCE. АЛГОРИТМ КЛАСТЕРИЗАЦІЇ»

з дисципліни «Оброблення надвеликих масивів даних»

Виконав:

студент 5-го курсу

кафедри АПЕПС ТЕФ

групи ТР-13мп

Шукюров Р. Р.

Викладач:

проф. Федорова Н.В.

Київ – 2021

**Мета роботи** - створення моделі кластеризації даних за допомогою технології MapReduce на основі програмної платформи Apache Hadoop.

**Завдання роботи:**

Постановка задачі кластеризації для всіх алгоритмів полягає в наступному:

Дано: X – простір об’єктів;

X l ={x i } l i=1 – вибірка елементів;

d: X × X -> [0; ∞) – функція відстані між об‘єктами.

Створити модель кластеризації даних за допомогою технології MapReduce.

Знайти: Y – множину кластерів і відображення а: Х → Y – алгоритм кластеризації такий, що кожен кластер складається з близьких між собою об‘єктів, а об‘єкти різних кластерів суттєво відрізняються.

**Теоретичні відомості:**

MapReduce — модель розподілених обчислювань у комп’ютерних кластерах, представлена компанією Google. Згідно з цією моделлю, додаток розділяється на значну кількість однакових елементарних завдань, що виконуються на вузлах кластера і потім, природнім шляхом зводяться у кінцевий результат.

Apache Hadoop — вільна програмна платформа для організації розподіленої обробки великих обсягів даних (що вимірюється у петабайтах) з використанням технології MapReduce, при якій завдання ділиться на багато дрібніших фрагментів, кожен з яких може бути запущений на окремому вузлі кластера.

До складу Hadoop входить реалізація розподіленої файлової системи Hadoop Distributed Filesystem (HDFS), яка автоматично забезпечує резервування даних і оптимізована для роботи MapReduce-задач у розподіленому середовищі. Для спрощення доступу до даних в сховищі Hadoop розроблена база даних HBase та SQL-подібна мова Hive, яка є свого роду SQL для MapReduce і запити якої можуть бути розпаралелені і оброблені кількома Hadoop-платформами

Кластеризація – задача групування множини об’єктів на підмножини(кластери) таким чином, щоб об’єкти із одного кластера були більш схожі одне на одного, ніж об’єкти із других кластерів по будь-якому критерію.

**Хід роботи**

Алгоритм кластеризації буде виконуватися з використанням технології MapReduce на локальному кластері Hadoop, розгорнутому за допомогою Docker`a. Мова програмування Python.

**Крок 1.** Архітектура базується на основі master-layer. В даному випадку необхідно створити конфігурацію 1 master 3 layer.

Для цього потрібно створити файл docker-compose.yml (рисунок 1) з налаштуванням всіх компонентів.

Рисунок 1. Файл docker-compose.yml

**Крок 2.** Запустимо цей файл через консоль за допомогою команди

*docker-compose up -d* (рисунок 2)



Рисунок 2. Запуск кластерів Hadoop

**Крок 3.**Реалізація алгоритму кластеризації на мові програмування Python -*claster.py*. На вхід подається двох вимірний масив точок X, на виході одновимірний масив з номерами кластерів Y. І алгоритм генерації вхідних даних *generate.py*.

**Крок 4.**Копіювання локального *claster.py* і *generate.py* в *master* (рисунок 4)



Рисунок 3. Копіювання скриптів

**Крок 5.**Вхід в контейнер *master* за допомогою команди *docker exec -it master bash* (рисунок 4)



Рисунок 4. Контейнер *master*

**Крок 6.**Створення вхідних даних за допомогою скрипту

*generate.py* (рисунок 5)

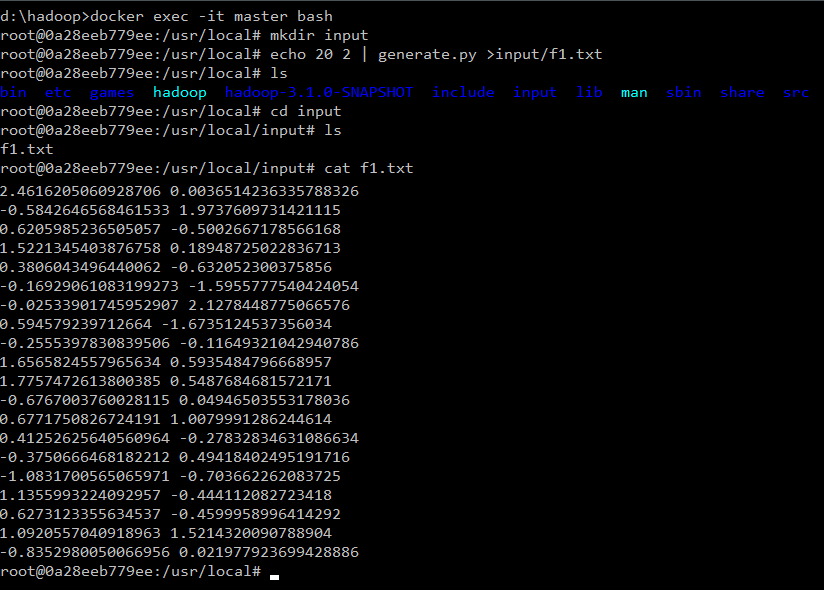


Рисунок 5. Створення вхідних даних

**Крок 7.**Програма MapReduce отримує доступ до файлів із розподіленої файлової системи Hadoop (HDFS). Перенесення вхідного каталогу і файлів в HDFS (рисунок 6).

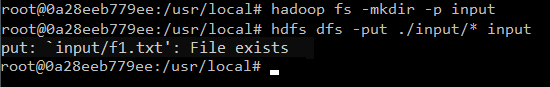


Рисунок 6. Перенесення каталогу

**Крок 8.**Запуск програми і вихідний результат (рисунок 7).

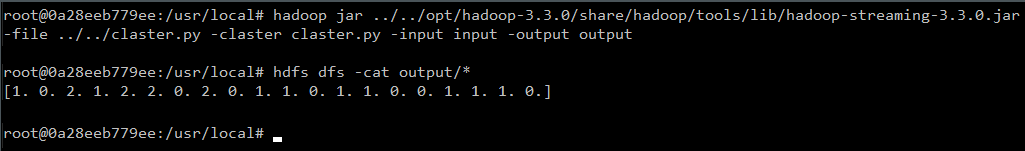


Рисунок 7. Вихідний результат

**Лістинг програми:**

#!/usr/bin/env python

"""claster.py"""

import sys

import numpy as np

num\_cluster = 3

iterations = 3

x = []

for line in sys.stdin:

line = line.strip()

x.append([float(i) for i in line.split()])

x = np.array(x)

N = len(x)

y = np.zeros(N)

for t in range(iterations):

if t == 0:

index\_ = np.random.choice(range(N),num\_cluster,replace=False)

mean = x[index\_]

else:

for k in range(num\_cluster):

mean[k] = np.mean(x[y==k], axis=0)

for i in range(N):

dist = np.sum((mean - x[i])\*\*2, axis=1)

pred = np.argmin(dist)

y[i] = pred

print(y)

#!/usr/bin/env python

"""generate.py"""

import sys

import numpy as np

for line in sys.stdin:

line = line.strip()

n, m = [int(i) for i in line.split()]

x = np.random.randn(n, m)

for xi in x:

print(xi[0], xi[1])

**Висновок:** Під час виконання лабораторної роботи я ознайомився з поняттями BigData, Hadoop, MapReduce. Встановив програмний засіб Apache Hadoop та написав програму на мові програмування Python для кластеризації даних з використанням технології MapReduce.