

Tematy projektów HDiPA semestr letni 2021/2022

Zamawiający: PKO BP (Warszawa)

Repozytorium taryfikatorów opłat bankowych

Zadania:

- zaprojektowanie systemu gromadzenia i przeszukiwania taryfikatorów opłat bankowych dla wielu banków
- funkcjonalność
 - pobieranie taryfikatorów różnych banków
 - parsowanie taryfikatora
 - przekształcenie danych taryfikatora do postaci tabelarycznej i składowanie w bazie danych
 - przeszukiwanie taryfikatora po słowach kluczowych

Technologie:

- Google Cloud Platform

Zamawiający: PKO BP (Warszawa)

Cel zadania: Zebranie i porównanie informacji o jednostkach samorządu terytorialnego (JST) z różnych źródeł internetowych oraz identyfikacja rozbieżności.

Opis:

W ramach zadania należy pobrać i przeanalizować bazę teleadresową JST, która jest udostępniana przez MSWiA (<http://administracja.mswia.gov.pl/adm/baza-jst/843,Samorzad-terytorialny-w-Polsce.html>). Następnie wszystkie dane (pola) zawarte w tej bazie należy zestawzić z informacjami dostępnymi na stronach internetowych każdej JST. Dane adresowe/geograficzne należy porównać z bazą TERYT oraz z danymi Poczty Polskiej (baza PNA). Zestawiane dane powinny być dołączone do bazy publikowanej przez MSWiA, dodatkowo rekordy, w których zidentyfikowano niezgodność danych należy odpowiednio oznaczyć w zakresie zgodności ze stroną internetową, bazą TERYT i PNA. Uwaga! Przed dokonaniem porównania należy zadbać o wystandaryzowanie zapisu informacji w poszczególnych polach.

Zamawiający: Szpital CM HCP

Parser dokumentów PDF i składnica danych medycznych

Zadania:

- zaprojektowanie schematu bazy danych do przechowania danych z dokumentów PDF, dokumentów skanowanych i obrazów medycznych
- zaimplementowanie (lub zaadoptowanie istniejącego) parsera dokumentów PDF
- translacja dokumentu PDF do postaci tabelarycznej
- zaimplementowanie interfejsu użytkownika obsługi parsowania, składowania i przeszukiwania dokumentów
- zaimplementowanie mechanizmu wgrywania skanowanych plików i obrazów medycznych z płyt
- wyszukiwanie danych i prezentowanie na osi czasu

Technologie:

- baza danych: PostgreSQL
- interfejs użytkownika: najlepiej Python Django
- backend: najlepiej Python + SQL
- Docker

Rodzaj dokumentów PDF: badania laboratoryjne krwi (format plików produkowany przez ALAB) – morfologia, biochemia, badanie moczu

Zamawiający: Szpital CM HCP

Projekt i implementacja internetowego konta pacjenta

Zadania:

- zaprojektowanie schematu bazy danych do przechowania podstawowych danych medycznych pacjentów
- funkcjonalność
 - dostęp do dokumentacji technicznej obrazowej i laboratoryjnej na osi czasu
 - wgrywanie wyników badań
 - weryfikacja wgranych badań przez lekarza
 - rejestracja na badania
 - umówienie się na wizytę
 - powiadomienia o badaniach

Technologie:

- baza danych: PostgreSQL
- interfejs użytkownika: najlepiej Python Django
- backend: najlepiej Python + SQL
- Docker

Zamawiający: IBM Software Lab (Kraków)

Apache Arrow Flight SQL service with ODBC

The goal of the project is to develop and assess an adapter service which externalizes Apache Arrow Flight SQL interfaces for data sources via ODBC interface.

Expected tasks:

- prior art search
- construct a prototype of the server with ApacheArrowFlightSQL interfaces
- means to configure ODBC connections
- transform row format into batch of columnar format
- performance assessment for light operational queries
- performance assessment for heavy unload/scanning queries

ref:

<https://arrow.apache.org/blog/2022/02/16/introducing-arrow-flight-sql/>

Zamawiający: IBM Software Lab (Kraków)

Integration library for Airbyte connectors and Apache Arrow format

The goal of the project is to develop and assess an integration layer between Airbyte connectors and Apache Arrow format

Expected tasks:

- prior art search
- construct a prototype of libraries which transform from Airbyte proprietary protocol to Arrow (Flight)
- performance assessment for transformation overhead for loading data
- performance assessment for transformation overhead for unloading data

ref:

<https://arrow.apache.org/blog/2022/02/16/introducing-arrow-flight-sql/>

<https://airbyte.com/>

Both projects means to be published at github as open source.

Zamawiający: Santander (Poznań)

Data Vault w BigData

Zadania:

- utworzenie nowego modelu danych zgodnego z Data Vault (najlepiej DV 2.0)
- jako źródło posłużyć może model credit_card składający się z 22 tabel dotyczących wniosków
- do ustalenia zakres danych testowych (ddl-ki i przykładowe jasony z danymi)
- modelowanie można zrealizować zarówno w chmurze jak i na on-prem
- do ustalenia: środowisko w Snowflake, Azure i dane i model w chmurze
- stworzymy zarówno model raw Data Vaulta jak i model biznesowy (albo sam widok Business Vault albo pełny Information Mart)
- zbudowanie warstwy do raportów, które można podpiąć jako źródło do dowolnego narzędzia raportowego
- dodatkowo można rozpoznać Wherescape do wspomagania modelowania (opcja)

Zamawiający: Santander (Poznań)

Wizualizacja danych streamingowych - porównanie narzędzi

Opis problemu:

Mamy dostęp do coraz większej liczby źródeł danych, które są aktualizowane 'na bieżąco'. Aby zapewnić użytkownikom biznesowym jak najszybszy dostęp do aktualnych danych chcemy przygotować dashboardy, które będą się odświeżać możliwie często. Potrzebujemy analizy dot. możliwych dostępnych rozwiązań i rekomendacji dot. najlepszego rozwiązania.

Narzędzia:

- PowerBI
- 2 inne narzędzia do wizualizacji zaproponowane przez studentów (np. Python, R, narzędzie 'z pudełka')

Zadania:

- przygotowanie źródła danych streamowanych (dowolne, ważne, żeby się odświeżały co jakiś czas)
- przygotowanie wizualizacji danych ze źródła - wizualizacja powinna się odświeżać tak, żeby prezentować możliwie najnowsze dane
- przygotowanie podsumowania:
 - opis rozwiązania technicznego
 - wady i zalety każdego z rozwiązań
 - rekomendacja
 - w szczególności zależy mi analizie następujących cech:
 - performance
 - koszty/licencji
 - nakład pracy i kompetencje: przygotowanie pierwszego dashboardu oraz późniejsze utrzymanie/modyfikacje/rozwój
 - dostępne wizualizacje w ramach narzędzia