

JAROSŁAW NOCUŃ

Inżynieria hurtowni danych

Grupa ćwiczeniowa IE_cw_SI



Uniwersytet
Ekonomiczny
w Katowicach

Spis treści

| | |
|---|-----------|
| Budowa hurtowni danych w rozwiązaniach Big Data. | 2 |
| Cel Realizacji Projektu | 9 |
| Opis Źródeł Danych | 10 |
| Model logiczny hurtowni danych w oparciu o schemat gwiazdy | 13 |
| Opis Procesów ETL | 18 |
| 1. Analiza wyników sprzedażowych mężczyzn i kobiet w latach 2001-2004 | 18 |
| 2. Analiza sprzedaży poszczególnych kategorii produktów na przestrzeni lat | 20 |
| 3. Analiza sprzedaży w podziale na kwartały i państwa | 23 |
| 4. Analiza największych wartości sprzedaży danych produktów, sprzedanych przez poszczególnych pracowników..... | 26 |
| 5. Podsumowanie | 29 |
| Bibliografia | 31 |

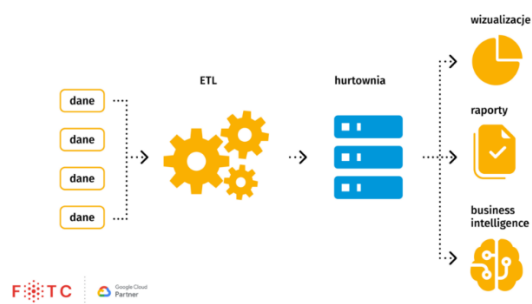
Budowa hurtowni danych w rozwiązaniach Big Data.

Hurtownie danych

Hurtownia danych (ang. data warehouse) to termin, który można interpretować w szerszym lub węższym ujęciu. W szerszym ujęciu hurtownia danych to całokształt systemu, który ma na celu nie tylko pobieranie danych z zasobów informacyjnych przedsiębiorstwa, oprócz tego jego zadaniem jest umieszczanie tych danych w nowej bazie, udostępniając je decydantom. W węższym ujęciu hurtownia jest bazą danych, której podstawowym zadaniem jest przechowywanie danych w celu zapewnienia do nich sprawnego dostępu aby ułatwić proces podejmowania decyzji.

Popularna jest także definicja hurtowni danych B. Inmona z 1996:

„Hurtownia danych to tematyczna baza danych, która trwale przechowuje zintegrowane dane opisane wymiarem czasu.”



Rysunek 1 Schemat przebiegu danych w hurtowni danych

Hurtownia danych posiada strukturę rozróżniającą dwa typy tabel. Są to: tabele faktów oraz tabele wymiarów

Tabela faktów zawiera wystąpienia konkretnych zdarzeń w rzeczywistości, które są przedmiotem analizy (np. wielkość sprzedaży). Tabela wymiarów zawiera informacje referencyjne według których dokonywane są analizy (np. czas, produkt). Zwyczajowo w strukturze hurtowni danych stosowany jest przedrostek Dim (od ang. dimension, wymiar) dla wymiarów oraz Fact (od ang. fact table, tabela faktów) dla tabel faktów.

Właściwości hurtowni danych mające wspomóc proces podejmowania decyzji według B. Inmona:

1. Zorientowanie tematyczne – zebrane dane dotyczą określonego tematu, a nie działań. Hurtownia danych koncentruje się na analizie konkretnego zagadnienia, w przeciwieństwie do bazy operacyjnej, która obsługuje procesy.
2. Nieulotność – dane umieszczone w hurtowni zazwyczaj nie ulegają zmianie. Użytkownicy mają pewność, że zapytanie o te same parametry początkowe zawsze zwróci identyczny wynik, niezależnie od liczby uruchomień czy czasu od ostatniego wykonania zapytania.
3. Zróżnicowanie czasowe – zbierane są dane historyczne, co oznacza, że mają charakter przyrostowy. W hurtowniach danych, w przeciwieństwie do baz operacyjnych, wszelkie zmiany (większość z nich) skutkują dodaniem nowych danych do bazy, a nie ich modyfikacją.
4. Zintegrowanie – dane są spójne. Nie chodzi tylko o spójność logiczną, która również musi być zachowana, ale przede wszystkim o spójność formatu. Dane przedstawiające te same informacje powinny mieć identyczny format, sposób kodowania oraz postać. Bez względu na źródło, z którego pochodzą, zostaną one przechowywane w tym samym formacie.

Big Data

Definicja Big Data według Parlamentu Europejskiego:

Big data odnosi się do zbiorów danych, które są tak duże i złożone, że do przetwarzania wymagają nowych technologii, takich jak sztuczna inteligencja. Dane pochodzą z wielu różnych źródeł. Często są to dane tego samego typu, np. dane GPS z milionów telefonów komórkowych są wykorzystywane do unikania korków drogowych. Dane mogą też być mieszane - np. dokumentacja zdrowotna i dane z aplikacji dla pacjentów. Technologia umożliwia bardzo szybkie gromadzenie danych (w czasie zbliżonym do rzeczywistego) i analizowanie ich w celu uzyskania nowych wniosków.

Definicja Big Data według Oracle:

Big Data to zbiory danych cechujących się większą różnorodnością i docierających do przedsiębiorstw w coraz większych ilościach i z większą szybkością. Wymienione trzy cechy uznaje się za kluczowy wyróżnik tego rodzaju zbiorów.

Kluczowe cechy Big Data:

Wyżej wymienione trzy kluczowe cechy Big Data to kolejno ilość, szybkość i różnorodność.

Ilość

Ilość danych ma znaczenie. W Big Data przetwarza się duże ilości nieustrukturyzowanych danych o małej gęstości. Mogą to być dane o nieznanej wartości, takie jak strumienie kliknięć na stronie internetowej lub w aplikacji mobilnej bądź dane z urządzeń z czujnikami. Czasami mogą to być dziesiątki terabajtów danych. A w przypadku dużych przedsiębiorstw — setki petabajtów.

Szybkość

Szybkość oznacza szybkie tempo odbierania, przetwarzania i wykorzystywania danych do dalszych działań. Zwykle dane, które docierają najszybciej są przekazywane bezpośrednio do pamięci, zamiast być zapisywane na dysku. Istnieją także inteligentne produkty z dostępem do Internetu, które działają w czasie rzeczywistym lub zbliżonym do rzeczywistego.

Wymagają one oceny i podejmowania działań w czasie rzeczywistym.

Różnorodność

Różnorodność oznacza dostępność wielu typów danych. Tradycyjne typy danych miały uporządkowaną strukturę i mogły być zapisane w relacyjnej bazie danych. Kiedy pojawiło się Big Data zaczęto też gromadzić nowe, nieustrukturyzowane typy danych.

Nieustrukturyzowane oraz częściowo ustrukturyzowane typy danych, do których zalicza się tekst, dźwięk i wideo, wymagają dodatkowego przetwarzania wstępnego aby móc z nich wydobyć ich znaczenie oraz obsłużyć metadane.

Apache Hive

Zgodnie z definicją ze strony hive.apache.org „Apache Hive to rozproszony, odporny na awarie system hurtowni danych, który umożliwia analitikę na masową skalę. Hive Metastore (HMS) zapewnia centralne repozytorium metadanych, które można łatwo analizować w celu podejmowania świadomych decyzji opartych na danych, dlatego jest kluczowym elementem wielu architektur jezior danych. Hive jest zbudowany na bazie Apache Hadoop i obsługuje przechowywanie na S3, adls, gs itp. poprzez hdfs. Hive umożliwia użytkownikom odczytywanie, zapisywanie i zarządzanie petabajtami danych przy użyciu języka SQL.”

Dzięki programowi Hive możliwe jest projektowanie struktury danych w znacznym stopniu bez konieczności ścisłej definicji struktury. Po zdefiniowaniu struktury, język HiveQL pozwala na przeprowadzanie zapytań dotyczących danych.



Rysunek 2 Logo Apache Hive

Hadoop Distributed File System

Hadoop Distributed File System (HDFS) to system plików o rozproszonej architekturze, który został zaprojektowany do funkcjonowania na standardowym sprzęcie komputerowym.

Chociaż wykazuje wiele cech podobnych do istniejących rozproszonych systemów plików, to jednak wyróżnia się istotnymi różnicami. HDFS charakteryzuje się wysoką odpornością na awarie i jest dedykowany do uruchamiania na niedrogim sprzęcie. Jego główną zaletą jest zapewnienie wysokiej przepustowości dostępu do danych aplikacji, co czyni go idealnym rozwiązaniem dla zastosowań wymagających przetwarzania dużych zbiorów danych.

Ponadto, HDFS adaptuje kilka standardów POSIX, aby umożliwić efektywny strumieniowy dostęp do danych systemu plików. Początkowo powstał jako infrastruktura wspierająca projekt wyszukiwarki internetowej Apache Nutch, a obecnie stanowi integralną część projektu Apache Hadoop Core.

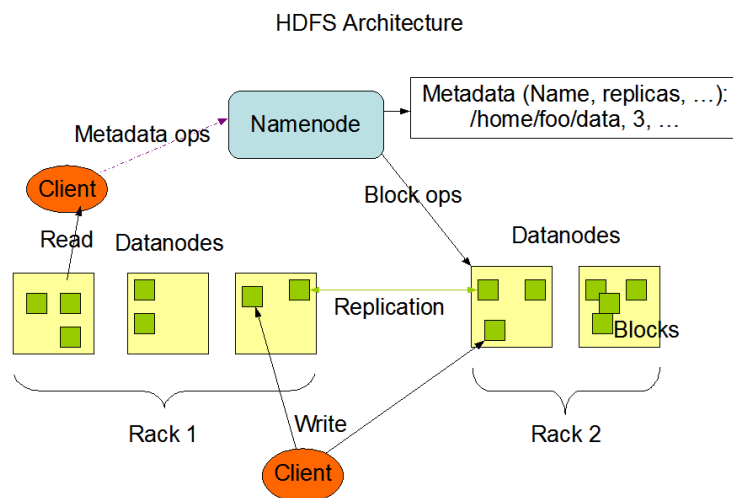
Programy operujące w ramach systemu HDFS często manipulują ogromnymi zbiorami danych. Standardowy plik w HDFS może osiągać rozmiary od gigabajtów do nawet terabajtów, co wymaga specjalnego dostrojenia systemu do obsługi tak dużych plików. Kluczowe jest zapewnienie wysokiej agregowanej przepustowości danych oraz zdolność do skalowania do setek węzłów w jednym klastrze. Dodatkowo, system powinien być w stanie obsłużyć dziesiątki milionów plików w jednej instancji, aby sprostać wymaganiom aplikacji działających w środowisku HDFS.

Architektura HDFS

Architektura HDFS oparta jest na modelu master/slave. W skład klastra HDFS wchodzi jeden węzeł NameNode, pełniący rolę serwera głównego, który zarządza przestrzenią nazw systemu plików oraz kontroluje dostęp klientów do plików. Dodatkowo, klastr składa się z pewnej liczby węzłów DataNodes, zazwyczaj po jednym na węzeł w klastrze, które zarządzają przestrzenią dyskową przyłączoną do węzłów, na których działają.

HDFS zapewnia abstrakcję przestrzeni nazw systemu plików i umożliwia przechowywanie danych użytkownika w postaci plików. Wewnętrznie pliki są dzielone na jeden lub więcej bloków, a te bloki są przechowywane w zestawie DataNodes. NameNode odpowiada za operacje związane z przestrzenią nazw systemu plików, takie jak otwieranie, zamykanie i zmiana nazw plików oraz katalogów. Dodatkowo, NameNode odpowiada za mapowanie bloków na konkretne DataNodes.

DataNodes są odpowiedzialne za obsługę żądań odczytu i zapisu klientów systemu plików. Ponadto, są one odpowiedzialne za tworzenie, usuwanie i replikację bloków zgodnie z instrukcjami otrzymanymi od NameNode'a.



Rysunek 3 Infrastruktura Hadoop Distributed File System

Proces ELT w Hadoop

ELT to skrót od Extract, Load i Transform.

W przeciwieństwie do tradycyjnego procesu ładowania przekształconych danych bezpośrednio do systemów docelowych, proces ELT polega na załadowaniu wszystkich danych do jeziora danych. Dzięki temu osiąga się szybszy czas ładowania. Dodatkowo, proces ładowania może opcjonalnie obejmować podstawowe zasady walidacji i oczyszczania danych. Następnie dane są przekształcane zgodnie z wymaganiami raportowania analitycznego. Choć proces ELT jest stosowany od pewnego czasu, to zyskuje na popularności szczególnie wraz z rozwojem technologii Hadoop.

Poniższy schemat przedstawia typowy proces ELT w kontekście Hadoopa.



Rysunek 4 Schemat przebiegu procesu ELT w Hadoop

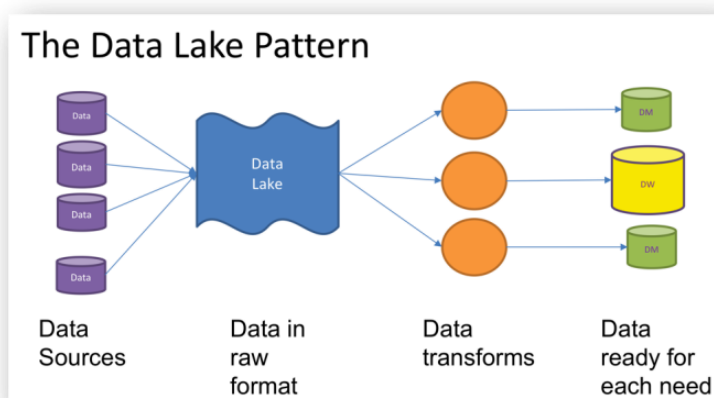
Data Lakes

Jezioro danych można określić jako centralne miejsce, w którym gromadzone są zarówno strukturyzowane, jak i nieustrukturyzowane dane, a także jako sposób organizacji różnorodnych danych pochodzących z różnych źródeł.

W dzisiejszych czasach jeziora danych nabierają coraz większego znaczenia, zwłaszcza w kontekście biznesowym i technologicznym, gdzie istnieje potrzeba przeprowadzania obszernych analiz oraz odkrywania wzorców w danych. Centralizacja danych w jednym miejscu znacznie ułatwia przeprowadzanie tych operacji.

W zależności od wykorzystywanej platformy, jezioro danych może znacząco ułatwić procesy związane z gromadzeniem i przetwarzaniem danych. Może obsługiwać różnorodne struktury danych, włączając w to dane niestrukturyzowane oraz multistrukturyzowane. Dodatkowo, umożliwia eksploatację danych w celu uzyskania korzyści biznesowych.

Jezioro danych wykorzystuje procesy ELT. Dane pobiera się z wielu różnych źródeł, następnie ładuje do jeziora danych. Dopiero w momencie kiedy dane są pobierane w celach analitycznych lub w celu ich przetworzenia wtedy poddawane są transformacjom.



Rysunek 5 Schemat przebiegu danych w jeziorze danych

Jezioro danych może być wykorzystane na wiele sposobów. Ma również dużo platform podrzędnych. Najczęściej używaną platformą jest wcześniej wspomniany Hadoop.

Cel Realizacji Projektu

Celem projektu jest wykorzystanie narzędzie SAS Data Integration Studio aby użyć hurtownię danych oraz stworzyć schemat gwiazdy w celu analizy procesów sprzedaży dla Adventure Works Cycles.

Adventure Works Cycles to fikcyjna, duża, międzynarodowa firma produkcyjna, która produkuje i dystrybuuje rowery na rynki komercyjne w Ameryce Północnej, Europie i Azji. Siedziba Adventure Works Cycles mieści się w Bothell w stanie Waszyngton. Firma zatrudnia wielu pracowników. Ponadto firma Adventure Works Cycles zatrudnia regionalnych zespołów sprzedaży na całym swoim rynku.

Kolejno opisane zostały następujące analizy:

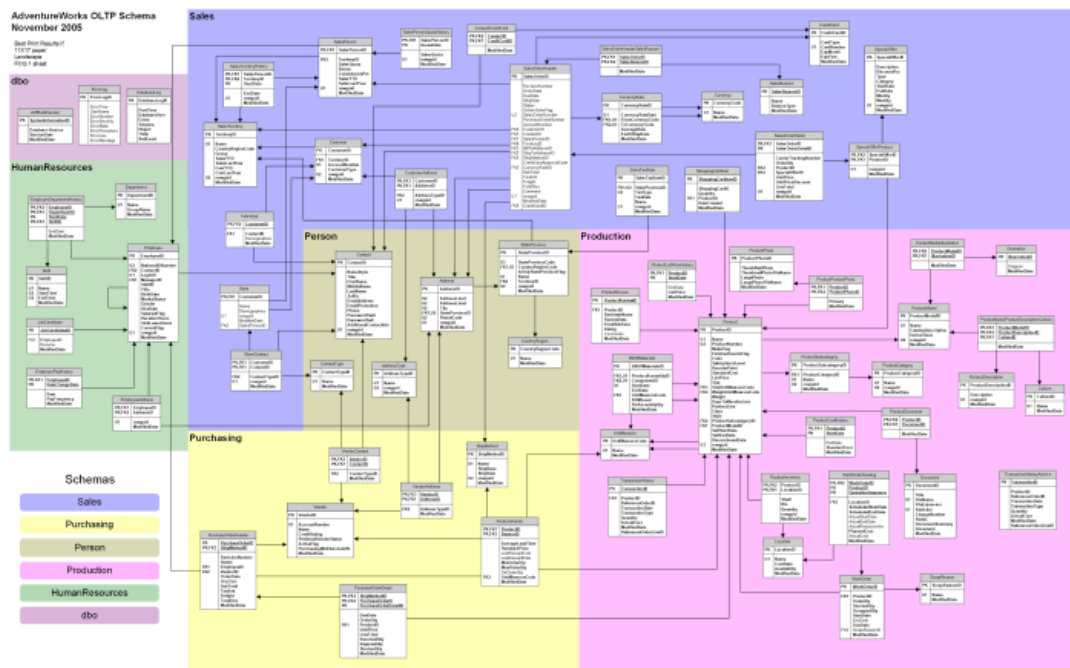
1. Analiza wyników sprzedażowych mężczyzn i kobiet w latach 2001-2004
2. Analiza sprzedaży poszczególnych kategorii produktów na przestrzeni lat
3. Analiza sprzedaży w podziale na kwartały i państwa
4. Analiza największych wartości sprzedaży danych produktów, sprzedanych przez poszczególnych pracowników

Na wstępie przedstawiono schemat użytej bazy danych, czyli Adventure Works, a także dokładnie opisano struktury tabel używanych oraz źródła danych wraz z ich zakresem i pochodzeniem. Następnie na podstawie tych informacji przeprowadzono procesy ETL, mające na celu utworzenie tabel wymiarów i faktu, co skutkowało stworzeniem schematu gwiazdy. Głównym tematem hurtowni danych są dane sprzedażowe związane z oferowanymi przez przedsiębiorstwo produktami, takimi jak rowery i akcesoria pochodne. W ramach analizy uzyskano istotne informacje dotyczące ogólnej sprzedaży, wyników sprzedażowych poszczególnych pracowników, a także zależności między ilością zamówień a czasem i miejscem.

Opis Źródeł Danych

W ramach realizacji projektu skorzystano z bazy danych Adventure Works. Baza ta obejmuje szczegółowe dane na temat oferowanych produktów, zasobów ludzkich oraz informacje bezpośrednio związane z procesami sprzedaży, zakupu i produkcji. Tabele zawarte w tej bazie zostały podzielone na 5 obszarów:

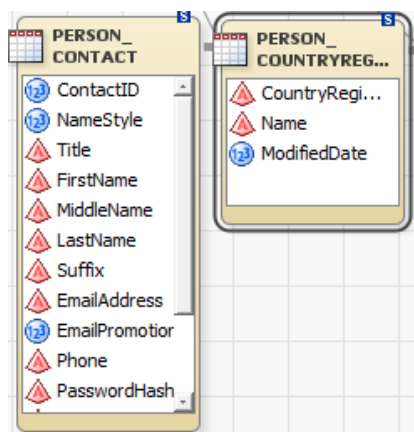
1. HumanResources – zawiera dane dotyczące poszczególnych pracowników takie jak dział w którym są zatrudnieni, zajmowane przez nich stanowisko, godziny rozpoczęcia i zakończenia zmiany oraz inne dane wykorzystywane przez firmę.
2. Person – zawiera dane dotyczące ludzi, ich dane osobowe, informacje kontaktowe. Te osoby to zarówno klienci jak i pracownicy.
3. Production – zawiera dane związane z procesem produkcji. Są to rozległe informacje o wielkości produkcji, jej planowaniu, zarządzaniu produkcją oraz o produktach, które powstają.
4. Purchasing – obejmuje dane związane z zakupem, dostawcami, informacjami odnośnie poszczególnych zamówień, sposobu dostawy oraz dane o fakturach zakupowych.
5. Sales – zawiera dane o sprzedaży, klientach, produktach, zamówieniach oraz inne dane związane z procesem sprzedaży.



Rysunek 6- Schemat bazy Adventure Works

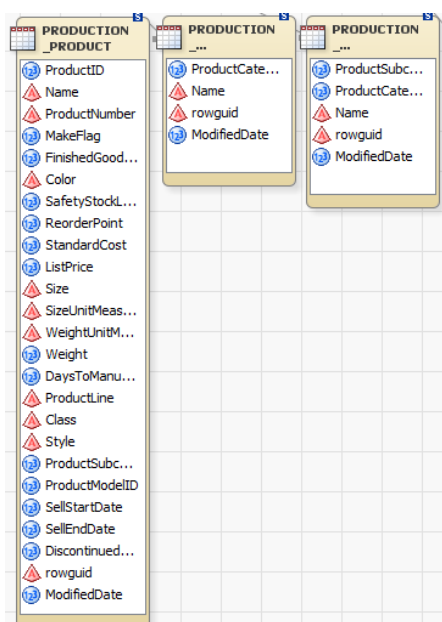
Podczas realizacji projektu zostały wykorzystane dane z tych obszarów: Person, Production, Sales oraz HumanResources.

Z obszaru Person zostały wykorzystane tabele Person_Contact oraz Person_CountryRegion.



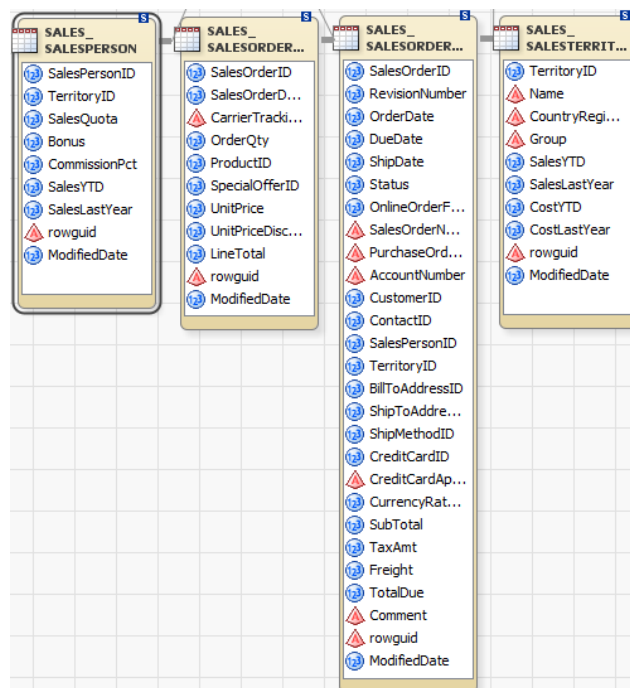
Rysunek 7 Wykorzystane tabele z obszaru Person

Z obszaru Production wykorzystano tabele Production_Product, Production_ProductCategory oraz Production_ProductSubCategory.



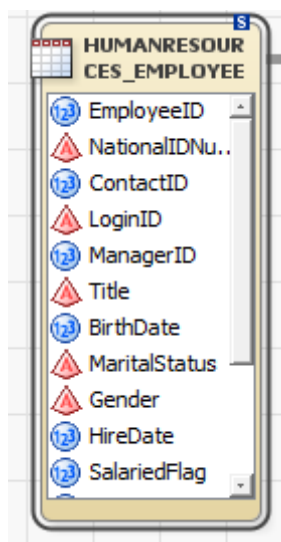
Rysunek 8 Wykorzystane tabele z obszaru Production

Z obszaru Sales wykorzystane zostały tabele Sales_SalesOrderDetail, Sales_SalesOrderHeader, Sales_SalesTerritory oraz Sales_SalesPerson.



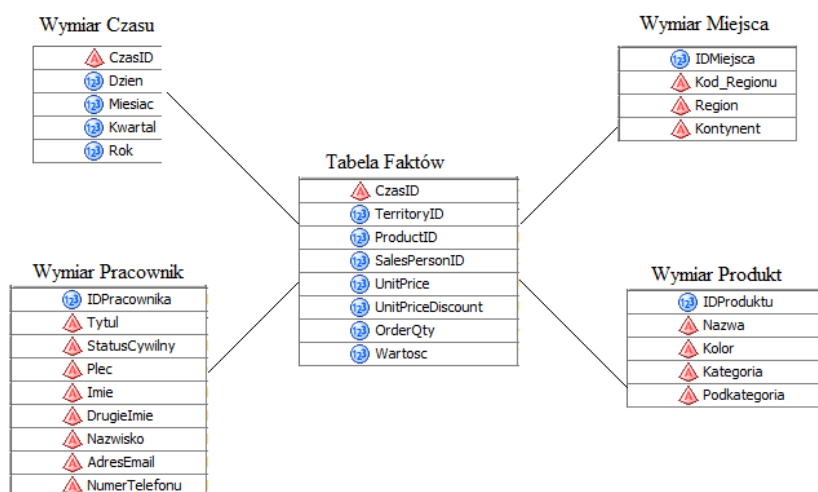
Rysunek 9 Wykorzystane tabele z obszaru Sales

Z obszaru HumanResources wykorzystano tabelę HumanResources_Employee



Rysunek 10 Wykorzystane tabele z obszaru HumanResources

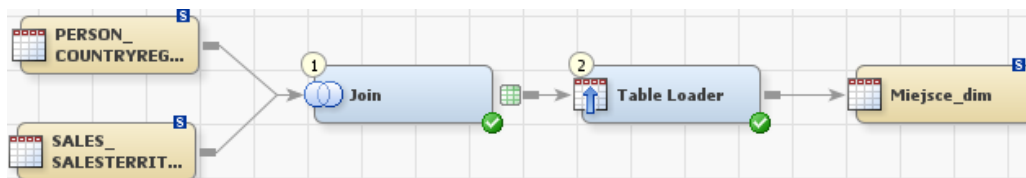
Model logiczny hurtowni danych w oparciu o schemat gwiazdy



Rysunek 11 Schemat gwiazdy

Wymiar Miejsca

To kolejne kryterium uwzględnione z modelu gwiazdy. Pozwala na analizę danych w kontekście lokalizacji. W celu przygotowania tego wymiaru wykorzystane zostały tabele `Person_CountryRegion` oraz `Sales_SalesTerritory`.



Rysunek 12 Proces tworzenia wymiaru miejsca

W węźle Join zostały utworzone kolumny `IDMiejsca`, `Kod_Regionu`, `Panstwo` oraz `Region`. Zostały one przeładowane do tabeli `Miejsce_dim`.

| # | IDMiejsca | Kod_Regionu | Panstwo | Region |
|----|-----------|-------------|------------------|----------------|
| 1 | 9 | AU | Australia | Pacific |
| 2 | 6 | CA | Canada | North Ameri... |
| 3 | 8 | DE | Germany | Europe |
| 4 | 7 | FR | France | Europe |
| 5 | 10 | GB | United Kingdo... | Europe |
| 6 | 1 | US | United States | North Ameri... |
| 7 | 5 | US | United States | North Ameri... |
| 8 | 4 | US | United States | North Ameri... |
| 9 | 3 | US | United States | North Ameri... |
| 10 | 2 | US | United States | North Ameri... |

Rysunek 13 Tabela Miejsce_dim

Wymiar Czasu

To jedno z kryteriów uwzględnionych w modelu gwiazdy. Pozwala na analizę danych w kontekście czasowym. W celu przygotowania tego wymiaru wykorzystana została tabela Sales_SalesOrderHeader.



Rysunek 14 Proces tworzenia wymiaru czasu

W węźle Join zostały utworzone kolumny CzasID, Dzień, Miesiąc, Kwartał oraz Rok. Te kolumny zostaną przeładowane do tabeli docelowej Czas_dim. Jednak aby było to możliwe najpierw trzeba przygotować dla każdej z tych kolumn odpowiednie wyrażenie (expression). Oto poszczególne wyrażenia:

Dla CzasID: `compress(substr(put(YEAR(DATEPART(WDEZZZ2.OrderDate)),4.),3,2)||'0'||put(MONTH(DATEPART(WDEZZZ2.OrderDate)),2.)||'0'||put(DAY(DATEPART(WDEZZZ2.OrderDate)),2.))`

Dla Dzień: `day(datepart(WDEZZZ2.OrderDate))`

Dla Miesiąc: `month(datepart(WDEZZZ2.OrderDate))`

Dla Kwartał: `qtr(datepart(WDEZZZ2.OrderDate))`

Dla Rok: `year(datepart(WDEZZZ2.OrderDate))`

| # | | Column | Colu... | Expression | Type | Length | Informat | Format |
|---|--|---------|---------|---------------------------------------|-----------|--------|----------|--------|
| 1 | | CzasID | | compress(substr(put(YEAR(DATEPART(... | Character | 8 | \$8. | (None) |
| 2 | | Dzien | | day(datepart(WDEZZZ2.OrderDate)) | Numeric | 8 | 8. | (None) |
| 3 | | Miesiac | | month(datepart(WDEZZZ2.OrderDate)) | Numeric | 8 | 8. | (None) |
| 4 | | Kwartal | | qtr(datepart(WDEZZZ2.OrderDate)) | Numeric | 8 | (None) | (None) |
| 5 | | Rok | | year(datepart(WDEZZZ2.OrderDate)) | Numeric | 8 | (None) | (None) |

Rysunek 15 Wyrażenia wykorzystane do utworzenia wymiaru czasu

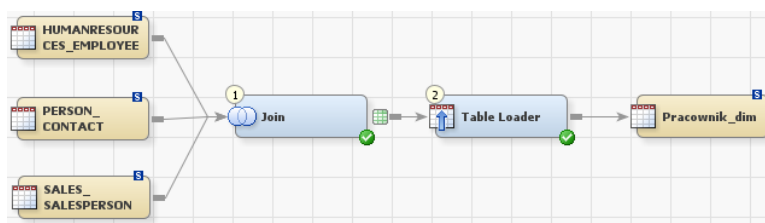
W ten sposób powstała tabela Czas_dim

| # | CzasID | Dzien | Miesiac | Kwartal | Rok |
|----|----------|-------|---------|---------|------|
| 1 | 0101001 | 1 | 10 | 4 | 2001 |
| 2 | 01010010 | 10 | 10 | 4 | 2001 |
| 3 | 01010011 | 11 | 10 | 4 | 2001 |
| 4 | 01010012 | 12 | 10 | 4 | 2001 |
| 5 | 01010013 | 13 | 10 | 4 | 2001 |
| 6 | 01010014 | 14 | 10 | 4 | 2001 |
| 7 | 01010015 | 15 | 10 | 4 | 2001 |
| 8 | 01010016 | 16 | 10 | 4 | 2001 |
| 9 | 01010017 | 17 | 10 | 4 | 2001 |
| 10 | 01010018 | 18 | 10 | 4 | 2001 |

Rysunek 16 Tabela Czas_dim

Wymiar Pracownik

To kolejne z kryteriów uwzględnionych w modelu gwiazdy. Pozwala na analizę danych w kontekście osób zatrudnionych. W celu przygotowania tego wymiaru wykorzystane zostały tabele HumanResources_Employee, Person_Contact oraz Sales_SalesPerson.



Rysunek 17 Proces tworzenia wymiaru pracownik

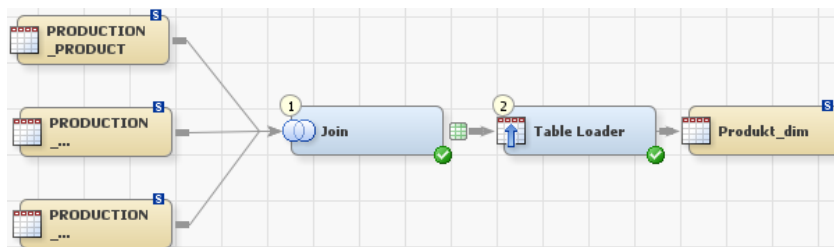
W węźle Join zostały utworzone kolumny IDPracownika, Tytuł, StatusCywilny, Płeć, Imię, DrugieImię, Nazwisko, AdresEmail oraz NumerTelefonu.

| # | IDPracownika | Tytuł | StatusCywilny | Płeć | Imię | DrugieImię | Nazwisko | AdresEmail | NumerTelefonu |
|----|--------------|---------------|---------------|------|---------|------------|-----------------|---------------------|---------------|
| 1 | 268 | North Ame... | M | M | Stephen | Y | Jiang | stephen0@adve... | 238-555-0197 |
| 2 | 275 | Sales Repr... | S | M | Michael | G | Blythe | michael9@adve... | 257-555-0154 |
| 3 | 276 | Sales Repr... | M | F | Linda | C | Mitchell | linda3@adventu... | 883-555-0116 |
| 4 | 277 | Sales Repr... | S | F | Jillian | | Carson | jillian0@adventu... | 517-555-0117 |
| 5 | 278 | Sales Repr... | M | M | Garrett | R | Vargas | garrett1@adve... | 922-555-0165 |
| 6 | 279 | Sales Repr... | M | M | Tsvi | Michael | Reiter | tsvi0@adventur... | 664-555-0112 |
| 7 | 280 | Sales Repr... | S | F | Pamela | O | Ansman-Wolfe... | pamela0@adve... | 340-555-0193 |
| 8 | 281 | Sales Repr... | M | M | Shu | K | Ito | shu0@adventur... | 330-555-0120 |
| 9 | 282 | Sales Repr... | M | M | José | Edvaldo | Saraiva | josé1@adventu... | 185-555-0169 |
| 10 | 283 | Sales Repr... | S | M | David | R | Campbell | david8@advent... | 740-555-0182 |

Rysunek 18 Tabela Pracownik_dim

Wymiar Produkt

To kolejne z kryteriów uwzględnionych w modelu gwiazdy. Pozwala na analizę danych w kontekście produktów. W celu przygotowania tego wymiaru wykorzystane zostały tabele Production_Product, Production_ProductCategory oraz Production_ProductSubCategory.



Rysunek 19 Proces tworzenia wymiaru produkt

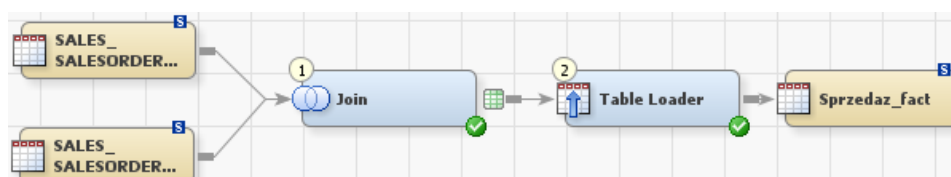
W węźle Join zostały utworzone kolumny IDProduktu, Nazwa, Kolor, Kategoria oraz Podkategoria

| # | IDProduktu | Nazwa | Kolor | Kategoria | Podkategoria |
|---|------------|-------------------|-------|-------------|--------------|
| 1 | 842 | Touring-Pan... | Grey | Accessories | Panniers |
| 2 | 843 | Cable Lock ... | | Accessories | Locks |
| 3 | 844 | Minipump ... | | Accessories | Pumps |
| 4 | 845 | Mountain Pu... | | Accessories | Pumps |
| 5 | 846 | Taillights - B... | | Accessories | Lights |

Rysunek 20 Tabela Produkt_dim

Tabela Faktów

To ostatnie z kryteriów uwzględnionych w modelu gwiazdy. Ta tabela jest połączona z pozostałymi za pomocą kluczy. Jest ona centralnym elementem schematu gwiazdy. W celu przygotowania tabeli faktów wykorzystane zostały tabele Sales_SalesOrderHeader oraz Sales_SalesOrderDetail.



Rysunek 21 Proces tworzenia tabeli faktów

W węźle Join zostały utworzone kolumny CzasID, IDMiejsca, IDProduktu, IDSprzedawcy, CenaJednostkowa, Zniżka, Ilość oraz Wartość.

Do utworzenia CzasID zostało wykorzystane wyrażenie:

```
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.))

Natomiast do utworzenia Wartość zostało wykorzystane wyrażenie:

(SALES_SALESORDERDETAIL."UnitPrice"n -
SALES_SALESORDERDETAIL."UnitPriceDiscount"n)
*SALES_SALESORDERDETAIL."OrderQty"n

| # |  CzasID |  IDMiejsca |  IDProduktu |  IDPracownika |  CenaJednostkowa |  Znizka |  Ilosc |  Wartosc |
|----|--|---|--|--|---|--|---|---|
| 1 | 010701 | 5 | 776 | 279 | \$2,024.99 | \$0.00 | 1 | 2024.994 |
| 2 | 010701 | 5 | 777 | 279 | \$2,024.99 | \$0.00 | 3 | 6074.982 |
| 3 | 010701 | 5 | 778 | 279 | \$2,024.99 | \$0.00 | 1 | 2024.994 |
| 4 | 010701 | 5 | 771 | 279 | \$2,039.99 | \$0.00 | 1 | 2039.994 |
| 5 | 010701 | 5 | 772 | 279 | \$2,039.99 | \$0.00 | 1 | 2039.994 |
| 6 | 010701 | 5 | 773 | 279 | \$2,039.99 | \$0.00 | 2 | 4079.988 |
| 7 | 010701 | 5 | 774 | 279 | \$2,039.99 | \$0.00 | 1 | 2039.994 |
| 8 | 010701 | 5 | 714 | 279 | \$28.84 | \$0.00 | 3 | 86.5212 |
| 9 | 010701 | 5 | 716 | 279 | \$28.84 | \$0.00 | 1 | 28.8404 |
| 10 | 010701 | 5 | 709 | 279 | \$5.70 | \$0.00 | 6 | 34.2 |

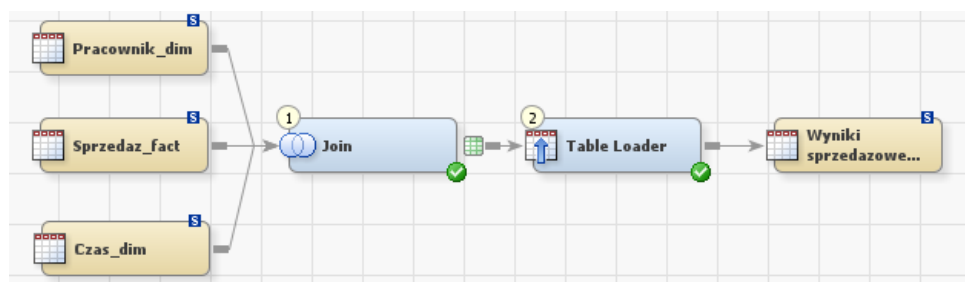
Rysunek 22 Tabela Sprzedaz_fact

Opis Procesów ETL

1. Analiza wyników sprzedażowych mężczyzn i kobiet w latach 2001-2004

Jako pierwszą przeprowadzono analizę wyników sprzedażowych mężczyzn i kobiet w latach 2001-2004. Dane te pozwolą na porównanie która z płci osiągnęła wyższe wyniki sprzedażowe w poszczególnych latach oraz o ile wyższe były te wyniki.

Do utworzenia tabeli wynikowej wykorzystany został następujący proces:



Rysunek 23 Proces - wyniki sprzedażowe kobiet i mężczyzn w latach 2001-2004

W tabeli wynikowej powstały 3 kolumny: Rok, Płeć oraz Wartość. Do utworzenia kolumny wartość zostało zastosowane wyrażenie SUM(Sprzedaz_fact."Wartosc"*n).

| Target table: Join (W66D318) | | | | | | | | | |
|------------------------------|--|---------|-------------|---------------------------------|-----------|--------------|------------|------------|-------------|
| # | | Column | Column D... | Expression | Type | Length | Informat | Format | Is Nulla... |
| 1 | | Rok | | | Numeric | 8 (None) | (None) | (None) | Yes |
| 2 | | Plec | Gender | | Character | 1 \$1. | \$1. | \$1. | Yes |
| 3 | | Wartosc | Wartosc | SUM(Sprzedaz_fact."Wartosc"*n) | Numeric | 8 DOLLAR21.2 | DOLLAR21.2 | DOLLAR21.2 | Yes |




Rysunek 24 Select - tworzenie kolumn do tabeli docelowej

Następnie wyniki zostały pogrupowane po roku oraz płci.

| Group by columns | | |
|------------------|-------------|---------------|
| Table name | Column name | Column ref... |
| Join | Rok | Name |
| Join | Plec | Name |

Rysunek 25 Grupowanie - zgrupowanie wyników po roku oraz płci

Po tak przeprowadzonym procesie powstała następująca tabela wynikowa.

| # |  Rok |  Płeć |  Wartość |
|---|---|--|---|
| 1 | 2001 | F | \$2,999,241.99 |
| 2 | 2001 | M | \$5,070,504.47 |
| 3 | 2002 | F | \$10,876,782.92 |
| 4 | 2002 | M | \$13,451,525.04 |
| 5 | 2003 | F | \$15,171,327.69 |
| 6 | 2003 | M | \$17,345,049.18 |
| 7 | 2004 | F | \$7,409,993.57 |
| 8 | 2004 | M | \$8,688,744.11 |

Rysunek 26 Tabela wynikowa - wyniki sprzedażowe kobiet i mężczyzn

Podsumowanie wyników:

Wartość sprzedaży w poszczególnych latach i płciach ukazuje pewne istotne trendy. Oto główne obserwacje:

Różnice w wartości sprzedaży między płciami:

W każdym z badanych lat wartość sprzedaży mężczyzn była wyższa niż kobiet.

W 2001 roku wartość sprzedanych przez mężczyzn produktów była o 2 071 262,48 \$ większa niż wartość produktów sprzedanych przez kobiety.

W 2002 roku wartość sprzedanych przez mężczyzn produktów była o 2 574 742,12 \$ większa niż wartość produktów sprzedanych przez kobiety. Była to największa różnica między wartościami produktów sprzedanych przez mężczyzn a wartościami produktów sprzedanych przez kobiety w latach 2001-2004.

W 2003 roku różnica między wartością produktów sprzedanych przez mężczyzn a wartością produktów sprzedanych przez kobiety wynosiła 2 173 721,49 \$.

W 2004 roku różnica między wartością produktów sprzedanych przez mężczyzn a wartością produktów sprzedanych przez kobiety była najmniejsza i wynosiła 1 278 750,54 \$.

Najlepszy rok sprzedaży:

2003 rok był najbardziej korzystny pod względem wartości sprzedaży zarówno dla mężczyzn, jak i kobiet.

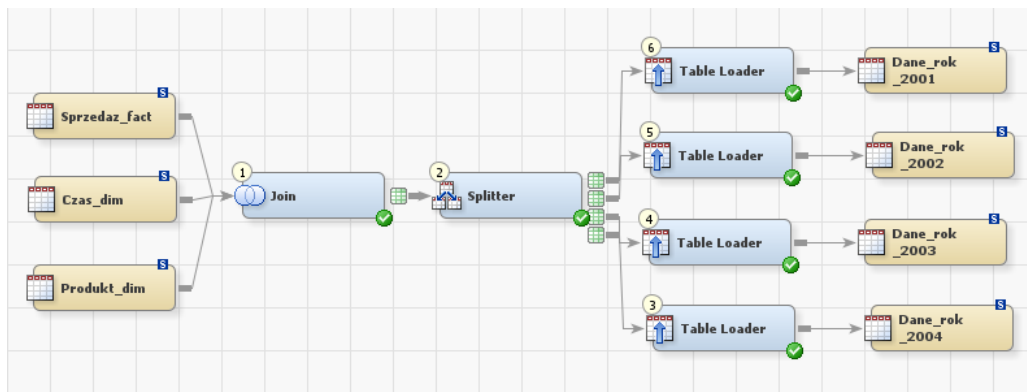
Wartość sprzedaży osiągniętej przez mężczyzn w 2003 roku wynosi: 17 345 049,18 \$.

Wartość sprzedaży osiągniętej przez kobiety w 2003 roku wynosi: 15 171 327,69 \$

2. Analiza sprzedaży poszczególnych kategorii produktów na przestrzeni lat

Kolejną przeprowadzoną analizą jest analiza sprzedaży poszczególnych kategorii produktów w latach 2001-2004. Ta analiza pozwoli zobrazować które kategorie produktów były najczęściej sprzedawane w każdym roku oraz jaką wartość miały sprzedane produkty danej kategorii.

Do utworzenia tabeli wynikowej posłużył poniższy proces.



Rysunek 27 Proces - analiza sprzedaży kategorii produktów na przestrzeni lat

W tabelach wynikowych powstały po 4 kolumny: Rok, ŁącznaIlość, ŁącznaWartość oraz Kategoria.

Do utworzenia kolumny ŁącznaIlość zostało wykorzystane wyrażenie:
count(Sprzedaz_fact."Ilosc"n).

Do utworzenia kolumny ŁącznaWartość zostało wykorzystane wyrażenie:
sum(Sprzedaz_fact."Wartosc"n).

| Target table: Join (W2NFPZW) | | | | | | | | | |
|------------------------------|--|---------------|-------------|--------------------------------|-----------|--------------|------------|------------|-------------|
| # | | Column | Column D... | Expression | Type | Length | Informat | Format | Is Nulla... |
| 1 | | Rok | | | Numeric | 8 (None) | (None) | (None) | Yes |
| 2 | | ŁącznaIlosc | | count(Sprzedaz_fact."Ilosc"n) | Numeric | 8 (None) | (None) | (None) | Yes |
| 3 | | ŁącznaWartosc | | sum(Sprzedaz_fact."Wartosc"n) | Numeric | 8 DOLLAR21.2 | DOLLAR21.2 | DOLLAR21.2 | Yes |
| 4 | | Kategoria | Kategoria | | Character | 1024 \$1024. | \$1024. | \$1024. | Yes |

Rysunek 28 Join - tworzenie kolumn do tabeli docelowej

Wyniki zostały pogrupowane według kolumny Rok oraz Kategoria.

| Group by columns | | |
|------------------|-------------|---------------|
| Table name | Column name | Column ref... |
| Join | Rok | Name |
| Join | Kategoria | Name |

Rysunek 29 Grupowanie - zgrupowanie wyników po roku oraz kategorii

Na skutek uruchomienia procesu powstały 4 tabele docelowe.

| # | Rok | ŁącznaIlość | ŁącznaWartość | Kategoria |
|---|------|-------------|-----------------|-------------|
| 1 | 2001 | 346 | \$20,239.44 | Accessories |
| 2 | 2001 | 3356 | \$10,665,950.75 | Bikes |
| 3 | 2001 | 614 | \$34,454.96 | Clothing |
| 4 | 2001 | 835 | \$615,474.98 | Components |

| # | Rok | ŁącznaIlość | ŁącznaWartość | Kategoria |
|---|------|-------------|-----------------|-------------|
| 1 | 2002 | 1186 | \$93,734.60 | Accessories |
| 2 | 2002 | 9654 | \$26,664,247.61 | Bikes |
| 3 | 2002 | 3482 | \$489,635.71 | Clothing |
| 4 | 2002 | 5031 | \$3,611,033.57 | Components |

| # | Rok | ŁącznaIlość | ŁącznaWartość | Kategoria |
|---|------|-------------|-----------------|-------------|
| 1 | 2003 | 17377 | \$594,793.48 | Accessories |
| 2 | 2003 | 15643 | \$35,198,897.49 | Bikes |
| 3 | 2003 | 9349 | \$1,024,023.00 | Clothing |
| 4 | 2003 | 8868 | \$5,489,723.20 | Components |

| # | Rok | ŁącznaIlość | ŁącznaWartość | Kategoria |
|---|------|-------------|-----------------|-------------|
| 1 | 2004 | 22285 | \$569,673.29 | Accessories |
| 2 | 2004 | 11378 | \$22,615,807.29 | Bikes |
| 3 | 2004 | 7949 | \$592,606.87 | Clothing |
| 4 | 2004 | 3964 | \$2,091,549.97 | Components |

Rysunek 30 Tabela wynikowa - analiza sprzedaży kategorii produktów na przestrzeni lat

Podsumowanie wyników:

Accessories

Najlepszy Rok Sprzedaży:

W 2004 roku Accessories odnotowały największą sprzedaż, osiągając 22 285 jednostek o wartości 569 673,29 \$.

Najgorszy Rok Sprzedaży:

W 2001 roku Accessories miały najmniejszą sprzedaż, osiągając 346 jednostek o wartości 20 239,44 \$.

Bikes

Najlepszy Rok Sprzedaży:

W 2003 roku Bikes osiągnęły najwyższą wartość sprzedaży, sprzedając 15 643 jednostki za 35 198 897,49 \$.

Najgorszy Rok Sprzedaży:

W 2001 roku Bikes odnotowały najniższą sprzedaż, wynoszącą 3 356 jednostek o wartości 10 665 950,75 \$.

Clothing

Najlepszy Rok Sprzedaży:

W 2002 roku Clothing uzyskały najwyższą wartość sprzedaży, wynoszącą 3 482 jednostki za 489 635,71 \$.

Najgorszy Rok Sprzedaży:

W 2001 roku Clothing miały najmniejszą sprzedaż, osiągając 614 jednostek o wartości 34 454,96 \$.

Components

Najlepszy Rok Sprzedaży:

W 2003 roku Components odnotowały największą sprzedaż, sprzedając 8 868 jednostek za 5 489 723,20 \$.

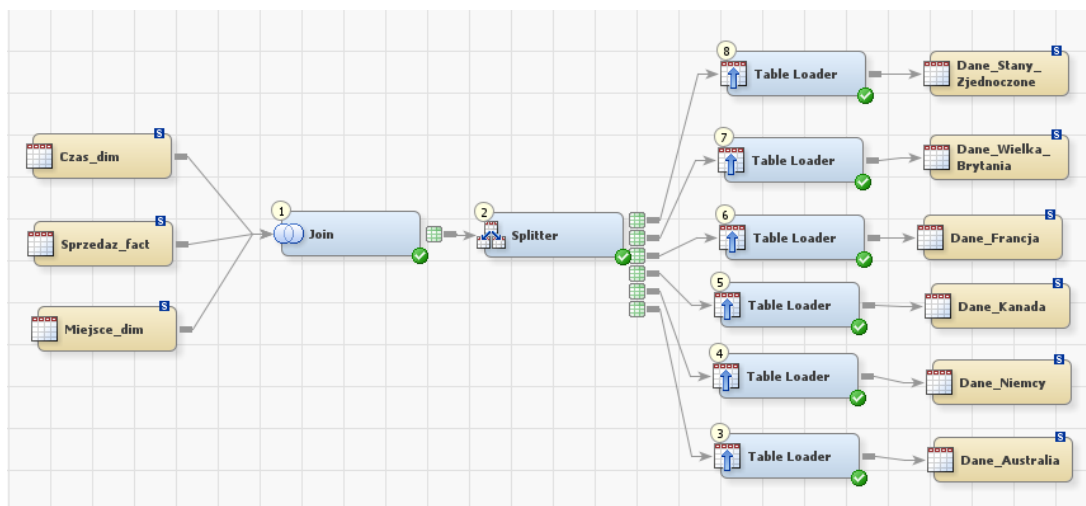
Najgorszy Rok Sprzedaży:

W 2001 roku Components miały najniższą sprzedaż, wynoszącą 835 jednostek o wartości 615 474,98 \$.

3. Analiza sprzedaży w podziale na kwartały i państwa

Kolejną analizą jest analiza sprzedaży w podziale na kwartały i państwa. Ta analiza pozwoli pokazać w którym kwartale którego roku oraz w którym państwie sumaryczna wartość sprzedaży była największa a w którym najmniejsza.

Do utworzenia tabeli wynikowej posłużył poniższy proces.



Rysunek 31 Proces - analiza sprzedaży w podziale na kwartały i regiony

W każdej z tabel wynikowych powstało 5 kolumn: Rok, Kwartał, Państwo, Ilość oraz Wartość. Kolumny Ilość oraz Wartość powstały dzięki wykorzystaniu tych samych wyrażeń co w poprzedniej analizie.

| Target table: Join (W2918NY) | | | | | | | | | |
|------------------------------|--|---------|-------------|------------------------------|-----------|--------|----------|--------|-------------|
| # | | Column | Column D... | Expression | Type | Length | Informat | Format | Is Nulla... |
| 1 | | Rok | | | Numeric | 8 | (None) | (None) | Yes |
| 2 | | Kwartał | | | Numeric | 8 | (None) | (None) | Yes |
| 3 | | Państwo | Name | | Character | 50 | \$50. | \$50. | Yes |
| 4 | | Ilość | OrderQty | count(Sprzedaz_fact."Ilość") | Numeric | 8 | 6. | 6. | Yes |
| 5 | | Wartość | Wartosc | sum(Sprzedaz_fact."Wartosc") | Numeric | 8 | (None) | (None) | Yes |

Rysunek 32 Join - tworzenie kolumn do tabel wynikowych

Wyniki zostały pogrupowane w następujący sposób.

| Group by columns | | |
|------------------|-------------|---------------|
| Table name | Column name | Column ref... |
| Join | Rok | Name |
| Join | Kwartał | Name |
| Join | Państwo | Name |

Rysunek 33 Grupowanie – zgrupowanie wyników po roku, kwartale oraz państwie

Po wykonaniu tego procesu powstało 6 tabel wynikowych.

The figure consists of six screenshots of Excel spreadsheets, each displaying quarterly sales data for a different country. The spreadsheets are titled 'View Data: Dane_Stanów_Zjednoczonych (13 rows)', 'View Data: Dane_Kanada (13 rows)', 'View Data: Dane_Wielka_Brytania (13 rows)', 'View Data: Dane_Australia (13 rows)', 'View Data: Dane_Francja (13 rows)', and 'View Data: Dane_Niemcy (13 rows)'. Each spreadsheet has columns for Rank, Year, Quarter, Country, Sales, and Value. The data is organized by quarter and year, with the highest sales values typically occurring in the third and fourth quarters of each year.

Rysunek 34 Tabela wynikowa - analiza sprzedaży w podziale na kwartały i regiony

Podsumowanie wyników:

Analiza Sprzedaży w Kwartałach dla Stanów Zjednoczonych:

Największa sprzedaż: W 2003 roku, w trzecim kwartale, sprzedaż w Stanach Zjednoczonych osiągnęła wartość 7 293 127,59 \$.

Najmniejsza sprzedaż: W 2001 roku, w trzecim kwartale, sprzedaż w Stanach Zjednoczonych wyniosła 2 988 281,22 \$.

Analiza Sprzedaży w Kwartałach dla Wielkiej Brytanii:

Największa sprzedaż: W 2003 roku, w czwartym kwartale, sprzedaż w Wielkiej Brytanii osiągnęła wartość 1 284 548,27 \$.

Najmniejsza sprzedaż: W 2001 roku, w czwartym kwartale, sprzedaż w Wielkiej Brytanii wyniosła 136 869,10 \$.

Analiza Sprzedaży w Kwartałach dla Francji:

Największa sprzedaż: W 2003 roku, w trzecim kwartale, sprzedaż we Francji osiągnęła wartość 1 280 104,79 \$.

Najmniejsza sprzedaż: W 2001 roku, w trzecim kwartale, sprzedaż we Francji wyniosła 94 661,73 \$.

Analiza Sprzedaży w Kwartalach dla Kanady:

Największa sprzedaż: W 2003 roku, w czwartym kwartale, sprzedaż w Kanadzie osiągnęła wartość 1 812 805,93 \$.

Najmniejsza sprzedaż: W 2001 roku, w trzecim kwartale, sprzedaż w Kanadzie wyniosła 706 870,45 \$.

Analiza Sprzedaży w Kwartalach dla Niemiec:

Największa sprzedaż: W 2003 roku, w czwartym kwartale, sprzedaż w Niemczech osiągnęła wartość 945 612,21 \$.

Najmniejsza sprzedaż: W 2001 roku, w trzecim kwartale, sprzedaż w Niemczech wyniosła 98 532,30 \$.

Analiza Sprzedaży w Kwartalach dla Australii:

Największa sprzedaż: W 2004 roku, w pierwszym kwartale, sprzedaż w Australii osiągnęła wartość 1 533 890,81 \$.

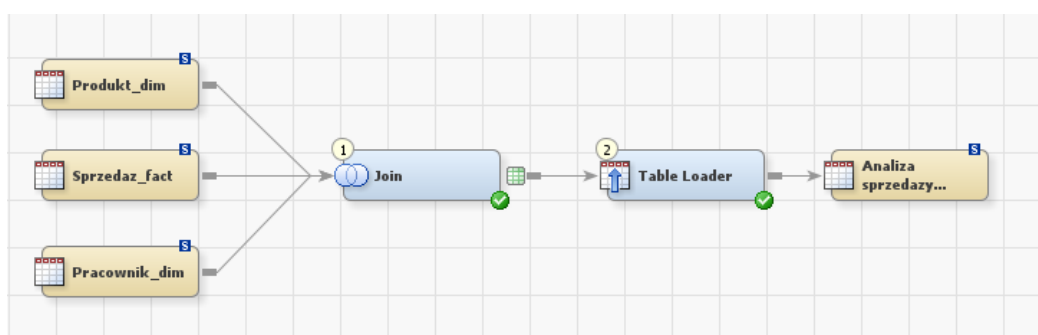
Najmniejsza sprzedaż: W 2001 roku, w trzecim kwartale, sprzedaż w Australii wyniosła 606 184,71 \$.

4. Analiza największych wartości sprzedaży danych produktów, sprzedanych przez poszczególnych pracowników

Ostatnią analizą jest analiza najlepiej sprzedających się produktów z podziałem na sprzedawców. Ta analiza pozwoli pokazać który z pracowników sprzedając dany produkt wygenerował największy obrót.

Do utworzenia tabeli wynikowej posłużyły dwa procesy. Najpierw opisany został pierwszy a potem drugi.

Proces 1.



Rysunek 35 Proces - analiza najlepiej sprzedających się produktów z podziałem na pracowników cz.1

W tabeli wynikowej powstało 6 kolumn: Imię, Nazwisko, Tytuł, Nazwa, Ilość, Wartość.

Kolumny Ilość oraz Wartość powstały dzięki zastosowaniu tych samych wyrażeń co w analizie drugiej.

| Target table: Join (W2WBDZL) | | | | | | | | | |
|------------------------------|--|----------|-------------|--------------------------------|-----------|--------------|----------|------------|-------------|
| # | | Column | Column D... | Expression | Type | Length | Informat | Format | Is Nulla... |
| 1 | | Imie | FirstName | | Character | 1024 | \$1024. | \$1024. | Yes |
| 2 | | Nazwisko | LastName | | Character | 1024 | \$1024. | \$1024. | Yes |
| 3 | | Tytuł | Title | | Character | 1024 | \$1024. | \$1024. | Yes |
| 4 | | Nazwa | Nazwa | | Character | 1024 | \$1024. | \$1024. | Yes |
| 5 | | Ilosc | OrderQty | count(Sprzedaz_fact."Ilosc"n) | Numeric | 8 6. | | 6. | Yes |
| 6 | | Wartosc | Wartosc | sum(Sprzedaz_fact."Wartosc"n) | Numeric | 8 DOLLAR21.2 | | DOLLAR21.2 | Yes |

Rysunek 36 Join - tworzenie kolumn do tabeli wynikowej cz.1

Poniżej przedstawiono fragment powstałej tabeli wynikowej.

| # | Imie | Nazwisko | Tytuł | Nazwa | Ilosc | Wartosc |
|----|------|----------|--------------|------------------|-------|------------|
| 1 | Amy | Alberts | European ... | AWC Logo ... | 8 | \$259.41 |
| 2 | Amy | Alberts | European ... | Bike Wash - ... | 3 | \$52.47 |
| 3 | Amy | Alberts | European ... | Cable Lock ... | 1 | \$90.00 |
| 4 | Amy | Alberts | European ... | Chain ... | 6 | \$182.16 |
| 5 | Amy | Alberts | European ... | Classic Vest... | 3 | \$1,046.85 |
| 6 | Amy | Alberts | European ... | Classic Vest... | 7 | \$1,666.13 |
| 7 | Amy | Alberts | European ... | Front Brake... | 7 | \$1,597.50 |
| 8 | Amy | Alberts | European ... | Front Deraill... | 6 | \$933.20 |
| 9 | Amy | Alberts | European ... | Full-Finger ... | 5 | \$766.60 |
| 10 | Amy | Alberts | European ... | Full-Finger ... | 4 | \$364.70 |

Rysunek 37 Tabela wynikowa - analiza najlepiej sprzedających się produktów z podziałem na pracowników cz.1

Aby móc przedstawić łączną wartość sprzedanych produktów o tej samej nazwie przez różnych pracowników powstał kolejny proces.



Rysunek 38 Proces - analiza najlepiej sprzedających się produktów z podziałem na pracowników cz.2

W tej tabeli wynikowej powstało 6 kolumn o tych samych nazwach co w poprzedniej tabeli.

| Target table: Produkty o łącznej wartości sprzedaż >= 250000 (Najlepiej sprzedające się prod) | | | | | | | | |
|---|--|----------|-------------|------------|-----------|--------|------------|------------|
| # | | Column | Column D... | Expression | Type | Length | Informat | Format |
| 1 | | Imię | FirstName | | Character | 1024 | \$1024. | \$1024. |
| 2 | | Nazwisko | LastName | | Character | 1024 | \$1024. | \$1024. |
| 3 | | Tytuł | Title | | Character | 1024 | \$1024. | \$1024. |
| 4 | | Nazwa | Nazwa | | Character | 1024 | \$1024. | \$1024. |
| 5 | | Ilość | OrderQty | | Numeric | 8 | 6. | 6. |
| 6 | | Wartość | Wartosc | | Numeric | 8 | DOLLAR21.2 | DOLLAR21.2 |

Rysunek 39 Join - tworzenie kolumn do tabeli wynikowej cz.2

Dodany został także nowy warunek do Where o następującej treści.

| # | Boolean | (| Operand | Operator | Operand |
|---|---------|---|--|----------|---------|
| 1 | | | Analiza_sprzedaży_produktyw_prze."Wartosc" | >= | 250000 |

Rysunek 40 Where - warunek, który sprawia, że w tabeli wynikowej wyświetlają się tylko rekordy w których łączna wartość sprzedaży jest większa bądź równa 250000\$

Wyniki zostały posortowane malejąco według wartości.

| Table name | Column ... | Sort order | Column... |
|--|------------|------------|-----------|
| Produkty o łącznej wartości sprzedaż >= 250000 | Wartosc | Descending | Name |

Rysunek 41 Order by - sortowanie wyników malejąco według wartości

W ten sposób powstała tabela końcowa.

| # | Imię | Nazwisko | Tytuł | Nazwa | Ilość | Wartość |
|----|---------|----------|--------------------------|----------------------------|-------|--------------|
| 1 | Jae | Pak | Sales Representative ... | Mountain-200 Black, 38 ... | 126 | \$556,062.46 |
| 2 | Linda | Mitchell | Sales Representative ... | Mountain-200 Black, 38 ... | 79 | \$469,252.31 |
| 3 | Linda | Mitchell | Sales Representative ... | Mountain-200 Black, 42 ... | 73 | \$445,828.48 |
| 4 | Jae | Pak | Sales Representative ... | Mountain-200 Black, 42 ... | 98 | \$431,195.83 |
| 5 | Linda | Mitchell | Sales Representative ... | Mountain-200 Silver, 42... | 67 | \$403,703.27 |
| 6 | Linda | Mitchell | Sales Representative ... | Mountain-200 Silver, 46... | 70 | \$381,538.19 |
| 7 | Linda | Mitchell | Sales Representative ... | Mountain-200 Silver, 38... | 73 | \$379,947.58 |
| 8 | Jae | Pak | Sales Representative ... | Mountain-200 Silver, 38... | 80 | \$377,425.64 |
| 9 | Jae | Pak | Sales Representative ... | Mountain-200 Silver, 42... | 77 | \$371,720.14 |
| 10 | Linda | Mitchell | Sales Representative ... | Mountain-200 Black, 46 ... | 71 | \$366,365.42 |
| 11 | Jae | Pak | Sales Representative ... | Mountain-200 Silver, 46... | 69 | \$358,225.88 |
| 12 | Michael | Blythe | Sales Representative ... | Mountain-200 Black, 38 ... | 71 | \$354,034.99 |
| 13 | Jillian | Carson | Sales Representative ... | Road-250 Black, 44 ... | 68 | \$327,068.36 |
| 14 | Michael | Blythe | Sales Representative ... | Mountain-200 Black, 42 ... | 69 | \$323,642.76 |
| 15 | Michael | Blythe | Sales Representative ... | Mountain-200 Silver, 38... | 65 | \$309,668.95 |
| 16 | Michael | Blythe | Sales Representative ... | Road-250 Black, 44 ... | 63 | \$297,843.69 |
| 17 | Jae | Pak | Sales Representative ... | Mountain-200 Black, 46 ... | 68 | \$293,939.04 |
| 18 | Michael | Blythe | Sales Representative ... | Road-250 Black, 48 ... | 62 | \$290,950.63 |
| 19 | Tsvi | Reiter | Sales Representative ... | Mountain-200 Black, 38 ... | 78 | \$287,335.80 |
| 20 | Tsvi | Reiter | Sales Representative ... | Mountain-100 Black, 38 ... | 35 | \$270,267.08 |
| 21 | Jillian | Carson | Sales Representative ... | Mountain-200 Black, 38 ... | 55 | \$268,563.00 |
| 22 | Jillian | Carson | Sales Representative ... | Road-350-W Yellow, 48 ... | 54 | \$261,031.39 |
| 23 | Jillian | Carson | Sales Representative ... | Mountain-200 Black, 42 ... | 59 | \$256,120.88 |
| 24 | Michael | Blythe | Sales Representative ... | Mountain-200 Silver, 42... | 58 | \$253,541.77 |

Rysunek 42 Tabela wynikowa - analiza najlepiej sprzedających się produktów z podziałem na pracowników cz.2

Podsumowanie wyników:

Analiza 5 największych wartości sprzedaży poszczególnych produktów, które zostały sprzedane przez konkretnych pracowników:

1. Łączna Wartość Sprzedaży: 556 062,46 \$

Nazwa produktu: Mountain-200 Black, 38

Sprzedający: Jae Pak

Liczba sprzedanych sztuk: 126

2. Łączna Wartość Sprzedaży: 469 252,31 \$

Nazwa produktu: Mountain-200 Black, 38:

Sprzedający: Linda Mitchell

Liczba sprzedanych sztuk: 79

3. Łączna Wartość Sprzedaży: 445 828,48 \$

Nazwa produktu: Mountain-200 Black, 42:

Sprzedający: Linda Mitchell

Liczba sprzedanych sztuk: 73

4. Łączna Wartość Sprzedaży: 431 195,83 \$

Nazwa produktu: Mountain-200 Black, 42:

Sprzedający: Jae Pak

Liczba sprzedanych sztuk: 98

5. Łączna Wartość Sprzedaży: 403 703,27 \$

Nazwa produktu: Mountain-200 Silver, 42:

Sprzedający: Linda Mitchell

Liczba sprzedanych sztuk: 67

5. Podsumowanie

Schemat gwiazdy:

Przeprowadzona praktyczna część projektu, skupiająca się na analizie sprzedaży w firmie Adventure Works w latach 2001-2004, obejmuje szczegółowe zbadanie istotnych elementów działalności przedsiębiorstwa. Analizowane były aspekty czasowe, geograficzne, produktowe oraz związane z pracownikami. Zdecydowano się na stworzenie schematu gwiazdy, co poskutkowało konstrukcją tabeli faktów oraz tabel wymiarów dotyczących czasu, miejsca, produktu i pracownika. Kolejnym krokiem było przeprowadzenie adekwatnych procesów ETL (Extract, Transform, Load) w celu pozyskania pożądanych danych.

Podsumowanie analizy wyników:

Wyniki przeprowadzonych analiz w kontekście zbudowanego schematu gwiazdy dla Adventure Works dostarczają istotnych wskazówek i informacji kluczowych dla dalszych działań przedsiębiorstwa. Poniżej przedstawiam analizę uzyskanych wyników oraz sugestie dotyczące potencjalnych raportów i analiz na podstawie zgromadzonych danych.

1. Analiza sprzedaży według płci:

Analiza wyników sprzedaży ze względu na płeć pracowników ujawniła ciekawe trendy w poszczególnych latach. Różnice między wynikami osiąganymi przez mężczyzn i kobiety sugerują potrzebę bliższego przyjrzenia się dynamice zespołu sprzedażowego. Przygotowanie cyklicznych raportów, które dokładniej monitorują wyniki sprzedaży, biorąc pod uwagę także liczbę pracowników danej płci oraz średnie obroty generowane przez pracowników, pozwoli na bieżąco dostosowywać strategie marketingowe, aby zwiększyć efektywność zespołu.

2. Analiza sprzedaży według kategorii produktów:

Analiza kategorii produktów pozwala zidentyfikować, które z nich cieszą się największym zainteresowaniem klientów. Potencjalne raporty obejmują miesięczne zestawienia popularności poszczególnych kategorii, analizę trendów sezonowych w sprzedaży, oraz raporty porównawcze między różnymi latami.

3. Analiza sprzedaży kwartalnej i regionalnej:

Analiza kwartalna i regionalna pozwala obserwować wyniki sprzedaży w poszczególnych okresach i obszarach geograficznych. Przydatne raporty obejmują analizę kwartalnych zmian w sprzedaży oraz analizę sezonowości sprzedaży poszczególnych kategorii produktów w różnych regionach. Takie podejście umożliwia dostosowywanie strategii marketingowych do odpowiedniego sezonu oraz lokalizacji.

4. Analiza najlepiej sprzedających się produktów z podziałem na sprzedawców:

Analiza efektywności sprzedaży poszczególnych pracowników może posłużyć jako podstawa do rozwijania talentów w zespole. Cykliczne raporty ukazujące najnowsze wyniki sprzedaży, analizę produktów generujących największe zyski, oraz zestawienia indywidualnych osiągnięć pracowników, mogą pozytywnie wpływać na rozwój konkurencji w zespole.

Sugestie dla przyszłych analiz i raportów:

Adventure Works może dalej rozwijać swoje analizy i raporty, dodając następujące elementy:

Analiza wpływu rabatów na sprzedaż: raporty miesięczne dotyczące efektywności zniżek mogą posłużyć do zidentyfikowania tych, które generują największe zyski.

Analiza lojalności klienta: badanie tendencji zakupowych stałych klientów w odniesieniu do nowych klientów, pozwoli lepiej dostosować strategię marketingową, ceny produktów i zniżki.

Prognozy sprzedażowe: Wykorzystanie danych historycznych do stworzenia prognoz sprzedażowych, co pozwoli przewidzieć wyniki sprzedażowe.

Bibliografia

1. mgr inż. Tomasz Libera, mgr Piotr Ziuziański „CHARAKTERYSTYKA BUDOWY HURTOWNI DANYCH I MOŻLIWOŚCI IMPLEMENTACJI WYMIARÓW RÓŻNEGO TYPU”
2. <https://www.oracle.com/pl/big-data/what-is-big-data/>
3. <https://www.europarl.europa.eu/news/pl/headlines/society/20210211STO97614/big-data-definicja-korzysci-wyzwania-infografika>
4. <https://hive.apache.org/>
5. https://en.wikipedia.org/wiki/Apache_Hive
6. <https://www.bitwiseglobal.com/en-us/traditional-etl-vs-elt-on-hadoop/>
7. https://miro.medium.com/v2/resize:fit:715/0*xTPlycyxOuaE99FJ.png
8. <https://fotc.com/app/uploads/2022/02/schemat-budowy-hurtowni-danych-1024x533.png>