Sprawozdanie - Hurtownia na podstawie danych sklepu internetowego

Hurtownie Danych
Prowadzący: dr inż. Dariusz Gąsior

Bączyński Konrad, 211147
Charewicz Jan, 212825

Spis treści

1. Wybór i opis źródeł

Do naszego projektu, wybraliśmy znalezione w internecie źródło danych, będące sztandarową reprezentacją sklepu internetowego - w tym przypadku - Sklepu Dell'a na terenie Stanów Zjednoczonych. Znaleziony plik .sql - na różnych forach, był po wielokrotnie podawany jako nawet dobry materiał na hurtownie danych. W pliku znaleziono zarówno tworzenie tabel, jak i liczne polecenia INSERT (...). Baza zawierała generowane (nierzadko o rozkładzie zbliżonym do rozkładu równomiernego) informacje na temat klientów, zamówień, pozycji na zamówieniach, produktach.

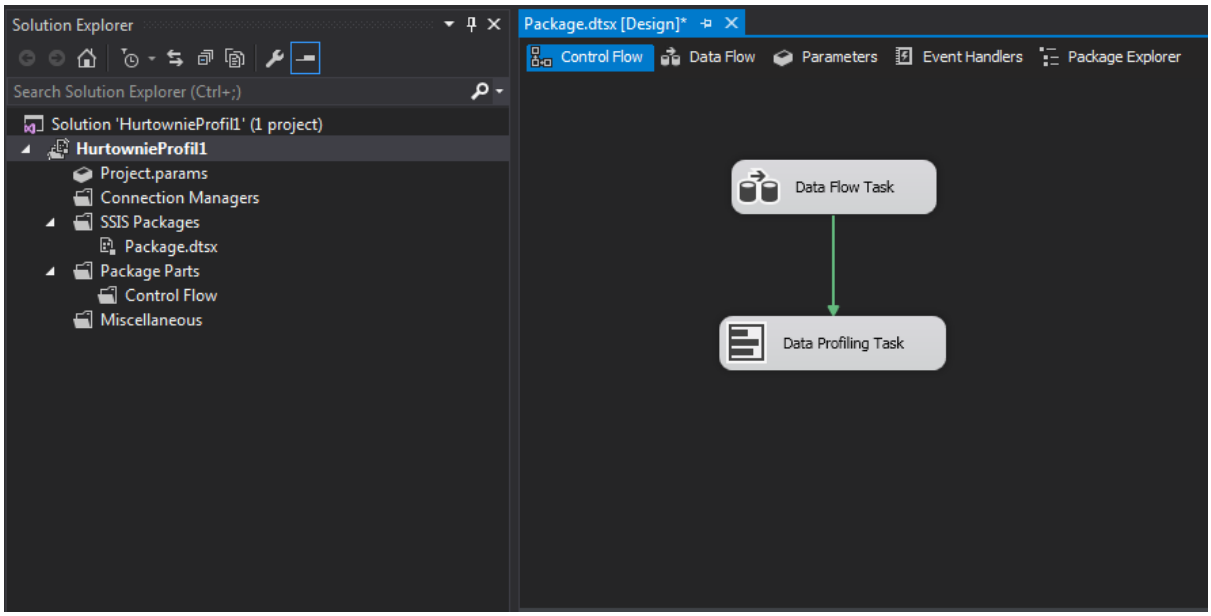
Motywację do wyboru źródła odnaleźliśmy w rekomendacjach tego źródła na forach, a także w bardzo przyjaznym formacie pliku. Niestety, jako, że język użyty w pliku, okazał się PostgreSQL'em, musieliśmy dokonać małych transformacji formatowania źródła, a by lepiej się pracowało w następnych etapach dane umieszczono (przy użyciu tak przeformatowanego i akceptowalnego przez MS pliku .sql) w bazie MS SQL SERVER i to ją wykorzystywaliśmy dalej jako źródło.

W opisie źródła danych uzyskaliśmy również częściowe informacje na temat dziedzin danych, na przykład w tabeli customer (opis klientów) pole email było polem sklejenia automatycznego pola imienia i nazwiska z dodaniem @dellstore.com. Dane były generowane, pole data jest z przedziału roku 2004. Opis podawał również informacje, że nazwiska i imiona zostały zahashowane, by nie ujawniać danych klientów, a pole numeru kategorii nie miało rzeczywistego odniesienia i było numerowane od 1 do 16 (Bo na tyle kategorii sklep dzielił swoje produkty).

```
CREATE TABLE customer (
    customerid integer NOT NULL,
    firstname character varying(50) NOT NULL,
    lastname character varying(50) NOT NULL,
    address1 character varying(50) NOT NULL,
    address2 character varying(50),
    city character varying(50) NOT NULL,
    state character varying(50),
    zip integer,
    country character varying(50) NOT NULL,
    region smallint NOT NULL,
    email VARCHAR(320),
    phone character varying(50),
    creditcardtype integer NOT NULL,
    creditcard character varying(50) NOT NULL,
    creditcardexpiration character varying(50) NOT NULL,
    age smallint,
    income integer,
    gender character varying(1)
);

CREATE TABLE inventory(
    prod_id integer NOT NULL,
    quan_in_stock integer NOT NULL,
    sales integer NOT NULL
);

CREATE TABLE orderlines (
    orderlineid integer NOT NULL,
    orderid integer NOT NULL,
    prod_id integer NOT NULL,
    quantity smallint NOT NULL,
    orderdate date NOT NULL
);

CREATE TABLE orders (
    orderid integer NOT NULL,
    orderdate date NOT NULL,
    customerid integer,
    netamount numeric(12,2) NOT NULL,
    tax numeric(12,2) NOT NULL,
    totalamount numeric(12,2) NOT NULL
);

CREATE TABLE products (
    prod_id integer NOT NULL,
    category integer NOT NULL,
    title character varying(50) NOT NULL,
    actor character varying(50) NOT NULL,
    price numeric(12,2) NOT NULL,
    special smallint
);

CREATE TABLE customer_hist (
    customerid integer NOT NULL,
    orderid integer NOT NULL,
    prod_id integer NOT NULL
);

CREATE TABLE dates(
    date_id integer NOT NULL,
    day integer,
    month integer,
    year integer
);
```

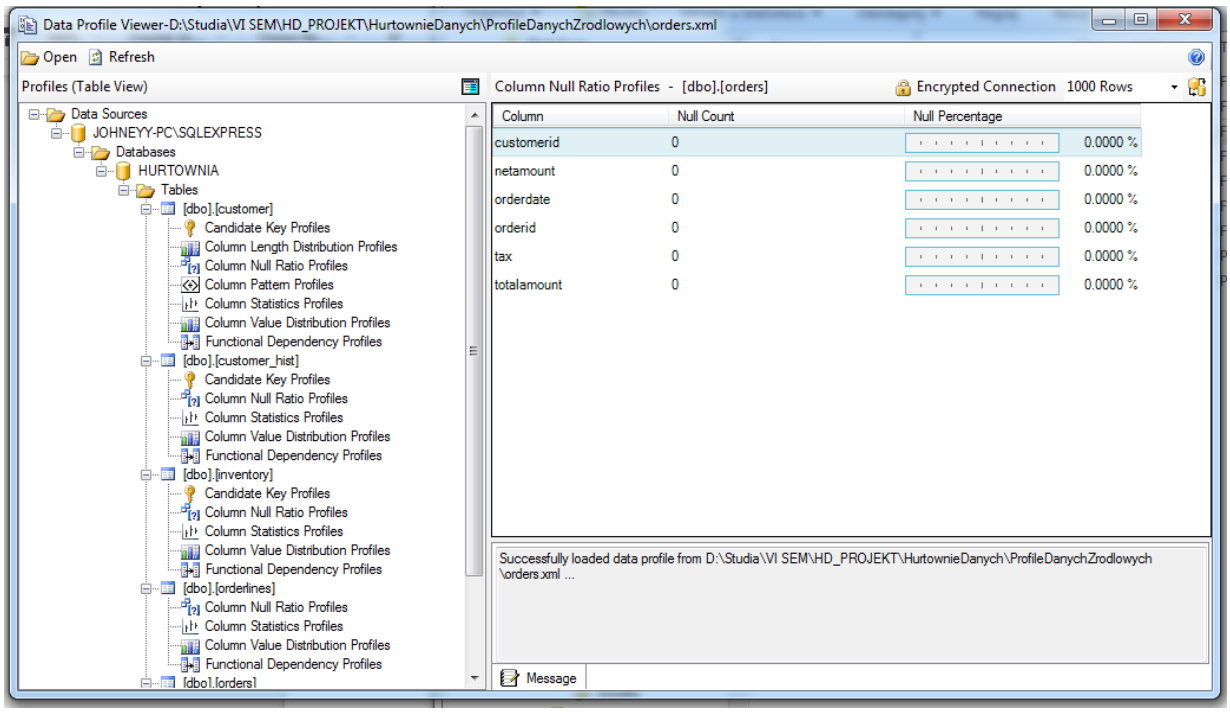
Rysunek 1 Przedstawia źródło danych

2. Jakość danych



Rysunek 2 Data Profiling Task w Visual Studio 2015.

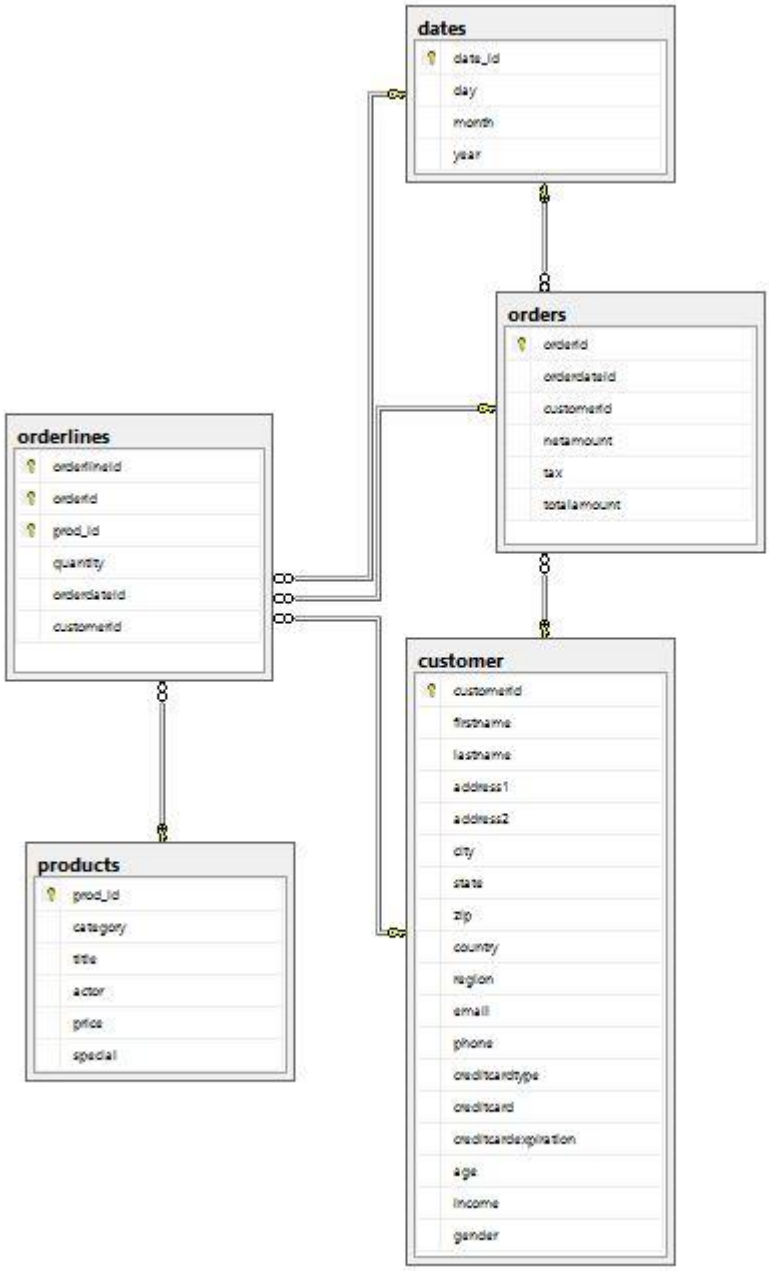
W celu wykonania oceny jakości danych, tworzy się profile na podstawie danych źródłowych, całość eksportowana jest do pliku .xml które to parsuje Data Profile Viewer (rys. poniżej). Całe zadanie opiewało na odpowiednią konfigurację zadań i połączeń w Projekcie Analysis Services.



Rysunek 3 Data Profile Viewer

Jak pokazuje [Rysunek 1] źródło faktycznie okazało się bardzo w porządku, wskaźniki pól o wartości NULL w całej bazie wynosiły okrągłe 0. Z profilowania już można było się całkiem sporo dowiedzieć, np. to, że minimalny wiek klientów wynosił 18, pochodzili oni w około połowie z jednego kraju i częściej robili mniejsze zamówienia.

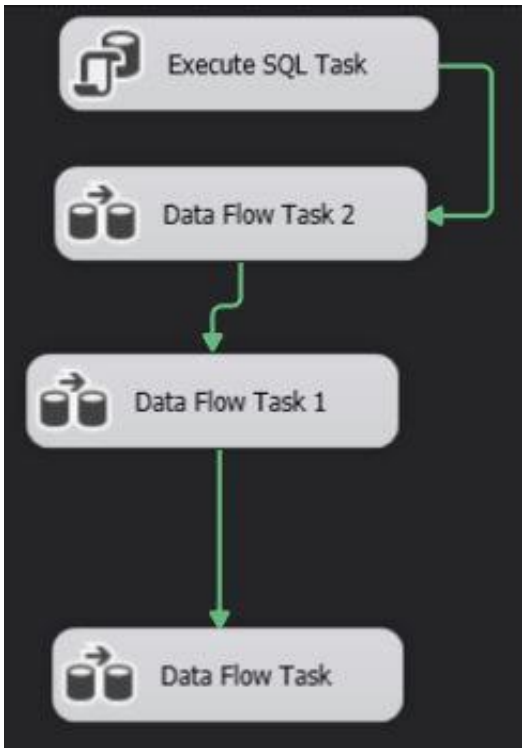
3. Projekt bazy analitycznej



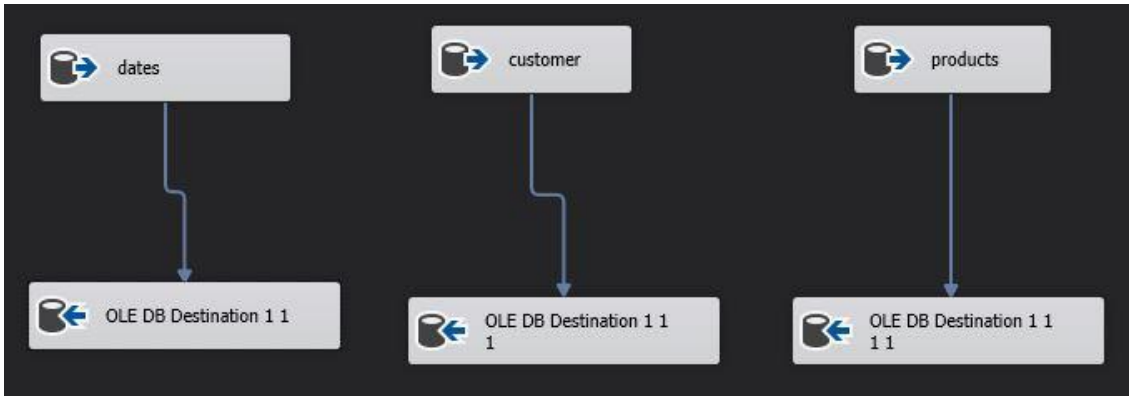
Rysunek 4 Diagram tabel bazy analitycznej.

Powyższy diagram [Rysunek 3] przedstawia zaproponowany po konsultacji z prowadzącym schemat docelowej bazy analitycznej. W niektórych miejscach należało więc zastosować spłaszczenia, by ze schematu z dwoma potencjalnymi tabelami faktów, skupić się na pozycjach na zamówieniu jako głównym wpisie w hurtowni.

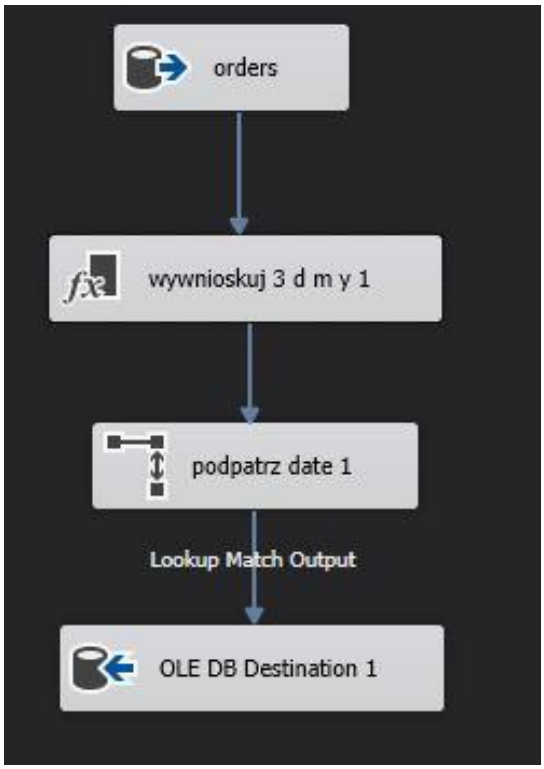
4. Proces ETL-owy *(Extract, Transform and Load)*



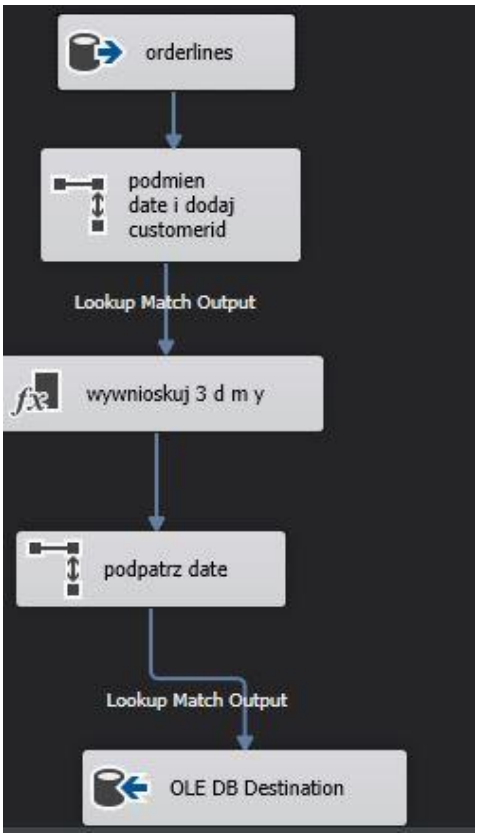
Rysunek 5 Control flow.



Rysunek 6 Data Flow Task 2.



Rysunek 7 Data Flow Task 1

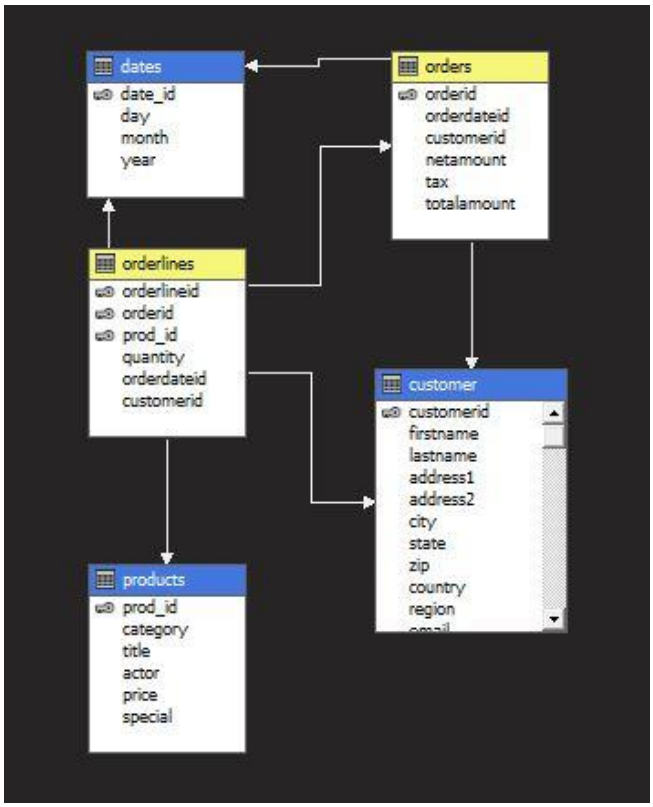


Rysunek 8 Data Flow Task.

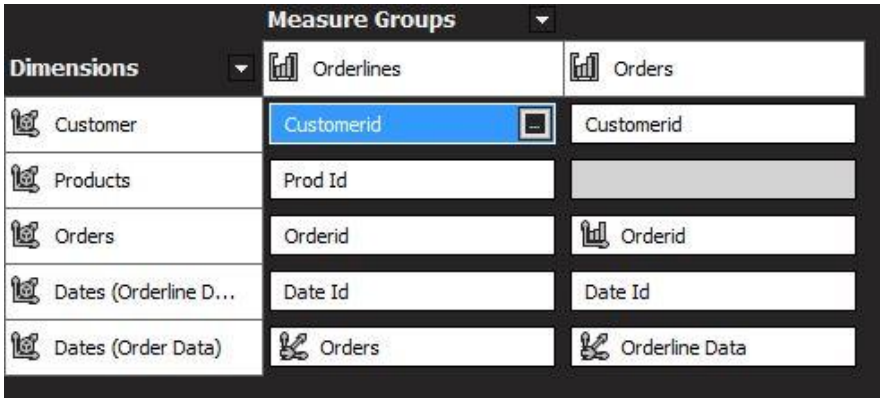
Powyższe rysunki [4-7] przedstawiają realizację procesu ETL. Dzięki niemu uzyskujemy bazę wypełnioną przetworzonymi już częściowo danymi. Na przykład dodajemy dodatkową kolumnę do tabeli i wnioskujemy ją na podstawie innej kolumny w innej tabeli. Ważnym elementem tego procesu jest przygotowanie najpierw tabel pobocznych, potencjalnie wykorzystywanych w bloczkach typu *lookup*, a dopiero w następnym etapie procesowanie tabeli faktów.

5. Kostka

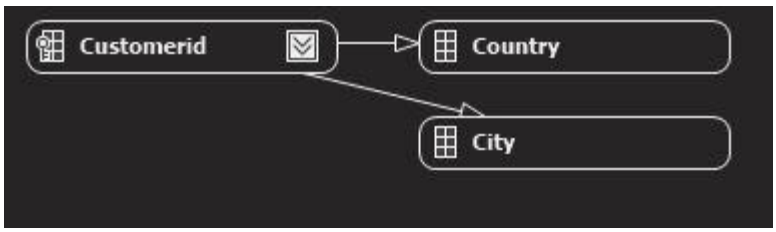
Zamieszczone obok rysunki przedstawiają proces tworzenia i konfiguracji kostki. [Rysunek 8] przedstawia strukturę kostki. Kostka pozwala na szybką analizę danych. [Rysunek 9] przedstawia wykorzystane wymiary do budowy kostki oraz dwie tabele faktów. [Rysunek 10] przedstawia wymiar klienta i relacje między jego atrybutami.



Rysunek 9 Struktura kostki.



Rysunek 10 Wykorzystane wymiary.



Rysunek 11 Relacje między atrybutami Klienta.

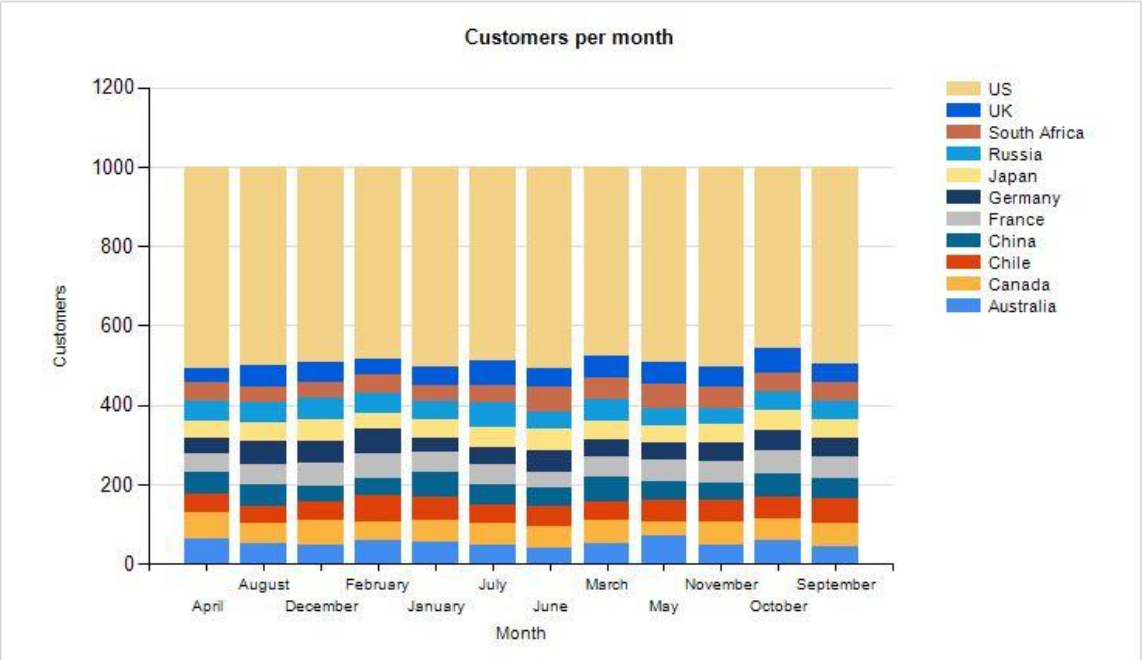
6. Raporty

- Raport 1 – Liczba klientów w perspektywie poszczególnych krajów i miesięcy.

Clients in countries per month

Country	Month	Number of customers
Australia		
	January	56
	February	59
	March	53
	April	62
	May	69
	June	40
	July	48
	August	50
	September	45
	October	59
	November	49
	December	49
Canada		
	January	52
	February	47
	March	58
	April	67
	May	38
	June	54
	July	54
	August	50
	September	56
	October	54

Tabela 1 Liczba klientów w poszczególnych krajach z podziałem na miesiące.

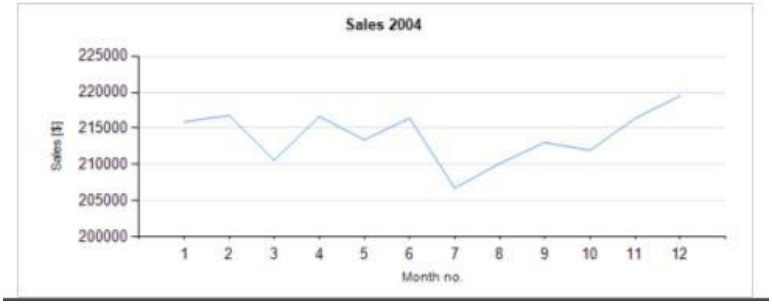


Rysunek 12 Liczba klientów w poszczególnych krajach z podziałem na miesiące.

- Raport 2 – Całkowita miesięczna sprzedaż w roku 2004 z podziałem na miesiące.

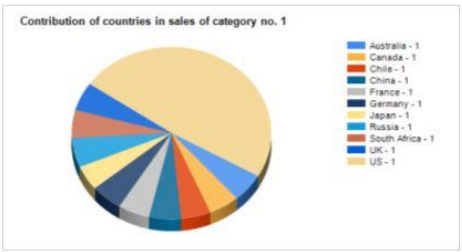
Monthly sales

No.	Month	Total sales
1	January	$215898.76
2	February	$216792.09
3	March	$210564.40
4	April	$216642.01
5	May	$213395.19
6	June	$216412.59
7	July	$206707.38
8	August	$210115.92
9	September	$213034.71
10	October	$211952.17
11	November	$216373.17
12	December	$219498.40



Rysunek 13 Całkowita miesięczna sprzedaż w poszczególnych miesiącach.

- Raport 3 – Udział krajów w rozkładzie liczby sprzedanych produktów różnych kategorii.



Country \ Category no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Australia	201	202	219	199	205	184	202	204	204	241	188	211	196	212	204	229
Canada	186	240	202	229	200	193	181	198	205	187	182	217	202	207	199	200
Chile	181	200	209	182	192	190	174	201	214	200	201	201	192	188	206	194
China	195	216	181	231	191	185	201	200	172	220	165	184	181	188	193	194
France	193	188	170	200	195	189	198	209	203	182	186	185	201	168	188	215
Germany	194	197	193	204	180	227	168	170	177	187	192	161	172	168	161	176
Japan	173	162	162	182	164	189	184	184	171	164	156	171	158	166	169	172
Russia	208	166	193	168	195	174	156	178	208	178	177	180	187	192	199	198
South Africa	198	203	194	209	185	177	197	180	201	214	191	199	193	176	183	173
UK	200	177	204	192	170	198	172	187	203	165	168	170	203	163	186	197
US	1829	1961	1888	1935	1921	1983	1716	1914	1962	1829	1768	1899	1785	1847	1834	1819

Rysunek 14 Liczba produktów sprzedanych w poszczególnych krajach w danych kategoriach.

Powyższe rysunki przedstawiają raporty wygenerowane dzięki analizie danych. Dzięki nim możemy zorientować się jak wielu klientów miał nasz sklep w każdym z krajów na przestrzeni całego roku. Możemy również dowiedzieć się jaka była całkowita miesięczna sprzedaż. Oraz jak rozkładała się sprzedaż w poszczególnych krajach z podziałem na kategorie.

7. Podsumowanie

Hurtownie danych to dość specyficzna baza danych, pozwalająca dokładniej przyjrzeć się danemu wycinkowi rzeczywistości. Wszystkie wymienione w sprawozdaniu etapy są niezbędne, by uzyskać zdolność wnioskowania. W przypadku wybranego źródła danych, już przy profilowaniu danych uzyskaliśmy obraz danych o dość równomiernym, poniekąd „sztucznym", czy „nienaturalnym" rozkładzie. Nie obniżyło to wartości edukacyjnych wyniesionych z całego procesu budowania i pracy z hurtownią danych. Przygotowane raporty są, naszym zdaniem, bardzo trafnymi jeśli chodzi o tematykę wybranego źródła, gdyż przy rzeczywistych danych, można by mówić o tym jak ów sklep się rozwijał, czy można zauważyć progres w poziomie sprzedaży, scharakteryzować w którym kraju wartoby pomyśleć nad otworzeniem nowego oddziału, czy odpowiedzieć na pytanie produkty których kategorii jak się sprzedają (nawet w poszczególnych krajach), a znając ich charakterystykę szukać argumentacji z czego może to wynikać.