Hurtownie danych – Projekt – Wersja końcowa

PWr. Wydział Informatyki i Telekomunikacji Data: 14.06.2024

Student	----------	**Ocena** Deklaracja Studenta
Indeks	260465	Proszę wskazać: 3.0 | 3.5 | 4.0 | 4.5 | 5.0
Imię	Jan	
Nazwisko	Czop	

Terminarz przekazania dokumentacji i odbioru produktu (Hurtownia Danych + Analiza Danych)

Terminy zerowe (promocyjne):

- Data przekazania dokumentacji: **16.06.2024 r.**
- **Odbiór projektu:**
  - **Na ocenę dostateczną - Poniedziałek**
  - **Na ocenę co najmniej 3.5 - Środa**

Termin normalny:

- Data przekazania dokumentacji: środa **19.06.2024 r**.
- **Odbiór projektu**:
  - **21.06.2024 r.**
  - Poprawa + test zaliczeniowy: **25.06.2024 r**.

1. Tytuł projektu (wybranego do realizacji)

Analiza przestępstw (morderstw) w USA

- **Uzasadnienie wyboru**

Wybrałem ten temat, ponieważ po pierwsze wydaje mi się dużo ciekawszym tematem do analizy i bardziej stabilnym niż piłka nożna. Sądzę, że na podstawie statystyk piłkarskich nie da się za wiele wywnioskować, odpowiedzieć na pytania badawcze z pewną dozą predykcji, która pomoże nam rozpoznać i przebadać pewne zjawiska. Morderstwa, czy przestępstwa z kolei mogą być kopalnią wiedzy do analizy ludzkich zachowań. A nie ma chyba nic ciekawszego na Ziemi niż ludzie…

Oryginalny temat dotyczący przestępstw w Los Angeles porzuciłem przede wszystkim z powodu brakujących danych na temat sprawców. Brak możliwości analizy sprawcy wg mnie jest bardzo uciążliwym czynnikiem dla analizy wycinka rzeczywistości. Zmieniłem temat na pokrewny, morderstwa w USA, dotyczy on specyficznego rodzaju przestępstw co w mojej opinii jest zaletą, bo

nie trzeba się rozdrabniać na wiele podtematów dotyczących różnych rodzajów przestępstw. Analiza drobnych kradzieży kieszonkowych będzie na pewno inną sprawą niż sprawy takie jak strzelanina lub przypadku kanibalizmu.

2. Charakterystyka dziedziny problemowej

Analiza zjawiska morderstw w USA stanowi istotny obszar badań zarówno dla organów ścigania, jak i nauki społecznej. Dziedzina problemowa dotycząca morderstw obejmuje szeroki zakres czynników, które wpływają na występowanie, rozwiązanie oraz prewencję tego rodzaju przestępstw. Morderstwa stanowią jedno z najpoważniejszych przestępstw kryminalnych, generując duże zainteresowanie społeczne oraz wymagając skutecznej reakcji ze strony organów ścigania i wymiaru sprawiedliwości. W USA, kwestie związane z morderstwami są regulowane przez różne przepisy prawne na poziomie federalnym, stanowym oraz lokalnym. Tego typu czyny mogą mieć różne motywacje i konteksty, obejmując m.in. przestępstwa z nienawiści, przestępstwa związane z przestępczością zorganizowaną, przestępstwa z zemsty, czy też przypadki zabójstw w ramach konfliktów rodzinnych. Istnieją także różnice kulturowe, społeczne i ekonomiczne, które mogą wpływać na dynamikę zjawiska morderstw w różnych regionach USA.

2.1 Opis obszaru analizy (wybrany fragment dziedziny, przeznaczony do szczegółowej analizy i opracowania hurtowni danych)

W ramach projektu hurtowni danych, obszarem szczegółowej analizy jest zjawisko morderstw w Stanach Zjednoczonych Ameryki. Stanowi ono centralny punkt zainteresowania ze względu na swoje znaczenie społeczne, kryminalistyczne i prawnicze. Analiza tego obszaru ma na celu zrozumienie głębszych mechanizmów, które leżą u podstaw tego zjawiska oraz identyfikację czynników wpływających na jego występowanie, rozwiązanie i prewencję. Morderstwa to zjawisko wielowymiarowe, które odzwierciedla różnorodne aspekty społeczeństwa. Analiza tego obszaru może dostarczyć wglądu w strukturę społeczną, relacje międzygrupowe, dynamikę przestępczości oraz skuteczność działań prewencyjnych i śledczych. Projekt hurtowni danych może przyczynić się do identyfikacji trendów w występowaniu morderstw w różnych regionach i grupach społecznych, co może być kluczowe dla opracowywania polityk publicznych i działań prewencyjnych. Projekt hurtowni danych mógłby przyczynić się do lepszego zrozumienia mechanizmów śledczych i procesów sądowych związanych z morderstwami. Analiza danych dotyczących rozwiązania spraw może ujawnić czynniki wpływające na skuteczność dochodzeń i procesów sądowych oraz identyfikację najlepszych praktyk w tym zakresie, w rezultacie analiza może przyczynić się

do rozwoju skuteczniejszych strategii prewencyjnych, poprawy efektywności działań policyjnych i śledczych oraz zwiększenia bezpieczeństwa społeczności.

2.2 Problemy

- Identyfikacja trendów przestępczych
- Wzorce zbrodni w zależności od lokalizacji
- Skuteczność działań policyjnych
- Czynniki demograficzne ofiar i sprawców
- Motywacje i relacje między ofiarami a sprawcami
- Wpływ innych czynników kryminalistycznych

2.3 Cel przedsięwzięcia

2.3.1 Oczekiwania i potrzeby w zakresie wsparcia podejmowania decyzji

- Identyfikacja obszarów wysokiej i niskiej stawki przestępczości
- Analiza trendów przestępczości
- Badanie związków między demografią a przestępczością
- Opracowanie strategii prewencyjnych i interwencyjnych

2.3.2 Zakres analizy – pytania badawcze

- Czy istnieją określone trendy w popełnianiu morderstw, takie jak wzrost lub spadek liczby przestępstw w określonych regionach lub czasie?
- Czy istnieją różnice w typach morderstw w różnych stanach?
- Czy niektóre obszary są bardziej narażone na określone rodzaje przestępstw niż inne?
- Czy można zidentyfikować pewne wzorce związane z wiekiem, płcią, rasą lub grupą etniczną ofiar i sprawców morderstw?
- Czy analiza relacji między ofiarami a sprawcami może rzucić światło na motywacje popełnienia przestępstwa?

2.3.3 Potencjalni użytkownicy

- Organizacje ścigania.
- Decydenci policyjni.
- Badacze naukowi.
- Organizacje pozarządowe i społeczne.
- Działy administracyjny USA.

3. Dane źródłowe

3.1. Źródła danych

Charakterystyka pliku zawierający danę źródłowe przeznaczone do stworzenia tematycznej hurtowni danych jest przedstawiona w tab. 1.

Tabela 1. Zbiory danych źródłowych

Lp.	Plik	Typ	Liczba rek.	Rozmiar[MB]	Opis
1.	database.csv	csv	Ok. 393 000	118.1	Zawiera rekordy morderstw

3.2. Lokalizacja, dostępność danych źródłowych

Dane dostępne są za darmo dla wszystkich użytkowników w formacie CSV pod linkiem: https://www.kaggle.com/datasets/murderaccountability/homicide-reports

3.3. Słownik danych – interpretacja

Interpretacja oraz wyjaśnienie znaczeń pojęć dziedzinowych zostały zawarte w tab.2.

Tabela 2. Słownik atrybutów

Plik: lap_times				
Lp.	**Atrybut**	**Typ danych**	**Znaczenie**	**Uwagi**
1.	Record ID	Numeryczny	Identyfikator morderstwa	>0
2.	Agency Code	Numeryczny	Kod agencji policji	
3.	Agency Name	Tekstowy	Nazwa agencji policji	
4.	Agency Type	Tekstowy	Rodzaj agencji policji	Enumerator
5.	City	Tekstowy	Miasto	
6.	State	Tekstowy	Stan	
7.	Year	Numeryczny	Rok	
8.	Month	Numeryczny	Miesiąc	
9.	Incident	Numeryczny	Ilość incydentów	
10.	Crime Type	Tekstowy	Rodzaj morderstwa	Enumerator
11.	Crime Solved	Boolean	Czy rozwiązano sprawę?	Enumerator
12.	Victim Sex	Tekstowy	Płeć ofiary	Enumerator
13.	Victim Age	Numeryczny	Wiek ofiary	
14.	Victim Race	Tekstowy	Rasa ofiary	Enumerator
15.	Victim Ethnicity	Tekstowy	Grupa eniczna ofiary	Enumerator
16.	Perpetrator Sex	Tekstowy	Płeć sprawcy	Enumerator
17.	Perpetrator Age	Numeryczny	Wiek sprawcy	
18.	Perpetrator Race	Tekstowy	Rasa sprawcy	Enumerator
19.	Perpetrator Ethnicity	Tekstowy	Grupa eniczna sprawcy	Enumerator

20.	Victim count	Numeryczny	Ilość ofiar	
21.	Perpetrator count	Numeryczny	Ilość sprawców	
22.	Relationship	Teskstowy	Relacje ofiara-sprawca	
23.	Weapon	Tesktowy	Broń użyta	
24.	Recorded source	Tekstowy	Źródło rekordu	Enumerator

3.4. Ocena jakościowa danych

Wynik analizy jakościowej przeprowadzonej za pomocą programu Tableau oraz profilu danych SSIS został przedstawiony w tab. 3.

Tabela 3. Ocena jakościowa danych

Plik: database.csv				
Lp.	**Atrybut**	**Typ danych**	**Zakres wartości**	**Uwagi - ocena jakości danych**
1.	Record ID	Numeryczny	638k	Nieprzydatna do celów analitycznych kolumna
2.	Agency Code	Numeryczny	12k unikalnych rek.	~~Nieprzydatna do celów analitycznych kolumna~~
3.	Agency Name	Tekstowy	9216k unikalnych rek.	
4.	Agency Type	Tekstowy	7 rodzajów rek.	
5.	City	Tekstowy	1782 unikalnych rek.	
6.	State	Tekstowy	51 stanów	
7.	Year	Numeryczny	1980-2014	
8.	Month	Numeryczny	12 miesięcy	
9.	Incident	Numeryczny	0-999	
10.	Crime Type	Tekstowy	2 rodzaje rek.	
11.	Crime Solved	Boolean	Yes albo No	

12.	Victim Sex	Tekstowy	3 rodzaje rek.	
13.	Victim Age	Numeryczny	0-100	
14.	Victim Race	Tekstowy	5 rodzajów rek.	
15.	Victim Ethnicity	Tekstowy	3 rodzaje rek.	
16.	Perpetrator Sex	Tekstowy	3 rodzaje rek.	
17.	Perpetrator Age	Numeryczny	0-100	
18.	Perpetrator Race	Tekstowy	5 rodzajów rek.	
19.	Perpetrator Ethnicity	Tekstowy	3 rodzaje rek.	
20.	Relationship	Tekstowy	28 unikalnych rek.	
21.	Weapon	Tekstowy	16 unikalnych rek.	
22.	Victim count	Numeryczny	0-10	
23.	Perpetrator count	Numeryczny	0-10	
24.	Recorded source	Tekstowy	2 rodzaje rek.	

W bazie nie ma praktycznie nulli.

4. Analityczne modele wielowymiarowe

4.1. Fakty **podlegające analizie oraz ich miary**

Analizie będzie podlegał zbiór zarejestrowanych zdarzeń (tab. 4.)

Tabela 4. Fakty podlegające analizie

Lp.	Fakty	Miary	Uwagi
1.	**Morderstwo**	Incident, Victim Count,Pepetrator count, Crime solved	**Crime_solved to BIT**

4.2. Kontekst analizy faktów

Ustalony kontekst analizy faktów został przedstawiony w tab. 5.

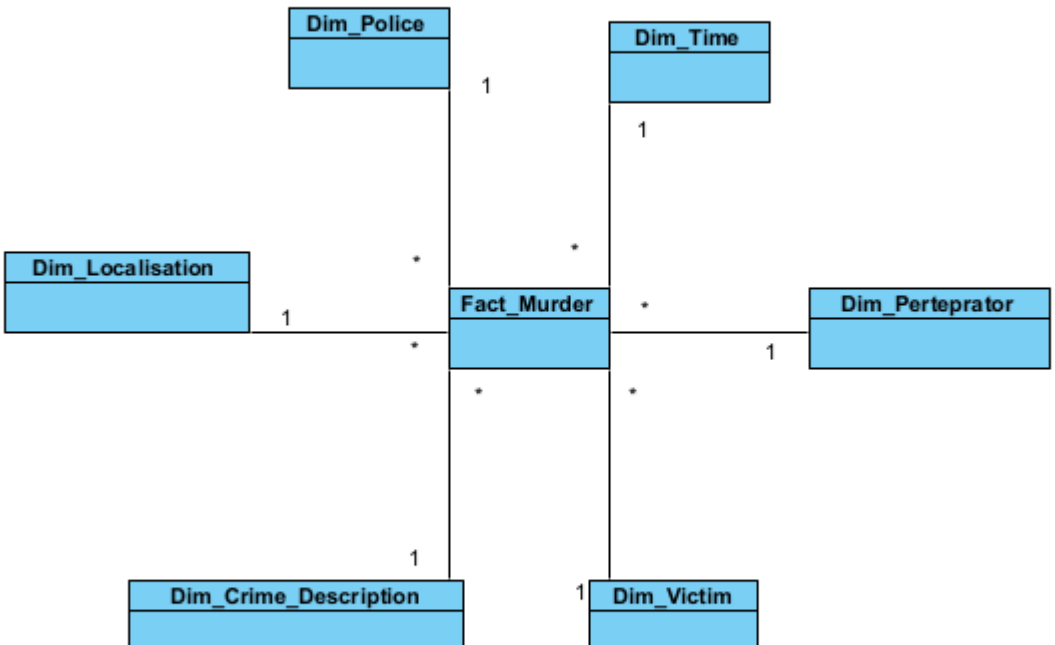
Tabela 5. Wymiary analizy faktów

Lp.	Wymiar	Atrybuty	Uwagi
1.	Policja	Agency Name, Agency Type, Recorded Source	
2.	Lokalizacja	City, State	
3.	Czas	Year, Month	
4.	Opis zbrodni	Crime Type, Relationship, Weapon	
5.	Ofiara	Victim Sex, Victim Age, Victim Race, Victim Ethnicity	
6.	Sprawca	Perpetrator Sex, Perpetrator Age, Perpetrator Race, Perpetrator Ethnicity	

4.3. Modele wielowymiarowe (UML)

Po przeanalizowaniu atrybutów źródła danych oraz ustalonego faktu/faktów i kontekstu analizy zaproponowano wielowymiarowy model konceptualny (rys. 1.). Składa się on z faktów: oraz … wymiarów. Model ten reprezentowany jest w postaci schematu na rys. 1. ?.

Rysunek 1. Wielowymiarowy model analityczny przedstawiony na poziomie konceptualnym



Model na rysunku 1. To tzw. Płatek śniegu (ang. Snowflake).

5. Projekt procesu ETL

5.1. Schemat bazy danych HD (skrypt SQL)

Stworzyłem tabele, za pośrednictwem Tableau Prep, która przygotowała mi wszystkie kolumny mojej tabeli źródłowej dla mojego pliku z danymi. Używam go jako tabeli-poczekalni.

```
CREATE TABLE [dbo].[Addon_main](
    [Record_ID] [bigint] NULL,
    [Agency_Code] [nvarchar](4000) NULL,
    [Agency_Name] [nvarchar](4000) NULL,
    [Agency_Type] [nvarchar](4000) NULL,
    [City] [nvarchar](4000) NULL,
    [State] [nvarchar](4000) NULL,
    [Year] [bigint] NULL,
    [Month] [nvarchar](4000) NULL,
    [Incident] [bigint] NULL,
    [Crime_Type] [nvarchar](4000) NULL,
    [Crime_Solved] [nvarchar](4000) NULL,
    [Victim_Sex] [nvarchar](4000) NULL,
```

```
    [Victim_Age] [bigint] NULL,

    [Victim_Race] [nvarchar](4000) NULL,

    [Victim_Ethnicity] [nvarchar](4000) NULL,

    [Perpetrator_Sex] [nvarchar](4000) NULL,

    [Perpetrator_Age] [bigint] NULL,

    [Perpetrator_Race] [nvarchar](4000) NULL,

    [Perpetrator_Ethnicity] [nvarchar](4000) NULL,

    [Relationship] [nvarchar](4000) NULL,

    [Weapon] [nvarchar](4000) NULL,

    [Victim_Count] [bigint] NULL,

    [Perpetrator_Count] [bigint] NULL,

    [Record_Source] [nvarchar](4000) NULL

)
```

Stworzyłem tabele przechowującą dane gotowe do załadowania i obróbki w kostce:

```
CREATE TABLE Dim_Police (
    Agency_Code IDENTITY(1,1) INT PRIMARY KEY,
    Agency_Name NVARCHAR(255), NOT NULL
    Agency_Type NVARCHAR(255), NOT NULL
    Record_Source NVARCHAR(255) NOT NULL
);
CREATE TABLE DimLocation (
    City_ID INT PRIMARY KEY IDENTITY,
    City NVARCHAR(255),
    State NVARCHAR(255)
);
CREATE TABLE DimTime (
    Time_ID INT IDENTITY(1,1 PRIMARY KEY IDENTITY,
    Year INT,
    Month NVARCHAR(50)
);
CREATE TABLE Dim_Crime_Description (
    Crime_Description_ID INT IDENTITY(1,1 PRIMARY KEY IDENTITY,
    Crime_Type NVARCHAR(255),
    Relationship NVARCHAR(255),
    Weapon NVARCHAR(255)
);
CREATE TABLE Dim_Victim (
    Victim_ID INT IDENTITY(1,1 PRIMARY KEY IDENTITY,
    Victim_Sex NVARCHAR(50),
    Victim_Age INT,
    Victim_Race NVARCHAR(50),
```

```
    Victim_Ethnicity NVARCHAR(50)
);
CREATE TABLE Dim_Perpetrator (
    Perpetrator_ID INT IDENTITY(1,1 PRIMARY KEY IDENTITY,
    Perpetrator_Sex NVARCHAR(50),
    Perpetrator_Age INT,
    Perpetrator_Race NVARCHAR(50),
    Perpetrator_Ethnicity NVARCHAR(50)
);
CREATE TABLE Fact_Murder (
    Record_ID INT IDENTITY(1,1 PRIMARY KEY,
    Incident INT,
    Victim_Count INT,
    Perpetrator_Count INT,
    Crime_Solved BIT,
    Police_ID INT,
    City_ID INT,
    Time_ID INT,
    Crime_Description_ID INT,
    Victim_ID INT,
    Perpetrator_ID INT,
    FOREIGN KEY (Police_ID) REFERENCES Dim_Police(Police_ID),
    FOREIGN KEY (City_ID) REFERENCES Dim_Location(City_ID),
    FOREIGN KEY (Time_ID) REFERENCES Dim_Time(Time_ID),
    FOREIGN KEY (Crime_Description_ID) REFERENCES
Dim_Crime_Description(Crime_Description_ID),
    FOREIGN KEY (Victim_ID) REFERENCES Dim_Victim(Victim_ID),
    FOREIGN KEY (Perpetrator_ID) REFERENCES Dim_Perpetrator(Perpetrator_ID)
       );
```

Stworzyłem podobne tabele na tymczasowe dane, które po opróżnieniu z powtórek danych będą wstawiane do powyższych i czyszczone.

Schematy tabel są identyczne co powyżej, poza oczywiście nazwą zaopatrzoną w postfikx ‘_Temp’ np.: Dim_Police_Temp albo Fact_Murder_Temp.

Wyjątek: wyjątkiem jest tabela Fact_Murder_Temp gdzie początkowo atrybut Crime_solved jest NVARCHAR(16) i za pomocą odpowiedniego skryptu (będzie wyszczególniony w następnym punkcie) zmieniany jes na BIT.

5.2. Specyfikacja procesów ETL (Control Flow + Data Flow)

Rysunek 2. Ogólny widok procesów ETL

Control Flow podzielony jest na 3 etapy realizowane dla każdego wymiaru oraz faktu:

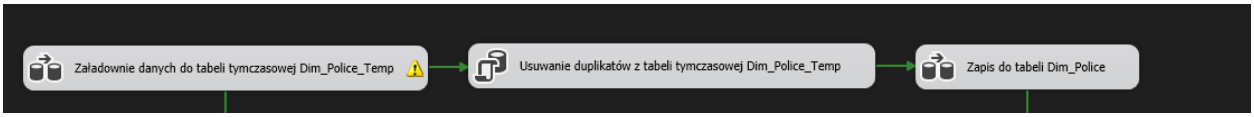
- Extract: czyli pozyskanie danych ze źródła

- Transform: czyli przekształcanie danych na takie, które nam „pasują”

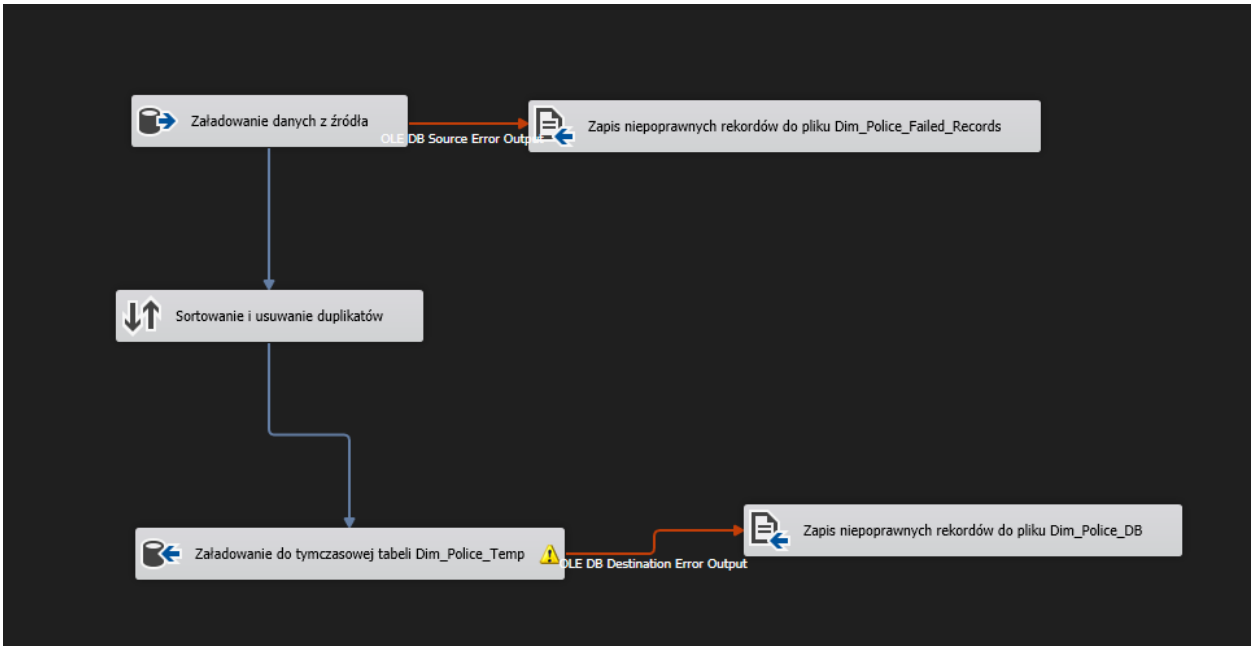
- Load: czyli załadowanie do końcowej bazy danych (w tym wypadku hurtowni danych)

Każdy wymiar ma zasadniczo identyczną ścieżkę procesów do przebycia, także przedstawię jedynie jeden wymiar jako reprezentację:

Rysunek 3. Ścieżka procesów ETL pojedycznego wymiaru (tutaj Dim_Police).



Rysunek 4. Załadowanie danych do tabeli tymczasowej Dim_Police_Temp.



- Zaczynymy od załadowania z tabeli poczekalni i zmapowania odpowiednich kolumn.

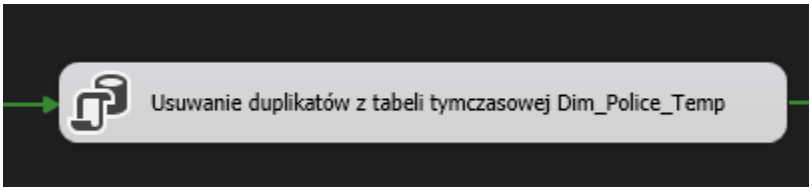
- Następnie sortujemy i usuwamy duplikaty z kolumn wyszczególnionych dla danego wymiaru.

Uwaga: Sortowanie jest bardzo uciążliwe dla komputera (szczególnie dla mojego starego laptopa) i ma problemy przy większych ilościach danych (u mnie powyżej 50 000 rekordów na raz). Można byłoby użyć alternatywnie odpowiedniej funkcji okienkowej. Przyczyną dlaczego tego nie zrobiłem jest fakt, że Visual Studio SSIS jeśli zamieniamy komponent (w tym wypadku Sort) połączenia zostawiają w sobie śmieci i uporządkowanie tego i doprowadzenie do braku błędów i zawirowań w trakcie procesów czasem prowadzi do usunięcia wszystkiego w projekcie i zrobienia tego od nowa. Ja swoje dane po prostu przetwarzalem partiami po 50k rekordów.

- Ładujemy dane do tabeli tymczasowej odpowiedniej dla danego wymiaru.

- Podczas wczytywania ze źródła oraz zapisywania w tabeli tymczasowej mam również zdefiniowane obsługi błędów, które (rekordy wraz z opisami kodów błędów) trafiają do opisywanych plików. W ten sposób pozbywamy się „intruzów”, w razie błędów wczytywania rekordów oraz ich zapisywania (błędnych).

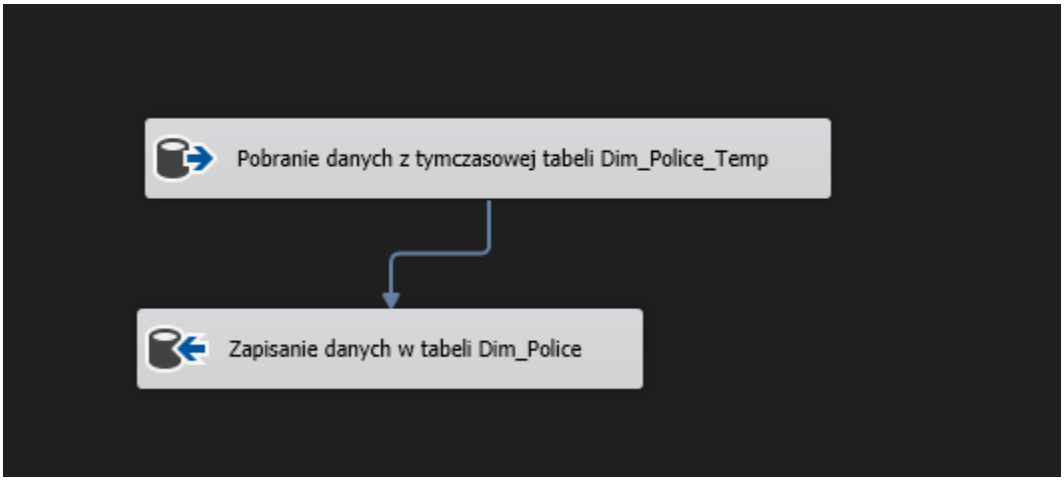
Rysunek 5. Usuwanie duplikatów z tabeli tymczasowej Dim_Police_Temp.



Kolejnym etapem jest usunięcie rekordów, które już istnieją (mają takie same dane atrybutów, bo klucz jest autoinkrementowany dla każdego faktu i wymiaru). W tym celu przygotowałem prosty skrypt odpowiedni dla każdego wymiaru:

```
DELETE tmp FROM [dbo].[Dim_Police_Temp] tmp INNER JOIN [dbo].[Dim_Police] p ON
p.Agency_Code = tmp.Agency_Code AND p.Agency_Name = tmp.Agency_Name AND
p.Agency_Type = tmp.Agency_Type AND p.Record_Source = tmp.Record_Source;
```

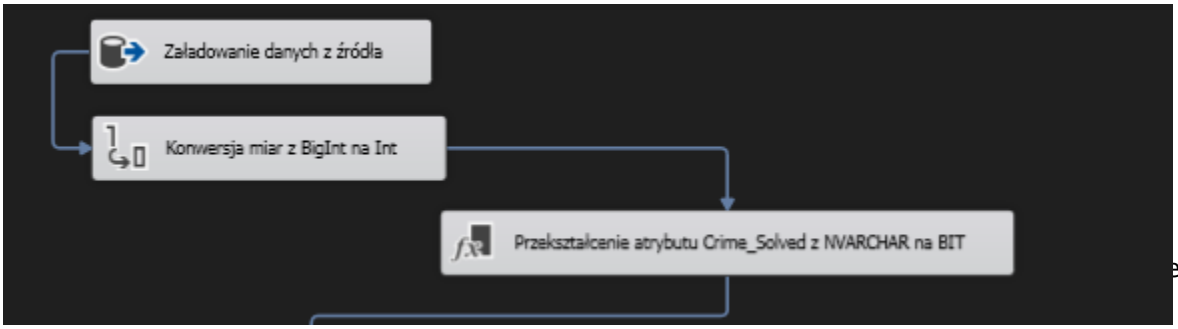
Rysunek 5. Zapis do tabeli Dim_Police.



Tutaj po prostu zasilamy już poprawnymi danymi tabele docelową zawartością tej tymczasowej.

Dla faktu mój proces ETL przebiega w bardzo podobnych warunkach poza pierwszym procesem czyli Extracem:

Rysunek 6. Pierwsza część procesu wczytywania danych i załadowania do tabeli tymczasowej Fact_Murder_Temp.



BigInt
ej z

- Następnie przekształcam atrybut Crime_Solved z NVARCHAR(16) na BIT za pomocą tego małego skryptu:

```
(DT_BOOL)(Crime_Solved == "Yes" ? 1 : 0)
```

Przekształcenie jest podyktowane specyfiką danych które posiadają jedynie rekordy „Yes” oraz „No” w swoim zbiorze.

Rysunek 7. Druga część procesu wczytywania danych i załadowania do tabeli tymczasowej Fact_Murder_Temp.

- Następnie używam Lookupów do wyłuskania odpowiednich ID dla konkretnych wymiarów z gotowych już tabel wymiar (tych docelowych).

- Ostatnią czynnością jest sorotwanie faktów i usuwanie powtórek.

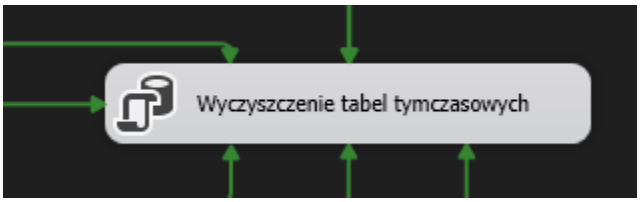
- Gotowe dane wrzucamy do tabeli tymczasowej Fact_Murder_Temp.

- Dla każdego Lookupa mam odpowiedni transfer rekordów, które nie znalazły swojego przedstawiciela do odpowiednich plików. Co do tej kwestii mam kilka zastrzeżeń w Punkcie PROBLEMY.

- Ochrona przeciw złym rekordom i zapisom w bazie danych tak jak dla wymiarów (opisane powyżej).

Przepływ danych od wymiarów do faktu jest oczywiście podyktowany specyfiką budowy faktu, który potrzebuje gotowych informacji na temat budowanych wymiarów.

Rysunek 7. Wyczyszczenie tabel tymczasowych.



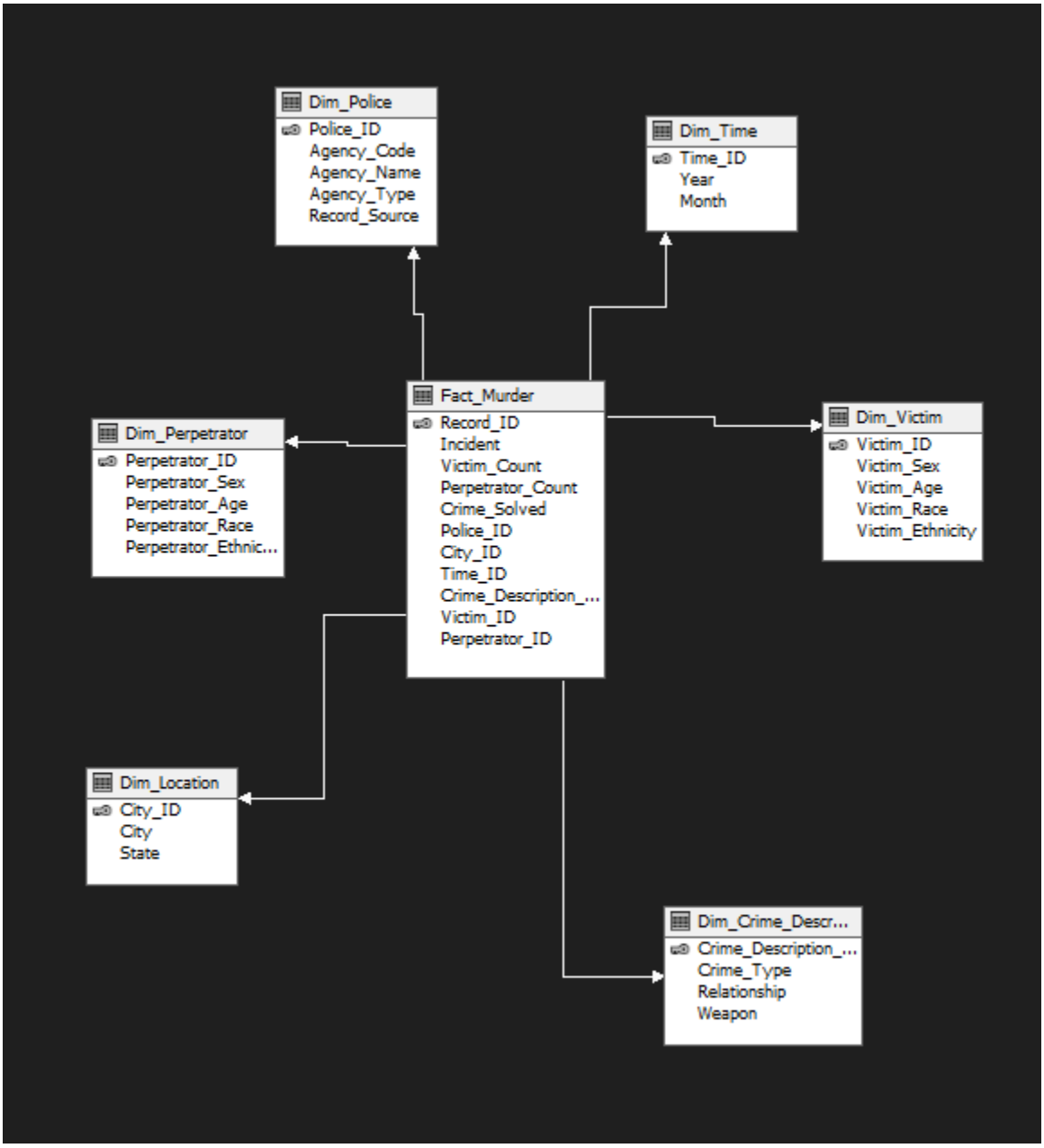
Na sam koniec usuwam dane z tabel tymczasowych skryptem żeby żadne niepotrzebne rekordy nie dostały się nieopatrznie do „krwiobiegu hurtowni”:

```
DELETE FROM Dim_Police_Temp

DELETE FROM Dim_Location_Temp;

DELETE FROM Dim_Time_Temp;

DELETE FROM Dim_Crime_Description_Temp;

DELETE FROM Dim_Victim_Temp;

DELETE FROM Dim_Perpetrator_Temp;

DELETE FROM Fact_Murder_Temp;
```

6. Implementacja modeli wielowymiarowych

6.1. Widok danych (Data Source View)

Rysunek 8. Widok danych (Data Source View).



6.2. Wymiary (kopia ekranu: Atrybuty + Hierarchie)

Nie używałem hierarchii w analizie danych z kostki.

6.3. Modele wielowymiarowe – Kostki (kopia ekranu: Miary + Wymiary + Diagram)

Rysunek 9. Kostka.