

Inżynieria Hurtowni Danych

Temat: Zastosowanie ROLAP, MOLAP, HOLAP w analityce biznesowej.

Karol Wolny

28.01.2024

Spis treści

[Część teoretyczna 2](#_ssoc6verp58t)

[**Wstęp 2**](#_wb5zbebsanlh)

[**1. OLAP 3**](#_rukm1c5fqoac)

[**2. Kostka OLAP 4**](#_usd7y8ymgm00)

[**3. Reguły budowania kostek OLAP 5**](#_48b1udgvxzak)

[**4. Typy kostek OLAP 6**](#_jjm9xa158d91)

[Część praktyczna 8](#_x7f3qr73h7yg)

[**5. Cele Projektu 8**](#_q3c70soxr5uj)

[**6. Opis źródeł danych 9**](#_vwfiroe0szv7)

[**7. Schemat logiczny hurtowni danych na podstawie schematu gwiazdy. 11**](#_gjtm0mc88gtx)

[Wymiar czasu 11](#_t2b431pm05e3)

[Wymiar miejsca 13](#_920tesfxtjfq)

[Wymiar produktu 14](#_o5fqf77watg5)

[Wymiar pracowników 15](#_kiuqdx1zjhar)

[Tabela faktów sprzedaży 17](#_jlwj9e74vlcu)

[**8. Opis procesów ETL 19**](#_qahbncujekpv)

[1. Analiza sprzedaży ze względu na miejsce docelowej wysyłki. 19](#_ghyh7bdasfmh)

[2. Analiza sprzedaży produktów z podziałem na kategorie. 23](#_t1d2ynmqpuh3)

[3. Analiza czynników sprzedażowych rowerów. 26](#_a1d39zn3gk49)

[4. Analiza sprzedaży na przestrzeni czasu. 29](#_li1p05zlhve)

[5. Analiza sprzedaży pracowników. 32](#_axbuum8nx2pf)

[**9. Analiza uzyskanych wyników 35**](#_46mpkzqok1rw)

[**10. Źródła danych 37**](#_5b2ozgxpqkio)

# Część teoretyczna

# Wstęp

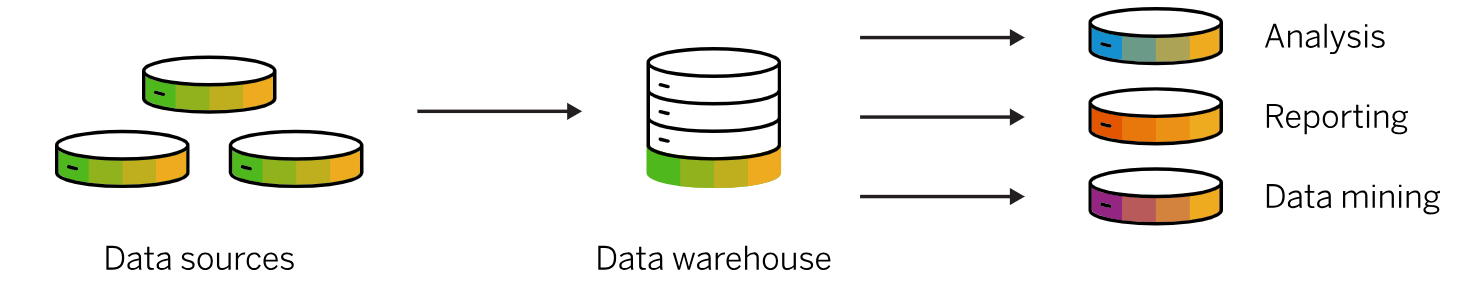
Czym są hurtownie danych?

Przytoczę definicję Williama H. Inmona jednego z pionierów w kształtowaniu terminologii oraz metodologii hurtowni danych. W 1996 roku w swojej książce „Building The Data Warehouse” pisał on tak: Hurtownia danych to tematyczna baza danych, która trwale

przechowuje zintegrowane dane opisane wymiarem czasu.

Hurtownia danych to zaawansowany system przechowywania, który łączy i standaryzuje ogromne zbiory danych z wielu różnych źródeł. Pełni rolę absolutnego źródła informacji dla rozwiązań Business Intelligence, narzędzi analitycznych oraz raportowania. Umożliwia to firmom nie tylko przechowywanie, ale również przekształcanie tych danych w precyzyjne analizy. Analizy te z kolei pozwalają na podejmowanie decyzji z większą ilością wiedzy, przez co zwiększają się szanse na trafność tejże decyzji. W hurtowni danych zawierają się zarówno aktualne, jak i historyczne dane, stając się dzięki temu cennym źródłem informacji dla potrzeb organizacyjnych.

Jak już wspomniałem wcześniej, dzięki hurtowniom danych firmy mogą podejmować lepsze decyzje. Jednakże, aby można było taką decyzje podejmować, dane w postaci surowej należy poddać obróbce. Mowa tutaj o tworzeniu zestawień statystycznych, wykresów i raportów, które podsumowują duże ilości danych i wyciągają z nich konkretne informacje. Proces ten nazywa się przetwarzaniem analitycznym danych, czyli **On-line Analytical Processing (OLAP).**



Rys. 1 schemat przetwarzania danych przez hurtownie Źródło: <https://www.sap.com/poland/products/technology-platform/datasphere/what-is-a-data-warehouse.html>

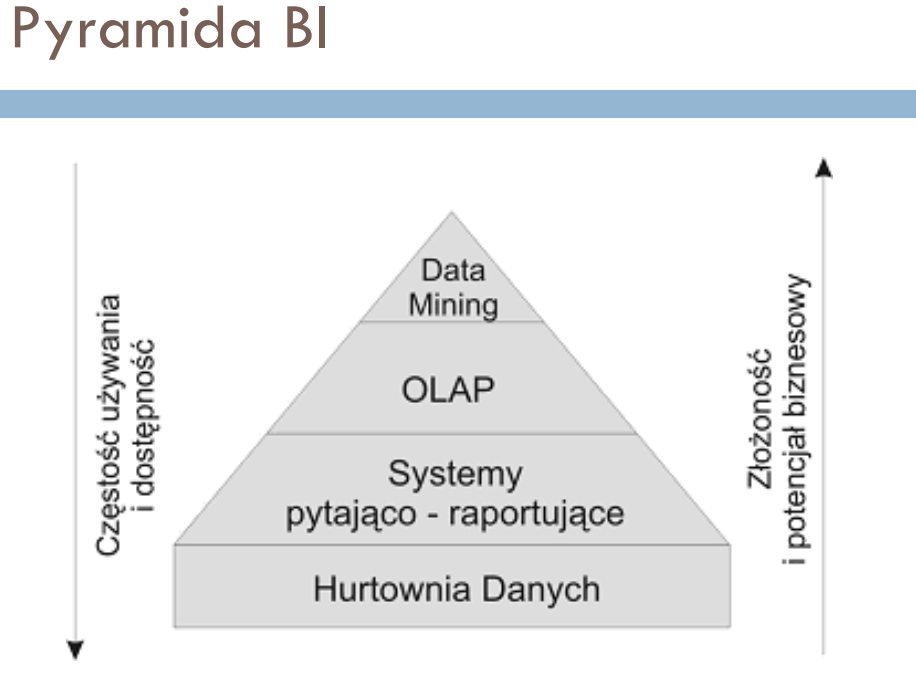
# OLAP

OLAP to rodzaj technologii używanej do przetwarzania danych w celu analizy biznesowej. Głównym zadaniem OLAP jest umożliwienie użytkownikom eksplorowania dużych zbiorów danych, identyfikowania trendów, wykrywania relacji i podejmowania decyzji na podstawie analizy danych.

Podstawą modelu OLAP są fakty i wymiary. Fakty to informacje podlegające analizie reprezentowane przez konkretne miary lub wskaźniki. Mówią one o takich informacjach jak np. liczba sprzedanych jednostek towarów lub ilość zwrotów, a do ich przedstawienia używa się wartości ciągłych lub numerycznych. Wymiary z kolei służą do organizowania i grupowania danych, dzięki czemu ustalają kontekst analizy. Mogą być one różne dla każdego modelu. Często używane wymiary to choćby czas i miejsce. Wymiar czasu pozwala na analizę danych w kontekście zmiany w czasie, może obejmować takie kategorie jak dzień, miesiąc, kwartał czy rok. Wymiar miejsca umożliwia precyzyjne określanie, skąd pochodzą dane fakty, kontynentu, państwa, czy miasta.

Dzięki takiej konstrukcji system OLAP pozwala przede wszystkim na:

* Analizę zarówno bardzo szczegółowych danych, jak i tych uogólnionych, podsumowanych.
* Efektywne wyszukiwanie danych. Wcześniej przetworzone dane, pozwalają na generowanie szybszych odpowiedzi na pytania użytkownika.
* Wielowymiarową analizę danych, możliwość zmiany punktu widzenia danych dzięki manipulacji wymiarami.



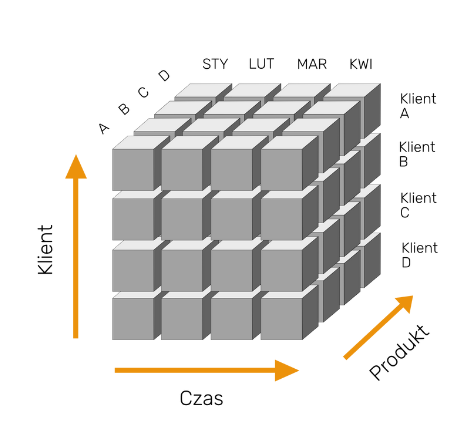
Rys. 2 piramida BI

Źródło: <https://www.mimuw.edu.pl/~son/datamining/DM2008/W12-olap.pdf>

# Kostka OLAP

Aby system OLAP mógł sprawnie dostarczać użytkownikowi tych wszystkich właściwości, o których pisałem wyżej, musi posiadać zaimplementowane narzędzie pozwalające najpierw na przechowywanie danych, a następnie ich analizę. Tym narzędziem, a właściwie sposobem organizacji, jest kostka OLAP.

Kostka OLAP to popularna struktura danych umożliwiająca analizę, opartą na modelu wielowymiarową. Są to jakby specjalne sześciany z uporządkowanymi danymi, ułożonymi w hierarchiczny sposób, które pokazują dokładne i wiarygodne informacje z danego okresu. Dane w kostce przechowywane są w sposób przypominający bardziej wielowymiarowe arkusze kalkulacyjne, niż standardową bazę danych. Dzięki zastosowaniu wymiarów kostka pozwala użytkownikowi na oglądanie danych z różnych punktów widzenia. Obejmuje też mechanizmy generowania raportów oraz wizualizacji danych.



Rys. 3 poglądowa budowa kostki OLAP

Źródło: <https://www.intense.pl/blog/baza-wiedzy,1095.html>

# Reguły budowania kostek OLAP

Istnieją reguły, które stanowią podstawę wyboru narzędzi OLAP w analityce danych. W roku 1993 brytyjski informatyk Edgar Frank Codd zaproponował 12 takich reguł:

1. Wielowymiarowy widok koncepcyjny
2. Przejrzystość
3. Dostępność
4. Stała wydajność raportowania
5. Architektura klient-serwer
6. Uniwersalna obsługa wymiarów
7. Dynamiczne obsługiwanie rzadkich macierzy
8. Wsparcie dla wielu użytkowników
9. Brak ograniczeń dla krzyżowych operacji między wymiarami
10. Intuicyjne manipulowanie danymi
11. Elastyczne tworzenie raportów
12. Bezlimitowe wymiary i poziomy agregacji

# 

# Typy kostek OLAP

Wyróżniamy 3 podstawowe modele danych używane w systemach analitycznych OLAP - ROLAP, MOLAP i HOLAP.

**ROLAP**

Model relacyjny, czyli ROLAP (ang. Relational OLAP) wykorzystuje relacyjne systemy zarządzania bazami danych, w którym dane są uporządkowane w tradycyjnych metodach, takich jak wiersze i kolumny w hurtowni danych. Raporty oparte na danych przechowywanych w modelu ROLAP bezpośrednio sięgają do tabel w hurtowni danych podczas wywoływania. Te tabele często przyjmują postać gwiazdy lub płatków śniegu, gdzie tabele faktów są połączone relacjami z tabelami wymiarów. W celu zwiększenia efektywności narzędzia ROLAP często wyposażone są w specjalistyczne silniki SQL, umożliwiające kompleksowe analizy wielowymiarowe.

**MOLAP**

Model wielowymiarowy MOLAP (ang. Muli dimensional OLAP) wykorzystują specjalne bazy danych i struktury danych, aby zarządzać, porządkować i analizować przechowywane informacje. MOLAP pobiera dane z głównego magazynu danych i innych ustalonych źródeł. Przedstawia te dane w postaci wielowymiarowych kostek, obliczając wcześniej wszystkie sumy na różnych wymiarach. Po przetworzeniu dane są zapisywane w bazie danych MOLAP. Użytkownicy biznesowi mogą również uzyskać dostęp do wielowymiarowych danych w celu dalszej analizy i raportowania. Zapytania w bazie danych MOLAP generalnie działają szybciej ze względu na zastosowane partie procesów obliczeń wstępnych. Architekturę MOLAP stosuje się najczęściej do prezentowania wstępnie agregowanych danych.

**HOLAP**

Model hybrydowy HOLAP (ang. Hybrid OLAP) próbuje połączyć pozytywne aspekty zarówno MOLAP, jak i ROLAP, aby wykorzystać możliwości przechowywania i przetwarzania danych. HOLAP przechowuje w swoim systemie dane (dane wielowymiarowe i dane relacyjne) wymagane przez użytkowników biznesowych.

Podsumowanie korzyści prezentowanych przez omawiane sposoby architektury danych:

Korzyści ROLAP:

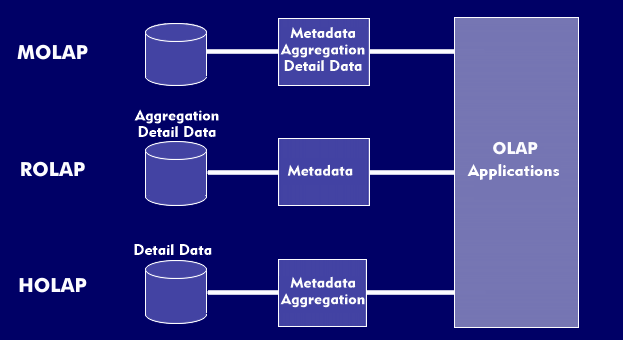
* Wykorzystanie funkcji relacyjnej bazy danych
* Obsługa dużych zbiorów danych

Korzyści MOLAP:

* Szybkie wyszukiwanie danych
* Szybkie przetwarzanie złożonych obliczeń

Korzyści HOLAP:

* Wyższa wydajność niż w MOLAP
* Możliwość drążenia danych jak w ROLAP



Rys. 4 typy kostek OLAP

Źródło: <https://www.itwissen.info/en/hybrid-OLAP-HOLAP-121971.html#gsc.tab=0>

# Część praktyczna

# Cele Projektu

Celem projektu jest wykorzystanie programu SAS Data Integration Studio do analizy danych firmy Adventure Works Cycles zgromadzonych w hurtowni danych. Firma Microsoft udostępniła bazę danych fikcyjnego przedsiębiorstwa Adventure Works Cycles, zawierającą obszerną gamę informacji o klientach, zamówieniach i produktach. Bazy użyjemy poprzez licencję MIT.



Rys. 5 logo firmy Adventure Works Cycles

Adventure Works Cycles to fikcyjne przedsiębiorstwo, które specjalizuje się w sprzedaży rowerów na rynkach Ameryki Północnej, Europie oraz Azji. Obecnie firma dąży do zwiększenia swojego udziału na rynkach światowych, poprawienia wyników sprzedażowych z poprzednich lat, lepszego dostosowania się do preferencji klientów oraz zrozumienia trendów towarzyszących wyborom konsumentów.

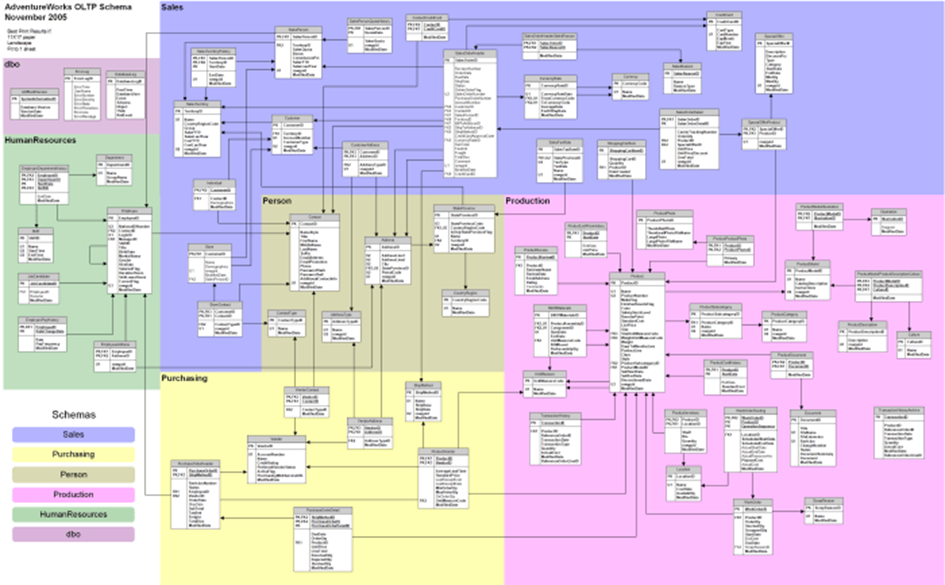
Głównym celem realizacji projektu było przeprowadzenie analizy, która odpowiadając na postawione pytania, pozwoli na podjęcie trafniejszych decyzji z perspektywy firmy. Tymi pytaniami są:

* Analiza sprzedaży ze względu na miejsce docelowej wysyłki.
* Analiza sprzedaży produktów z podziałem na kategorie.
* Analiza sprzedaży na przestrzeni czasu.
* Analiza sprzedaży pracowników.
* Analiza czynników sprzedażowych rowerów.

# Opis źródeł danych

W projekcie wykorzystano bazę Adventure Works, opublikowaną przez Microsoft i przechowującą dane o fikcyjnej firmie Adventure Works Cycles. Zawiera ona szczegółowe informacje na temat oferowanych produktów, zasobów ludzkich, a także danych bezpośrednio związanych z procesem sprzedaży, zakupu oraz produkcji. Baza składa się z 6 obszarów, które zaprezentowane zostały poniżej.

* Sales (dane sprzedażowe),
* Purchasing (dane zakupowe),
* Person (dane osobowe),
* Production (dane produkcyjne),
* HumanResources (dane o zasobach ludzkich),
* Dbo (dane systemowe).



Rys. 6 schemat bazy danych Adventure Works

W projekcie użyte zostały dane z obszarów Production, Sales, Human Resources oraz Person. Obszary te składają się z poszczególnych tabel, z których tylko część została wykorzystana w procesie analizy. Tabele wykorzystane w analizie to:

**Production**

Production\_Product, Production\_ProductSubcategory, Production\_ProductCategory

**Sales**

Sales\_SalesOrderHeader, Sales\_SalesTerritory, Sales\_SalesPerson, Sales\_SalesOrderDetail

**Human Resources**

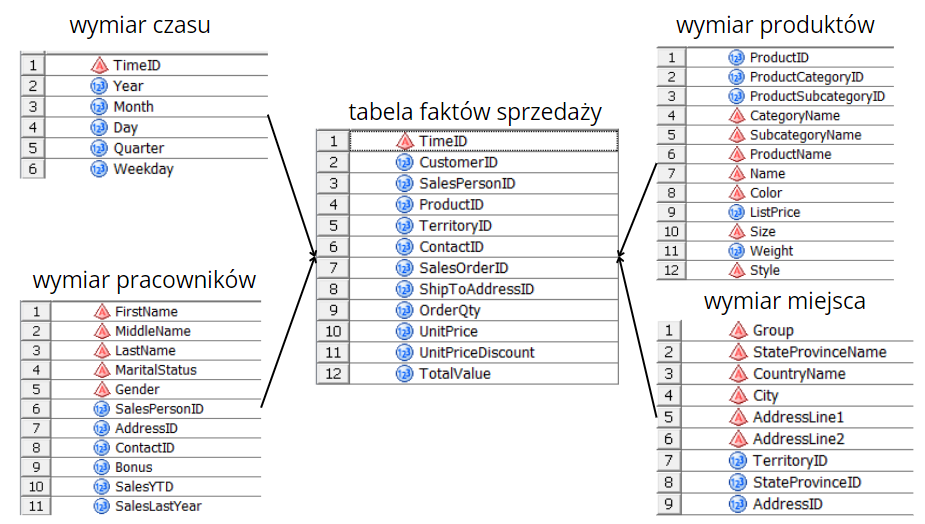
HumanResources\_Employee, HumanResources\_EmployeeAdress,

**Person**

Person\_Address, Person\_Contact

# Schemat logiczny hurtowni danych na podstawie schematu gwiazdy.

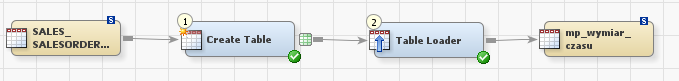
Tworząc hurtownię danych, za jej model zdecydowałem się przyjąć schemat gwiazdy. Jest to najprostszy z możliwych modeli, który składa się z tabeli faktów oraz czterech wymiarów. Wymiary odnoszą się do czasu, miejsca, klientów i produktów. Model ten pozwolił mi na przeprowadzenie procesów, które poddałem późniejszej analizie.



Rys. 7 schemat gwiazdy

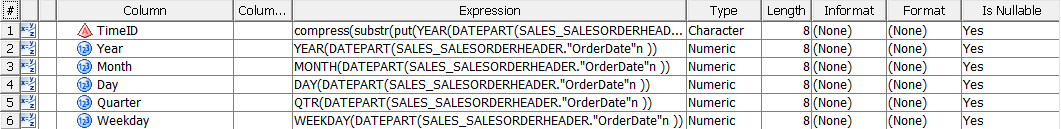
## Wymiar czasu

Pierwszym z wymiarów modelu gwiazdy obejmuje analizę danych w kontekście czasowym. Podczas tworzenia tego modelu skorzystano z informacji zawartych w tabeli Sales\_OrderHeader, gdzie kluczową rolę odgrywa zmienna OrderDate.



Rys. 8 proces tworzenia tabeli wymiaru czasu

Zamieszczony powyżej schemat ilustruje proces tworzenia wymiaru czasu. Tabelę, o której mowa, połączono z węzłem Create Table. W zakładce Result utworzono niezbędne kolumny i połączono je z atrybutem OrderDate, aby pozyskać odpowiednie informacje. W celu ich pozyskania do każdej nowo utworzonej kolumny dodano odpowiedni kod:



Rys. 9 funkcje tworzące wymiar czasu

Dla TimeID użyto kodu:

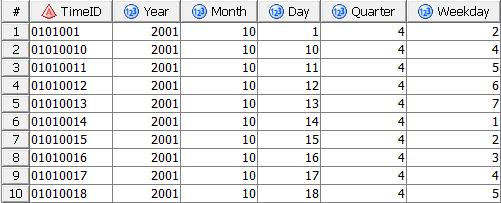
compress(substr(put(year(datepart(SALES\_SALESORDERHEADER."OrderDate"n)),4.),3,2) ||'0'|| put(month(datepart(SALES\_SALESORDERHEADER."OrderDate"n)),2.)||'0'||

put(day(datepart(SALES\_SALESORDERHEADER."OrderDate"n)),2.)).

Dla wierszy Year, Quarter, Month, Day i Weekday (XXX – day, year, …):

XXX(datepart(SALES\_SALESORDERHEADER."OrderDate"n)).

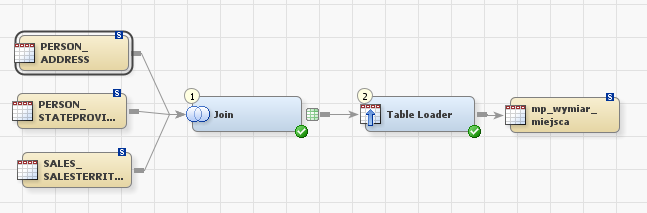
Pozwoli to na utworzenie tabeli z unikalnym ID dla każdej daty zamówienia.



Rys. 10 tabela wynikowa wymiaru czasu

## Wymiar miejsca

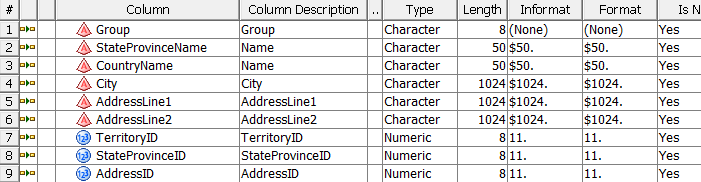
Wymiar miejsca odpowiada za informacje o terytoriach, na których odbywają się sprzedaże. Wymiar tworzą trzy tabele: Person\_Address, Sales\_SalesTerritory i Person\_StateProvince. Tabele połączone są za pomocą kluczy StateProvinceID i TerritorryID.

****

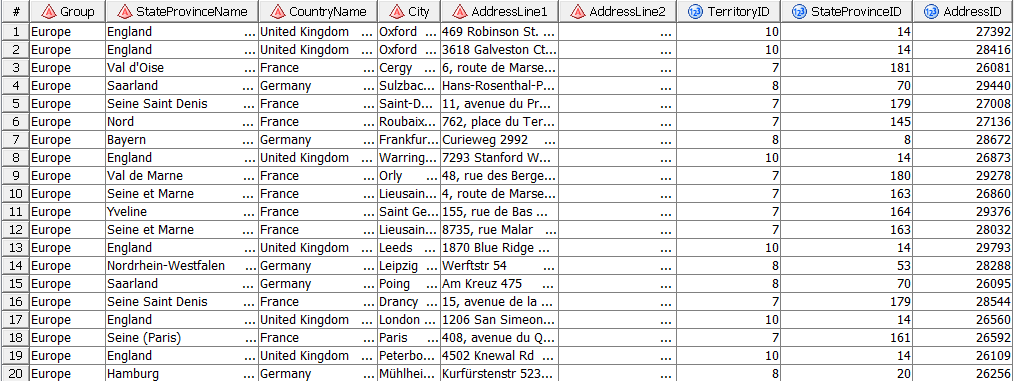
Rys. 11 proces tworzenia tabeli wymiaru miejsca



Rys. 12 klucze łączące tabele w węźle Join



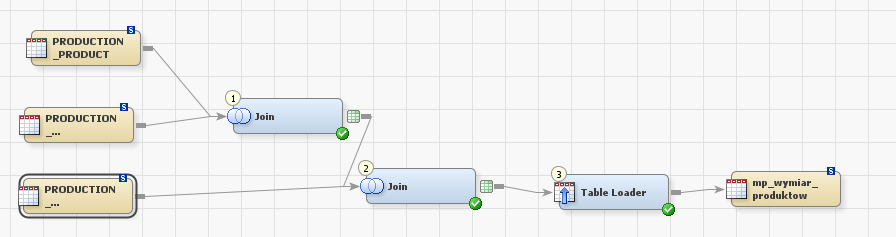
Rys. 13 kolumny tabeli wymiaru miejsca



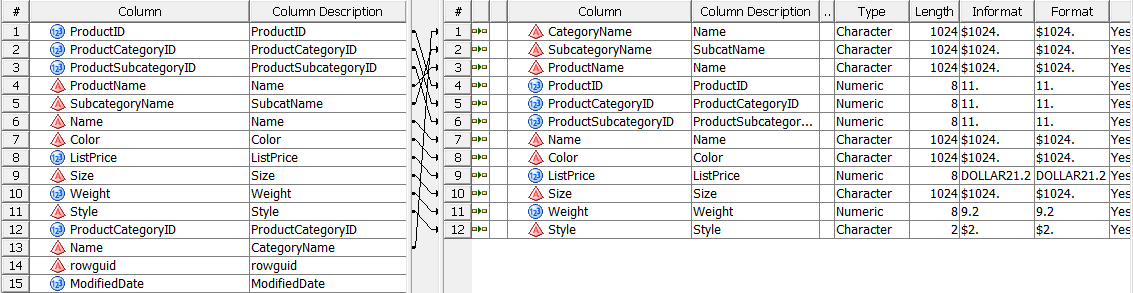
Rys.14 tabela wynikowa wymiaru miejsca

## Wymiar produktu

Wymiar produktu dostarcza informacji na temat szczegółów dotyczących produktów. Wymiar utworzony został z trzech tabel Production\_Product, Production\_ProductSubcategory Production\_ProductCategory. Tabele połączone zostały ze sobą za pomocą klucza ProductCategoryID.



Rys. 15 tworzenie tabeli wymiaru produktu



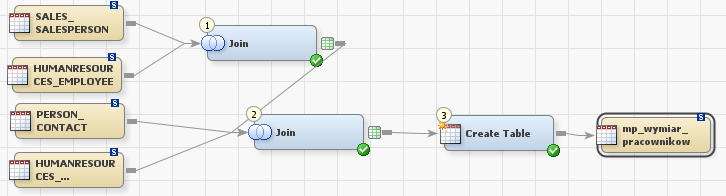
Rys. 16 mapowanie w węźle Join



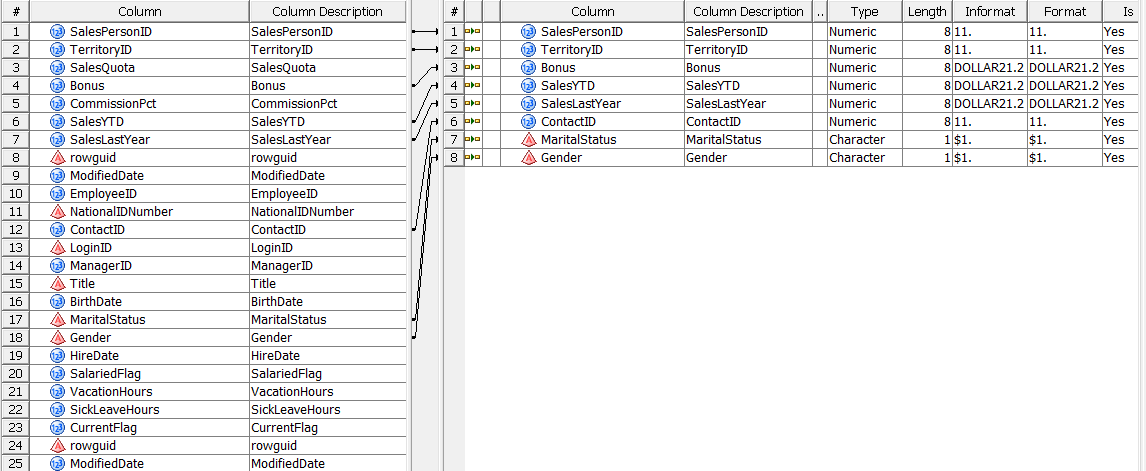
Rys. 17 tabela wynikowa wymiaru produktu

## Wymiar pracowników

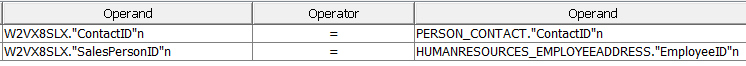
Wymiar pracowników odpowiedzialny za dane dotyczące pracowników sprzedających w firmie. Składa się on z czterech tabeli - Sales\_SalesPerson, HumanResources\_Employee, Person\_Contact oraz HumanResources\_EmployeeAdress. Tabele te zostały połaczone za pomoca kluczy: ContactID oraz SalesPersonID z EmployeeID. Tabela głowna została utworzona za pomoacą węzła Create Table.



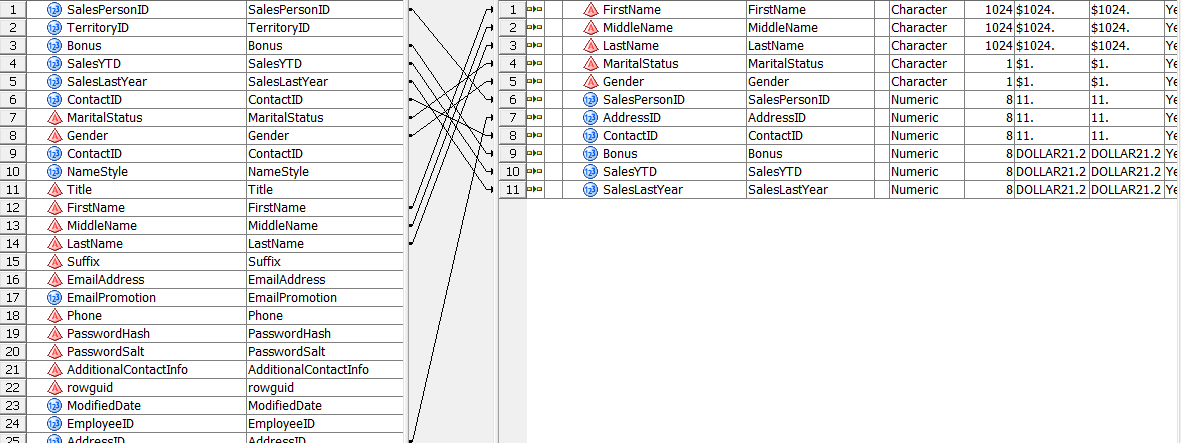
Rys. 18 tworzenie tabeli wymiaru pracowników



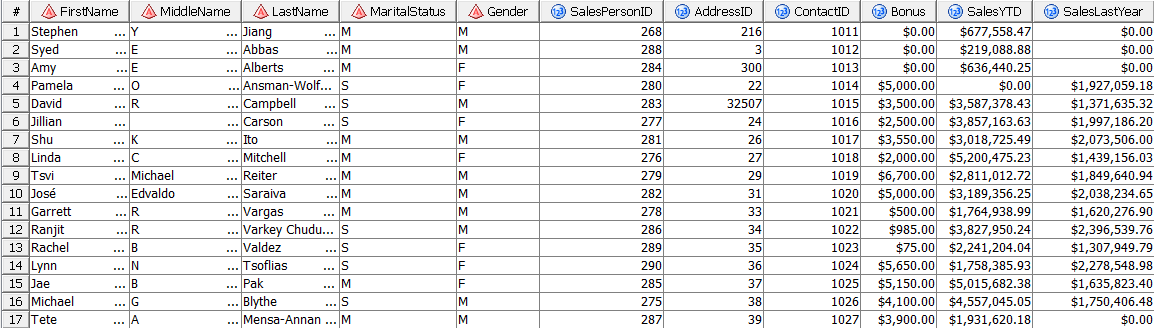
Rys. 19 mapowanie w pierwszym węźle Join



Rys. 20 klucze łączące tabele



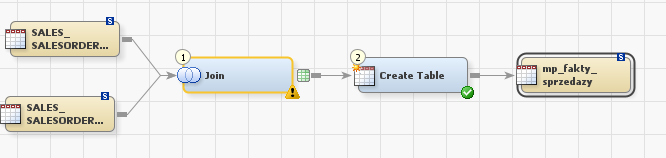
Rys. 21 mapowanie w drugim węźle Join



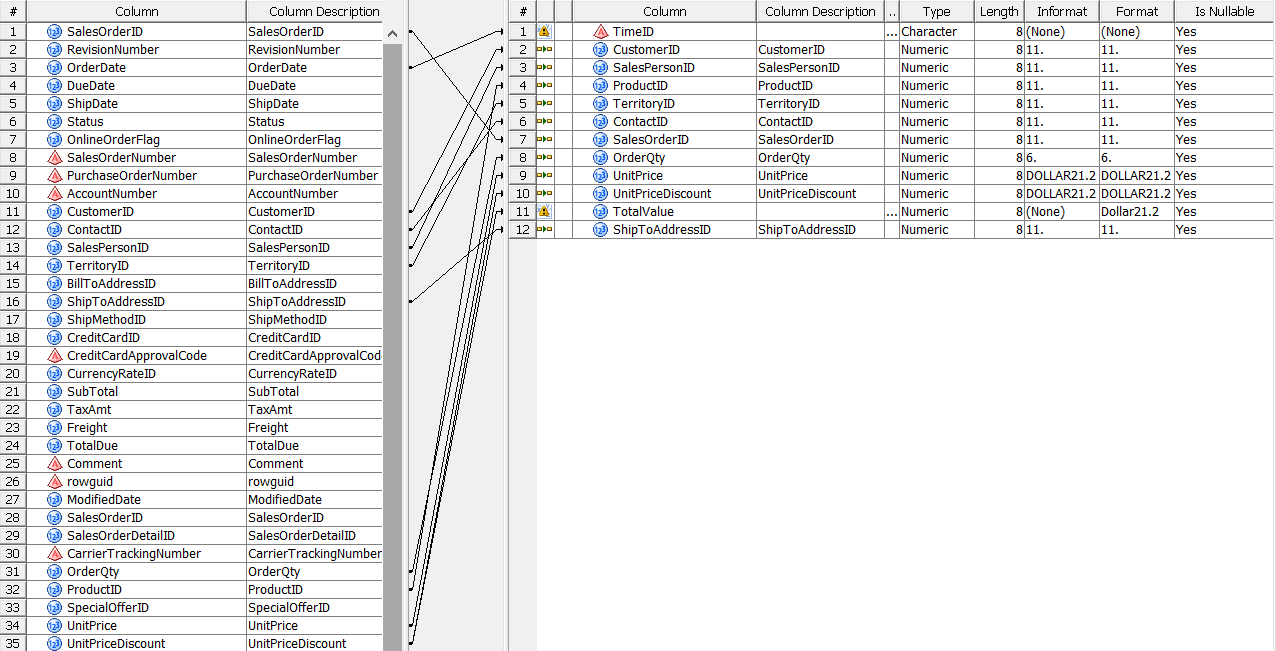
Rys. 22 tabela wynikowa wymiaru pracowników

## Tabela faktów sprzedaży

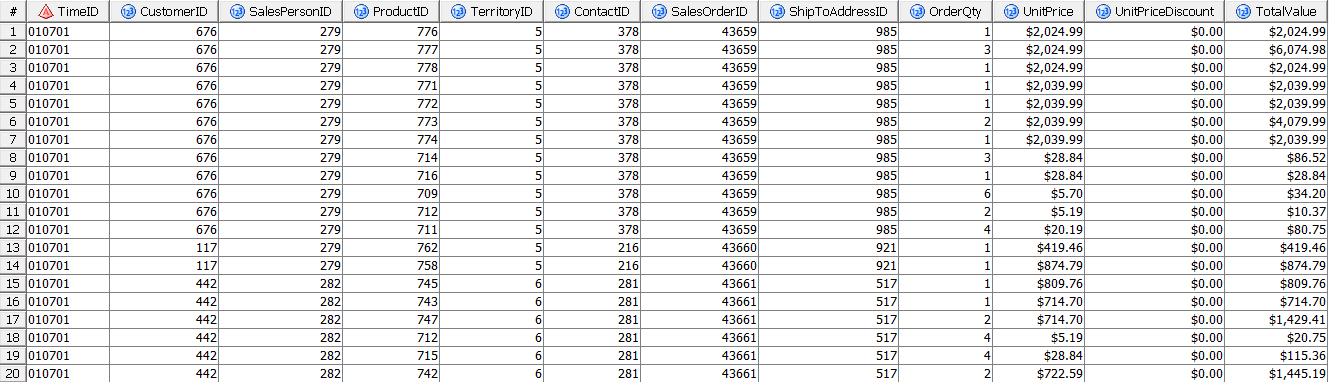
Tabela faktów sprzedaży to główna tabela schematu gwiazdy, to dzięki niej będą mogły być przeprowadzane wszystkie analizy biznesowe. Aby utworzyć tabelę faktów, wykorzystano tabele Sales\_SalesOrderDetail i Sales\_SalesOrderHeader, łącząc je standardowo za pomocą operacji Join oraz powiązując kluczem SalesOrderID. Główne kolumny dzięki, którym tabela faktów będzie mogła być łączona z wymiarami sprzedaży to: TimeID, ProductID, SalesPersonID i ShippAdressID. Bardzo ważne z perspektywy analiz są także tabele odpowiadające za dane dotyczące wartości i ilości sprzedaży, na czele z łączną wartością zamówienia — TotalValue.



Rys. 23 proces tworzenia tabeli faktów sprzedaży



Rys. 24 mapowanie w węźle Join



Rys. 25 tabela wynikowa faktów sprzedaży

# 

# Opis procesów ETL

Przy użyciu procesów ETL przeprowadziłem pięć analiz sprzedaży w firmie Adventure Works Cycles z lat 2001–2004. Badania obejmowały analizę wyników sprzedaży pod kątem różnych wymiarów, takich jak czas, lokalizacja, produkt oraz pracownik. Aby analizy były bardziej wartościowe, używałem w nich wskaźników sprzedażowych, które tworzyłem dla każdej z analizy. Były to wskaźniki:

* Łączna ilość sprzedanych produktów – Oblicza liczbę produktów sprzedanych w konkretnym mieście;
* Łączna wartość sprzedanych produktów – Wylicza całkowitą wartość sprzedaży produktów w danym mieście;
* Średnia wartość sprzedanego produktu – Określa średnią wartość pojedynczego sprzedanego produktu.

Wskaźniki w swojej budowie nie różnią się wiele w poszczególnych analizach, dlatego pozwoliłem sobie na jednorazowe umieszczenie tworzącego je kodu w sekcji Expression.

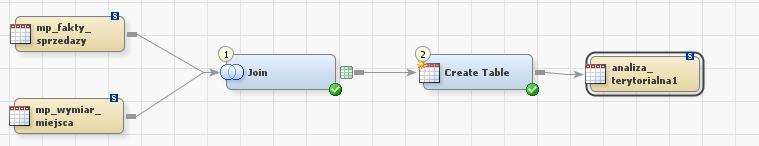


Rys. 26 kody tworzące wskaźniki

## 1. Analiza sprzedaży ze względu na miejsce docelowej wysyłki.

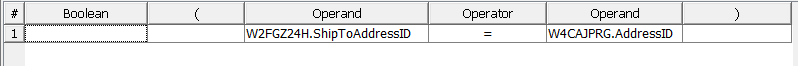
Celem realizacji procesu jest znalezienie informacji o rozkładzie terytorialnym sprzedaży, biorąc pod uwagę docelowe wysyłki produktów. Analiza podzielona zostanie na następujące pytania szczegółowe:

* Które państwa generują najwyższą wartość sprzedaży?
* Gdzie sprzedawane są średnio najdroższe produktu?
* Gdzie wysyła się najmniej produktów?



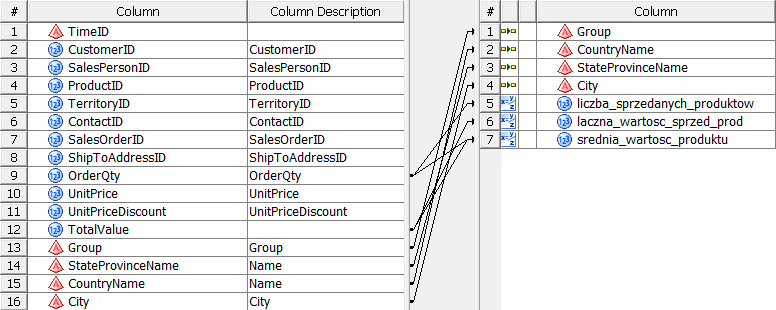
Rys. 27 Proces ETL tworzący tabelę poddawaną analizie

Tabela główna utworzona została poprzez połączenie tabeli faktu oraz wymiaru miejsca za pomocą węzła Join. Tabele zostały połączone ze sobą poprzez kolumny ShippToAdressID i AdressID.

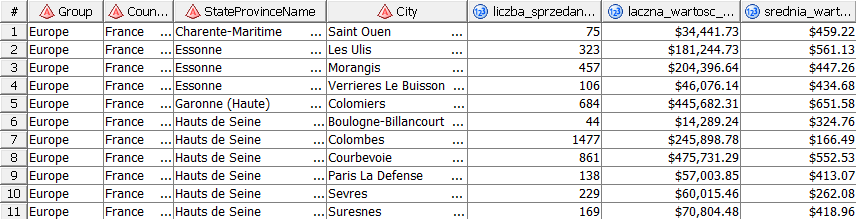


Rys. 28 klucz łączący tabele

W zakładce Select wybrane zostały potrzebne kolumny, a także dodane nowe, które umożliwia dokonanie analizy. Nowe kolumny to liczba sprzedanych produktów, łączna wartość sprzedanych produktów, średnia wartość sprzedanego produktu. Następnie dane zostały pogrupowane kolejno po kolumnach: Group, CountryName, StateProvinceName i City.



Rys. 29 mapowanie tabeli głównej i dodawanie nowych kolumn

****

Rys. 30 postać tabeli głównej

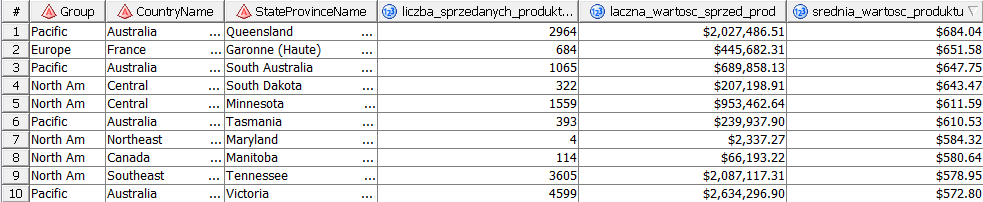
**Które państwa generują najwyższą wartość sprzedaży?**



Rys. 31 tabela analizy łącznej wartości produktów z podziałem na kraje

Najwyższą wartość sprzedanych produktów, bo 24 184 609 dolarów zanotowano w USA a dokładnie w ich południowo - zachodniej części. Następna w kolejności jest Kanada z wynikiem 16 355 770 dolarów. Z państw europejskich najwięcej przychodów generuje Wielka Brytania - 7 670 721 USD.

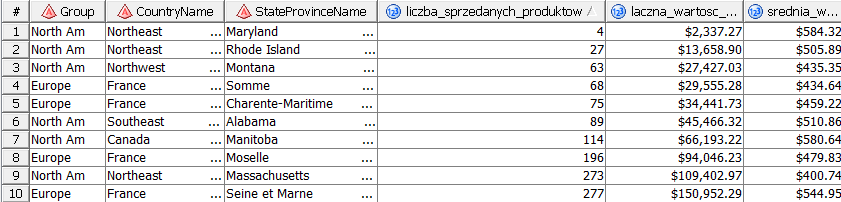
**Gdzie sprzedawane są średnio najdroższe produktu?**



Rys. 32 analizy średniej wartości produktów z podziałem na kraje i prowincje

Średnia najwyższa cena sprzedawanych produktów przypada na region Queensland w Australii i wynosi 684 dolary. Następny w kolejności jest region Garrone we Francji z wynikiem 651 USD. Najwyższą średnią wartość w przeliczeniu na jeden produkt w USA odnotowano w stanie Południowej Dakoty.

**Gdzie wysyła się najmniej produktów?**



Rys. 33 tabela analizy liczby sprzedanych produktów z podziałem na kraje i prowincje

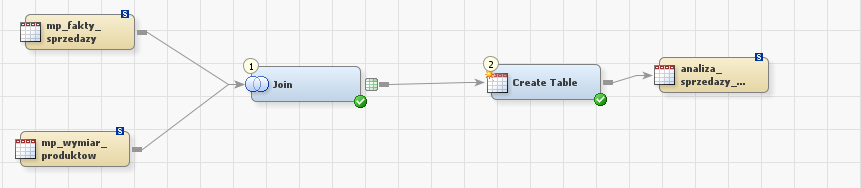
Najmniej produktów sprzedano do stanu Meryland w USA, bo jedynie 4. Następnie do stanów Rhode Island i Montana odpowiednio po 27 i 63 produkty.

## 2. Analiza sprzedaży produktów z podziałem na kategorie.

Kolejna analiza dotyczy rozkładu sprzedaży produktów w różnych kategoriach. Tabela główna, która posłuży jako podstawa analiz, utworzona została z tabel faktów sprzedaży i wymiaru produktów. W znalezienie interesujących wniosków pomogą następujące pytania szczegółowe:

* Produkty, jakiej kategorii sprzedają się najlepiej?
* Produkty, jakiej kategorii sprzedają się najgorzej?

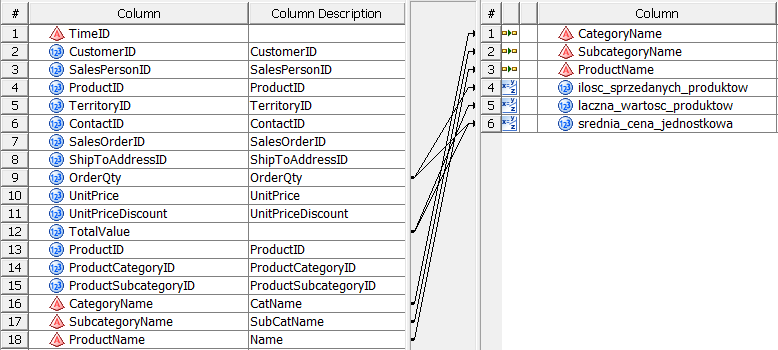
Do tabeli głównej analizy dodano kolumny zawierające wartości sprzedażowe dla każdego rodzaju produktu, są to liczba sprzedanych produktów, łączna wartość sprzedanych produktów, średnia wartość sprzedanego produktu.



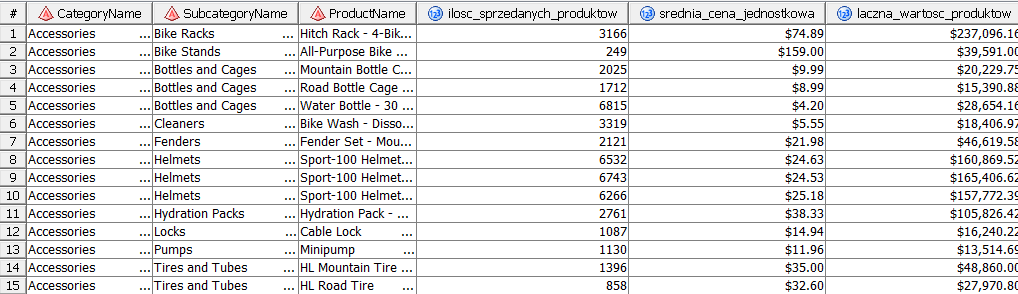
Rys. 34 Proces ETL tworzący tabelę główną



Rys. 35 klucz łączący tabele



Rys. 36 mapowanie tabeli głównej



Rys. 37 wygląd tabeli głównej

**Produkty, jakiej kategorii sprzedają się najlepiej?**

****

Rys. 38 tabela analizy najwyższych wartości sprzedażowych dla danej kategorii

Z analizy jasno wynika, że rowery są produktami, które przynosiły firmie największe zyski. Rowery dominują również pod względem ilości sprzedawanych sztuk jak i najwyższej średniej cenie jednostkowej. Drugą pod względem przyniesionych firmie dochodów kategoria są komponenty.

**Produkty, jakiej kategorii sprzedają się najgorzej?**



Rys. 39 tabela analizy najwyższych wartości sprzedażowych dla danej kategorii

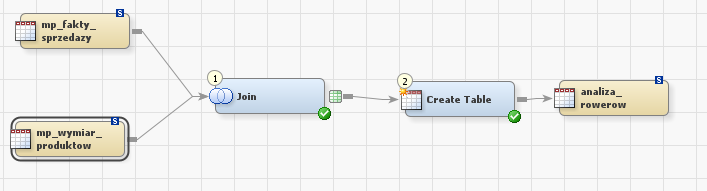
Najniższa sprzedaż zanotowały kategorie ubrań i akcesoriów. Były to również kategorie o najniższej średniej cenie jednostkowej. Co ciekawe, ubrania były drugimi najczęściej sprzedawanymi produktami.

## 

## 3. Analiza czynników sprzedażowych rowerów.

Celem realizacji procesu jest znalezienie informacji o rozkładzie czynników sprzedażowych rowerów. Analiza zostanie przeprowadzona na podstawie tabeli głównej, którą utworzą tabele faktów sprzedaży i wymiarów produktów. Tabele zostaną połączone ze sobą za pomocą klucza ProductTD, następnie nałożone zostanie ograniczenie filtrujące jedynie kategorię rowery. Dodane zostaną także wskaźniki wspomagające analizę, są to ilość sprzedanych produktów, łączna wartość sprzedaży i średnia wartość sprzedawanych produktów. Analiza odpowie na następujące pytania szczegółowe:

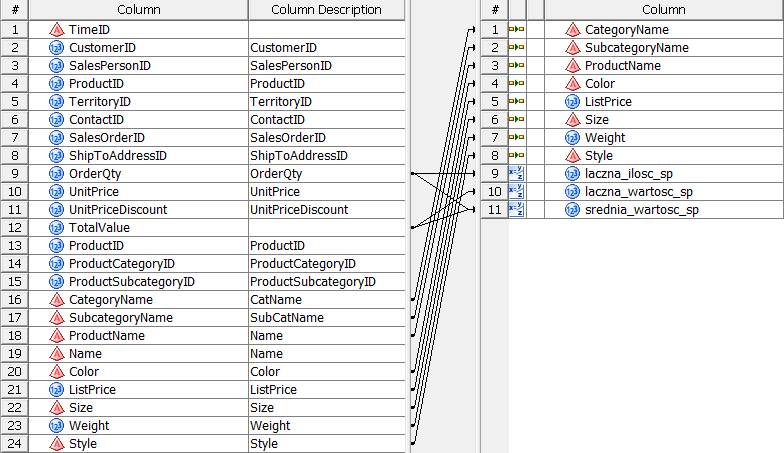
* Jaka podkategoria rowerów sprzedaje się najlepiej, a jaka najgorzej?
* Rowery, o jakich kolorach osiągnęły najwyższe wartości sprzedaży?
* Rowery, o jakich rozmiarach sprzedawały się najlepiej?



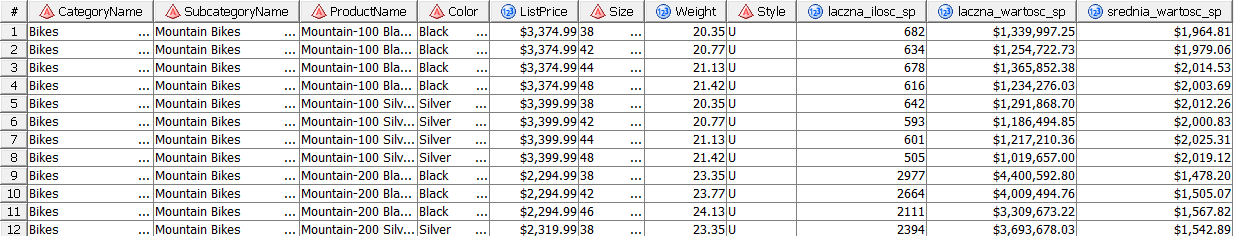
Rys. 40 proces ETL tworzący tabelę główną analizy



Rys. 41 klucz główny oraz ograniczenia w Where

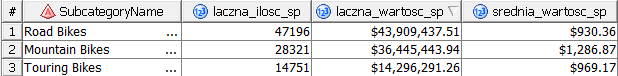


Rys. 42 mapowanie w węźle Join



Rys. 43 postać tabeli głównej

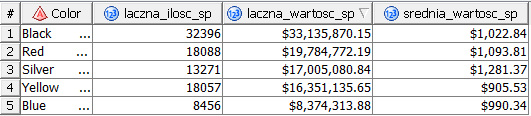
**Jaka podkategoria rowerów sprzedaje się najlepiej, a jaka najgorzej?**



Rys. 44 tabela analizy wartość sprzedaży poszczególnych rodzajów rowerów

Analiza dostarcza informacji, że najlepiej sprzedającymi się rowerami były rowery szosowe, które wygenerowały 43 909 437 dolarów dochodów. Były to także najliczniej sprzedające się rowery, o najniższej średniej cenie jednostkowej. O około 7,5 mln dolarów mniejsza sprzedaż zanotowały rowery górskie, które jednak odznaczały się najwyższą średnią ceną sprzedaży.

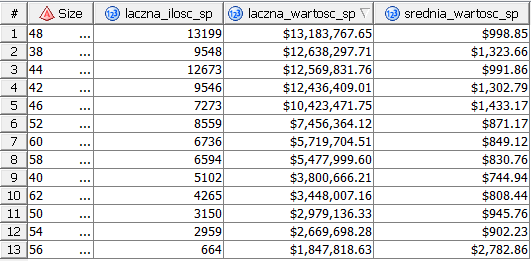
**Rowery, o jakich kolorach osiągnęły najwyższe wartości sprzedaży?**



Rys. 45 tabela analizy wartość sprzedaży poszczególnych kolorów rowerów

Analiza pokazuje, że największą wartość sprzedaży osiągnęły rowery o kolorze czarnym, a ich wynik znacząco przewyższał pozostałe kolory. Były to też rowery, które sprzedawały się najczęściej, o ponad 14 tys. sztuk niż następny w kolejności kolor czerwony. Najsłabiej sprzedającym się kolorem był niebieski. Najwyższa średnia jednostkowa cena przypada na rowery o kolorze srebrnym.

**Rowery, o jakich rozmiarach sprzedawały się najlepiej?**

****

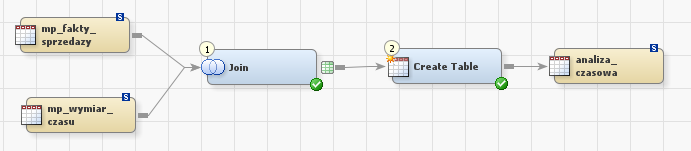
Rys. 46 tabela analizy wartość sprzedaży poszczególnych rozmiarów rowerów

Z analizy wynik, że firma najwięcej zarobiła na rowerach o rozmiarze 48. Był to także rozmiar, który sprzedawał się najczęściej. Najgorszy wynik sprzedaży uzyskały rowery o rozmiarze 56. Najwyższa średnia cena jednostkowa przypada na rowery o rozmiarze 56 i przewyższa ona bardzo znacząco inne rozmiary, o ponad 1000 dolarów na sztuce. Najtańszy średnio jest rozmiar 40.

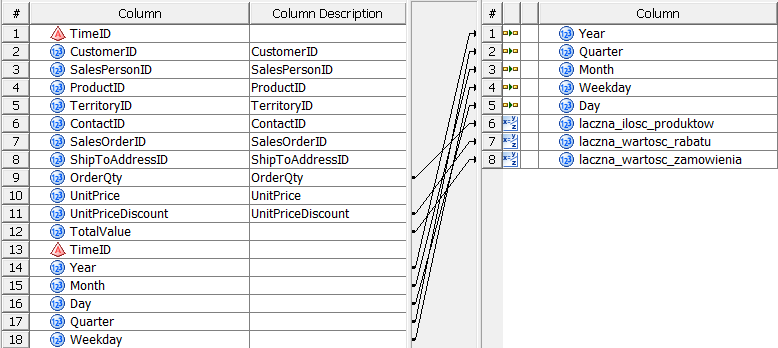
## 4. Analiza sprzedaży na przestrzeni czasu.

Ta analiza dostarczy informacji na temat rozkładu sprzedaży w czasie. Tabela główna, która posłuży do analizy, utworzona została z tabeli faktów sprzedaży oraz wymiaru czasu. Tabele połączone zostały za pomącą klucza TimeID oraz pogrupowane kolejno według roku, kwartału, miesiąca, dnia tygodnia i dnia miesiąca. Do tabeli głównej analizy dodano kolumny zawierające wartości sprzedażowe dla danej sprzedaży w czasie, są to liczba sprzedanych produktów, łączna wartość udzielonego rabatu oraz łączna wartość zamówienia. W analizie użyte zostaną następujące pytania szczegółowe:

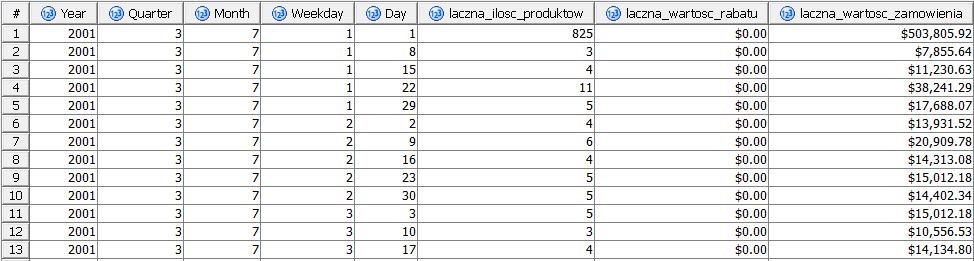
* Jak wyglądał rozkład sprzedaży na przestrzeni lat?
* Jak wyglądał rozkład sprzedaży w danych kwartałach?
* Jak wyglądał rozkład sprzedaży w danych miesiąca roku?



Rys. 47 Proces ETL tworzący tabelę główną

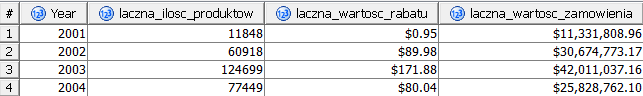


Rys. 48 mapowanie tabeli głównej w węźle Join



Rys. 49 wygląd tabeli głównej analizy

**Jak wyglądał rozkład sprzedaży na przestrzeni lat?**

****

Rys. 50 tabela analizy sprzedaży na przestrzeni lat

Jak wynika z analizy, najwyższą łączną wartość sprzedaży zanotowano w roku 2003, najniższą zaś w roku 2001. Na lata te odpowiednio przypada też najwyższa i najniższa liczba sprzedanych produktów. Taka sama sytuacja występuję w kontekście rabatu. Między rokiem 2003 a 2004 zanotowano istotny spadek sprzedaży.

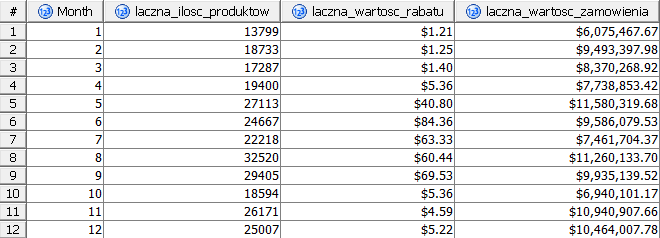
**Jak wyglądał rozkład sprzedaży w danych kwartałach?**



Rys. 51 tabela analizy sprzedaży w podziale na kwartały roku

W ciągu czterech lat firma notowała najwyższe wyniki sprzedaży w drugim kwartale każdego roku, natomiast najniższe osiągała w pierwszym kwartale. Pomimo tych różnic, warto zauważyć, że nie są one bardzo duże, co sugeruje, że obroty były stosunkowo stabilne i równomiernie rozłożone na poszczególne kwartały. To oznacza, że firma utrzymywała dość stałą dynamikę sprzedaży przez cały okres analizy.

**Jak wyglądał rozkład sprzedaży w danych miesiąca roku?**



Rys. 52 tabela analizy sprzedaży w podziale na miesiące

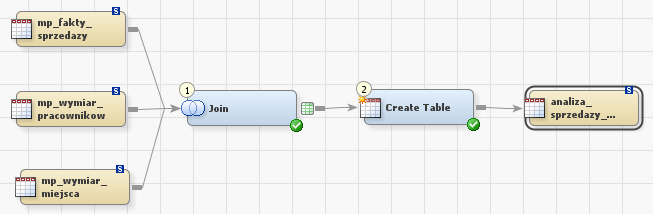
Najwyższe wartości sprzedaży firma odnotowywała w maju oraz w sierpniu, najniższe zaś w styczniu i w październiku. Jako że większość krajów, do których sprzedawano towary, leży na półkuli północnej, można stwierdzić, że najwyższe wartości sprzedażowe odnotowywano wiosną i latem. Największą ilość produktów sprzedano w sierpniu i wrześniu

## 

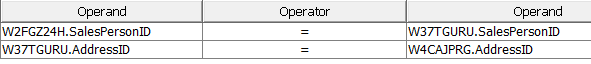
## 5. Analiza sprzedaży pracowników.

Ta analiza dostarczy informacji na temat wskaźników sprzedażowych wśród sprzedawców firmy. Tabela główna została utworzona na podstawie trzech tabel — faktów sprzedaży, wymiaru pracowników oraz wymiaru miejsca. Do tabel dodano wskaźniki wspomagające analizę, są to ilość sprzedanych produktów, łączna wartość sprzedaży, średnia wartość sprzedawanych produktów i łączna wartość udzielonego rabatu. W analizie odpowiedziano na następujące pytania szczegółowe:

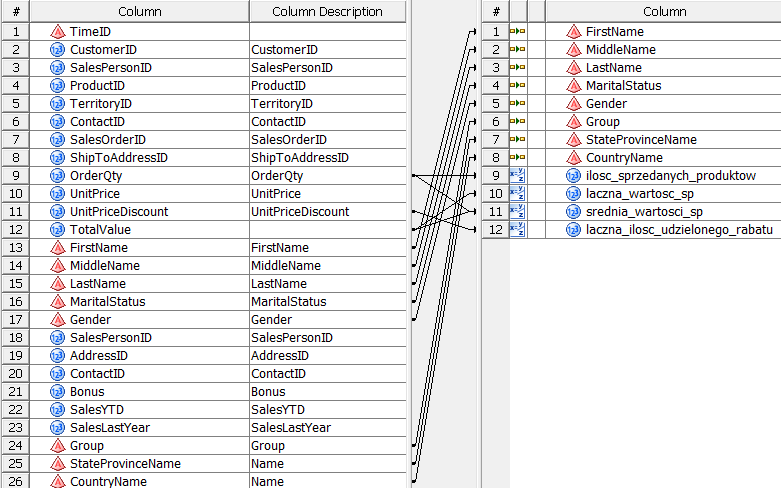
* Którzy pracownicy uzyskali najwyższą, a którzy najniższa wartość sprzedaży?
* Jak wygląda rozkład płci wśród sprzedawców?



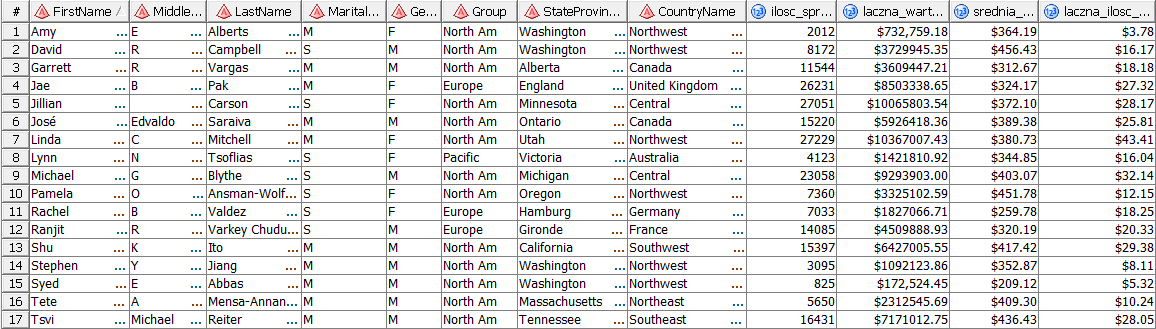
Rys. 53 Proces ETL tworzący tabelę główną



Rys. 54 klucze łączące tabele w węźle Join



Rys. 55 mapowanie tabeli głównej w węźle Join



Rys. 56 postać tabeli głównej używanej do analiz

**Którzy pracownicy uzyskali najwyższą, a którzy najniższa wartość sprzedaży?**

****

Rys. 57 tabela analizy pracowników pod kątem wartości sprzedaży

Z analizy wynika, że największymi wartościami sprzedaży mogą się poszczycić sprzedawcy z USA. Trzy pierwsze miejsca zajmują odpowiednio Linda Mitchell, Jillian Carson i Michael Blythe. Najwięcej produktów sprzedała Linda Mitchell, a produkty o najwyższej średniej wartości sprzedaje David Campbell.

Po drugiej stronie rankingu wartości sprzedanych towarów znajduje się Syed Abbas i Amy Alberts, oboje ze stanu Washington w USA. Syed Abbas sprzedawał również najmniej produktów, a do tego o najniższej średniej wartości.

**Jak wygląda rozkład płci wśród sprzedawców?**



Rys. 58 tabela analizy sprzedaży pod względem płci sprzedawców

Z przeprowadzonej analizy wynika, że mężczyźni osiągnęli łącznie wyższy wynik sprzedażowy niż kobiety. Jednakże jeżeli uwzględnimy, że kobiet na tym stanowisku jest o trzy mniej i uśrednimy wynik dla jednej osoby, to kobiety przewyższają mężczyzn. Łączna wartość sprzedanych produktów uśredniona dla kobiety to 5 177 555 dolarów a dla mężczyzny 4 244 815 dolarów.

# Analiza uzyskanych wyników

Praktyczna część projektu, która obejmowała analizę sprzedaży w firmie Adventure Works Cycles w latach 2001-2004, polegała na szczegółowym zbadaniu kluczowych elementów działalności przedsiębiorstwa. Obejmowały one wymiary czasowe, geograficzne, produktowe i pracownicze. Na początku stworzyłem schemat gwiazdy, co doprowadziło do utworzenia tabeli faktów oraz tabel wymiarów obejmujących czas, miejsce, produkt i pracownika. Następnie przeprowadziłem odpowiednie procesy ETL w celu pozyskania pożądanych danych. Podsumowując analizy przeprowadzonych wyników pod kątem możliwych raportów, jakie można dzięki nim przeprowadzić, pozwolę sobie podzielić to podsumowanie na części.

W analizie pierwszej zbadałem, jak terytorialnie rozkłada się sprzedaż firmy. Wnioski z niej płynące jednoznacznie dowodzę, że to USA jest największym odbiorcą produktów firmy. Kraj ten generuje znacznie wyższe przychody niż inne miejsca zbytu. Szczególnie ważne z perspektywy biznesowej jest południowo - zachodnia część kraju, która znacząco przewyższa resztę docelowych regionów. Informacje te mogą posłużyć do oceny potencjału poszczególnych rynków zbytu i być cenną wskazówką dla grupy zarządczej firmy.

Najwyższa łączna wartość sprzedaży nie oznacza jednak najwyższych średnich cen jednostkowych produktów. Najdroższe produkty kupowano w Australii i to tam firma mogłaby wysyłać więcej produktów typu exclusive. Analiza regionów o najniższej sprzedaży pozwoli zdecydować o np. zamknięciu nierentownych rynków zbytu.

Kolejna analiza dotycząca wyników sprzedażowych poszczególnych kategorii produktów, daje nam obraz najbardziej dochodowych towarów firmy. Z analizy jasno wynika, że produktami, które przynoszą najwyższe zyski, są rowery. Pozwala to zdecydować o całościowym rozwoju firmy na przyszłe lata. Z perspektywy firmy najbardziej racjonalne jest ulepszanie i rozwijanie tych produktów, które przynoszą najwyższe zyski. Jest to też powód do poszukiwania innowacji w tej kategorii produktów.

Analiza następna, czyli trzecia, jest nijako kontynuacją poprzedniej. Dostarcza nam ona cennych informacji na temat preferencji klientów odnośnie rowerów. Dane dowodzą, że najlepiej sprzedającym się typem rowerów był rower szosowy. Z badania wywnioskowywać można również obraz najbardziej dochodowych rowerów. Były to rowery czarne o umiarkowanych rozmiarach. Pozwala to na dostosowanie przez firmę aspektu marketingowego i skupienie się na pokazywaniu konsumentom obrazu produktów, które potencjalnie będą dla niego najbardziej pożądane. Z badania wynika, że najgorzej sprzedającym się kolorem rowerów był niebieski, można zatem zastanowić się nad usunięciem tego koloru z oferty sprzedażowej.

Analiza sprzedaży na przestrzeni czasu pozwala określić ogólny trend rozwojowy firmy, jak również zbadać intensywność sprzedażową w poszczególnych częściach roku. Z przeprowadzonego badania wynika, że w ostatnim roku, z którego pochodzą dane, firma zaliczyła spadek dochodów. Skłania to nad zastanowieniem się nad ewentualnymi problemami firmy i rozpoczęcie ich rozwiązywania. Kolejne części analizy dostarczają informacji na temat tego, w jakich kwartałach i miesiącach obrót firmy był najwyższy. Wynika z nich, że to wiosną i latem obroty firmy są najwyższe, choć trzeba przyznać, że na przestrzeni całego roku różnice pomiędzy kwartałami nie są duże. Zdecydowanie najmniej dochodów firma generuje w styczniu. Przeprowadzona analiza pozwala na wygenerowanie raportów, które pomogą na przykład w decyzji o rozplanowaniu terminów promocji i przecen, tak aby miały one najkorzystniejszy wpływ na ogólny dochód firmy.

Ostatnią przeprowadzoną analiza, jest analiza sprzedaży poszczególnych pracowników. Pozwala ona wygenerować raport określający, którzy pracownicy uzyskali najlepsze wyniki sprzedażowe. To z kolei jest podstawą do ewentualnych decyzji o awansie i premiach. Pozwala też w drugą stronę dowiedzieć się którzy pracownicy generują najniższe sprzedaże i to przeanalizować przyczyny tej sytuacji.

Reasumując, uważam, że przeprowadzone analizy dostarczają bardzo cennych informacji, które niewątpliwe pomogą w sprawnym funkcjonowaniu firmy. Zaś raporty, które dzięki nim można wygenerować, pozwolą na podejmowanie w pełni racjonalnych decyzji, mających swoje przyczyny w rzeczywistości.

# Źródła danych

Podczas pracy nad projektem korzystałem z następujących źródeł wiedzy:

* Chris Todman, „Projektowanie Hurtowni Danych. Zarządzanie Kontaktami z Klientami (CRM)”, 2003, Wydawnictwo Naukowo-Techniczne.
* Agnieszka Chodkowska-Gyurics, „Hurtownie danych”, 2020, Wydawnictwo Naukowe PWN.
* <https://www.sap.com/poland/products/technology-platform/datasphere/what-is-a-data-warehouse.html>
* <https://home.agh.edu.pl/~horzyk/lectures/db/BazyDanychAccess-HurtownieDanych.pdf>
* <https://www.sap.com/poland/products/technology-platform/datasphere/what-is-a-data-warehouse.html>
* <https://www.mimuw.edu.pl/~son/datamining/DM2008/W12-olap.pdf>
* <https://bpc-guide.pl/co-to-jest-olap-online-analytical-processing/>
* <https://mfiles.pl/pl/index.php/Systemy_OLAP>
* <https://www.studocu.com/pl/document/uniwersytet-ekonomiczny-w-katowicach/informatyka-i-ekonometria/olap/6604412>
* <https://www.educba.com/rolap-vs-molap-vs-holap/>
* <https://snehalshirole.blogspot.com/2012/04/olap-rolap-molap-and-holap.html>