**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 3](#_Toc453106888)

[2. Jakość danych 5](#_Toc453106889)

[3. Projekt bazy analitycznej 6](#_Toc453106890)

[4. Proces ETL-owy *(Extract, Transform and Load)* 7](#_Toc453106891)

[5. Kostka 9](#_Toc453106892)

[6. Raporty 10](#_Toc453106893)

[ Raport 1 – Liczba klientów w perspektywie poszczególnych krajów i miesięcy. 10](#_Toc453106894)

[ Raport 2 – Całkowita miesięczna sprzedaż w roku 2004 z podziałem na miesiące. 11](#_Toc453106895)

[ Raport 3 – Udział krajów w rozkładzie liczby sprzedanych produktów różnych kategorii. 1](#_Toc453106896)1

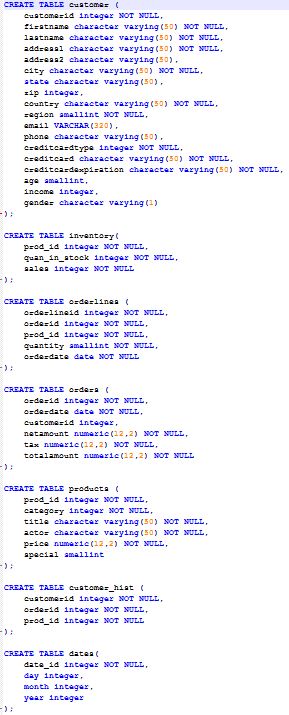
[7. Podsumowanie 1](#_Toc453106897)2

# 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.

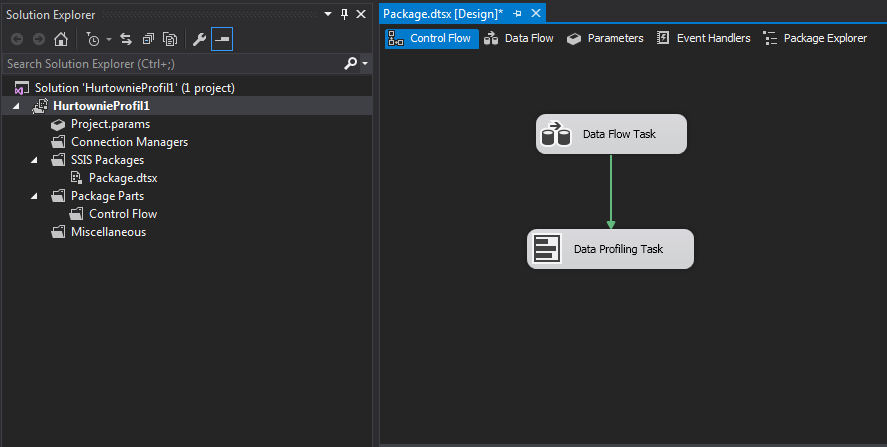
Motywację do wyboru źródła odnaleźliśmy w rekomendacjach tego źródła na forach, a także w bardzo przyjaznym formacie pliku. Niestety, jako, że język użyty w pliku, okazał się PostgreSQL’em, musieliśmy dokonać małych transformacji formatowania źródła, a by lepiej się pracowało w następnych etapach dane 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).



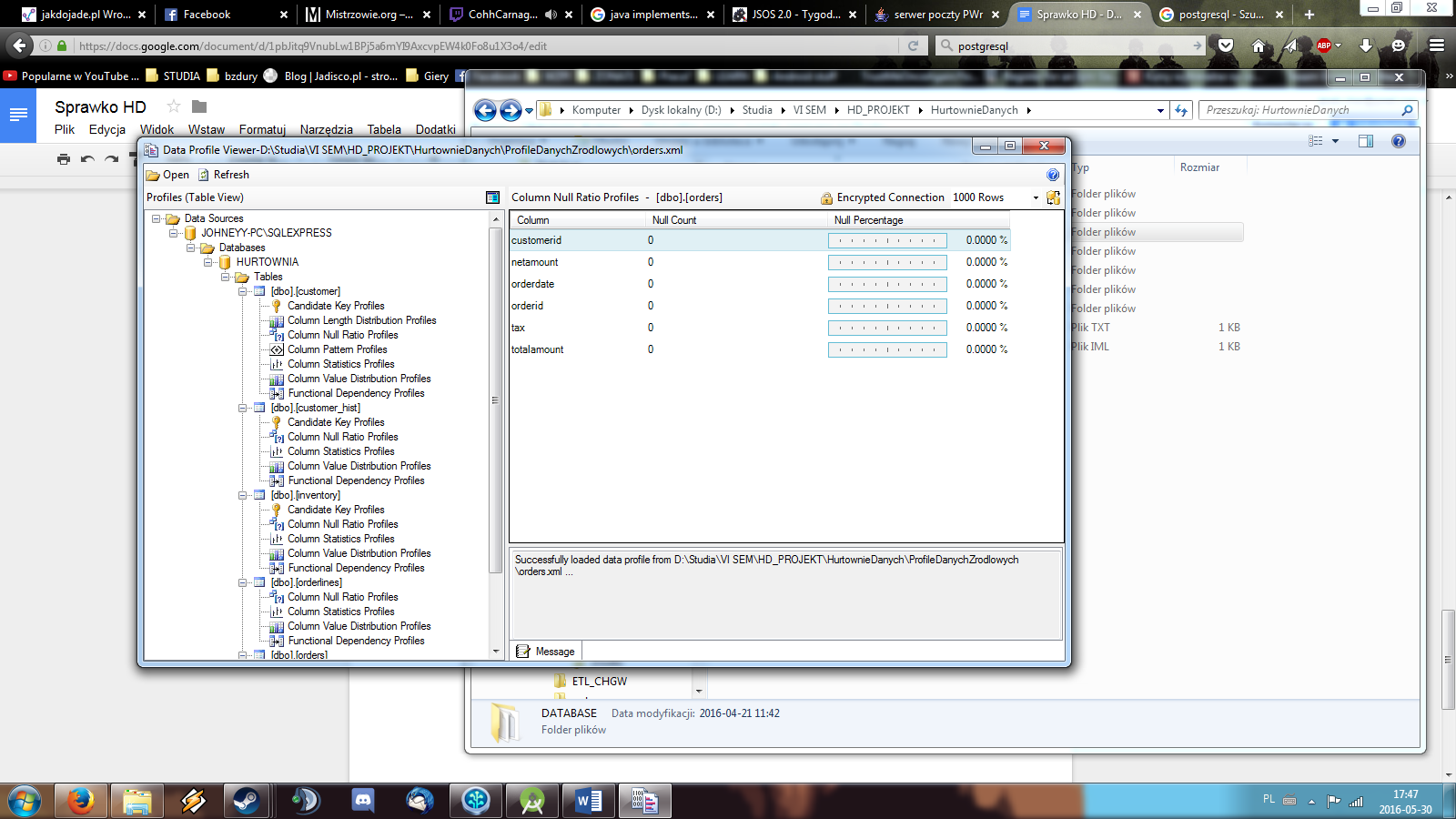
Rysunek 1 Przedstawia źródło danych

# Jakość danych



Rysunek 2 Data Profiling Task w Visual Studio 2015.

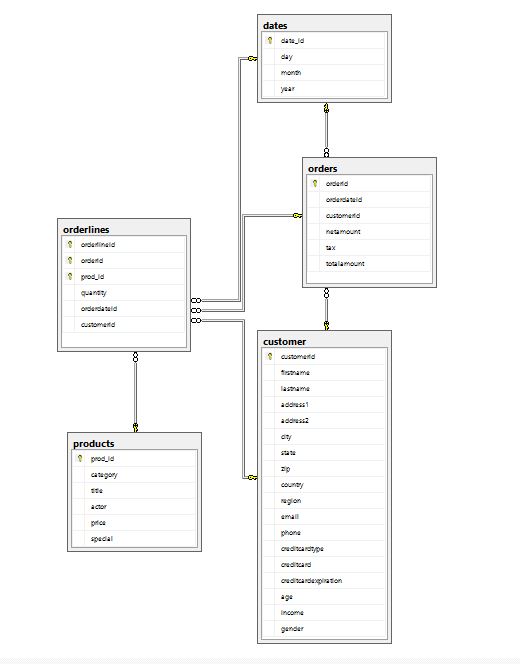
W celu wykonania oceny jakości danych, tworzy się profile na podstawie danych źródłowych, całość eksportowana jest do pliku .xml które to parsuje Data Profile Viewer (rys. poniżej). Całe zadanie opiewało na odpowiednią konfigurację zadań i połączeń w Projekcie Analysis Services.



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.

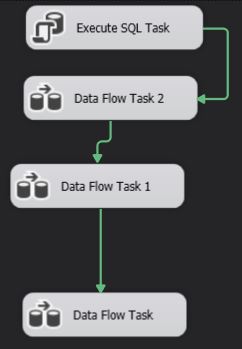
# Projekt bazy analitycznej



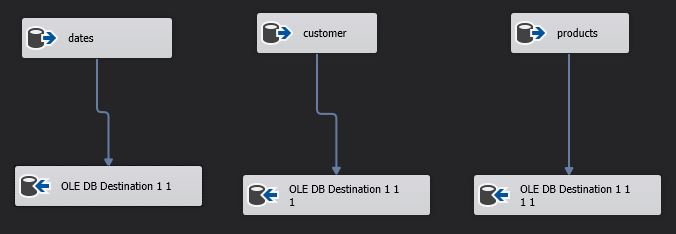
Rysunek 4 Diagram tabel bazy analitycznej.

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.

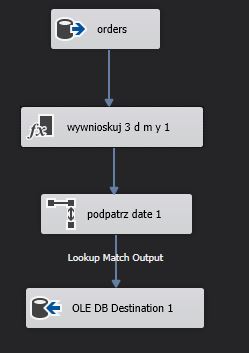
# Proces ETL-owy *(Extract, Transform and Load)*



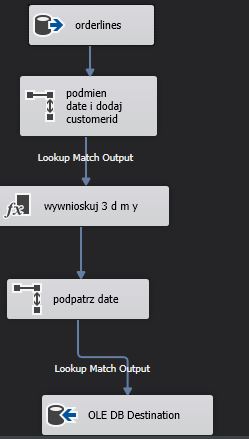
Rysunek 5 Control flow.



Rysunek 6 Data Flow Task 2.



Rysunek 7 Data Flow Task 1

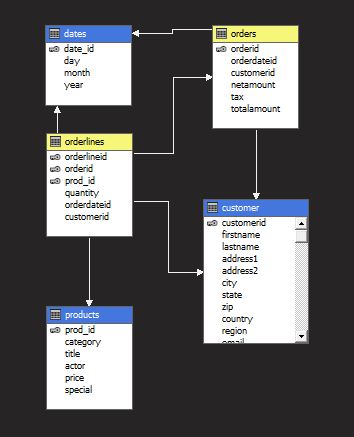


Rysunek 8 Data Flow Task.

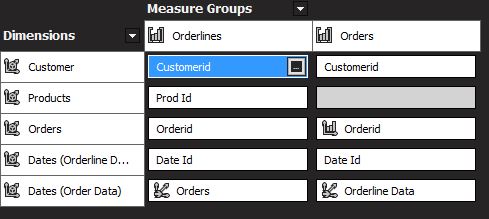
Powyższe rysunki [4-7] przedstawiają realizację procesu ETL. Dzięki niemu uzyskujemy bazę wypełnioną przetworzonymi już częściowo danymi. Na przykład dodajemy dodatkową kolumnę do tabeli i wnioskujemy ją na podstawie innej kolumny w innej tabeli. Ważnym elementem tego procesu jest przygotowanie najpierw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

# Kostka

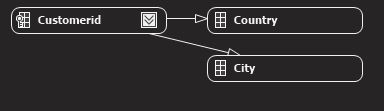
Zamieszczone obok rysunki przedstawiają proces tworzenia i konfiguracji kostki. [Rysunek 8] przedstawia strukturę kostki. Kostka pozwala na szybką analizę danych. [Rysunek 9] przedstawia wykorzystane wymiary do budowy kostki oraz dwie tabele faktów. [Rysunek 10] przedstawia wymiar klienta i relacje między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



Rysunek 9 Struktura kostki.



Rysunek 10 Wykorzystane wymiary.



Rysunek 11 Relacje między atrybutami Klienta.

# Raporty

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

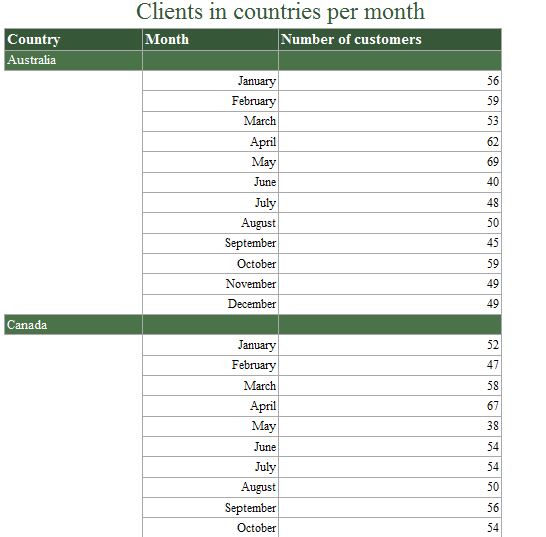
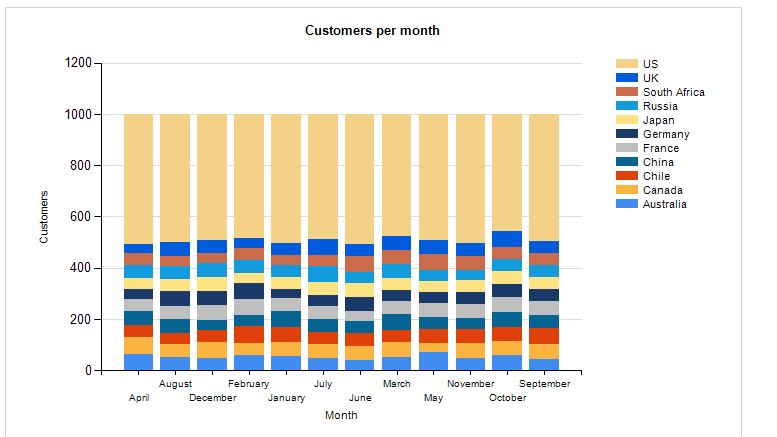
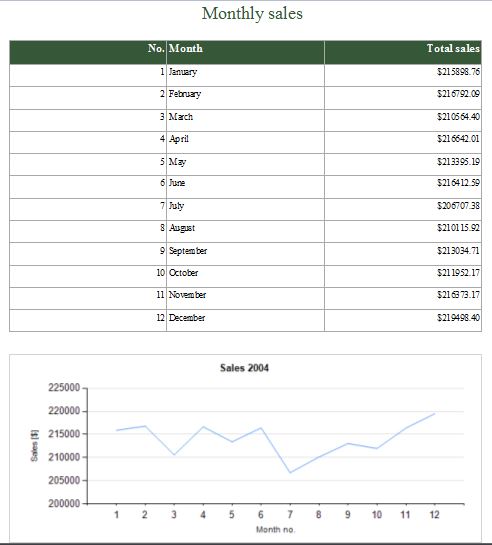


Tabela 1 Liczba klientów w poszczególnych krajach z podziałem na miesiące.



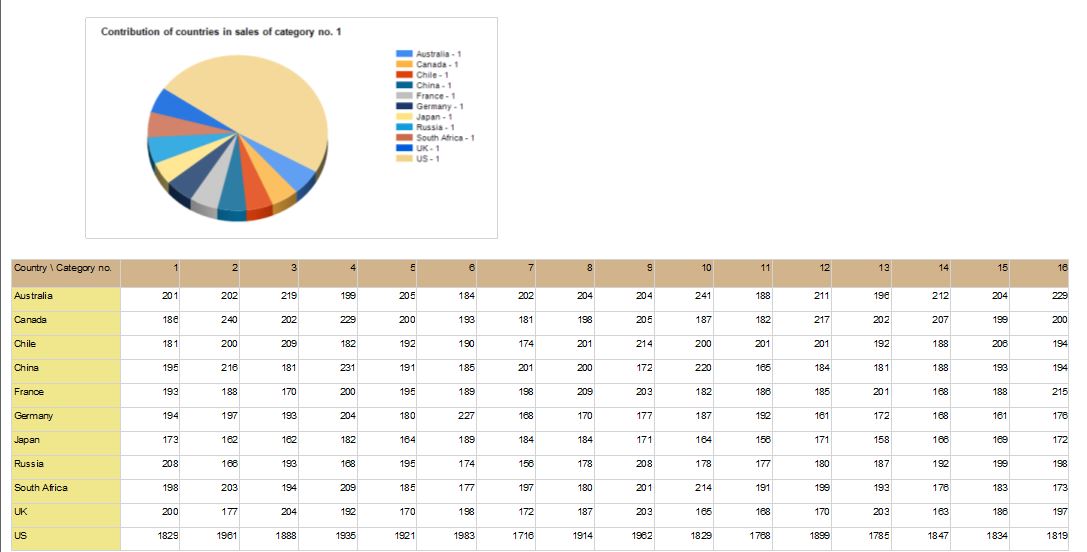
Rysunek 12 Liczba klientów w poszczególnych krajach z podziałem na miesiące.

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



Rysunek 13 Całkowita miesięczna sprzedaż w poszczególnych miesiącach.

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



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.

# 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.