Zestaw 3 – Crimes in Chicago

Pochodzenie danych to `https://www.kaggle.com/currie32/crimes-in-chicago`

Dane zawierają informacje o przestępstwach zarejestrowanych w Chicago w latach 2001-2017.

Zbiór danych

Wykorzystywane są dwa zbiory danych.

Pierwszy, główny, "strumieniowy" to zbiór plików mających format csv i następujące pola:

- `ID` – Unikalny identyfikator rekordu (przestępstwa)
- `Date` – Data zdarzenia
- `IUCR` – Kod *Illinois Uniform Crime Reporting*. Jest on bezpośrednio powiązany z podstawowym typem i opisem typu przestępstwa.
  Zobacz listę kodów IUCR pod adresem `https://data.cityofchicago.org/d/c7ck-438e`.
- `Arrest` – Wskazuje, czy dokonano aresztowania
- `Domestic` – Wskazuje, czy incydent był związany z miejscem zamieszkania, zgodnie z definicją zawartą w ustawie *Illinois Domestic Violence Act*.
- `District` – Wskazuje dzielnicę policji, w której miał miejsce incydent.
  Zobacz dzielnice pod adresem `https://data.cityofchicago.org/d/fthy-xz3r`.
- `ComArea` – Wskazuje obszar społeczności, w którym miał miejsce incydent. Chicago ma 77 takich obszarów.
  Zobacz obszary społeczności pod adresem `https://data.cityofchicago.org/d/cauq-8yn6`.
- `Latitude` – Szerokość geograficzna miejsca, w którym doszło do incydentu.
- `Longitude` – Długość geograficzna miejsca, w którym doszło do incydentu.

Drugi zbiór statyczny `Chicago_Police_Department_-_Illinois_Uniform_Crime_Reporting__IUCR__Codes.csv` zawiera następujące pola:

- `IUCR` – Kod *Illinois Uniform Crime Reporting*
- `PRIMARY DESCRIPTION` – podstawowa kategoria przestępstwa
- `SECONDARY DESCRIPTION` – szczegółowa kategoria przestępstwa
- `INDEX CODE` - Przestępstwa związane z indeksem (`I`) to przestępstwa rejestrowane w całym kraju przez program *Uniform Crime Reports* Federalnego Biura Śledczego (FBI) w celu dokumentowania trendów przestępczości w czasie (dane publikowane są co pół roku) i obejmują morderstwa, napaść na tle seksualnym, rozbój, napad z użyciem siły i pobicie, włamanie, kradzież, kradzież pojazdu mechanicznego i podpalenie. Przestępstwa niezwiązane z indeksem (`N`) to wszystkie inne rodzaje incydentów kryminalnych, w tym akty wandalizmu, naruszenia broni, naruszenia pokoju publicznego itp.

ETL – obraz czasu rzeczywistego

Utrzymywanie agregacji na poziomie miesiąca, głównej kategorii przestępstwa i dzielnicy. Wartości agregatów to:

- pełna liczba przestępstw
- liczba przestępstw zakończonych aresztowaniem
- liczba przestępstw zakończonych związanych z przemocą domową
- liczba przestępstw rejestrowanych przez FBI (patrz: INDEX CODE)

Wykrywanie "anomalii"

Wykrywanie "anomalii" ma polegać na wykrywaniu pojawienia znaczącego procenta liczby przestępstw rejestrowanych przez FBI w stosunku do liczby wszystkich przestępstw zarejestrowanych na poziomie poszczególnych dzielnic w podanym okresie czasu.
Program ma być parametryzowany przez:

- D - długość okresu czasu wyrażoną w dniach
- P – procent przestępstw rejestrowanych przez FBI (minimalna)

Wykrywanie anomalii ma być dokonywane każdego dnia.

Przykładowo, dla parametrów D=7, P=40 program każdego dnia będzie raportował te dzielnice, w których w ciągu ostatnich 7 dni procent liczby przestępstw rejestrowanych przez FBI w stosunku do liczby wszystkich przestępstw przekroczył 40%.

Raportowane dane mają zawierać

- analizowany okres - okno (start i stop)
- identyfikator dzielnicy
- liczbę przestępstw rejestrowanych przez FBI
- liczbę wszystkich przestępstw
- procent przestępstw rejestrowanych

Załóż, że dane mogą być nieuporządkowane – mogą być opóźnione o jeden dzień.