HURTOWNIE DANYCH

DOKUMENTACJA PROJEKTU

Paweł Wozignój

Spis treści

[1 Baza źródłowa 2](#_Toc188534690)

[1.1 Przedstawienie bazy źródłowej 2](#_Toc188534691)

[1.2 Opis przyjętej strategii wypełniania bazy źródłowej danymi 3](#_Toc188534692)

[2 Hurtownia danych 3](#_Toc188534693)

[2.1 Model pojęciowy 3](#_Toc188534694)

[2.2 Model logiczny hurtowni w schemacie gwiazdy 4](#_Toc188534695)

[2.3 Model logiczny hurtowni w schemacie płatka śniegu 5](#_Toc188534696)

[2.4 Skrypty SQL zakładające hurtownie danych 6](#_Toc188534697)

[3 ETL z bazy źródłowej do hurtowni 8](#_Toc188534698)

[3.1 Mapa danych 8](#_Toc188534699)

[3.2 Opis procesu ETL 9](#_Toc188534700)

[3.3 Użyte skrypty SQL 11](#_Toc188534701)

[3.4 Ekrany ilustrujące poprawne działanie projektów SSIS i ekrany przedstawiające wstawienie danych do hurtowni 12](#_Toc188534702)

[4 Kostki OLAP 21](#_Toc188534703)

[4.1 Ekrany pokazujące zawartość opublikowanych kostek OLAP 21](#_Toc188534704)

[4.2 Przykładowe analizy danych wykonane w Excel na podstawie kostek OLAP 23](#_Toc188534705)

# Baza źródłowa

Bazą danych źródłową była relacyjna baza danych Microsoft Access o nazwie **"Baza Finanse i Czas.accdb"**. W ramach projektu baza została rozszerzona oraz wzbogacona o przykładowe dane zgodnie z założeniami:

**Rozszerzenie funkcjonalności bazy danych**  
Do istniejącej bazy danych dodano funkcjonalność umożliwiającą ewidencję czasu spędzanego na różnorodnych aktywnościach. Zmiany obejmowały:

* Utworzenie tabeli dedykowanej przechowywaniu danych o czasie (np. "Czas\_spędzony"), zawierającej takie pola, jak: identyfikator użytkownika, data, czas rozpoczęcia i zakończenia aktywności, rodzaj aktywności.
* Powiązanie nowych tabel z istniejącą strukturą bazy danych w sposób zapewniający integralność relacji.
* Zoptymalizowanie bazy w celu efektywnego przechowywania danych i ich późniejszej analizy.

**Wprowadzenie przykładowych danych**  
Baza została wypełniona przykładowymi danymi, które obejmowały:

* **Przychody** – różne źródła przychodów, takie jak wynagrodzenie, premie czy dodatkowe zarobki.
* **Wydatki** – szczegółowe informacje dotyczące kosztów, w tym wydatków na zakupy, usługi czy inne potrzeby.
* **Czas spędzony** – dane o aktywnościach użytkowników, uwzględniające rodzaj aktywności (praca, nauka, hobby itp.), czas jej trwania oraz daty.

## Przedstawienie bazy źródłowej

Obraz zawierający tekst, diagram, Plan, zrzut ekranu

Opis wygenerowany automatycznie

## Opis przyjętej strategii wypełniania bazy źródłowej danymi

Wypełnianie bazy źródłowej danymi zostało zrealizowane przez firmę zajmującą się analizą opłacalności inwestycji. Dane zostały zgromadzone w celu oceny potencjalnej rentowności budowy nowej galerii handlowej w określonej lokalizacji.

Proces wypełniania bazy obejmował gromadzenie informacji o przychodach, wydatkach oraz czasie spędzanym przez potencjalnych klientów na aktywnościach w istniejących obiektach handlowych oraz w ich codziennym życiu

Okres, za który zgromadzono dane, obejmował przedział czasu od **30 czerwca 2019 r. do 30 czerwca 2020 r.**

# Hurtownia danych

## Model pojęciowy

**Przyjęte pojęcia**  
Kluczowe pojęcia to:

* **Typ Przychodu** – rodzaje przychodów, np. wynagrodzenie, premie.
* **Kategoria Wydatku** – podział wydatków, uwzględniający typy i kategorie (np. zakupy, usługi).
* **Czas** – wymiar czasowy obejmujący dni, miesiące, lata i nazwy dni tygodnia.
* **Miejscowość** – informacje o lokalizacjach, w tym typ miejscowości (np. miasto, wieś).
* **Forma Płatności** – rodzaje płatności, takie jak gotówka, karta kredytowa.
* **Osoba** – dane osobowe związane z analizowanymi rekordami.

**Planowane analizy**  
Hurtownia umożliwia realizację analiz takich jak:

* Analiza przychodów i wydatków w różnych okresach czasu.
* Porównanie preferencji zakupowych w zależności od lokalizacji i typu miejscowości.
* Analiza form płatności wykorzystywanych przez różne grupy osób.
* Wyznaczenie trendów sezonowych w przychodach, wydatkach i aktywnościach.

**Decyzje wspierane przez hurtownię**  
Hurtownia ma na celu wspomaganie decyzji dotyczących inwestycji, w tym:

* Czy warto wybudować centrum handlowe w analizowanej lokalizacji.
* Ocena rentowności inwestycji na podstawie analizy preferencji konsumenckich.
* Ustalenie najbardziej opłacalnych typów usług i produktów dla danej miejscowości.

## Model logiczny hurtowni w schemacie gwiazdy

Obraz zawierający tekst, diagram, Równolegle, Czcionka

Opis wygenerowany automatycznie

## Model logiczny hurtowni w schemacie płatka śniegu

Obraz zawierający tekst, Czcionka, diagram, Równolegle

Opis wygenerowany automatycznie

## Skrypty SQL zakładające hurtownie danych

CREATE DATABASE DWH

GO

USE DWH

GO

-- Tabela wymiaru: Typ Przychodu

CREATE TABLE dTypPrzychodu (

dTPr\_id INT IDENTITY(1,1) PRIMARY KEY,

dTPr\_nazwa VARCHAR(20) NOT NULL

);

-- Tabela wymiaru: Kategoria Wydatku

CREATE TABLE dKategoriaWydatku (

dKWy\_id INT IDENTITY(1,1) PRIMARY KEY,

dKWy\_nazwa VARCHAR(20) NOT NULL,

dKWy\_typ VARCHAR(20) NOT NULL

);

-- Tabela wymiaru: Czas

CREATE TABLE dCzas (

dCz\_id INT IDENTITY(1,1) PRIMARY KEY,

dCz\_dzien DATE NOT NULL,

dCz\_miesiac DATE NOT NULL,

dCz\_rok INT NOT NULL,

dCz\_dzien\_tygodnia INT NOT NULL,

dCz\_nazwa\_dt VARCHAR(50) NOT NULL

);

-- Tabela wymiaru: Miejscowość

CREATE TABLE dMiejscowosc (

dM\_id INT IDENTITY(1,1) PRIMARY KEY,

dM\_nazwa VARCHAR(30),

dM\_typ\_miejscowosci VARCHAR(20)

);

-- Tabela wymiaru: Forma Płatności

CREATE TABLE dFormaPlatnosci (

dFP\_id INT IDENTITY(1,1) PRIMARY KEY,

dFP\_nazwa VARCHAR(20)

);

-- Tabela wymiaru: Osoba

CREATE TABLE dOsoba (

dOs\_id INT IDENTITY(1,1) PRIMARY KEY,

dOs\_id\_indeks VARCHAR(10),

dOs\_nazwa VARCHAR(80),

dOs\_plec CHAR(1),

dOs\_data\_ur DATE,

dOs\_rocznik AS YEAR([dOs\_data\_ur]),

dOs\_wiek AS YEAR(GETDATE()) - YEAR([dOs\_data\_ur]),

dOs\_miejscowosc\_miesz VARCHAR(30),

dOs\_miejscowosc\_studia VARCHAR(30),

dOs\_typ\_miejscowosc\_miesz VARCHAR(30),

dOs\_typ\_miejscowosc\_studia VARCHAR(30)

);

-- Tabela wymiaru: Kategoria Czasu

CREATE TABLE dKategoriaCzasu (

dKCz\_id INT IDENTITY(1,1) PRIMARY KEY,

dKCz\_nazwa VARCHAR(20) NOT NULL,

dKCz\_typ VARCHAR(20) NOT NULL

);

-- Tabela faktów: Przychód

CREATE TABLE fPrzychod (

fPR\_id INT IDENTITY(1,1) PRIMARY KEY,

fPR\_dOs\_id INT,

fPR\_dTPR\_id INT,

fPR\_dCz\_id INT,

fPR\_kwota DECIMAL(9, 2),

fPR\_czas\_trwania DECIMAL(9, 2),

fPR\_liczba INT,

FOREIGN KEY (fPR\_dOs\_id) REFERENCES dOsoba(dOs\_id),

FOREIGN KEY (fPR\_dTPR\_id) REFERENCES dTypPrzychodu(dTPr\_id),

FOREIGN KEY (fPR\_dCz\_id) REFERENCES dCzas(dCz\_id)

);

-- Tabela faktów: Wydatek

CREATE TABLE fWydatek (

fW\_id INT IDENTITY(1,1) PRIMARY KEY,

fWy\_dCz\_id INT,

fWy\_dOs\_id INT,

fWy\_dKWy\_id INT,

fWy\_dMi\_id INT,

fWy\_dFP\_id INT,

fW\_kwota DECIMAL(9, 2),

fW\_liczba INT,

FOREIGN KEY (fWy\_dOs\_id) REFERENCES dOsoba(dOs\_id),

FOREIGN KEY (fWy\_dKWy\_id) REFERENCES dKategoriaWydatku(dKWy\_id),

FOREIGN KEY (fWy\_dFP\_id) REFERENCES dFormaPlatnosci(dFP\_id),

FOREIGN KEY (fWy\_dMi\_id) REFERENCES dMiejscowosc(dM\_id),

FOREIGN KEY (fWy\_dCz\_id) REFERENCES dCzas(dCz\_id)

);

-- Tabela faktów: Czas Spędzony

CREATE TABLE fCzasSpedzony (

fCS\_id INT IDENTITY(1,1) PRIMARY KEY,

fCS\_dOs\_id INT,

fCS\_dCz\_id INT,

fCS\_dMi\_id INT,

fCS\_dKCz\_id INT,

fCS\_czas\_trwania DECIMAL(5, 2),

FOREIGN KEY (fCS\_dOs\_id) REFERENCES dOsoba(dOs\_id),

FOREIGN KEY (fCS\_dCz\_id) REFERENCES dCzas(dCz\_id),

FOREIGN KEY (fCS\_dMi\_id) REFERENCES dMiejscowosc(dM\_id),

FOREIGN KEY (fCS\_dKCz\_id) REFERENCES dKategoriaCzasu(dKCz\_id)

);

# ETL z bazy źródłowej do hurtowni

## Mapa danych

Obraz zawierający tekst, linia, zrzut ekranu, numer

Opis wygenerowany automatycznie

Obraz zawierający tekst, linia, zrzut ekranu, numer

Opis wygenerowany automatycznie

Obraz zawierający tekst, linia, zrzut ekranu, numer

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu, linia, numer

Opis wygenerowany automatycznie

## Opis procesu ETL

Obraz zawierający zrzut ekranu, linia, lampa, design

Opis wygenerowany automatycznie

**Proces ETL - zasilenie wymiarów**

Proces zasilenia tabel wymiarów został zaimplementowany w pliku **ETL - wymiary**. Dane źródłowe są przechowywane w plikach Excel, które zostały wygenerowane jako eksport tabel z bazy danych Access. Proces ma na celu przygotowanie i załadowanie danych do tabel wymiarów z zastosowaniem mechanizmu oceny dopasowania \_Similarity.

**Kroki procesu:**

1. **Ekstrakcja danych:**
   * Dane są pobierane z plików Excel, które zawierają dane wyeksportowane z bazy danych Access.
   * Każdy plik Excel zawiera dane dla określonego wymiaru, np. dKategorieWydatku, dMiejscowosc, dKategoriaCzasu, dOsoba.
2. **Transformacja danych:**
   * Wczytane dane są oceniane pod kątem dopasowania do istniejących wartości w tabelach wymiarów:
     + Rekordy o \_Similarity >= 0.9 są uznawane za dopasowane i trafiają do tabel mappingowych.
     + Rekordy o \_Similarity < 0.9 są oznaczane jako niedopasowane i ładowane do tabel wymiarów jako nowe wpisy.
   * Proces zapewnia obsługę sytuacji, w których dane są niekompletne lub częściowo zgodne.
3. **Ładowanie danych:**
   * **Dopasowane dane (\_Similarity > 0.9):**  
     Ładowane do tabel mappingowych i wykorzystywane do powiązań w procesie zasilania faktów.
   * **Niedopasowane dane (\_Similarity ≤ 0.9):**  
     Ładowane bezpośrednio do tabel wymiarów jako nowe wpisy, aby zapewnić ich dostępność w hurtowni.

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

**Proces ETL - zasilenie faktów**

Proces zasilania tabel faktów został zaimplementowany w pliku **ETL - fakty**. Dane źródłowe są również przechowywane w plikach Excel, które zawierają informacje transakcyjne dotyczące przychodów, wydatków i czasu spędzonego. W procesie tym zastosowano mechanizm **Lookup**, aby powiązać dane z tabelami wymiarów.

**Kroki procesu:**

1. **Ekstrakcja danych:**
   * Dane transakcyjne (wydatki, przychody, czas spędzony) są pobierane z odpowiednich plików Excel.
2. **Transformacja danych:**
   * Dane transakcyjne są powiązywane z tabelami wymiarów za pomocą operacji **Lookup**, która wykorzystuje dopasowane wartości z mappingu:
     + **Rekordy zgodne:** Zostają poprawnie powiązane z kluczami wymiarów.
     + **Rekordy niezgodne:** W przypadku braku dopasowania dane są ładowane jako osobne rekordy.
3. **Ładowanie danych:**
   * Dane są ładowane do tabel faktów:
     + fWydatek – zawiera szczegóły dotyczące wydatków.
     + fPrzychod – zawiera szczegóły dotyczące przychodów.
     + fCzasSpedzony – zawiera szczegóły dotyczące czasu spędzonego w lokalizacji.

## Użyte skrypty SQL

-- 1 --

INSERT INTO dTypPrzychodu (dTPr\_nazwa)

VALUES

('Zwrot kosztów'),

('Wynajem'),

('Wynagrodzenie'),

('Stypendium'),

('Zwrot podatku'),

('Darowizna'),

('Darowizna'),

('Sprzedaż'),

('Inne');

-- 2 --

INSERT INTO dFormaPlatnosci (dFP\_nazwa)

VALUES

('gotówka'),

('przelew'),

('karta płatnicza'),

('polecenie zapłaty'),

('inna');

-- 3 --

DECLARE @data\_poczatek date;

DECLARE @data\_koniec date;

DECLARE @data date;

DECLARE @i int;

SET @data\_poczatek = '2019-06-30';

SET @data\_koniec = '2020-06-30';

SET @data = @data\_poczatek;

SET @i = 1;

WHILE @data <= @data\_koniec

BEGIN

INSERT INTO [dbo].[dCzas]

([dCz\_id]

,[dCz\_dzien]

,[dCz\_miesiac]

,[dCz\_rok]

,[dCz\_dzien\_tygodnia]

,[dCz\_nazwa\_dt])

VALUES

(@i

,@data

,cast(SUBSTRING ( CAST( @data AS varchar(10)) ,1, 8 )+'01' AS date)

,cast(SUBSTRING ( CAST( @data AS varchar(10)) ,1, 4 ) AS int)

,DATEPART(dw,@data)

,DATENAME(weekday,@data))

;

SET @data = DATEADD(d,1,@data);

SET @i = @i + 1;

END;

GO

SELECT \* FROM [dbo].[dCzas]

GO

SET IDENTITY\_INSERT dbo.dCzas ON

## Ekrany ilustrujące poprawne działanie projektów SSIS i ekrany przedstawiające wstawienie danych do hurtowni

**Zrzuty ekranu przedstawiające zawartość tabel hurtowni przed wykonaniem projektów SSIS:**

**Obraz zawierający tekst, zrzut ekranu, oprogramowanie

Opis wygenerowany automatycznie**

**Obraz zawierający zrzut ekranu, tekst, oprogramowanie, linia

Opis wygenerowany automatycznie**

**Obraz zawierający zrzut ekranu, tekst, linia, oprogramowanie

Opis wygenerowany automatycznie**

**Obraz zawierający tekst, oprogramowanie, zrzut ekranu, Ikona komputerowa

Opis wygenerowany automatycznie**

**Obraz zawierający zrzut ekranu, tekst, linia, oprogramowanie

Opis wygenerowany automatycznie**

**Obraz zawierający zrzut ekranu, tekst, linia, oprogramowanie

Opis wygenerowany automatycznie**

**Obraz zawierający zrzut ekranu, oprogramowanie, tekst

Opis wygenerowany automatycznie**

**Obraz zawierający tekst, zrzut ekranu, oprogramowanie, linia

Opis wygenerowany automatycznie**

**Obraz zawierający tekst, zrzut ekranu, linia, oprogramowanie

Opis wygenerowany automatycznie**

**Obraz zawierający tekst, zrzut ekranu, linia, oprogramowanie

Opis wygenerowany automatycznie**

**Zrzut ekranu przedstawiający zawartość tabeli mapping:**

**Obraz zawierający tekst, oprogramowanie, Ikona komputerowa, Oprogramowanie multimedialne

Opis wygenerowany automatycznie**

**Ekrany przedstawiające poprawne wykonanie projektów SSIS dla wymiarów:**

**Obraz zawierający tekst, zrzut ekranu, oprogramowanie, Oprogramowanie multimedialne

Opis wygenerowany automatycznie**

**dKategorieWydatku:**

**Obraz zawierający zrzut ekranu

Opis wygenerowany automatycznie**

**Obraz zawierający zrzut ekranu, czarne

Opis wygenerowany automatycznie**

**dMiejscowosc:**

**Obraz zawierający zrzut ekranu, Oprogramowanie multimedialne, Oprogramowanie graficzne

Opis wygenerowany automatycznie**

**Obraz zawierający zrzut ekranu, czarne

Opis wygenerowany automatycznie**

**dKategoriaCzasu:**

**Obraz zawierający Oprogramowanie multimedialne, Oprogramowanie graficzne, oprogramowanie, zrzut ekranu

Opis wygenerowany automatycznie**

**Obraz zawierający zrzut ekranu, czarne

Opis wygenerowany automatycznie**

**dOsoba:**

**Obraz zawierający zrzut ekranu, Oprogramowanie multimedialne, elektronika, oprogramowanie

Opis wygenerowany automatycznie**

**Obraz zawierający Oprogramowanie multimedialne, oprogramowanie, Oprogramowanie graficzne, zrzut ekranu

Opis wygenerowany automatycznie**

**Ekrany przedstawiające poprawne wykonanie projektów SSIS dla faktów:**