Projekt – raport

Analiza zależności między przestępczością a wypadkami drogowymi w Chicago

Autory:

Barbara Jurzysta

Viktoriia Kyryk

Styczeń 2025

Spis treści

[1 Cel projektu i potencjalne korzyści z wdrożenia 1](file:///C:\Users\vikak\OneDrive%20-%20Politechnika%20Warszawska\2024-2025\BigData\Przykładowy%20konspekt%20projektu.docx#__RefHeading___Toc75_3941455604)

[2 Dane 2](file:///C:\Users\vikak\OneDrive%20-%20Politechnika%20Warszawska\2024-2025\BigData\Przykładowy%20konspekt%20projektu.docx#__RefHeading___Toc77_3941455604)

[2.1 Dane dotyczące przestępczości 2](file:///C:\Users\vikak\OneDrive%20-%20Politechnika%20Warszawska\2024-2025\BigData\Przykładowy%20konspekt%20projektu.docx#__RefHeading___Toc79_3941455604)

[2.2 Dane o wypadkach drogowych 3](file:///C:\Users\vikak\OneDrive%20-%20Politechnika%20Warszawska\2024-2025\BigData\Przykładowy%20konspekt%20projektu.docx#__RefHeading___Toc81_3941455604)

[3.1 Składowanie 3](file:///C:\Users\vikak\OneDrive%20-%20Politechnika%20Warszawska\2024-2025\BigData\Przykładowy%20konspekt%20projektu.docx#__RefHeading___Toc83_3941455604)

3 Stos architektoniczny.........................................................................................................................4

[3.2 Przepływ danych 4](file:///C:\\Users\\vikak\\OneDrive%20-%20Politechnika%20Warszawska\\2024-2025\\BigData\\Przykładowy%20konspekt%20projektu.docx" \l "__RefHeading___Toc85_3941455604)

[3.3 Warstwa analityczna 4](file:///C:\Users\vikak\OneDrive%20-%20Politechnika%20Warszawska\2024-2025\BigData\Przykładowy%20konspekt%20projektu.docx#__RefHeading___Toc87_3941455604)

[3.4 Pozostałe 4](file:///C:\Users\vikak\OneDrive%20-%20Politechnika%20Warszawska\2024-2025\BigData\Przykładowy%20konspekt%20projektu.docx#__RefHeading___Toc89_3941455604)

4 Opis zaimplementowanego rozwiązania...........................................................................................5

4.1 Diagram przepływu danych.........................................................................................................5

4.2 Opis..............................................................................................................................................5

4.3 Podsumowanie.............................................................................................................................7

5 Załączone pliki...................................................................................................................................7

6 Podział pracy......................................................................................................................................7

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

Celem projektu jest przeprowadzenie analizy zależności między przestępczością a wypadkami drogowymi w Chicago. Projekt ma na celu zbadanie, czy istnieje powiązanie między poziomem przestępczości w różnych dzielnicach Chicago, a częstotliwością oraz rodzajem wypadków drogowych w tych samych rejonach. Chcemy przeanalizować także, czy na przeciągu lat przestępczość/ilość wypadków drogowych zmalała. Jako potencjalne korzyści z wdrożenia widzimy, możliwość zwiększenia bezpieczeństwa mieszkańców poprzez poprawę działania służb porządkowych w danych obszarach, wyodrębnienie miejsc w których często dochodzi do danego rodzaju przestępstw lub wypadków.

# **2 Dane**

**2.1 Dane dotyczące przestępczości w Chicago**

Ten zestaw danych przedstawia zgłoszone przestępstwa w Chicago od 2001 roku do chwili obecnej, z wyłączeniem ostatnich siedmiu dni. Dane pochodzą z systemu CLEAR (Citizen Law Enforcement Analysis and Reporting) Departamentu Policji w Chicago, ale pobrałyśmy ten zbiór danych z Kaggle.

## Dane składają się z 6.99 millionów wierszy i 22 kolumn.

Dane zgłoszenie zawiera informacje o:

* Miejsca zajścia przestępstwa:

1. **latitude** - Szerokość geograficzna miejsca (Lokalizacja ta jest przesunięta względem rzeczywistego miejsca dla częściowej anonimizacji, ale znajduje się na tym samym bloku.);
2. **longitude** - Długość geograficzna(Lokalizacja ta jest przesunięta względem rzeczywistego miejsca dla częściowej anonimizacji, ale znajduje się na tym samym bloku.)
3. **community\_area** - Obszar społecznościowy (Chicago ma 77 takich obszarów);
4. **district** - Dystrykt policyjny;
5. **beat** - Sektor(najmniejszy obszar geograficzny patrolowany przez policję)
6. **block** - Częściowo zredagowany adres, na którym doszło do incydentu, wskazujący na ten sam blok, co rzeczywisty adres
7. **location\_description** - Opis miejsca, w którym doszło do incydentu
8. **location** - Miejsce incydentu w formacie umożliwiającym tworzenie map i analiz geograficznych
9. **x\_coordinate** - Współrzędna X miejsca incydentu w układzie State Plane Illinois East NAD 1983.
10. **y\_coordinate** - Współrzędna Y miejsca incydentu w układzie State Plane Illinois East NAD 1983
11. **ward** - Dzielnica (okręg Rady Miasta), w której doszło do incydentu

* Data i czas:

1. **date** – Data i czas zajścia zdarzenia (z dokładnością do sekund)
2. **updated\_on** - data i czas ostatniej aktualizacji rekordu
3. year - rok, w którym doszło do incydentu

* Klasyfikacja przestępstw:

1. **fbi\_code** - Klasyfikacja przestępstw zgodna z Narodowym Systemem Zgłaszania Incydentów FBI (NIBRS)

* Szczególy zdarenia:

1. **domestic**: Informacja, czy incydent był związany z przemocą domową, zgodnie z definicją Illinois Domestic Violence Act
2. **arrest**: Informacja, czy dokonano aresztowania

* Identyfikatory:
  1. **unique\_key**: Unikalny identyfikator rekordu
  2. **case\_number**: Numer RD (Records Division Number) Departamentu Policji w Chicago, unikalny dla danego incydentu
* Opis przestępstwa:
  1. **primary\_type**: Główna kategoria opisu kodu IUCR
  2. **description**: Dodatkowy opis kodu IUCR, będący podkategorią głównego opisu
  3. **iucr**: Kod Illinois Uniform Crime Reporting (IUCR). Jest bezpośrednio powiązany z **primary\_type** i **description**

**2.2 Dane o wypadkach drogowych w Chicago**

Dane o wypadkach drogowych w Chicago zostały dostarczone przez Departament Policji w Chicago i pochodzą z elektronicznego systemu raportowania wypadków (E-Crash). Zawierają informacje o wypadkach na ulicach Chicago. Dane są aktualizowane po zakończeniu raportu lub jego zmianach. Dane z systemu E-Crash są dostępne dla niektórych dzielnic policyjnych od 2015 roku, ale dane miejskie są dostępne dopiero od września 2017 roku. Zawierają wypadki, które miały miejsce od września 2017 roku, z wyjątkiem tych, w których Departament Policji w Chicago nie było agencją reagującą, takich jak wypadki na autostradach. Dane składają się z 48 kolumn i 895 tysięcy wierszy. Jeden wiersz to jedno zdarzenie, o którym mamy informację(ze względu na ilość kolumn zgrupowałyśmy kolumny na potrzeby wypisania):

1. Identyfikator zdarzenia
2. Data i godzina wypadku (Szacunkowa data wypadku, jeśli zgłoszenie nastąpiło po czasie, oraz faktyczna data i godzina zdarzenia), Data powiadomienia policji, Czas wypadku, dzień tygodnia, Miesiąc
3. Prędkość dopuszczalna
4. Informacje o urządzeniu sterującym ruchem
5. Warunki atmosferyczne oraz oświetleniowe
6. Typ kolizji, Typ drogi, Liczba pasów ruchu, Ustawienie drogi, Stan nawierzchni, Wady drogi
7. Typ raportu (np. na miejscu, w biurze)
8. Typ wypadku
9. Związek z skrzyżowaniem (określenie, czy skrzyżowanie miało wpływ na wypadek).
10. Określenie, czy wypadek wydarzył się poza publiczną drogą.
11. Ucieczka z miejsca wypadku
12. Uszkodzenia w wyniku wypadku (szacowane)
13. Przyczyny wypadku (główna i drugorzędna)
14. Adres miejsca wypadku, Rejon policyjny, Współrzędne geograficzne, Lokalizacja wypadku jako punkt, umożliwiająca analizę geograficzną w oprogramowaniu portalu danych.
15. Czy zostało zrobione zdjęcie, czy zostały zebrane zeznania, czy kolizja spowodowana przez otwarte drzwi pojazdu, które weszły w drogę rowerzyście, czy strefa robót drogowych, typ tej strefy oraz obecność pracowników
16. Liczba jednostek uczestniczących w wypadku (pojazdów, pieszych, rowerzystów)
17. Rodzaj najcięższych obrażeń, jakich doznała osoba w wypadku, Liczba ofiar, Liczba ofiar śmiertelnych, ofiar z poważnymi obrażeniami, drobnymi obrażeniami, obrażeniami zgłoszonymi, ale niewidocznymi, bez obrażeń, w nieznanym stanie

**3 Stos architektoniczny**

# **3.1 Składowanie**

1. **Hadoop Distributed File System (HDFS)**
   * Surowe dane dotyczące przestępczości i wypadków drogowych będą przechowywane w HDFS, co umożliwi skalowalne przechowywanie i dostęp do pełnych zbiorów danych.
2. **Apache Hive -** Dane zostaną załadowane do tabel Hive, co pozwoli na ich analizę przy użyciu języka HQL:
   * **Dane przestępczości**: ze względu na dużą liczbę rekordów (6,99 miliona wierszy), Hive umożliwi efektywne filtrowanie, agregowanie oraz wykonywanie operacji na danych z wieloma brakującymi wartościami.
   * **Dane wypadków drogowych**: ze względu na bardziej jednolitą strukturę (48 kolumn), Hive będzie używane do wykonywania złożonych zapytań i analiz czasowych, takich jak porównania danych z różnych lat.
3. **MongoDB**
   * Przechowywanie zagregowanych danych do wizualizacji, np. liczba wypadków w dzielnicach, najczęstsze typy przestępstw.

# **3.2 Przepływ danych**

**Apache NiFi**

* Dane przestępczości i wypadków drogowych będą automatycznie przetwarzane przy użyciu **Apache NiFi**.

# **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. Repozytorium dostępne pod adresem: https://github.com/vikakyr/Big-Data-Project

# **4 Opis zaimplementowanego rozwiązania**

**4.1 Diagram przepływu danych** Obraz zawierający tekst, diagram, linia, Plan

Opis wygenerowany automatycznie

**4.2 Opis**

Najpierw zapisałyśmy oba zbiory lokalnie, następnie skopiowałyśmy je na maszynę wirtualną. Do transformacji danych i zapisu wyniku w pliku CSV do systemu HDFS wykorzystałyśmy Apache NiFi. Na zbiorze danych o przestępstwach (Crimes) przeprowadziłyśmy następujące transformacje:

1. Rozdzieliłyśmy kolumnę Date, która zawierała datę i czas, na trzy oddzielne kolumny: Day, Month, Year.
2. Odfiltrowałyśmy rekordy, pozostawiając tylko te, które zostały zarejestrowane od 2015 roku.
3. Wybrałyśmy następujące kolumny:

* ID – id rekordu,
* Primary Type (jako Primary\_Type) – podsawowy typ przestępstwa,
* Description - opis przestępstwa,
* Latitude - szerokość geograficzna miejsca przestępstwa,
* Longitude - długość geograficzna miejsca przestępstwa,
* Arrest - czy dokonano aresztowania,
* Domestic - czy incydent był związany z przemocą domową, zgodnie z definicją ustawy Illinois o przemocy domowej.,
* Boundaries - ZIP Codes (jako District\_by\_zip) - Znalazłyśmy możliwość pobrania wraz z danymi numer dystryktu, gdy dystrykty są podzielone przy użyciu kodów pocztowych, dla naszych potrzeb te dystrykty traktujemy jako „dzielnice” Chicago

Na zbiorze o wypadkach drogowych (Traffic Crashes) przeprowadziłyśmy następujące transformacje:

1. Rozdzieliłyśmy kolumnę Date, która zawierała datę i czas, na trzy oddzielne kolumny: Day, Month, Year.
2. Odfiltrowałyśmy rekordy, pozostawiając tylko te, które zostały zarejestrowane od 2015 roku.
3. Wybrałyśmy następujące kolumny:

* CRASH\_RECORD\_ID – id rekordu,
* POSTED\_SPEED\_LIMIT – dozwolona prędkość na danej drodze,
* WEATHER\_CONDITION – warunki drogowe podczas wypadku,
* LIGHTING\_CONDITION - warunki oświetleniowe w momencie wypadku,
* ROADWAY\_SURFACE\_COND - stan nawierzchni drogi,
* CRASH\_TYPE – typ wypadku,
* NUM\_UNITS - liczba jednostek zaangażowanych w wypadek. Jednostka może być pojazdem mechanicznym, pieszym, rowerzystą lub innym użytkownikiem drogi, który nie jest pasażerem
* MOST\_SEVERE\_INJURY - najcięższe obrażenie odniesione przez osobę zaangażowaną w wypadek,
* INJURIES\_TOTAL - łączna liczba osób, które odniosły obrażenia śmiertelne, uniemożliwiające funkcjonowanie, nieuniemożliwiające funkcjonowanie oraz możliwe obrażenia,
* LATITUDE – szerokość geograficzna miejsca wypadku,
* LONGITUDE - długość geograficzna miejsca wypadku,
* Boundaries - ZIP Codes (jako District\_by\_zip) – Znalazłyśmy możliwość pobrania wraz z danymi numer dystryktu, gdy dystrykty są podzielone przy użyciu kodów pocztowych, dla naszych potrzeb te dystrykty traktujemy jako „dzielnice” Chicago

Dane z HDFS zostały przetworzone w następujący sposób:

* Załadowane do Jupyter Notebook przy użyciu PySpark i poddane analizie takiej jak: liczba przestępstw w zależności od roku z podziałem na to, czy doszło do aresztowania; liczba przestępstw w zależności od roku z podziałem na to, czy incydent był związany z przemocą domową; znalezienie miejsca z największą ilością przestępstw; liczba przestępstw względem dzielnicy, najwięcej wypadków w przy jakich warunkach; znalezienie miejsca z największą ilością wypadków; porównanie przestępstw i wypadków w dystryktach, sprawdzanie korelacji między typem wypadku a dystryktem; Sprawdzanie korelacji między typem przestępstw a dystryktem. (Wykresy oraz komentarze w pliku zawierającym prezentację)
* Dane zostały załadowane do tabel Hive, gdzie przeprowadzono prostą analizę naprzykład: liczba przestępstw względem roku, liczba wypadków w zależności od miesiąca, liczba wypadków w zależności od sezonu (wiosna, lato, jesień, zima)
* Dane z HDFS zostały zapisane lokalnie aby mogły zostać wczytane do tabel MongoDB przy użyciu MongoDB Compass. Gdzie oprócz wczytania i sprawdzenia czy prawidłowo dane zostały poprawnie załadowane. Następnie tworzymy dwie inne kolekcje poprzez agregacje danych:
  + traffic\_crash\_grouped – zawierającą zliczenie wypadków przy zgrupowaniu dystryktu, miesiąca i roku
  + crime\_crash\_grouped – zawierającą zliczenie przestępstw z aresztowaniem, przy zgrupowaniu po dystryktach

**4.3 Podsumowanie**

Z danych możemy wyciągnąć ciekawe wnioski, jednak nie udało określić czy przestępstwa są powiązane z wypadkami, dlatego, że ilość obu w danym dystrykcie zależy od powierzchni danego dystryktu a one bardzo różnią się wielkością. Ponadto widać spadek w obu zbiorach dla ilości incydentów w 2020 i 2021 roku. Można łatwo połączyć to w pandemią koronawirusa, dlatego ciężko wyciągać wnioski czy ilość wypadków/przestępstw ma tendencje do spadku czy wzrostu, gdy zmalała w tych latach a następnie wzrosła.

**5 Załączone pliki**

Oto lista załączonych jeszcze plików:

1. Przestępstwa i wypaki drogowe w Chicago.pptx – plik zawierający prezentacje, która zawiera także screeny oraz wykresy z PySpark
2. nifi\_template.xml – szablon NiFi
3. hive\_commands.docx – plik zawiera zżuty ekranu części projektu, gdzie stosujemy Hive
4. PySpark.ipynb – plik Jupyter Notebook zawierający analizę danych używając PySpark
5. MongoDB.docx – plik zawiera zżuty ekranu z załadowywania danych do kolekcji w MongoDB
6. Testy.docx – plik zawierający zżuty ekranu potwierdzający testy, które przeprowadziłyśmy

# **6 Podział pracy**

Podział pracy jest przedstawiony poniżej.

Obraz zawierający tekst, Czcionka, numer, paragon

Opis wygenerowany automatycznie