



POZNAN UNIVERSITY OF TECHNOLOGY

Tematy projektów Edycja 2020

Robert Wrembel
Poznan University of Technology
Institute of Computing Science
Robert.Wrembel@cs.put.poznan.pl
www.cs.put.poznan.pl/rwrembel



Temat 1

- **Stopień wykorzystania danych na platformie TeraData**
- **Klient: Santander**
- **Opis**
 - Dostarczenie danych do monitorowania używalności danych składowanych na platformie TeraData w oparciu o dane audytowe gromadzone po stronie silnika bazy
 - Udostępnienie szczegółowych danych statystycznych dla wybranych tabel (zakres tych danych zostanie podany w późniejszym etapie, przy uszczegóławianiu założeń)
 - Wyniki udostępniane w postaci raportów umożliwiających zidentyfikowanie tych danych, które przestały być używane, danych, które są bardzo mocno wykorzystywane, ustalenia częstotliwości używania obszarów danych



Temat 1 (cd.)

⇒ Opis (cd)

- Dedykowane oprogramowanie oparte o SQL, które będzie przygotowywało dane do dalszego prezentowania w formie raportów, wraz z przykładem takiego raportu (narzędzie dowolne, np. Tableau, Excel)

⇒ Opiekun: Łukasz Gientka (Santander)

⇒ L. osób: 4



Temat 2

⇒ Opracowanie modelowego środowiska ETL na potrzeby szkoleń dla użytkowników biznesowych

⇒ Klient: Santander

⇒ Opis

- opracowanie schematu modelowej bazy danych zawierającej tabele faktów i wymiarów wraz z metadanymi
- opracowanie zestawu danych imitujących wsad z systemów źródłowych wraz z metadanymi
- zaimplementowanie operacji typu ETL/ELT ładujących i transformujących dane
- zaimplementowanie prototypów zapytań analitycznych na modelowym środowisku

⇒ Opiekun: Agnieszka Kruś i Kamila Ziemia (Santander)

⇒ L. osób: 4



Temat 3

- ⇒ **Przedstawienie modelu logicznego baz danych wchodzących w skład korporacyjnej hurtowni danych**
- ⇒ **Klient: Santander**
- ⇒ **Opis**
 - korporacyjna hurtownia danych (KHD) składa się z wielu baz danych: baza podstawowa - model główny, dedykowane data marty, bazy MIS
 - brak fizycznych powiązań między tabelami (ograniczenia typu klucz podstawowy - klucz obcy) uniemożliwia automatyczne wygenerowanie modelu logicznego baz danych
 - celem projektu jest
 - przygotowanie automatu identyfikującego ograniczenia integralnościowe i/lub indeksy na podstawie obrazu tabel
 - przygotowanie skryptu DDL tworzących tabele z minimalną liczbą pól (indeksy, ogr. int.) w dedykowanej bazie danych



Temat 3 (cd.)

- ⇒ **Oczekiwane wyniki:**
 - lista ograniczeń integralnościowych i indeksów z obrazów tabel z aplikacji przechowującej zmiany do modelu KHD
 - przygotowanie DDL tworzącego tabele w bazie TeraData
 - prezentacja graficzna (dla dowolnie sformułowanej listy tabel powstaje Mapa zawierająca tabele z listy i związki między nimi)
- ⇒ **Opiekun: Włodzimierz Hoffmann i Jacek Majcherczak (Santander)**
- ⇒ **L. osób: 4**



Tematy PKO BP

- Każdy z tematów może być kontynuowany w ramach pracy mgr
- Możliwość odbycia płatnego stażu w banku (min. 27h/tydz)
- Praca w rzeczywistym projekcie



Temat 4

- **Hurtownia danych dla analizy ofert bankowych**
- **Klient: PKO BP**
- **Opis**
 - Na podstawie danych dot. ofertowych produktu pozyskiwanych w sposób zautomatyzowanych ze stron dostawców konkurujących podmiotów ustalenie:
 - parametrów ofert wskazanej listy konkurujących podmiotów (lista jest predefiniowana)
 - określenie pozycji oferty przedsiębiorstwa w odniesieniu do konkurencji
 - wyznaczenie miary obrazującej atrakcyjność ofert poszczególnych podmiotów, w tym własnej
- **Rodzaje ofert**
 - depozytowa
 - kredytowa (pożyczka, hipoteka)



Temat 4 (cd.)

⇒ Oczekiwane rezultaty:

- algorytm do automatycznego pozyskiwania informacji o ofercie konkurujących podmiotów
- algorytm wyznaczających miarę atrakcyjności i pozycjonowania oferty w stosunku do konkurencji
- baza danych gromadząca oferty
- odświeżanie ofert co 24h
- aplikacje przeszukiwania i porównywania ofert + wizualizacja wyników
- aplikacje analizy zmian w ofertach w zadany czas + wizualizacja wyników
- graficzna interpretacja atrakcyjności oferty

⇒ Opiekun: Mariusz Sienkiewicz (PKO BP)

⇒ L. osób: 5



Temat 5

⇒ Data lake: wzbogacanie danych o podmiocie gospodarczym ze źródeł otwartych (zewnętrznych)

⇒ Klient: PKO BP

⇒ Opis

- W oparciu o dane publikowane w publicznych źródłach cyfrowych, pozyskiwanie danych dotyczących wskazanych podmiotów:
 - identyfikacyjnych, kontaktowych, głównego profilu
 - działalności/branże, głównych kontrahentów, obroty/przychody (o ile dostępne), liczba zatrudnionych
 - na podstawie zgromadzonych danych, ustalenie aktualnych danych dot. podmiotu
- wykrycie utraty aktualności danych w data lake



Temat 5 (cd.)

⇒ Wyniki

- bot skanujący zasoby cyfrowe
- algorytm identyfikowania danych aktualnych
- aplikacja przeszukiwania data lake
- aplikacja analizująca zmiany aktualności danych + wizualizacja

⇒ Opiekun: Mariusz Sienkiewicz (PKO BP)

⇒ L. osób: 5



Temat 6

⇒ Monitor czasów wykonywania operacji push-down (ETL)

⇒ Klient: IBM Polska

⇒ Opis

- pomiar czasów wykonywania push-down selekcji, agregacji, połączenia
- współpraca z DB2, Oracle, SQLServer
- wykorzystanie w źródle
 - dedykowanych struktur danych
 - przetwarzania równoległego
- implementacja: C++, Python, Java, Scala, Groovy

⇒ Opiekun: Michał Bodziony (IBM Polska)

⇒ L. osób: 4



Temat 7

- ➔ **Głębokie sieci neuronowe w rozpoznawaniu tekstu niejednorodnego**
- ➔ **Klient: Syndigo**
- ➔ **Opis**
 - **Opracowanie architektury sieci neuronowej zdolnej do czytania trudnego tekstu znajdującego się na opakowaniach produktów spożywczych. Głównym wyzwaniem jest niejednorodna struktura tekstu:**
 - niektóre fragmenty tekstu posiadają formę swobodą, niektóre są umieszczone w tabeli, czasem łamana jest perspektywa tekstu (np. tekst pisany na półokręgu), fragmenty opakowań zawierają też tekst "malowany" (tj. niestandardowe czcionki graficzne)
 - opakowania spożywcze zawierają bardzo dużo dodatkowych znaków graficznych, które utrudniają przeprowadzenie procesu OCR
 - tekst jest pisany różnymi kolorami na wielokolorowym tle



Temat 7 (cd.)

- ➔ **Zadanie: przygotowanie przepływu, na który składają się:**
 - wstępna obróbka zdjęć (transformacja do white&black, rotacje zdjęcia, wyostanie krawędzi)
 - wybór systemu OCR (Google Tesseract, OCR EAST, Firebase OCR) i przetestowanie wybranego systemu na dostarczonych packshotach
 - dopasowanie parametrów systemu OCR do specyfiki opakowań spożywczych (rozpoznawanie bounding boxes)
 - uruchomienie czytania tekstu z analizowanych opakowań
- ➔ **Opiekun: prof. Mikołaj Morzy**
- ➔ **L. osób: 4**



Temat 7 (cd.)

- **Uwaga: projekt ma charakter badania wstępnego do dużego projektu badawczo-wdrożeniowego, który będzie realizowany przez Politechnikę Poznańską i Syndigo. Dobra realizacja postawionego zadania może wiązać się z możliwością zatrudnienia na 1.5 roku, i pogłębiania wiedzy na temat praktycznych zastosowań technik uczenia maszynowego do rozwiązywania rzeczywistych problemów**



Temat 8

- **Optymalizacja ETL z UDF: opening a black box**
- **Klient: IBM Polska**
- **Opis**
 - **zapoznanie się z technikami oceny funkcjonalności black box**
 - **opracowanie zestawu testów umożliwiających poznanie własności UDF**
 - **wykonanie eksperymentów**
- **Opiekun: Michał Bodziony (IBM Polska)**
- **L. osób: 2**