Projekt - konspekt

Adam Frej, Jan Gąska

Listopad 2022

Spis treści

1 Cel projektu i potencjalne korzyści z wdrożenia **1**

2 Dane **1**
2.1 Dane dotyczące przestępczości 2
2.2 Dane mieszkaniowe 2

3 Stos architektoniczny **2**
3.1 Składowanie 2
3.2 Przepływ danych 3
3.3 Warstwa analityczna 3
3.4 Pozostałe 3

4 Podział pracy **3**

1 Cel projektu i potencjalne korzyści z wdrożenia

Celem projektu jest stworzenie systemu do analizowania przestępczości w Nowym Jorku w kontekście zamożności obszarów, w których do niej dochodzi. Projekt na podstawie pojedynczych zgłoszeń policyjnych dotyczących wykroczeń i przestępstw w całym mieście na przestrzeni lat oraz raportów finansowych wspólnot mieszkaniowych i spółdzielni będzie analizował korelacje i zależności między danymi ze względu na lokalizację. Przykładowe analizy będą uwzględniały agregacje po dzielnicach i zmiany w czasie. System będzie umożliwiał składowanie i przetwarzanie danych wielkoskalowych. Dzięki temu możliwa jest analiza danych z całego Nowego Jorku przez wiele lat. System pozwoli znaleźć obszary miasta z najgorszą przestępczością z podziałem na kategorie zdarzeń i sprawdzi czy występuje zależność finansowa.

2 Dane

W projekcie posłużymy się dwoma zbiorami danych, udostępnionych przez administrację miasta Nowego Yorku na stronie opendata.cityofnewyork.us. Dane są udostępnione na podstawie tamtejszego "Open Data Law", które ustanawia, iż dane zbierane przez różne jednostki administracyjne Zarządu Miasta Nowego Yorku muszą być w sposób darmowy ogólnodostępne na oficjalnej stronie internetowej. Zatem korzystanie z danych oraz prowadzenie analiz jest dozwolone oraz w pełni legalne. W naszej analizie korzystamy z dwóch tabel pochodzących z dwóch źródeł: danych zgłoszeń na do Nowojorskiego Departamentu Policji oraz danych zestawiających finansowe statystyki bloków z mieszkaniami własnościowymi (tzw. condominium), które zobowiązane są do zbierania statystyk.
Dane można pozyskać przez oficjalne API lub pobrać bezpośrednio ze strony (bez logowania) w następujących formatach: CSV, TSV, RDF, XML, RSS. Ramki danych można znaleźć pod następującymi odnośnikami:

- Dane dotyczące przestępczości
- Dane mieszkaniowe

Dokładniejszy opis danych zostanie umieszczony w odpowiednich sekcjach.

2.1 Dane dotyczące przestępczości

Dane zostały udostępnione przez Departament Policji Miasta Nowego Yorku, który zbiera, przetwarza oraz przechowuje informacje dotyczące zgłaszanych przestępstw. Dane składają się z 35 kolumn oraz posiadają 7.83 miliona rekordów, gdzie każdy rekord odpowiada faktowi dokonania zgłoszenia. Planowane jest odświeżanie dzienne albo miesięczne. Dane zgłoszenie zawiera informacje o:

- Miejscu zajścia (dzielnica + szerokość i długość geograficzna danego miejsca)
- Czas zajścia z dokładnością do sekundy
- Klasyfikacja zajścia (czego dotyczyło zgłoszenie, klasa spośród możliwych)
- Rasa, płeć oraz przedział wiekowy napastnika
- Rasa, płeć oraz przedział wiekowy ofiary
- Jurysdykcja policyjna miejsca zajścia

2.2 Dane mieszkaniowe

Dane zostały dostarczone przez Departament Finansów Miasta Nowego Yorku, który zbiera i dokonuje administracji danych finansowych wspólnot mieszkaniowych. Dane składają się z 61 kolumn oraz 28.5 tysięcy wierszy, gdzie każdy wiersz stanowi daną wspólnotę wchodzącą w skład jednego condominium, zawiera jej dane finansowe oraz dane fizyczno-administracyjne (adres, pole powierzchni, rok założenia). **Uwaga**. Dane będzie należało poddać obróbce na poziomie pre-processingu ze względu na ich brak atomowości. Wynika to ze względu na ponowienie (poczwórne) ciągu danych zachodzących dla danego rekordu, z tą różnicą, iż każde powtórzenie odwołuje się do podobnej jednostki mieszkalnej w zakresie danego condominium. Na poziomie pre-processingu rozdzielimy te dane na 4 różne rekordy. Wedle opisu dostarczyciela danych, powielone opisy jednostek wchodzących w jeden rekord nie powielane są w następnych rekordach. Dane odświeżane są z częstotliwością roczną, a ich przedział czasowy obejmuje lata 2012 - 2021.
Ramka danych posiada dokładny opis w załączniku na stronie Dane mieszkaniowe, jednakże w skrócie opiszę jakie ważne informacje można uzyskać z danych zestawów kolumn, które zostaną zutylizowane podczas przetwarzania danych oraz ich agregacji z drugą ramką. Informacje, po zatomizowaniu, w rekordzie będą przechowywać:

- Dokładny adres danego lokalu (położenie w mieście oraz w danym bloku i przynależność do dzielnicy)
- Przychód przypadający na mieszkanie
- Przychód całościowy na stopę kwadratową powierzchni
- Całkowite pole powierzchni mieszkania
- Całkowita wartość rynkowa
- Rok budowy budynku
- Rok zgłoszenia przychodów

Dane pozwolą na przeprowadzenie analizy ze względu na typ danego bloku, średni przychód (odzwierciedlenie zamożności danego obszaru) oraz na geografię bloku (położenie w dzielnicy, adres).

3 Stos architektoniczny

3.1 Składowanie

- Apache Hive - wszystkie niezagregowane dane będą składowane w tabelach Hive. Dane mieszkaniowe mają bardziej jednolitą formę i w większości zawierają finansowe dane liczbowe, więc silnik HQL będzie tu bardziej przydatny do agregacji. Natomiast dane dotyczące przestępczości są mniej uporządkowane,

mają wiele braków i zawierają głównie zmienne kategoryczne, ale jest ich dużo więcej - są wielkoskalowe, więc Hive też jest tu potrzebny.

- MongoDB lub Apache HBase - platforma NoSQL będzie służyła do składowania obliczonych agregacji przygotowanych do wyświetlenia statystyk.

3.2 Przepływ danych

- Apache NiFi - wszystkie dane będą automatycznie pre-processowane przy pomocy NiFi z ewentualnym wsparciem skryptowym przy trudniejszych operacjach.

3.3 Warstwa analityczna

- Apache Spark - wielkoskalowe obliczenia i agregacje wykonywane będą na platformie Spark.
- Jupyter Notebook - finalne obliczenia i wizualizacje przedstawione będą w Pythonie w narzędziu Jupyter Notebook.

3.4 Pozostałe

Kod utrzymywane będzie przy pomocy GIT'a na GitHub'ie.
Rozwiązanie będzie zaimplementowane w lokalnym środowisku.

4 Podział pracy

Podział pracy jest przedstawiony poniżej.

Zadanie	Adam	Jan
Zaprojektowanie idei projektu	✓	✓
Utrzymanie projektu na GitHub	✓	✓
Dokumentacja Projektu	✓	✓
Pozyskanie danych		✓
Automatyzacja przepływu danych	✓	
Składowanie danych Hive	✓	
Składowanie agregacji NoSql		✓
Wsadowa analiza danych		✓
Testy funkcjonalne	✓	✓

Tabela 1: Podział obowiązków w projekcie