Sprawozdanie

Zestaw 3: Crimes in Chicago

Kafka Streams

Spis treści

Konfiguracja projektu

Przygotowanie środowiska

1. Uruchom klaster za pomocą poniższej komendy:

```
gcloud dataproc clusters create ${CLUSTER_NAME} \
--enable-component-gateway --region ${REGION} --subnet default \
--master-machine-type n1-standard-4 --master-boot-disk-size 50 \
--num-workers 2 --worker-machine-type n1-standard-2 --worker-boot-disk-size 50 \
--image-version 2.1-debian11 --optional-components DOCKER,ZOOKEEPER \
--project ${PROJECT_ID} --max-age=3h \
--metadata "run-on-master=true" \
--initialization-actions \
gs://goog-dataproc-initialization-actions-${REGION}/kafka/kafka.sh
```

2. Pobierz crime-in-chicago-result.zip z danymi z linku poniższego linku: https://www.cs.put.poznan.pl/kjankiewicz/bigdata/stream_project/crimes-in-chicago_result.zip
3. Wgraj plik zip do swojego bucketa. **Uwaga! Nie wgrywaj go do żadnego dodatkowego folderu, tylko bezpośrednio do bucketa.**
4. Uruchom 4 terminale SSH na masterze klastra.
5. Wrzuć plik zip z projektem za pomocą przycisku “Upload” w jednym z terminali.
6. Rozpakuj plik zip, zmień uprawnienia plików sh ze skryptami oraz upewnij się, że skrypty sh pasują do formatu UNIX za pomocą poniższych komend:

```
unzip projekt2.zip
chmod +x *.sh
sed -i 's/\r//' *.sh
```

7. Uruchom skrypt przygotowujący środowisko oraz tematy kafki:

```
export BUCKET_NAME=<nazwa Twojego bucketa>
./create_environment_and_topics.sh
```

Pamiętaj, żeby zmienić nazwę bucketa!

8. Uruchom skrypt tworzący bazę danych oraz połączenie z bazą:

```
./create_db_and_db_connection.sh
```

Uruchomienie produkcji, przetwarzania oraz zapisu strumienia danych

9. W pierwszym terminalu uruchom podgląd na temat z zagregowanymi danymi:

```
./lookup.sh
```

   Pozwoli to upewnić się, że dane zostały już przetworzone.

10. Uruchom w drugim terminalu skrypt włączający pobieranie danych z tematu do bazy danych:

```
./connect_db.sh
```

11. Uruchom w trzecim terminalu skrypt włączający producenta kafki:

```
./kafka_producer.sh
```

12. Uruchom w czwartym terminalu skrypt włączający przetwarzanie strumieni w trybie A:

```
./aggregator.sh
```

   Zamiast przetwarzania strumieni w trybie A, możesz także uruchomić przetwarzanie strumieni w trybie C:

```
./aggregatorC.sh
```

13. W momencie, w którym w pierwszym terminalu pojawią się dane, możesz wyłączyć podgląd tematu (1 terminal), producenta (2 terminal), przetwarzanie strumieni (3 terminal) i nasłuch bazy (4 terminal) za pomocą skrótu klawiszowego CTRL+C.

Odczyt wyników przetwarzania

14. Żeby odczytać wszystkie dane z bazy danych uruchom poniższy skrypt:

```
./all_data.sh
```

15. Żeby odczytać konkretne dane (dane dla konkretnej kategorii przestępstwa, dystryktu oraz miesiąca) użyj poniższego skryptu:

```
./read_db.sh "<kategoria>" <numer dystryktu> "<data w formacie RRRR-MM>"
```

   Możesz skorzystać z przykładowego wywołania skryptu:

```
./read_db.sh "THEFT" 14 "2001-04"
```

Reset

16. W razie potrzeby zresetuj środowisko oraz bazę danych za pomocą poniższego skryptu:

```
./reset.sh
```

Producent; skrypty inicjujące i zasilający

Skrypt inicjujący: create_environment_and_topics.sh

Skrypt pobiera dane z bucketa, rozpakowuje je, zmienia nazwę katalogu a następnie tworzy tematy Kafki.

Oto prawidłowy wynik działania skryptu:

Początek:

```
Copying gs://pbds-24-jp/crimes-in-chicago_result.zip...
\ [1 files][106.7 MiB/106.7 MiB]
Operation completed over 1 objects/106.7 MiB.
Archive:  crimes-in-chicago_result.zip
   creating: crimes-in-chicago_result/
  inflating: crimes-in-chicago_result/part-00000-10b00d71-feee-417e-b0bc-888b1e6afec7-c000.csv

  inflating: crimes-in-chicago_result/part-00001-10b00d71-feee-417e-b0bc-888b1e6afec7-c000.csv
```

Koniec:

```
  inflating: crimes-in-chicago_result/part-00098-10b00d71-feee-417e-b0bc-888b1e6afec7-c000.csv

  inflating: crimes-in-chicago_result/part-00099-10b00d71-feee-417e-b0bc-888b1e6afec7-c000.csv

Created topic chicago-data.
Created topic count.
Created topic json.
```

Skrypt zasilający: kafka_producer.sh

Skrypt włącza zasilanie tematu Kafki za pomocą producenta z pliku .jar.

Oto prawidłowy wynik działania skryptu:

```
log4j:WARN No appenders could be found for logger (org.apache.kafka.clients.producer.ProducerC
onfig).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
```

Oznacza to, że dane są obecnie produkowane na temat Kafki.

Skrypt resetujący: reset.sh

Uwaga ten skrypt resetuje zarówno środowisko jak i całą bazę danych!

W razie gdyby potrzebny był reset jedynie środowiska (zalecam jednak resetować całość), wystarczy usunąć lub zakomentować całą część kodu po:

```
#--------------------------------------------------------------------------
```

Oto prawidłowy wynik działania skryptu:

```
Created topic chicago-data.
Created topic count.
Created topic json.
mysql: [Warning] Using a password on the command line interface can be insecure.
mymysql
mymysql
Untagged: mysql:debian
Untagged: mysql@sha256:49f4fcb0087318aa1c222c7e8ceacbb541cdc457c6307d45e6ee4313f4902e33
Deleted: sha256:992367f7e447afa8958f0b52f60a32bf01362731ac144922e319c26d5bade895
Deleted: sha256:b68b443a975f1ffbedfced4560fab66e4fec6f348c8c8cc3a6a5ca0db954e297
Deleted: sha256:9f34d685d35342014f8ee981c7e494551d51ce9e63127e8ac2ef693eab31da2e
Deleted: sha256:c307b7c8ad780b99b02a69ec5960521219c17ad5c7e70749663e6beb581bbb71
Deleted: sha256:3685aad6248228f6e66607b8628bd0ffa2c8e89f3c38b745e0b719518f2e4f4b
Deleted: sha256:9b2e05fd79115847d2934598d249c19a837c7fd304ee4664d4ab7bf4dcda060e
Deleted: sha256:033f370c3dba845cbb8e201e80dd55b9529ee2b6420aa374619af1b5dc520a83
Deleted: sha256:776c135f9fa504e85077984834779l172900f0afee4e3c69e024c965695735ed
Deleted: sha256:b4bfd96abc5784e0cf976c777cd1be1bc9ad0324db9d5c7a7b06e71ccf49f339
Deleted: sha256:b053c8eb9234d6219a5206440bf15a5e223c333df48e8768d690868030349d3e
Deleted: sha256:0a813129f7ead3347c5e0bdd97b28b85b6409a0541970ca333ed62a90561ec49
Deleted: sha256:44f5f55f9d4e1513c62e98559de376a270cbd515f4f56792c5a1bff1b2dd4998
Deleted: sha256:8ce178ff9f343a37169f68dd0df03099524afb71a879551c5f17e493c7b1d3ec
Unable to find image 'mysql:debian' locally
debian: Pulling from library/mysql
1d5252f66ea9: Pull complete
b034716887fd: Pull complete
882448c7d618: Pull complete
c14ac4640b87: Pull complete
3d2978613e40: Pull complete
c33d24cb8c75: Pull complete
52fa14fc88fc: Pull complete
42deb22ec16d: Pull complete
c3d31d600a65: Pull complete
b0b9f5aba00d: Pull complete
98222a66de15: Pull complete
2ff8f0c3c67b: Pull complete
Digest: sha256:49f4fcb0087318aa1c222c7e8ceacbb541cdc457c6307d45e6ee4313f4902e33
Status: Downloaded newer image for mysql:debian
```

```
68e19fc7cce7af16a6038e4a917516e15c9062a7123b6d2e72b8852ff22a5a46
mysql: [Warning] Using a password on the command line interface can be insecure.
mysql: [Warning] Using a password on the command line interface can be insecure.
--2024-06-09 11:37:03--  https://repo1.maven.org/maven2/com/mysql/mysql-connector-j/8.0.33/mys
ql-connector-j-8.0.33.jar
Resolving repo1.maven.org (repo1.maven.org)... 199.232.192.209, 199.232.196.209, 2a04:4e42:4c:
:209, ...
Connecting to repo1.maven.org (repo1.maven.org)|199.232.192.209|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2481560 (2.4M) [application/java-archive]
Saving to: ‘mysql-connector-j-8.0.33.jar’

mysql-connector-j-8.0.3 100%[============================>]   2.37M  --.-KB/s    in 0.1s

2024-06-09 11:37:03 (19.6 MB/s) - ‘mysql-connector-j-8.0.33.jar’ saved [2481560/2481560]

--2024-06-09 11:37:03--  https://packages.confluent.io/maven/io/confluent/kafka-connect-jdbc/1
0.7.0/kafka-connect-jdbc-10.7.0.jar
Resolving packages.confluent.io (packages.confluent.io)... 18.66.233.42, 18.66.233.37, 18.66.2
33.59, ...
Connecting to packages.confluent.io (packages.confluent.io)|18.66.233.42|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 275798 (269K) [application/octet-stream]
Saving to: ‘kafka-connect-jdbc-10.7.0.jar’

kafka-connect-jdbc-10.7 100%[============================>] 269.33K  --.-KB/s    in 0.003s

2024-06-09 11:37:03 (83.2 MB/s) - ‘kafka-connect-jdbc-10.7.0.jar’ saved [275798/275798]

log4j.logger.org.reflections=ERROR
```

Utrzymanie obrazu czasu rzeczywistego - transformacje

Opis transformacji:

- Mapowanie danych wejściowych:

```
.map((key,value) -> {
   String newValue = String.format("IUCR: %s, Category: %s, District: %f, Date:
%s, Arrest: %b, Domestic: %b, IndexCode: %s", value.getIUCR(),
value.getCategory(), value.getDistrict(), value.getMonthYear(),
value.getArrest(), value.getDomestic(), value.getIndexCode());
   return KeyValue.pair(value.getCategory() + "-" + value.getDistrict(),
newValue);
})
```

Dane wejściowe są przekształcane, aby utworzyć nowe pary klucz-wartość. Klucz zawiera informacje o kategorii przestępstwa i dystrykcie, a wartość zawiera szczegółowe informacje o przestępstwie potrzebne do agregacji.

- Grupowanie danych:

```
.groupByKey()
```

Dane są grupowane na podstawie nowo utworzonych kluczy (kategoria przestępstwa i dystrykt).

- Definiowanie okien czasowych:

```
.windowedBy(TimeWindows.of(Duration.ofDays(30)).grace(gracePeriod))
```

Proces agregacji odbywa się w 30-dniowych oknach czasowych z okresem tolerancji 1 dzień.

- Agregacja:

```
.aggregate(
                    () -> {
                            String iucr = "";
                            String category = "0000-00";
                            Float district = 0f;
                            String date = "";
                            Long all = 0L;
                            Long arrest = 0L;
                            Long domestic = 0L;
                            Long fbi = 0L;
                            String aggregate = String.format("IUCR: %s,
Category: %s, District: %f, Date: %s, All: %d, Arrest: %d, Domestic: %d, FBI:
%d", iucr, category, district, date, all, arrest, domestic, fbi);
                            return aggregate;
                            },
                       (aggKey, newValue, aggregate) -> {
                            String iucr = "";
                            String category = "";
                            Float district = 0f;
                            String date = "0000-00";
                            Boolean arrestValue = false;
                            Boolean domesticValue = false;
                            String indexCode = "";
                            Long all = 0L;
                            Long arrest = 0L;
                            Long domestic = 0L;
                            Long fbi = 0L;

                            Matcher matcher = pattern.matcher(aggregate);
                            if (matcher.find()) {
                                all = Long.parseLong(matcher.group(4));
                                arrest = Long.parseLong(matcher.group(5));
                                domestic = Long.parseLong(matcher.group(6));
                                fbi = Long.parseLong(matcher.group(7));
                            }

                            Matcher matcher2 = pattern2.matcher(newValue);
                            if (matcher2.find()) {
                                iucr = matcher2.group(1);
                                category = matcher2.group(2);
                                district = Float.parseFloat(matcher2.group(3));
                                date = matcher2.group(4);
                                arrestValue =
Boolean.parseBoolean(matcher2.group(5));
                                domesticValue =
Boolean.parseBoolean(matcher2.group(6));
                                indexCode = matcher2.group(7);
                            }

                            if (category.equals("No matching category")) {
```

```
System.out.println("---------------------------------Error---------------------
---------------------------------");
                                System.out.println("IUCR:" + iucr);
                                System.out.println("Category: " + category);
                                System.out.println("District: " + district);
                                System.out.println("Aggregate: " + aggregate);
                                System.out.println("newValue: " + newValue);
                            }

                            all=all+1;

                            if(arrestValue == Boolean.TRUE) {
                                arrest=arrest+1;
                            }
                            if(domesticValue == Boolean.TRUE) {
                                domestic=domestic+1;
                            }
                            if(indexCode.equals("I")) {
                                fbi=fbi+1;
                            }

                            aggregate = String.format("Category: %s, District:
%f, Date: %s, All: %d, Arrest: %d, Domestic: %d, FBI: %d", category, district,
date, all, arrest, domestic, fbi);
                            return aggregate;
                        },
                        Materialized.with(stringSerde, stringSerde)
                )
```

Dla każdej grupy w oknie czasowym inicjalizowany jest stan “zerowy” agregatu. Jest to początkowy zbiór wartości, które będą modyfikowane w miarę napływu nowych danych.

Przy każdej nowej wartości, która pasuje do grupy i okna czasowego, stan agregatu jest aktualizowany. Proces ten obejmuje:

- Parsowanie aktualnego stanu agregatu w celu wyciągnięcia istniejących wartości.
- Parsowanie nowej wartości przestępstwa w celu wyciągnięcia jej szczegółów.
- Aktualizowanie liczników i innych wartości w stanie agregatu na podstawie nowych danych (np. zwiększenie liczby aresztowań, przestępstw domowych, itp.).

Zaktualizowany stan agregatu jest następnie zapisywany. Proces ten jest powtarzany dla każdej nowej wartości, która napływa do systemu.

Utrzymanie obrazu czasu rzeczywistego – obsługa trybu A

```
.windowedBy(TimeWindows.of(Duration.ofDays(30)).grace(gracePeriod))
```

Kafka Streams emituje częściowe wyniki agregacji na bieżąco, ale tylko po zamknięciu okna czasowego. Oznacza to, że konsument otrzyma wynik po zamknięciu okna oraz aktualizacje w okresie tolerancji.

Utrzymanie obrazu czasu rzeczywistego – obsługa trybu C

```
.suppress(Suppressed.untilWindowCloses(Suppressed.BufferConfig.unbounded()));
```

Dodanie modyfikacji .suppress(Suppressed.untilWindowCloses(Suppressed.BufferConfig.unbounded())) w Kafka Streams wprowadza mechanizm tłumienia wyników pośrednich, co zmienia sposób przetwarzania i emitowania danych. Tłumienie powoduje, że częściowe wyniki agregacji są zatrzymywane i nie są natychmiast emitowane. Zamiast tego, są one buforowane do momentu zamknięcia okna czasowego oraz okresu tolerancji, po czym emitowany jest tylko końcowy wynik.

Wykrywanie anomalii - BRAK

Program przetwarzający strumienie danych; skrypt uruchamiający

Skrypt uruchamiający tryb A: aggregator.sh

Po uruchomieniu pojawia się opis topologii:

```
      <-- KSTREAM-FILTER-0000000001
    Processor: KSTREAM-MAP-0000000003 (stores: [])
      --> KSTREAM-FILTER-0000000007
      <-- KSTREAM-MAPVALUES-0000000002
    Processor: KSTREAM-FILTER-0000000007 (stores: [])
      --> KSTREAM-SINK-0000000006
      <-- KSTREAM-MAP-0000000003
    Sink: KSTREAM-SINK-0000000006 (topic: KSTREAM-AGGREGATE-STATE-STORE-0000000004-repartition
)
      <-- KSTREAM-FILTER-0000000007

  Sub-topology: 1
    Source: KSTREAM-SOURCE-0000000008 (topics: [KSTREAM-AGGREGATE-STATE-STORE-0000000004-repar
tition])
      --> KSTREAM-AGGREGATE-0000000005
    Processor: KSTREAM-AGGREGATE-0000000005 (stores: [KSTREAM-AGGREGATE-STATE-STORE-0000000004
])
      --> KTABLE-TOSTREAM-0000000009, KTABLE-TOSTREAM-0000000012
      <-- KSTREAM-SOURCE-0000000008
    Processor: KTABLE-TOSTREAM-0000000009 (stores: [])
      --> KSTREAM-MAP-0000000010
      <-- KSTREAM-AGGREGATE-0000000005
    Processor: KTABLE-TOSTREAM-0000000012 (stores: [])
      --> KSTREAM-MAP-0000000013
      <-- KSTREAM-AGGREGATE-0000000005
    Processor: KSTREAM-MAP-0000000010 (stores: [])
      --> KSTREAM-SINK-0000000011
      <-- KTABLE-TOSTREAM-0000000009
    Processor: KSTREAM-MAP-0000000013 (stores: [])
      --> KSTREAM-SINK-0000000014
      <-- KTABLE-TOSTREAM-0000000012
    Sink: KSTREAM-SINK-0000000011 (topic: count)
      <-- KSTREAM-MAP-0000000010
    Sink: KSTREAM-SINK-0000000014 (topic: json)
      <-- KSTREAM-MAP-0000000013
```

Po jakimś czasie ukazuje się taki widok:

```
IUCR:0840
Category: No matching category
District: 14.0
Aggregate: Category: No matching category, District: 14.000000, Date: 2001-10, All: 4, Arrest:
 0, Domestic: 0, FBI: 0
newValue: IUCR: 0840, Category: No matching category, District: 14.000000, Date: 2001-10, Arre
st: false, Domestic: false, IndexCode: null
---------------------------------Error-------------------------------------------------------
IUCR:0841
Category: No matching category
District: 17.0
Aggregate: Category: No matching category, District: 17.000000, Date: 2001-10, All: 2, Arrest:
 0, Domestic: 0, FBI: 0
newValue: IUCR: 0841, Category: No matching category, District: 17.000000, Date: 2001-10, Arre
st: false, Domestic: false, IndexCode: null
---------------------------------Error-------------------------------------------------------
IUCR:0840
Category: No matching category
District: 20.0
Aggregate: Category: No matching category, District: 20.000000, Date: 2001-10, All: 4, Arrest:
 0, Domestic: 0, FBI: 0
newValue: IUCR: 0840, Category: No matching category, District: 20.000000, Date: 2001-10, Arre
st: false, Domestic: false, IndexCode: null
---------------------------------Error-------------------------------------------------------
IUCR:0840
Category: No matching category
District: 20.0
Aggregate: Category: No matching category, District: 20.000000, Date: 2001-10, All: 5, Arrest:
 0, Domestic: 0, FBI: 0
newValue: IUCR: 0840, Category: No matching category, District: 20.000000, Date: 2001-10, Arre
st: false, Domestic: false, IndexCode: null
---------------------------------Error-------------------------------------------------------
IUCR:0840
Category: No matching category
District: 20.0
Aggregate: Category: No matching category, District: 20.000000, Date: 2001-10, All: 6, Arrest:
 0, Domestic: 0, FBI: 0
newValue: IUCR: 0840, Category: No matching category, District: 20.000000, Date: 2001-10, Arre
st: false, Domestic: false, IndexCode: null
```

Wynika to z faktu, że dane dostarczone w csv były niekompletne.

Z tego względu wyświetlam wszystkie numery IUCR, które nie występują w pliku Chicago_Police_Department_-_Illinois_Uniform_Crime_Reporting__IUCR__Codes.csv.

Przestępstwa zgłoszone z tymi numerami IUCR są kategoryzowane jako kategoria “No matching category”.

Na powyższym zdjęciu ekranu widać, że nie znaleziono w pliku csv IUCR o wartości 840 oraz 841.

Oto lista wszystkich IUCR, które udało mi się odkryć, że nie występują w pliku csv:

- 840
- 841
- 842
- 499
- 5005
- 5008
- 9901

Skrypt uruchamiający tryb C: aggregatorC.sh

Po uruchomieniu pojawia się opis topologii:

```
    Processor: KSTREAM-MAP-0000000003 (stores: [])
      --> KSTREAM-FILTER-0000000007
      <-- KSTREAM-MAPVALUES-0000000002
    Processor: KSTREAM-FILTER-0000000007 (stores: [])
      --> KSTREAM-SINK-0000000006
      <-- KSTREAM-MAP-0000000003
    Sink: KSTREAM-SINK-0000000006 (topic: KSTREAM-AGGREGATE-STATE-STORE-0000000004-repartition
)
      <-- KSTREAM-FILTER-0000000007

  Sub-topology: 1
    Source: KSTREAM-SOURCE-0000000008 (topics: [KSTREAM-AGGREGATE-STATE-STORE-0000000004-repar
tition])
      --> KSTREAM-AGGREGATE-0000000005
    Processor: KSTREAM-AGGREGATE-0000000005 (stores: [KSTREAM-AGGREGATE-STATE-STORE-0000000004
])
      --> KTABLE-SUPPRESS-0000000009
      <-- KSTREAM-SOURCE-0000000008
    Processor: KTABLE-SUPPRESS-0000000009 (stores: [KTABLE-SUPPRESS-STATE-STORE-0000000010])
      --> KTABLE-TOSTREAM-0000000011, KTABLE-TOSTREAM-0000000014
      <-- KSTREAM-AGGREGATE-0000000005
    Processor: KTABLE-TOSTREAM-0000000011 (stores: [])
      --> KSTREAM-MAP-0000000012
      <-- KTABLE-SUPPRESS-0000000009
    Processor: KTABLE-TOSTREAM-0000000014 (stores: [])
      --> KSTREAM-MAP-0000000015
      <-- KTABLE-SUPPRESS-0000000009
    Processor: KSTREAM-MAP-0000000012 (stores: [])
      --> KSTREAM-SINK-0000000013
      <-- KTABLE-TOSTREAM-0000000011
    Processor: KSTREAM-MAP-0000000015 (stores: [])
      --> KSTREAM-SINK-0000000016
      <-- KTABLE-TOSTREAM-0000000014
    Sink: KSTREAM-SINK-0000000013 (topic: count)
      <-- KSTREAM-MAP-0000000012
    Sink: KSTREAM-SINK-0000000016 (topic: json)
      <-- KSTREAM-MAP-0000000015
```

Po jakimś czasie ukazuje się taki widok:

```
IUCR:0841
Category: No matching category
District: 25.0
Aggregate: Category: No matching category, District: 25.000000, Date: 2007-03, All: 10, Arrest
: 0, Domestic: 0, FBI: 0
newValue: IUCR: 0841, Category: No matching category, District: 25.000000, Date: 2007-03, Arre
st: false, Domestic: false, IndexCode: null
---------------------------------Error-------------------------------------------------------
IUCR:0841
Category: No matching category
District: 25.0
Aggregate: Category: No matching category, District: 25.000000, Date: 2007-03, All: 11, Arrest
: 0, Domestic: 0, FBI: 0
newValue: IUCR: 0841, Category: No matching category, District: 25.000000, Date: 2007-03, Arre
st: false, Domestic: false, IndexCode: null
---------------------------------Error-------------------------------------------------------
IUCR:0842
Category: No matching category
District: 4.0
Aggregate: Category: No matching category, District: 4.000000, Date: 2007-03, All: 19, Arrest:
 0, Domestic: 0, FBI: 0
newValue: IUCR: 0842, Category: No matching category, District: 4.000000, Date: 2007-03, Arres
t: false, Domestic: false, IndexCode: null
---------------------------------Error-------------------------------------------------------
IUCR:0840
Category: No matching category
District: 19.0
Aggregate: Category: No matching category, District: 19.000000, Date: 2007-03, All: 5, Arrest:
 0, Domestic: 0, FBI: 0
newValue: IUCR: 0840, Category: No matching category, District: 19.000000, Date: 2007-03, Arre
st: false, Domestic: false, IndexCode: null
---------------------------------Error-------------------------------------------------------
IUCR:0840
Category: No matching category
District: 19.0
Aggregate: Category: No matching category, District: 19.000000, Date: 2007-03, All: 6, Arrest:
 0, Domestic: 0, FBI: 0
newValue: IUCR: 0840, Category: No matching category, District: 19.000000, Date: 2007-03, Arre
st: false, Domestic: false, IndexCode: null
```

Wynika to z faktu, że dane dostarczone w csv były niekompletne.

Z tego względu wyświetlam wszystkie numery IUCR, które nie występują w pliku Chicago_Police_Department_-_Illinois_Uniform_Crime_Reporting__IUCR__Codes.csv.

Przestępstwa zgłoszone z tymi numerami IUCR są kategoryzowane jako kategoria “No matching category”.

Na powyższym zdjęciu ekranu widać, że nie znaleziono w pliku csv IUCR o wartości 841, 842 oraz 841.

Oto lista wszystkich IUCR, które udało mi się odkryć, że nie występują w pliku csv:

- 840
- 841
- 842
- 499
- 5005
- 5008
- 9901

Miejsce utrzymywania obrazów czasu rzeczywistego – skrypt tworzący

Skrypt tworzący: create_db_and_db_connection.sh

Oto prawidłowy wynik działania skryptu:

```
Unable to find image 'mysql:debian' locally
debian: Pulling from library/mysql
1d5252f66ea9: Pull complete
b034716887fd: Pull complete
882448c7d618: Pull complete
c14ac4640b87: Pull complete
3d2978613e40: Pull complete
c33d24cb8c75: Pull complete
52fa14fc88fc: Pull complete
42deb22ec16d: Pull complete
c3d31d600a65: Pull complete
b0b9f5aba00d: Pull complete
98222a66de15: Pull complete
2ff8f0c3c67b: Pull complete
Digest: sha256:49f4fcb0087318aa1c222c7e8ceacbb541cdc457c6307d45e6ee4313f4902e33
Status: Downloaded newer image for mysql:debian
d33d107b0d99eb0555e35d00ccd8fe2467ce193573708493c0fda0d2d734d5aa
mysql: [Warning] Using a password on the command line interface can be insecure.
mysql: [Warning] Using a password on the command line interface can be insecure.
--2024-06-09 11:29:00--  https://repo1.maven.org/maven2/com/mysql/mysql-connector-j/8.0.33/mys
ql-connector-j-8.0.33.jar
Resolving repo1.maven.org (repo1.maven.org)... 199.232.192.209, 199.232.196.209, 2a04:4e42:4c:
:209, ...
Connecting to repo1.maven.org (repo1.maven.org)|199.232.192.209|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2481560 (2.4M) [application/java-archive]
Saving to: 'mysql-connector-j-8.0.33.jar'

mysql-connector-j-8.0.3 100%[============================>]   2.37M  --.-KB/s    in 0.1s

2024-06-09 11:29:01 (19.1 MB/s) - 'mysql-connector-j-8.0.33.jar' saved [2481560/2481560]

--2024-06-09 11:29:01--  https://packages.confluent.io/maven/io/confluent/kafka-connect-jdbc/1
0.7.0/kafka-connect-jdbc-10.7.0.jar
Resolving packages.confluent.io (packages.confluent.io)... 18.66.233.22, 18.66.233.59, 18.66.2
33.42, ...
```

```
Connecting to packages.confluent.io (packages.confluent.io)|18.66.233.22|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 275798 (269K) [application/octet-stream]
Saving to: 'kafka-connect-jdbc-10.7.0.jar'

kafka-connect-jdbc-10.7 100%[============================>] 269.33K  --.-KB/s    in 0.1s

2024-06-09 11:29:01 (2.52 MB/s) - 'kafka-connect-jdbc-10.7.0.jar' saved [275798/275798]

log4j.logger.org.reflections=ERROR
```

Skrypt uruchamiający nasłuch: connect_db.sh

Oto prawidłowy wynik działania skryptu:

```
[2024-06-09 11:30:33,175] WARN The configuration 'offset.flush.interval.ms' was supplied but i
sn't a known config. (org.apache.kafka.clients.admin.AdminClientConfig)
[2024-06-09 11:30:33,175] WARN The configuration 'key.converter.schemas.enable' was supplied b
ut isn't a known config. (org.apache.kafka.clients.admin.AdminClientConfig)
[2024-06-09 11:30:33,175] WARN The configuration 'offset.storage.file.filename' was supplied b
ut isn't a known config. (org.apache.kafka.clients.admin.AdminClientConfig)
[2024-06-09 11:30:33,176] WARN The configuration 'value.converter.schemas.enable' was supplied
 but isn't a known config. (org.apache.kafka.clients.admin.AdminClientConfig)
[2024-06-09 11:30:33,176] WARN The configuration 'plugin.path' was supplied but isn't a known
config. (org.apache.kafka.clients.admin.AdminClientConfig)
[2024-06-09 11:30:33,176] WARN The configuration 'value.converter' was supplied but isn't a kn
own config. (org.apache.kafka.clients.admin.AdminClientConfig)
[2024-06-09 11:30:33,176] WARN The configuration 'key.converter' was supplied but isn't a know
n config. (org.apache.kafka.clients.admin.AdminClientConfig)
[2024-06-09 11:30:33,718] WARN The configuration 'offset.flush.interval.ms' was supplied but i
sn't a known config. (org.apache.kafka.clients.admin.AdminClientConfig)
[2024-06-09 11:30:33,718] WARN The configuration 'key.converter.schemas.enable' was supplied b
ut isn't a known config. (org.apache.kafka.clients.admin.AdminClientConfig)
[2024-06-09 11:30:33,718] WARN The configuration 'offset.storage.file.filename' was supplied b
ut isn't a known config. (org.apache.kafka.clients.admin.AdminClientConfig)
[2024-06-09 11:30:33,718] WARN The configuration 'value.converter.schemas.enable' was supplied
 but isn't a known config. (org.apache.kafka.clients.admin.AdminClientConfig)
[2024-06-09 11:30:33,718] WARN The configuration 'plugin.path' was supplied but isn't a known
config. (org.apache.kafka.clients.admin.AdminClientConfig)
[2024-06-09 11:30:33,718] WARN The configuration 'value.converter' was supplied but isn't a kn
own config. (org.apache.kafka.clients.admin.AdminClientConfig)
[2024-06-09 11:30:33,718] WARN The configuration 'key.converter' was supplied but isn't a know
n config. (org.apache.kafka.clients.admin.AdminClientConfig)
Jun 09, 2024 11:30:34 AM org.glassfish.jersey.internal.inject.Providers checkProviderRuntime
WARNING: A provider org.apache.kafka.connect.runtime.rest.resources.RootResource registered in
 SERVER runtime does not implement any provider interfaces applicable in the SERVER runtime. D
ue to constraint configuration problems the provider org.apache.kafka.connect.runtime.rest.res
ources.RootResource will be ignored.
Jun 09, 2024 11:30:34 AM org.glassfish.jersey.internal.inject.Providers checkProviderRuntime
WARNING: A provider org.apache.kafka.connect.runtime.rest.resources.ConnectorsResource registe
red in SERVER runtime does not implement any provider interfaces applicable in the SERVER runt
ime. Due to constraint configuration problems the provider org.apache.kafka.connect.runtime.re
st.resources.ConnectorsResource will be ignored.
Jun 09, 2024 11:30:34 AM org.glassfish.jersey.internal.inject.Providers checkProviderRuntime
WARNING: A provider org.apache.kafka.connect.runtime.rest.resources.LoggingResource registered
 in SERVER runtime does not implement any provider interfaces applicable in the SERVER runtime
. Due to constraint configuration problems the provider org.apache.kafka.connect.runtime.rest.
resources.LoggingResource will be ignored.
Jun 09, 2024 11:30:34 AM org.glassfish.jersey.internal.Errors logErrors
WARNING: The following warnings have been detected: WARNING: The (sub)resource method listLogg
ers in org.apache.kafka.connect.runtime.rest.resources.LoggingResource contains empty path ann
otation.
WARNING: The (sub)resource method createConnector in org.apache.kafka.connect.runtime.rest.res
ources.ConnectorsResource contains empty path annotation.
WARNING: The (sub)resource method listConnectors in org.apache.kafka.connect.runtime.rest.reso
urces.ConnectorsResource contains empty path annotation.
WARNING: The (sub)resource method listConnectorPlugins in org.apache.kafka.connect.runtime.res
t.resources.ConnectorPluginsResource contains empty path annotation.
WARNING: The (sub)resource method serverInfo in org.apache.kafka.connect.runtime.rest.resource
s.RootResource contains empty path annotation.

[2024-06-09 11:30:34,771] WARN The configuration 'metrics.context.connect.kafka.cluster.id' wa
s supplied but isn't a known config. (org.apache.kafka.clients.consumer.ConsumerConfig)
```

Miejsce utrzymywania obrazów czasu rzeczywistego – cechy

Wybór MySQL jako bazy danych:

- Znajomość i Dojrzałość Technologii:

  MySQL jest dobrze znanym, dojrzałym systemem zarządzania bazami danych (DBMS), który jest szeroko stosowany w różnych aplikacjach. Dzięki temu mamy dostęp do dużej ilości dokumentacji, zasobów edukacyjnych oraz wsparcia społeczności.

- Wsparcie dla JDBC Connector:

  MySQL jest w pełni kompatybilny z JDBC, co umożliwia łatwe połączenie z Apache Kafka za pomocą Kafka Connect JDBC Sink Connector. To zapewnia bezproblemową integrację i transfer danych z tematów Kafki do bazy danych MySQL.

- Elastyczność i Skalowalność

  MySQL może być uruchamiany zarówno lokalnie, jak i w kontenerach Docker, co pozwala na elastyczne i łatwe skalowanie w zależności od potrzeb. Możliwość uruchomienia MySQL w kontenerze Docker sprawia, że środowisko jest łatwe do konfiguracji i zarządzania.

Wybór Kafka Connect:

- Integracja z Apache Kafka:

  Kafka Connect jest narzędziem, które umożliwia łatwą integrację różnych systemów źródłowych i docelowych z Apache Kafka. Wybór Kafka Connect pozwala na efektywne przesyłanie danych między Kafką a bazą danych MySQL.

Konsument: skrypt odczytujący wyniki przetwarzania

Skrypt odczytujący wszystkie dane z bazy danych: all_data.sh

Skrypt pozwala na odczytanie wszystkich danych z bazy danych.

Oto fragment przykładowego wyniku:

```
2543    THEFT   25      2001-05 229     43      11      229
2544    BATTERY 8       2001-05 313     65      26      53
2545    OTHER OFFENSE   4       2001-05 85      3       32      0
2546    BATTERY 1       2001-05 56      21      9       5
2547    LIQUOR LAW VIOLATION    10      2001-05 5       5       0       0
2548    BATTERY 6       2001-05 315     61      86      58
2549    THEFT   24      2001-05 150     29      3       150
2550    THEFT   19      2001-05 331     62      1       331
2551    MOTOR VEHICLE THEFT     3       2001-05 72      6       0       72
2552    BATTERY 5       2001-05 341     65      135     72
2553    OTHER OFFENSE   10      2001-05 63      20      20      0
2554    THEFT   10      2001-05 144     34      6       144
2555    MOTOR VEHICLE THEFT     10      2001-05 76      10      0       76
2556    THEFT   16      2001-05 194     23      1       194
2557    OTHER OFFENSE   7       2001-05 77      9       31      0
2558    MOTOR VEHICLE THEFT     11      2001-05 75      9       1       75
2559    BURGLARY        10      2001-05 43      1       1       43
```

Skrypt odczytujący wynik dla określonych parametrów: read_db.sh

Skrypt pozwala na odczytanie wyniku dla określonej kategorii, dystryktu oraz miesiąca określonych podczas wywoływania.

Oto przykładowy wynik:

```
Category: THEFT
District: 14
Date: 2001-04
Number of all crimes: 341
Number of crimes that ended up with an arrest: 35
Number of domestic crimes: 4
Number of crimes reported by FBI: 341
```

W przypadku, kiedy nie zostaną znalezione dane dla określonej kategorii, dystryktu oraz miesiąca wyświetli się następujący komunikat:

```
No data found for category THEFT and district 14 on date 2020-04
```