**Chmura Obliczeniowa i Technologie Big Data**

*Moduł „PyPork” realizujący mechanizmy logiki rozmytej dla platformy Apache Pig*

|  |  |
| --- | --- |
| Kierownik | dr hab. inż. Dariusz Mrozek, Prof. PŚ |
| Autorzy | Jan Skowronek  Zuzanna Lempa  Paweł Kudzia  Wojciech Wątroba |
| Kierunek | Informatyka SSM |
| Specjalizacja | IDSI |
| Technologia | Apache Pig |

Spis treści

[1. Wstęp 3](#_Toc77010074)

[2.1. Cel projektu 3](#_Toc77010075)

[2.2. Założenia projektu 3](#_Toc77010076)

[2. Stos technologiczny 3](#_Toc77010077)

[2.1. Przygotowanie środowiska 4](#_Toc77010078)

[2.2. Przykład użycia biblioteki „PyPork” 6](#_Toc77010079)

[2.3. Architektura biblioteki „PyPork” 7](#_Toc77010080)

[3. Interfejs Pig Latin 8](#_Toc77010081)

[4. Interfejs Python 8](#_Toc77010082)

[4.1. Rozmyte operatory logiczne 9](#_Toc77010083)

[4.1.1. Funkcja F\_OR 9](#_Toc77010084)

[4.1.2. Funkcja F\_AND 9](#_Toc77010085)

[4.1.3. Funkcja F\_NOT 9](#_Toc77010086)

[4.2. Funkcje przynależności 10](#_Toc77010087)

[4.2.1. Funkcja triangle\_membership 10](#_Toc77010088)

[4.2.2. Funkcja trapezoid\_membership 10](#_Toc77010089)

[4.2.3. Funkcja l\_class\_membership 11](#_Toc77010090)

[4.2.4. Funkcja y\_class\_membership 11](#_Toc77010091)

[4.2.5. Funkcja gaussian\_membership 12](#_Toc77010092)

[5. Podsumowanie 12](#_Toc77010093)

# 1. Wstęp

Big Data to termin odnoszący się do dużych i różnorodnych zbiorów danych. Przetwarzanie tego typu danych jest zadaniem wyjątkowo trudnym i wymagającym, ale z drugiej strony analiza ogromnych zbiorów okazuje się być wartościowa, ponieważ pozwala zdobyć nowy rodzaj wiedzy. Z tego też powodu na przestrzeni lat powstały narzędzia i rozbudowane platformy programistyczne umożliwiające m.in. efektywne gromadzenie, przechowywanie i przetwarzanie dużych ilości informacji. Warto zaznaczyć, że dane mogą pochodzić z wielu różnych źródeł, co niesie za sobą konieczność użycia specjalistycznego oprogramowania, za pomocą którego można usprawnić proces przetwarzania. Korzyści z używania narzędzi Big Data są niezaprzeczalne, dlatego istotne jest, aby istniała możliwość rozszerzania o nowe funkcjonalności i tak już zaawansowanych platform.

W ramach kursu *„Chmura Obliczeniowa i Technologie Big Data”* należy zrealizować własny moduł rozszerzający dla jednej z przykładowych platform programistycznych stosowanych w dziedzinie Big Data.

## 2.1. Cel projektu

Celem projektu jest wzbogacenie przykładowej platformy programistycznej o moduł, który umożliwi realizację rozmytego przetwarzania danych. Moduł ten należy opracować dla technologii Apache Pig będącej częścią ekosystemu Hadoop.

## 2.2. Założenia projektu

Nowy moduł dla Apache Pig powinien dostarczać podstawowe funkcjonalności rozmytego przetwarzania danych, do których należy zaliczyć: rozmyte filtrowanie z wykorzystaniem zmiennych lingwistycznych, grupowanie rozmyte względem zmiennej lingwistycznej, czy złączenia. Bibliotekę należy zaimplementować wykorzystując do tego UDF (*ang. User-defined function*), czyli funkcje rozszerzające dany program lub środowisko.

Ogromną zaletą ekosystemu Hadoop jest to, iż funkcje mogą zostać zaimplementowane w wybranych językach programowania, wspieranych przez aplikację Apache Pig. Jednym z nich jest Python i to właśnie język ten posłużył do wykonania projektu.

# 2. Stos technologiczny

W niniejszym rozdziale przedstawiony został stos technologiczny, który posłużył do realizacji projektu. Opisane zostały zagadnienia związane z przygotowaniem środowiska testowego, jak i również architektura autorskiej biblioteki *„PyPork”*.

## 2.1. Przygotowanie środowiska

W celu rozpoczęcia pracy z platformą Apache Pig należy przygotować środowisko, które zazwyczaj przyjmuje postać jakiegoś systemu operacyjnego z rodziny Linux. Z reguły wiele narzędzi programistycznych jest tworzonych właśnie pod ten system. Nie inaczej jest w przypadku technologii Hadoop i Apache Pig, w związku z czym pierwszym krokiem do rozpoczęcia pracy było skonfigurowanie przykładowej maszyny wirtualnej. Autorzy projektu zdecydowali, że użyją jedną z popularnych dystrybucji Linuksa, którą niewątpliwie jest Ubuntu. Instalacja systemu jest prosta i tak właściwie sprowadza się do postępowania według wskazówek wyświetlanych na ekranie. W przypadku, gdy zdecydowano się wybrać wersję desktop Ubuntu, podczas instalacji systemu warto skorzystać z opcji *„minimal”*, dzięki której finalnie zainstalowane zostaną tylko niezbędne pakiety i narzędzia.

Po prawidłowym wdrożeniu systemu na maszynie wirtualnej można przejść do kolejnego kroku, który dotyczy instalacji Hadoop i Apache Pig. Aby zminimalizować mogące się pojawić problemy i przy okazji uprościć ręczną konfigurację, autorzy projektu utworzyli przykładowy skrypt powłoki Bash. Skrypt ten realizuje następujące zadania:

* Instalacja środowiska uruchomieniowego Java,
* Pobranie i rozpakowanie paczek Hadoop i Apache Pig,
* Przeniesienie plików Hadoop i Apache Pig do odpowiednich katalogów,
* Utworzenie zmiennych środowiskowych,
* Sprawdzenie poprawności wykonanej instalacji za pomocą polecenia wypisującego informacje o zainstalowanej wersji danego oprogramowania.

Na listingu 3.1.1. zaprezentowany został opisywany skrypt instalacyjny. Należy pamiętać o tym, aby uruchomić skrypt z uprawnieniami administratora systemu.

Listing 2.1.1. Skrypt instalacyjny

#!/bin/bash

sudo apt install openjdk-8-jdk

sudo apt install openjdk-8-jre

java -version

# main directory where Hadoop and Pig will be installed

MAIN\_DIRECTORY=/usr/local

cd $MAIN\_DIRECTORY

# download hadoop

sudo wget http://apache.claz.org/hadoop/common/hadoop-2.10.1/hadoop-2.10.1.tar.gz

sudo tar xzf hadoop-2.10.1.tar.gz

sudo mkdir hadoop

sudo mv hadoop-2.10.1/\* hadoop

sudo rm -r hadoop-2.10.1

# download pig

sudo wget http://apache.claz.org/pig/pig-0.17.0/pig-0.17.0.tar.gz

sudo tar xzf pig-0.17.0.tar.gz

sudo mkdir pig

sudo mv pig-0.17.0/\* pig

sudo rm -r pig-0.17.0

# home directories for Hadoop and Pig

MY\_HADOOP\_HOME=$MAIN\_DIRECTORY/hadoop

MY\_PIG\_HOME=$MAIN\_DIRECTORY/pig

# create environmental variables

echo export HADOOP\_HOME=$MY\_HADOOP\_HOME >> ~/.bashrc

echo export HADOOP\_INSTALL=$MY\_HADOOP\_HOME >> ~/.bashrc

echo export HADOOP\_MAPRED\_HOME=$MY\_HADOOP\_HOME >> ~/.bashrc

echo export HADOOP\_COMMON\_HOME=$MY\_HADOOP\_HOME >> ~/.bashrc

echo export HADOOP\_HDFS\_HOME=$MY\_HADOOP\_HOME >> ~/.bashrc

echo export HADOOP\_COMMON\_LIB\_NATIVE\_DIR=$MY\_HADOOP\_HOME/lib/native >> ~/.bashrc

echo export HADOOP\_OPTS="-Djava.library.path=$MY\_HADOOP\_HOME/lib/nativ" >> ~/.bashrc

echo export YARN\_HOME=$MY\_HADOOP\_HOME >> ~/.bashrc

# if you installed java 7 change name here

echo export JAVA\_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64 >> ~/.bashrc

# add Hadoop and Pig to PATH

HADOOP\_PATH=$MY\_HADOOP\_HOME/sbin:$MY\_HADOOP\_HOME/bin

PIG\_PATH=$MY\_PIG\_HOME/bin

echo export PATH=$PATH:$HADOOP\_PATH:$PIG\_PATH >> ~/.bashrc

source ~/.bashrc

hadoop version

pig -version

echo "Script completed."

Następnie można zweryfikować poprawności instalacji Apache Pig. Najprostszy sposób polega na uruchomieniu aplikacji w trybie lokalnym korzystając z komendy przedstawionej na listingu 2.1.2.

pig -x local

Listing 2.1.2. Uruchomienie Apache Pig w trybie lokalnym

Jeśli instalacja odbyła się w sposób prawidłowy, aplikacja powinna uruchomić się w trybie interaktywnym, w którym możliwe jest wpisywanie poleceń języka Pig Latin. Aby zapoznać się z podstawowymi funkcjami można np. wczytać plik CSV i wyświetlić jego zawartość.

## 2.2. Przykład użycia biblioteki „PyPork”

W celu sprawdzenia poprawności opisywanej biblioteki należy skorzystać z dostarczonego pliku skryptowego w języku Pig Latin, który umieszczony został w katalogu pig\_examples. W przykładzie pokazano użycie operacji rozmytych na danych pochodzących z pliku CSV.

Ponadto należy pamiętać również o tym, aby na początku własnych skryptów zarejestrować główny plik biblioteki main\_fuzzy.py. Plik ten musi zostać umieszczony w miejscu, do którego jest łatwy dostęp, jak np. główny katalog użytkownika lub jego pulpit.

Rejestracja pliku biblioteki odbywa się w pliku o rozszerzeniu .pig za pomocą operatora REGISTER w podobny sposób, jak prezentuje to listing 2.2.1. Ponadto należy pamiętać o tym, aby w miejscu, w którym znajduje się skrypt Pig Latin umieszczony został również plik konfiguracyjny pattern.json oraz odpowiedni plik CSV przechowujące dane używane w skrypcie użytkownika.

Listing 2.2.1. Wskazanie lokalizacji skryptu main\_fuzzy.py

REGISTER '/home/piggy/PyPork/main\_fuzzy.py' using jython as fuzzy

Uruchomienie skryptu dla języka Pig Latin odbywa się z wykorzystaniem komendy z listingu 2.2.2. W oknie konsoli systemowej powinny zostać wypisane zarówno logi generowane przez narzędzie Hadoop oraz wyniki poszczególnych operacji zdefiniowanych w skrypcie języka Pig Latin.

Listing 2.2.2. Uruchomienie skryptu linguistic.pig

pig -x local linguistic.pig

## 2.3. Architektura biblioteki „PyPork”

Projekt składa się z kilku plików skryptowych języka Python, w których zdefiniowane zostały funkcje odpowiedzialne za realizację określonych zadań.

W katalogu pig\_examples umieszczono przykłady użycia rozszerzenia *„PyPork”*. Można tutaj znaleźć:

* linguistic.pig – zawiera kilka przykładów pokazujących działanie autorskich funkcji,
* user\_data.csv – przechowuje analizowane dane,
* pattern.json – plik konfiguracyjny, w którym zdefiniowane są zmienne lingwistyczne.

W głównym katalogu znajduje się skrypt main\_fuzzy.py zawierający wszystkie potrzebne funkcje do prawidłowego działania biblioteki. W pliku tym dostępne są m.in. funkcje operujące na wartościach lingwistycznych, definicje funkcji przynależności (np. trójkątna, trapezowa), czy funkcje realizujące działanie rozmytych operatorów AND, OR i NOT. Każda z funkcji została udokumentowana bezpośrednio w kodzie źródłowym (standardowe komentarze dokumentacji), dzięki czemu podczas pracy z kodem możliwe jest uzyskanie informacji pomocniczych bezpośrednio w edytorze kodu. Skrypt ten jest wymagany do pracy z Apache Pig, co zresztą zaprezentowano w poprzednim podrozdziale.

Ponadto główny katalog repozytorium zawiera kilka skryptów, które weryfikują poprawność działania poszczególnych funkcjonalności:

* fequal\_test.py – sprawdza działanie funkcji fequal,
* linguistic\_test.py – zawiera testy sprawdzające obsługę wartości rozmytych,
* membership\_test.py – przechowuje przykładowe testy, które weryfikują poprawność zaimplementowanych funkcji przynależności,
* operators\_test.py – sprawdza działanie rozmytych operatorów AND, OR i NOT.

Powyższe skrypty należy uruchomić z poziomu konsoli systemowej w podobnym sposób, jak pokazano na listingu 2.3.1.

Listing 2.3.1. Uruchamianie skryptu Python z poziomu konsoli systemowej

python operators\_test.py

Wynik uruchomienia pliku operators\_test.py zaprezentowany został na listingu 2.3.2.

Listing 2.3.2. Wynik działania testów ze skryptu operators\_test.py

Tests: F\_OR, F\_AND, F\_NOT operators

========== test 1 ==========

values[0]: 1

values[1]: 0.5

result: {'F\_OR': 1, 'F\_AND': 0.5, 'F\_NOT': {'values[0]': 0, 'values[1]': 0.5}}

========== test 2 ==========

values[0]: 0.75

values[1]: 0.1

result: {'F\_OR': 0.75, 'F\_AND': 0.1, 'F\_NOT': {'values[0]': 0.25, 'values[1]': 0.9}}

========== test 3 ==========

values[0]: 0.1

values[1]: 0.4

result: {'F\_OR': 0.4, 'F\_AND': 0.1, 'F\_NOT': {'values[0]': 0.9, 'values[1]': 0.6}}

========== test 4 ==========

values[0]: 0.6

values[1]: 1

result: {'F\_OR': 1, 'F\_AND': 0.6, 'F\_NOT': {'values[0]': 0.4, 'values[1]': 0}}

========== test 5 ==========

values[0]: 0.2

values[1]: 0.3

result: {'F\_OR': 0.3, 'F\_AND': 0.2, 'F\_NOT': {'values[0]': 0.8, 'values[1]': 0.7}}

Poniżej w sposób graficzny zaprezentowane zostało ułożenie modułu.

=obrazek=

Rysunek 2.3.1. Schemat architektury biblioteki „PyPork”

# 3. Interfejs Pig Latin

Rozdział zawiera opis interfejsu dostępnego w języku Pig Latin z wykorzystaniem kilku przykładów. Do każdego scenariusza użycia dołączone zostały również otrzymane wyniki.

# 4. Interfejs Python

Niniejszy rozdział dotyczy dokumentacji technicznej interfejsu dostępnego w języku Python. Opisane zostały wszystkie funkcje, z których korzysta biblioteka, jak i również klasy utworzone w celu przeprowadzenia testów poszczególnych funkcjonalności.

## 4.1. Rozmyte operatory logiczne

Dostępnych jest pięć rozmytych operatorów logicznych.

### 4.1.1. Funkcja F\_OR

Funkcja realizuje działanie rozmytego operatora OR.

Wejście: Wartości numeryczne, które mają zostać przetestowane. Każda wartość musi być z zakresu od 0 do 1 (włącznie).

Wyjście: Maksymalna wartość z podanych na wejściu wartości.

Przykład użycia w Python: F\_OR(1, 0.5) lub F\_OR(\*[1, 0.5]).

Listing 4.1.1.1. Funkcja F\_OR

def F\_OR(\*values):

    return max(values)

### 4.1.2. Funkcja F\_AND

Funkcja realizuje działanie rozmytego operatora AND.

Wejście: Wartości numeryczne, które mają zostać przetestowane. Każda wartość musi być z zakresu od 0 do 1 (włącznie).

Wyjście: Minimalna wartość z podanych na wejściu wartości.

Przykład użycia w Python: F\_AND(1, 0.5) lub F\_AND(\*[1, 0.5]).

Listing 4.1.2.1. Funkcja F\_AND

def F\_AND(\*values):

    return min(values)

### 4.1.3. Funkcja F\_NOT

Funkcja realizuje działanie rozmytego operatora NOT.

Wejście: Wartość numeryczna, która ma zostać przetestowana. Wartość musi być z zakresu od 0 do 1 (włącznie).

Wyjście: Wynik wykonania operacji rozmytego operatora NOT.

Przykład użycia w Python: F\_NOT(1).

Listing 4.1.3.1. Funkcja F\_NOT

def F\_NOT(value):

    return 1 - value

## 4.2. Funkcje przynależności

Dostępnych jest pięć funkcji przynależności.

### 4.2.1. Funkcja triangle\_membership

Funkcja realizuje działanie trójkątnej funkcji przynależności.

Wejście: Wartość numeryczna x oraz parametry funkcji: a, b, c.

Wyjście: Wartość numeryczna z zakresu od 0 do 1 (włącznie).

Przykład użycia w Python: triangle\_membership(4, 5, 10, 12).

Listing 4.2.1.1. Funkcja triangle\_membership

def triangle\_membership(x, a, b, c):

    if x <= a:

        return 0

    elif x <= b:

        return (x - a) / (b - a)

    elif x <= c:

        return (c - x) / (c - b)

    else:

        return 0

### 4.2.2. Funkcja trapezoid\_membership

Funkcja realizuje działanie trapezowej funkcji przynależności.

Wejście: Wartość numeryczna x oraz parametry funkcji: a, b, c, d.

Wyjście: Wartość numeryczna z zakresu od 0 do 1 (włącznie).

Przykład użycia w Python: trapezoid\_membership(0.5, 1, 11, 14, 15).

Listing 4.2.2.1. Funkcja trapezoid\_membership

def trapezoid\_membership(x, a, b, c, d):

    if x <= a:

        return 0

    elif x <= b:

        return (x - a) / (b - a)

    elif x <= c:

        return 1

    elif x <= d:

        return (d - x) / (d - c)

    else:

        return 0

### 4.2.3. Funkcja l\_class\_membership

Funkcja realizuje działanie funkcji przynależności klasy L.

Wejście: Wartość numeryczna x oraz parametry funkcji: a, b.

Wyjście: Wartość numeryczna z zakresu od 0 do 1 (włącznie).

Przykład użycia w Python: l\_class\_membership(5, 24, 30).

Listing 4.2.3.1. Funkcja l\_class\_membership

def l\_class\_membership(x, a, b):

    if x <= a:

        return 1

    elif x <= b:

        return (b - x) / (b - a)

    else:

        return 0

### 4.2.4. Funkcja y\_class\_membership

Funkcja realizuje działanie funkcji przynależności klasy Y.

Wejście: Wartość numeryczna x oraz parametry funkcji: a, b.

Wyjście: Wartość numeryczna z zakresu od 0 do 1 (włącznie).

Przykład użycia w Python: y\_class\_membership(5, 24, 30).

Listing 4.2.4.1. Funkcja y\_class\_membership

def y\_class\_membership(x, a, b):

    if x <= a:

        return 0

    elif x <= b:

        return (x - a) / (b - a)

    else:

        return 1

### 4.2.5. Funkcja gaussian\_membership

Funkcja realizuje działanie gaussowskiej funkcji przynależności.

Wejście: Wartość numeryczna x oraz parametry funkcji: mean, sd.

Wyjście: Wartość numeryczna z zakresu od 0 do 1 (włącznie).

Przykład użycia w Python: gaussian\_membership(1, 2, 5).

Listing 4.2.5.1. Funkcja gaussian\_membership

def gaussian\_membership(x, mean, sd):

    return exp(((x - mean) \*\* 2) / 2 \* (sd \*\* 2))

# 5. Podsumowanie

Wykonanie projektu pt. *„PyPork”* pozwoliło autorom zapoznać się z platformą Apache Pig, która znajduje zastosowanie w analizie i przetwarzaniu dużych zbiorów danych. Platforma ta udostępnia proceduralny język Pig Latin, dzięki któremu możliwe jest pisanie zarówno bardzo prostych, jak i bardziej zaawansowanych skryptów umożliwiających przetwarzanie danych. Pig Latin oprócz udostępniania typowych mechanizmów charakterystycznych dla języków programowania tj. różne typy danych (tuple, bag, map), czy funkcje (np. LOAD, STORE), daje również możliwość definiowania własnych funkcji, co pozwala wprowadzić dodatkowe operacje w potoku przetwarzania. Opcja rozbudowy istniejących rozwiązań niewątpliwie jest jedną z kluczowych zalet platform Hadoop i Apache Pig, ponieważ dzięki temu standardowy zestaw narzędzi (który swoją drogą i tak jest obszerny) może być rozbudowany. W kontekście projektu *„PyPork”* Apache Pig wzbogacono o wybrane elementy logiki rozmytej, która jak się okazuje może być niezwykle przydatna w ciekawej dziedzinie informatyki jaką jest Big Data.

Zaletą technologii Apache Pig jest przede wszystkim to, iż do definiowanie UDF możliwe jest z wykorzystaniem kilku języków programowania. Implementacja biblioteki w języku Python okazała się bardzo wygodna.

Autorzy modułu są zadowoleni z uzyskanych wyników i osiągniętego celu, choć nie da się ukryć, że aktualne rozwiązanie mogłoby zostać bardziej usprawnione. Jeśli dodatkowym celem byłoby udostępnienie biblioteki *„PyPork”* szerzej publiczności, wtedy koniecznym krokiem jest usprawnienie repozytorium w serwisie GitHub. Ponadto przydałaby się rozbudowana i jednocześnie czytelna dokumentacja biblioteki, aby każda zainteresowana osoba mogła z tej biblioteki skorzystać. Tego typu zmiana z pewnością podniosłaby ogólną jakość projektu, a sama popularyzacja tym samym mogłaby przyczynić się do powstania społeczności, której członkowie rozwijaliby istniejące rozwiązanie.

Podsumowując, realizacja projektu *„PyPork”* poruszającego dosyć niestandardową tematykę w ramach kursu *„Chmura Obliczeniowa i Technologie Big Data”*, zdecydowanie poszerzyła wiedzę dotyczącą technik pracy z językiem Python, zbiorów rozmytych i co najważniejsze specjalistycznych narzędzi służących do analizy, gromadzenia i przetwarzania dużych zbiorów danych. Nie da się ukryć, że wiedza o istnieniu narzędzi tj. Hadoop, czy Apache Pig jest niezwykle przydatna podczas projektowania komercyjnych produktów informatycznych, które główny nacisk kładą m.in. na sprawne przekazywanie i analizowanie dużych zbiorów informacji.