INŻYNIERIA HURTOWNI DANYCH

ANALIZA SPRZEDAŻY FIRMY ADVENTURE WORKS

NORBERT KABZIŃSKI

UNIWERSYTET EKONOMICZNY W KATOWICACH

Spis treści

1 Zastosowanie wybranego narzędzia w profilowaniu danych dla potrzeb hurtowni danych

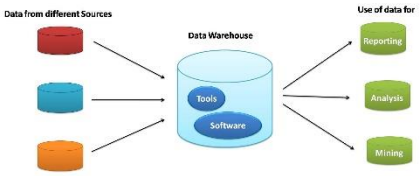
1.1 Wstęp i pojęcia

Dane to informacje, które są zebrane, zapisane i przechowywane w formie, która umożliwia ich późniejsze wykorzystanie. Mogą one przyjmować różne formy, takie jak tekst, liczby, obrazy, dźwięki czy video. Dane są fundamentalnym elementem w dzisiejszym świecie i odgrywają kluczową rolę w wielu dziedzinach życia.

Gromadzenie danych zachodzi nieustannie, każdej sekundy, a samo źródło informacji jest wszędzie. Zbieramy dane dotyczące klimatu, śledzimy statystyki sprzedaży, a nawet gry komputerowe gromadzą niezliczone ilości informacji o swoich graczach. Platformy mediów społecznościowych mają dostęp do różnorodnych aspektów naszego życia, tworząc obraz naszych preferencji i zachowań. W dzisiejszych czasach niemal każda działalność online pozostawia ślad cyfrowy, a dzięki powszechnej dostępności smartfonów ślad ten obejmuje nie tylko aktywność, lecz także lokalizacje, materiały wideo oraz audio. Wszystko to stanowi cenne źródło danych, które mogą być później wykorzystane do stworzenia zbioru danych.

Zbiór danych to zorganizowany zbiór informacji, które są zebrane, zgromadzone lub zarejestrowane w celu analizy, interpretacji lub przetwarzania. Zbiór danych może obejmować różnorodne elementy, takie jak liczby, tekst, obrazy, dźwięki czy inne rodzaje informacji. Jednym ze sposób użytkowania ich w celu analiz jest hurtownia danych.

Hurtownia danych to system zarządzania informacjami, który został zaprojektowany z myślą o ułatwieniu i wsparciu działań związanych z analizami biznesowymi. Gromadzą ogromne ilości danych historycznych z różnych źródeł co umożliwia systematyzację informacji z różnych dziedzin.



Rysunek 1-1
https://editor.analyticsvidhya.com/uploads/16664data%20warehouse%20image.jpg

1.2 Profilowanie danych

Praca na tak ogromnych źródłach danych jakie znajdują się w typowej hurtowni danych niesie za sobą ryzyko działania na brudnych danych. Brudne dane to takie dane, które w pewien sposób zawierają błędy lub niedoskonałości. Mogą posiadać duplikaty, być niekompletne, zawierać wykluczające się informację czy też literówki. Dane mogą stać się brudne na wielu etapach od momentu wprowadzenia przez przechowywanie do nieprawidłowego użytkowania. Często wynikają z pomyłek człowieka lub problemów technicznych. Brudne dane mogą zniekształcać obrazy rzeczywistości, prowadząc do wniosków opartych na fałszywych założeniach w konsekwencji doprowadzając do utraty czasu oraz zasobów.

Brak wiedzy odnośnie charakterystyki danych w danym zbiorze danych wywołuje konieczność przeprowadzenia procesu oceny ich jakości, mającego na celu uniknięcie potencjalnych konsekwencji wynikających z niedoskonałej jakości danych. W tym kontekście niezbędne staje się zastosowanie procesu profilowania danych.

Profilowanie danych to szczegółowe badanie danych, którego celem wyciągnięcie informacji dotyczących ich struktury i jakości. Proces ten umożliwia identyfikację potencjalnych problemów związanych z danymi, dostarczając istotnych wiadomości na temat charakterystyki zbioru danych. Jest to kluczowe dla oceny danych oraz podjęcia decyzji dotyczących dalszych działań, takich jak konieczność przeprowadzenia procesu oczyszczania danych. Wśród możliwych do wykrycia błędów znajdują się:

1. Puste wartości (reprezentujące nieznane lub brakujące dane).
2. Wartości niezgodne z kryteriami.
3. Wartości charakteryzujące się nieproporcjonalnie wysoką lub niską częstością występowania.
4. Dane, którym brakuje zastosowania.
5. Wartości wykraczające poza ustalone normy zakresu danych.

Przeprowadzenie profilowania danych umożliwia identyfikację i korektę tych nieprawidłowości, co przyczynia się do zwiększenia integralności oraz rzetelności analiz opartych na danych.

Etapy profilowania danych można podzielić na:

1. Analizę kompletności.
2. Analizę unikatowości.

3. Analizę rozkładu wartości.
4. Analizę zakresu.
5. Analizę wzorców.
6. Analizę powiązań.

Analiza kompletności danych to proces oceny danych obecnych w zbiorze, skoncentrowany na sprawdzeniu stopnia wypełnienia rekordów. Pozwala zbadać, w których miejscach, gdzie powinny być dane, brakuje ich, oraz w których miejscach, gdzie nie powinno być danych, takowe się znajdują.

Analiza unikatowości danych stanowi procedurę oceny stopnia niepowtarzalności informacji oraz identyfikacji potencjalnych duplikatów. Proces ten skupia się na sprawdzeniu, w jakim zakresie dane występują bez powtórzeń, a także ocenie uzasadnienia istnienia ewentualnych zdublowanych wpisów. Wartościowa ocena wysokiej unikalności danych sugeruje, że informacje te są prawdopodobnie czyste oraz jednoznaczne.

Analiza rozkładu wartości to proces, który ma na celu zbadanie, w jaki sposób różne wartości zmiennej rozkładają się pod względem częstości występowania w zbiorze danych. Jedną z powszechnie stosowanych metod w tego typu badaniach jest analiza Benforda.

Analiza zakresu to proces, który ma na celu najmniejszych i największych oraz najczęściej występujących wartości w kontekście danej zmiennej. Celem tego procesu jest identyfikacja potencjalnych nieprawidłowości w wartościach zmiennych oraz zrozumienie, jak dalece odbiegają one od wartości typowych. W ramach analizy zakresu możliwe jest również wykrycie wartości, które, choć znajdują się teoretycznie w przedziale możliwych wartości, istotnie wpływają na parametry rozkładu.

Analiza wzorców jest procesem mającym na celu zbadanie ewentualnych nieprawidłowości w formacie danych w danym zbiorze. Ocenia zgodność typów danych z ich oczekiwanymi typami. Sprawdza prawidłowe przypisanie wartości liczbowych do zmiennych numerycznych czy odpowiednia reprezentacja dat. Analiza wzorców umożliwia identyfikację przypadków, w których typy danych nie różnią się od oczekiwanych, co w dalszej perspektywie pozwala na ich poprawienie lub uzupełnienie.

Analiza powiązań to proces, którego celem jest zbadanie wzajemnych zależności między daną zmienną a innymi zmiennymi w zbiorze danych. Występują zależności:

1. Logiczne, które dotyczą sytuacji, kiedy elementy typu A zawierają się w zbiorze B.
2. Funkcyjne, gdzie wartość jednej zmiennej może być funkcją wartości innej lub wielu zmiennych, co oznacza, że wartość jednej wartości może wpływać na wartości innych zmiennych.
3. Częściowe, które obejmują sytuacje, w których zależności dotyczą niemal wszystkich jednostek, z wyjątkiem pewnych określonych przypadków.
4. Warunkowe, czyli zależności występują pod pewnym warunkiem.

1.3 Biblioteka ydata-profiling w języku Python oraz jej przekładowe zastosowanie w procesie profilowania danych.

Jednym z narzędzi powszechnie wykorzystywanych do analizy danych jest język programowania Python, który cieszy się popularnością ze względu na dostępność licznych bibliotek o charakterze open-source. Każda z tych bibliotek jest zaprojektowana w celu rozwiązania konkretnych problemów związanych z przetwarzaniem danych. Przykładem takiej biblioteki jest ydata-profiling, która została opracowana w odpowiedzi na potrzebę efektywnego i szybkiego profilowania danych, oferując wizualne prezentacje informacji dotyczących analizowanego zbioru danych oraz dostarcza statystyczną analizę tych informacji, co jest istotnym elementem procesu profilowania danych.

Kaggle jest platformą, skupiającą profesjonalistów i pasjonatów z zakresu analizy danych, nauki danych oraz uczenia maszynowego. Kaggle Notebook to narzędzie, stworzone przez firmę Kaggle. Umożliwia użytkownikom tworzenie, udostępnianie i współpracę nad projektem danych w chmurze.

Zestaw danych wykorzystany do przeprowadzenia profilowania jest zatytułowany „Tweets and User Engagement” oraz pochodzi z platformy Kaggle.com. Zbiór ten zawiera informacje dotyczące tweetów oraz interakcji między użytkownikami na platformie społecznościowej Twitter. Warto zaznaczyć, że w ramach tego zbioru danych dostępny jest także wskaźnik Klout score, który reprezentuje poziom wpływów użytkowników publikujących wpisy na wspomnianej platformie. Zbiór ten stanowi cenne źródło informacji, lecz jednocześnie charakteryzuje się znacznym

rozmiarem. Proces profilowania danych jest kluczowym narzędziem, które pozwoli na identyfikację potencjalnych problemów.

Celem tej części pracy jest ukazanie przykładowego procesu profilowania danych, który mógłby być wykorzystany przed lub w trakcie projektowania hurtowni danych. Czynności zastosowane tutaj jako przykłady znalazłyby swoje zastosowanie podczas badania danych przed utworzeniem schematu logicznego.

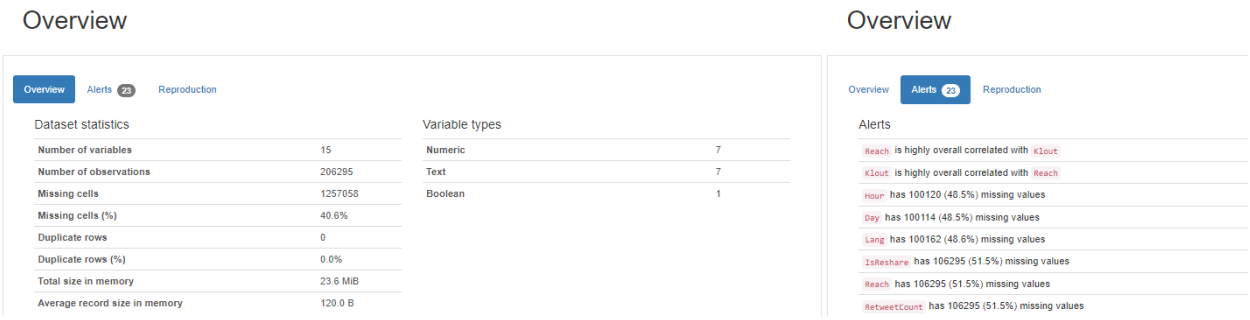
```
import numpy as np
import pandas as pd
from ydata_profiling import ProfileReport


df = pd.read_csv('/kaggle/input/tweets-and-user-engagement/Twitterdatainsheets.csv',
low_memory=False)
profile = ProfileReport(df, title="Pandas Profiling Report")
profile
```

Rysunek 1-2 Pierwszy człon kodu do profilowania danych.

Kod w języku Python w zeszycie Kaggle Notebook znajduje się na rysunku 1-2. Importuje niezbędne biblioteki, czyli:

1. NumPy, która dostarcza wsparcie do obliczania operacji na dużych tablicach i macierzach numerycznych.
2. Pandas, która dostarcza narzędzia analityczne do analizy danych.
3. Ydata_profiling, która dostarcza funkcje i narzędzia odpowiedzialne za proces i wizualizację profilowania danych.

Dalej korzystając z funkcji biblioteki Pandas wczytuje dane z pliku CSV do obiektu DataFrame *df* znajdującego się w chmurze, a *low_memory=False* oznacza, że nie Pandas nie będzie próbować optymalizować zużycia pamięci podczas wczytywania danych. W następnym kroku stosuje klasę *ProfileReport* dostarczoną przez bibliotekę ydata_profiling oraz wyświetla efekt końcowy.



Rysunek 1-3 Część przeglądowa

Analiza wyników raportu profili danych ujęta jest w dwie sekcje, obejmujące ogólny przegląd oraz szczegółowe wyniki analizy dla każdej z kolumn. W pierwszej sekcji zawarte są informacje dotyczące całego zbioru danych, bez szczegółowego podziału na poszczególne kolumny. Dzięki tej części raportu uzyskujemy podstawowe informacje i ostrzeżenia dotyczące analizowanego zbioru.

Raport wskazuje, że w całym zestawie danych nie występują zduplikowane rekordy. Jednakże, istotnym aspektem jest fakt, że brakujące wartości występują w 40,6% komórkach. Taka informacja jest kluczowa dla zrozumienia kompletności danych, a brakujące informacje mogą wpływać na rzetelność analizy.

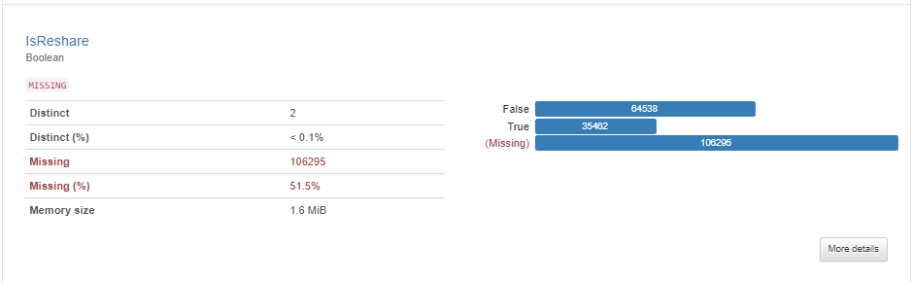
Podsumowując ogólną sekcję raportu, mamy do czynienia z 206 295 obserwacjami oraz 15 zmiennymi. Spośród tych zmiennych, 7 to zmienne numeryczne, 7 to zmienne tekstowe, a jedna zmienna jest typu Boolean. Ta informacja jest istotna dla zrozumienia różnorodności danych i rodzaju informacji zawartych w analizowanym zbiorze.

W sekcji "Alerts" raportu przedstawiane są najczęściej występujące możliwe problemy związane z analizowanymi kolumnami. Dzięki tej zakładce uzyskujemy istotne informacje dotyczące potencjalnych problemów w danych. Na przykład, raport zwraca uwagę na wysoką korelację pomiędzy zmiennymi "Reach" a "Klout". Korelacja ta może sugerować istnienie pewnego stopnia współzależności między tymi dwiema zmiennymi, co może być istotne przy dalszych analizach.

W części dotyczącej analizy poszczególnych zmiennych w raporcie, dostępne są istotne statystyki charakteryzujące każdą z kolumn w analizowanym zbiorze danych. Informacje te obejmują unikalność rekordów, wartości ekstremalne, odsetek negatywnych wartości oraz zer, liczbę brakujących komórek i średnie dla poszczególnych zmiennych. Po prawej stronie raportu znajdują się graficzne reprezentacje tych statystyk, takie jak wykresy i histogramy, co ułatwia

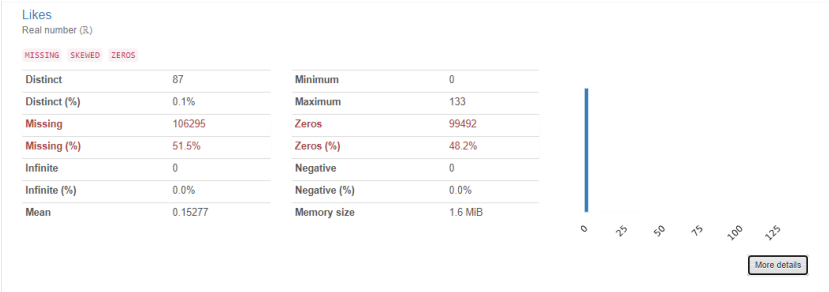
wizualizację charakterystyki poszczególnych kolumn. Warto zaznaczyć, że opcja "More details" dostarcza dodatkowych informacji statystycznych, co pozwala na bardziej zaawansowaną analizę danych w kontekście poszczególnych zmiennych. Dzięki tej sekcji raportu możliwe jest kompleksowe zrozumienie właściwości i dynamiki każdej zmiennej w badanym zbiorze danych.

Poniżej przeprowadzono analizę trzech wybranych kolumn z analizowanego zestawu danych.



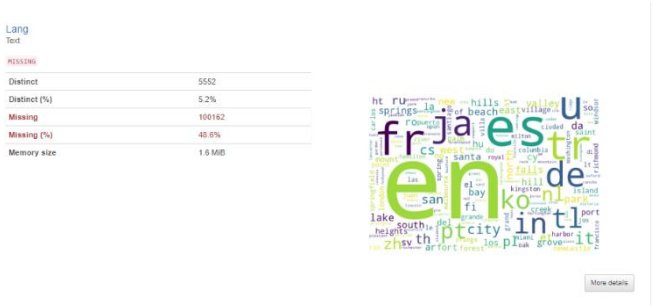
Rysunek 1-4 Raport szczegółowy zmiennej IsReshare

Zmienna IsReshare, której dane zostały zwizualizowane na rysunku 1-4 to zmienna typu boolean, co oznacza, że przyjmuje tylko dwie wartości: prawda (True) lub fałsz (False). Po prawej stronie raportu można zaobserwować graficzne przedstawienie ilości wartości True, False oraz brakujących. Jest to istotne, aby zrozumieć jakie są proporcje między wystąpieniem prawdziwych a fałszywych wartości w analizowanym zbiorze danych. Należy zauważyć, że 51,5% komórek tej zmiennej nie posiada danych. Brakujące dane mogą wpłynąć na wiarygodność analizy, dlatego konieczne jest uwzględnienie tego aspektu przy pracy z tym zestawem danych.



Rysunek 1-5 Raport szczegółowy zmiennej Likes

Zmienna Likes, przedstawiona na rysunku 1-5, jest zmienną numeryczną. Przedstawia ilość polubieni pod tweedami. Na wykresie po prawej stronie zauważalna jest dominująca przewaga wartości "0" nad pozostałymi. W tej kategorii występuje 87 unikatowych wartości, a ich średnia wynosi 0,15277. Wartość minimalna to 0, natomiast maksymalna wynosi 133. Należy zauważyć, że aż 51,5% wartości w komórkach jest pustych, co może mieć istotny wpływ na analizę danych.



Rysunek 1-6 Raport szczegółowy zmiennej Lang

Zmienna Lang, zilustrowana na rysunku 1-6, jest zmienną tekstową, reprezentującą kody języków używanych w tweetach. W analizie tej zmiennej można zauważyć, że występują aż 5552 unikatowe wartości, co sugeruje znaczną różnorodność języków używanych do pisania tweetów. Podobnie jak w przypadku poprzednich kolumn, również dla zmiennej Lang można zauważyć, że brakuje aż 48,6% wartości .

2 Wstęp

2.1 Cel realizacji projektu

Adventure Works Cycles to nazwa międzynarodowej i fikcyjnej firmy z siedzibą w Bothell w stanie Waszyngton. Specjalizuje się produkcją rowerów, sprzętu rowerowego oraz akcesoriów. Firma ta prowadzi sprzedaż w sześciu krajach na trzech różnych kontynentach:

1. Ameryce Północnej (USA oraz Kanada).
2. Europie (Niemcy, Francja oraz Wielka Brytania).
3. Oceanię (Australia).

Adventure Works to przykładowa baza danych w różnych produktach i narzędziach Microsoft. Należy podkreślić, że jest to jedynie fikcyjna firma stworzona w celu ilustrowania różnych funkcji, a nie rzeczywista, funkcjonująca organizacja.

SAS Data Integration Studio to zaawansowane narzędzie do integracji danych, rozwijane przez firmę SAS Institute. Pozwala zarządzać danymi z różnych źródeł. Oprogramowanie to umożliwia organizację pracy nad danymi stosując procesy ETL (Extract, Transform, Load), które są kluczowe do skutecznego zarządzania danymi.

Rysunek 2-2 Logo firmy SAS Institute
Źródło:https://upload.wikimedia.org/wikipedia/commons/1/10/SAS_logo_horiz.svg

Rysunek 2-1 Logo firmy Adventure Works Cycles
Źródło:https://i0.wp.com/blog.jpries.com/wp-content/uploads/2015/12/AdventureWorks-Logo_blog.jpg

Celem realizacji projektu było przeprowadzenie wybranych analiz sprzedaży firmy Adventure Works, korzystając z narzędzia SAS Data Integration Studio. Analizy obejmują:

1. Wartość sprzedanych produktów w danym kraju z podziałem na pory roku.
2. Ilość sprzedanych modeli rowerów do danego kraju.
3. Udział procentowy kart kredytowych wykorzystywanych do dokonywania płatności z podziałem na kraje.
4. Wartość sprzedaży danego pracownika z podziałem na kraje zamówień.
5. Najgorzej sprzedające się produkty w Stanach Zjednoczonych.
6. Odsetek jaki stojaki na rowery stanowią na tle całkowitej sprzedaży w danym kraju.

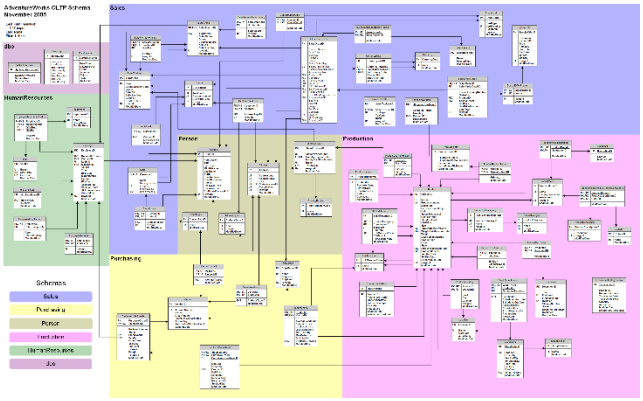
Tematyką hurtowni danych została sprzedaż. W ramach tych badań przywiązano szczególną wagę do aspektów terytorialnych dokładnie analizując zmiany i trendy w dziedzinie handlu występujące w różnych krajach.

Żeby zrealizować projekt, konieczne było w pierwszym etapie (zgodne z wytycznymi) opracowanie opisu źródeł danych. Kolejnym krokiem było sporządzenie projektu bazy danych, wykorzystując schemat gwiazdy, a także szczegółowe omówienie każdego z procesów. W dalszej fazie realizacji projektu przeprowadzono procesy ETL konieczne do osiągnięcia zamierzonych rezultatów, a na zakończenie dokonano podsumowania wniosków.

2.2 Ogólny schemat bazy Adventure Works

Baza danych Adventure Works to przykładowa baza danych, która pierwotnie została opublikowana przez firmę Microsoft w celu przedstawienia, jak zaprojektować bazę danych SQL Server. Struktura tej bazy danych obejmuje podział na obszary skupiające się wokół:

1. Danych systemowych (Dbo).
2. Danych o zasobach ludzkich (HumanResources).
3. Danych osobowych (Person).
4. Danych produkcyjnych (Production).
5. Danych o sprzedaży (Sales).
6. Danych o zakupach (Purchasing).



Rysunek 2-3 Schemat bazy Adventure Works
Źródło: https://i0.wp.com/improveandrepeat.com/wp-content/uploads/2018/12/AdvWorksOLTPSchemaVisio.png?ssl=1

Dane w analizowanej bazie obejmują okres od lata 2001 roku do lata 2004 roku. Przedstawiając szczegółowe informacje dotyczące różnorodnych aspektów funkcjonowania przedsiębiorstwa. Zakres tych danych obejmuje obszary od zatrudnienia pracowników, poprzez produkcję, aż po transakcje handlowe.

W celu przeprowadzenia analiz skupiono się głównie na schemacie danych dotyczących sprzedaży. Skorzystano również z tabel z trzech obszarów tematycznych: danych osobowych, danych o zasobach ludzkich oraz danych produkcyjnych.

2.3 Obszar danych sprzedażowych

#	Name	Description
1	SalesOrderID	SalesOrderID
2	RevisionNumber	RevisionNumber
3	OrderDate	OrderDate
4	DueDate	DueDate
5	ShipDate	ShipDate
6	Status	Status
7	OnlineOrderFlag	OnlineOrderFlag
8	SalesOrderNumber	SalesOrderNumber
9	PurchaseOrderNumber	PurchaseOrderNum...
10	AccountNumber	AccountNumber
11	CustomerID	CustomerID
12	ContactID	ContactID
13	SalesPersonID	SalesPersonID
14	TerritoryID	TerritoryID
15	BillToAddressID	BillToAddressID
16	ShipToAddressID	ShipToAddressID
17	ShipMethodID	ShipMethodID
18	CreditCardID	CreditCardID
19	CreditCardApprovalCode	CreditCardApproval...
20	CurrencyRateID	CurrencyRateID
21	SubTotal	SubTotal
22	TaxAmt	TaxAmt
23	Freight	Freight
24	TotalDue	TotalDue
25	Comment	Comment
26	rowguid	rowguid
27	ModifiedDate	ModifiedDate

Rysunek 2-4 Kolumny tabeli SALES_SALESORDERHEADER

Tabela SALES_SALESORDERHEADER stanowi centrum w strukturze danych dotyczących sprzedaży. Jej kluczem głównym jest SalesOrderID. Tabela ta pełni kluczową rolę w gromadzeniu niezbędnych informacji dotyczących zamówień, zawierając istotne szczegóły, takie jak daty transakcji, status zamówienia, wartość podatku, koszty dostawy oraz całkowite zobowiązania finansowe klienta. Dodatkowo, tabela SALES_SALESORDERHEADER zawiera klucze obce, które odnoszą się do innych tabel. Te klucze obejmują informacje dotyczące rodzaju dostawy, użytej karty kredytowej do dokonania płatności, danych kontaktowych podmiotu składającego zamówienie, podmiotu złożonego zamówienia, terytorium oraz sprzedawcy obsługującego dane zamówienie.

Pozostałe tabele w analizowanej bazie danych to SALES_SALESPERSON, SALES_SALESORDERDETAIL, SALES_CREDITCARD, SALES_CONTACTCREDITCARD oraz SALES_INDIVIDUAL. Kluczami głównymi tych tabel są klucze obce pochodzące z tabeli SALES_SALESORDERHEADER, które stanowią powiązanie między tymi strukturami

#	Name	Description
1	SalesOrderID	SalesOrderID
2	SalesOrderDetailID	SalesOrderDetailID
3	CarrierTrackingNumber	CarrierTrackingNum...
4	OrderQty	OrderQty
5	ProductID	ProductID
6	SpecialOfferID	SpecialOfferID
7	UnitPrice	UnitPrice
8	UnitPriceDiscount	UnitPriceDiscount
9	LineTotal	LineTotal
10	rowguid	rowguid
11	ModifiedDate	ModifiedDate

#	Name	Description
1	CreditCardID	CreditCardID
2	CardType	CardType
3	CardNumber	CardNumber
4	ExpMonth	ExpMonth
5	ExpYear	ExpYear
6	ModifiedDate	ModifiedDate

#	Name	Description
1	CustomerID	CustomerID
2	ContactID	ContactID
3	Demographics	Demographics
4	ModifiedDate	ModifiedDate

#	Name	Description
1	SalesPersonID	SalesPersonID
2	TerritoryID	TerritoryID
3	SalesQuota	SalesQuota
4	Bonus	Bonus
5	CommissionPct	CommissionPct
6	SalesYTD	SalesYTD
7	SalesLastYear	SalesLastYear
8	rowguid	rowguid
9	ModifiedDate	ModifiedDate

Rysunek 2-5 Kolumny tabeli SALES_SALEORDERDETAIL, SALES_CREDITCARD, SALES_INDIVIDUAL, SALES_SALESPERSON

2.4 Obszar danych osobowych

W obszarze danych osobowych wykorzystanym oraz najważniejszym elementem struktury możemy nazwać tabele PERSON_CONTACT, gdzie kluczem głównym jest CONTACTID. Tabela ta pełni centralną rolę w gromadzeniu szczegółowych informacji dotyczących przedstawicieli klientów. Zawiera ona istotne dane, takie jak tytuł, imię, nazwisko, adres e-mail, numer. Tabela PERSON_CONTACT stanowi zatem istotne źródło informacji kontaktowych.

#	Name	Description
1	ContactID	ContactID
2	NameStyle	NameStyle
3	Title	Title
4	FirstName	FirstName
5	MiddleName	MiddleName
6	LastName	LastName
7	Suffix	Suffix
8	EmailAddress	EmailAddress
9	EmailPromotion	EmailPromotion
10	Phone	Phone
11	PasswordHash	PasswordHash
12	PasswordSalt	PasswordSalt
13	AdditionalContactInfo	AdditionalContactInfo
14	rowguid	rowguid
15	ModifiedDate	ModifiedDate

Rysunek 2-6 Kolumny tabeli PERSON_CONTACT

W obszarze danych osobowych znajdują się również tabele PERSON_ADDRESS, PERSON_STATEPROVINCE oraz PERSON_COUNTRYREGION. Te struktury przechowują istotne informacje dotyczące adresów klientów, regionów, w jakich się znajdują oraz krajów, do których są przypisani.

#	Name	Description
1	AddressID	AddressID
2	AddressLine1	AddressLine1
3	AddressLine2	AddressLine2
4	City	City
5	StateProvinceID	StateProvinceID
6	PostalCode	PostalCode
7	rowguid	rowguid
8	ModifiedDate	ModifiedDate

#	Name	Description
1	StateProvinceID	StateProvinceID
2	StateProvinceCode	StateProvinceCode
3	CountryRegionCode	CountryRegionCode
4	IsOnlyStateProvinceFlag	IsOnlyStateProvinc...
5	Name	Name
6	TerritoryID	TerritoryID
7	rowguid	rowguid
8	ModifiedDate	ModifiedDate

#	Name	Description
1	CountryRegionCode	CountryRegionCode
2	Name	Name
3	ModifiedDate	ModifiedDate

Rysunek 2-7 Kolumny tabel PERSON_ADDRESS, PERSON_STATEPROVINCE oraz PERSON_COUNTRYREGION.

2.5 Obszar danych o zasobach ludzkich

#	Name	Description
1	EmployeeID	EmployeeID
2	NationalIDNumber	NationalIDNumber
3	ContactID	ContactID
4	LoginID	LoginID
5	ManagerID	ManagerID
6	Title	Title
7	BirthDate	BirthDate
8	MaritalStatus	MaritalStatus
9	Gender	Gender
10	HireDate	HireDate
11	SalariedFlag	SalariedFlag
12	VacationHours	VacationHours
13	SickLeaveHours	SickLeaveHours
14	CurrentFlag	CurrentFlag
15	rowguid	rowguid
16	ModifiedDate	ModifiedDate

Rysunek 2-8 Kolumny tabeli HUMANRESOURCES_EMPLOYEE

Jeżeli chodzi o obszar danych o zasobach ludzkich skorzystałem wyłącznie z tabeli HUMANRESOURCES_EMPLOYEE, która zawiera szczegółowe informacje na temat pracowników przedsiębiorstwa. Kluczowym elementem identyfikacyjnym jest EmployeeID. Tabela ta obejmuje istotne dane, takie jak imię, nazwisko, informacje o przełożonym pracownika, datę zatrudnienia, płeć oraz stan cywilny.

2.6 Obszar danych o produkcji

#	Name	Description
1	ProductID	ProductID
2	Name	Name
3	ProductNumber	ProductNumber
4	MakeFlag	MakeFlag
5	FinishedGoodsFlag	FinishedGoodsFlag
6	Color	Color
7	SafetyStockLevel	SafetyStockLevel
8	ReorderPoint	ReorderPoint
9	StandardCost	StandardCost
10	ListPrice	ListPrice
11	Size	Size
12	SizeUnitMeasureCode	SizeUnitMeasureCode
13	WeightUnitMeasureCode	WeightUnitMeasure...
14	Weight	Weight
15	DaysToManufacture	DaysToManufacture
16	ProductLine	ProductLine
17	Class	Class
18	Style	Style
19	ProductSubcategoryID	ProductSubcategor...
20	ProductModelID	ProductModelID
21	SellStartDate	SellStartDate
22	SellEndDate	SellEndDate
23	DiscontinuedDate	DiscontinuedDate
24	rowguid	rowguid
25	ModifiedDate	ModifiedDate

Rysunek 2-9 Kolumny tabeli PRODUCTION_PRODUCT.

Tabela PRODUCTION_PRODUCT zawiera wszystkie szczegółowe informacje dotyczące konkretnych produktów. Klucz główny tej tabeli to ProductID, ale pojawiają się w niej klucze obce, które stanowią odniesienie do innych tabel. W jej zakresie znajdują się istotne dane, takie jak nazwa produktu, kolor, standardowy koszt, cena, rozmiar, kategoria i podkategoria. Tabela PRODUCTION_PRODUCT pełni kluczową rolę jako źródło danych na temat oferowanych przez przedsiębiorstwo produktów.

#	Name
1	ProductCategoryID
2	Name
3	rowguid
4	ModifiedDate

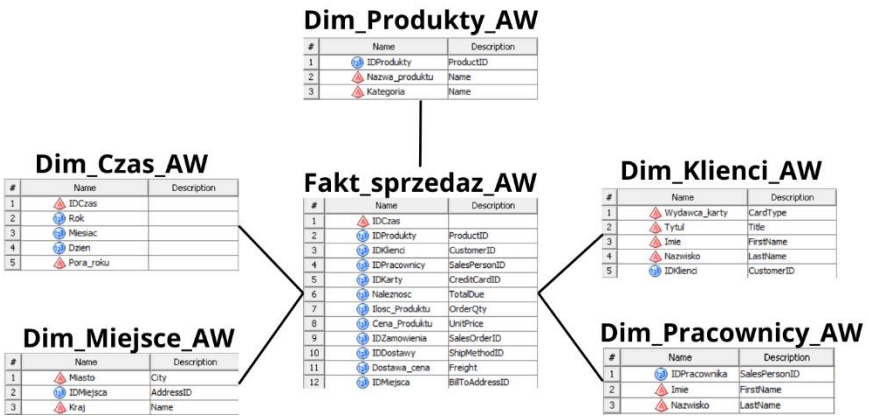
#	Name
1	ProductSubcategoryID
2	ProductCategoryID
3	Name
4	rowguid
5	ModifiedDate

Rysunek 2-10 Kolumny tabel PRODUCT_PRODUCTCATEGORY i PRODUCT_PRODUCTSUBCATEGORY

W projekcie wykorzystałem również tabele PRODUCTION_PRODUCTCATEGORY (zawierająca informację o kategorii) oraz PRODUCTION_PRODUCTSUBCATEGORY (zawierająca informację o podkategorii). Ich klucze główne to ProductSubcategoryID oraz ProductCategoryID.

3 Model logiczny hurtowni danych w oparciu o schemat gwiazdy

3.1 Schemat gwiazdy



Rysunek 3-1 Schemat gwiazdy

Chcąc dokonać efektywnych analiz, konieczne jest zaprojektowanie bazy danych o schemacie gwiazdy. Baza ta będzie składać się z tabeli faktów połączonej z czterema tabelami wymiarów. Podejście to znacząco ułatwia ekstrakcję danych umożliwiając wyeliminowanie zbędnych informacji na skutek czego wzrasta wydajność przeprowadzanych procesów. Schemat dla tego projektu znajduje się w rysunku 4-1.

3.2 Wymiar czas

Tabela wymiaru czasu zawiera informację o datach z zakresu Styczeń 2001 do Grudnia 2005. Każdy wiersz daty posiada dwie kolumny typu tekstowego – IDCzas i Pora_roku oraz trzy kolumny typu numerycznego – Rok, Miesiąc, Dzień. Proces tworzenia tabeli Dim_Czas_AW wymagał wykorzystania węzła User Written Code.

```
%let fromDate = 01Jan2000;
%let toDate = 31Dec2005;

data &_OUTPUT1;
    length Pora_Roku $8.;
    do data = "&fromDate"d to "&toDate"d;
      IDCzas = substr(put(year(data),4.),3,2) || put(month(data), z2.) || put(day(data), z2.);
      Rok = year(data);
      Miesiac = month(data);
      Dzien = day(data);
      if Miesiac in (3, 4, 5) then Pora_Roku = 'Wiosna';
        else if Miesiac in (6, 7, 8) then Pora_Roku = 'Lato';
        else if Miesiac in (9, 10, 11) then Pora_Roku = 'Jesien';
        else Pora_Roku = 'Zima';
        output;
    end;

    format data date9.;
    drop data;
run;
```

Rysunek 3-2 Kod generujący tabelę wymiaru czasu

Kod z rysunku 4-2 tworzy nowy zestaw danych o nazwie &_OUTPUT1 z informacjami dotyczącymi dat w określonym zakresie. ID generowane jest poprzez usunięcie dwóch pierwszych

cyfr z roku i dodanie zera do jednocyfrowych dni oraz miesięcy. Dodatkowo instrukcja warunkowa dopisuje odpowiednią porę roku zależnie od miesiąca. Tak wygenerowaną tabelę &_OUTPUT1 węzeł Table Loader wczytuje do tabeli wymiary czasu.

Rysunek 3-3 Schemat procesu tworzenia tabeli czasu

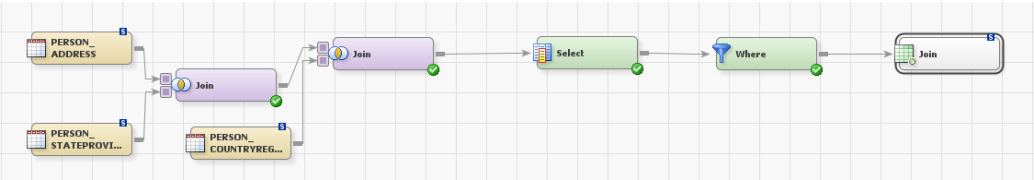
Dim_Czas_AW można nazwać jednym z najbardziej istotnych elementów schematu gwiazdy ze względu na możliwość analizy czynników sprzedażowych przy uwzględnieniu istotnych ograniczeń czasowych.

#	IDCzas	Rok	Miesiac	Dzien	Pora_roku
1	000101	2000	1	1	Zima
2	000102	2000	1	2	Zima
3	000103	2000	1	3	Zima
4	000104	2000	1	4	Zima
5	000105	2000	1	5	Zima

Rysunek 3-4 Wymiar czasu.

3.3 Wymiar miejsce

Wyjątkowo ważnym wymiarem dla analiz przeprowadzonych w tym projekcie był wymiar czasu – Dim_Miejsce_AW. Został użyty w każdej z pięciu analiz, odpowiadając za dane o kraju do którego przypisane są zamówienia.

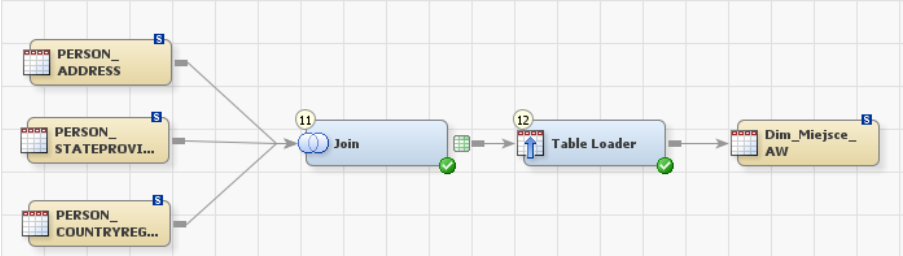


Rysunek 3-5 Schemat węzła Join w procesie tworzącym wymiar miejsca

Proces tworzenia przedstawionej tabeli wymagał połączenia trzech zbiorów danych: PERSON_ADDRESS, PERSON_STATEPROVINCE oraz PERSON_COUNTRYREGION, z wykorzystaniem węzła *Join*. Węzeł *Join* umożliwia stosowanie ograniczeń z języka SQL w ramach programu SAS Data Integration Studio. W jego zakresie znajdują się ograniczenia charakterystyczne dla tego języka, takie jak *Select, Where, Order by, Group by* i *Having*. W ramach ograniczenia *Select* zdefiniowano kolumny, a mianowicie:

1. Miasto (Kolumna City z tabeli PERSON_STATEPROVINCE).
2. Kraj (Kolumna Name z PERSON_COUNTRYREGION).

3. IDMiejsca (Kolumna AddressID z PERSON_ADDRESS).



Rysunek 3-6 Schemat procesu tworzącego wymiar miejsca.

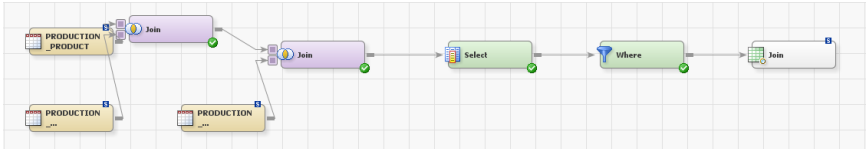
Analogicznie do wymiaru czasu węzeł Table Loader został użyty w celu ładowania danych do tabeli wymiaru.

#	Miasto	IDMiejsca	Kraj
1	Bothell ...	1	United States ...
2	Bothell ...	2	United States ...
3	Bothell ...	3	United States ...
4	Bothell ...	4	United States ...
5	Bothell ...	5	United States ...
6	Bothell ...	6	United States ...
7	Bothell ...	7	United States ...
8	Bothell ...	8	United States ...
9	Bothell ...	9	United States ...

Rysunek 3-7 Wymiar miejsca.

3.4 Wymiar produkt

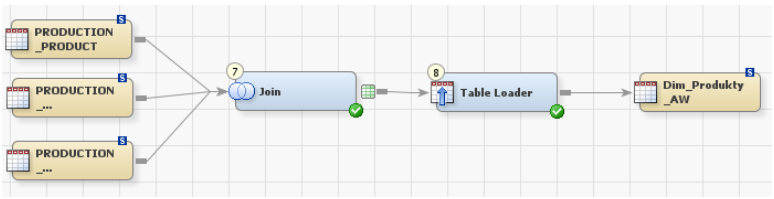
Równie istotne znaczenie miało utworzenie wymiaru produktów. W nim znajdowały się kluczowe, z punktu widzenia analiz, informacje dotyczące produktów, a mianowicie nazwa produktu oraz kategoria. W późniejszych analizach informacje te pomogły określić jakie produkty znajdowały się w konkretnych zamówieniach oraz określenia ich przynależności do konkretnych kategorii. Proces tworzenia przebiegł analogicznie do poprzedniego procesu.



Rysunek 3-8 Schemat węzła Join w procesie tworzącym wymiar produktów.

Połączone zostały trzy tabele z obszaru produkcji: PRODUCTION_PRODUCT, PRODUCTION_PRODUCTCATEGORY oraz PRODUCTION_PRODUCTSUBCATEGORY. W ograniczeniu select powstały kolumny:

1. IDProdukty (Kolumna ProductID z tabeli PRODUCTION_PRODUCT)
2. Nazwa_produktu (Kolumna Name z tabeli PRODUCTION_PRODUCT) .
3. Kategoria (Kolumna Name z tabeli PRODUCTION_PRODUCTSUBCATEGORY).



Rysunek 3-9 Schemat tworzacy wymiar produktów.

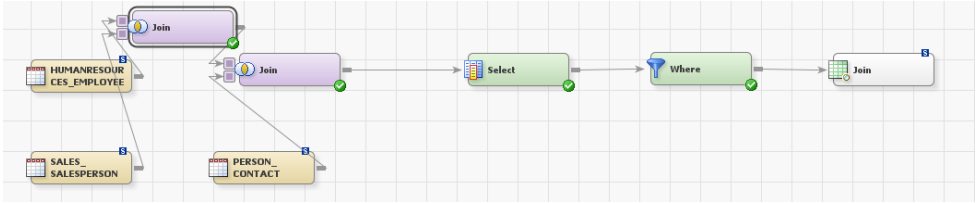
Następnie dane zostały wczytane do wymiaru produktów przy pomocy węzła Table Loader.

#	IDProdukty	Nazwa_produktu	Kategoria
1	680	HL Road Frame - Black, 58 ...	Road Frames ...
2	680	HL Road Frame - Black, 58 ...	Road Frames ...
3	680	HL Road Frame - Black, 58 ...	Road Frames ...
4	680	HL Road Frame - Black, 58 ...	Road Frames ...
5	706	HL Road Frame - Red, 58 ...	Road Frames ...
6	706	HL Road Frame - Red, 58 ...	Road Frames ...
7	706	HL Road Frame - Red, 58 ...	Road Frames ...
8	706	HL Road Frame - Red, 58 ...	Road Frames ...
9	707	Sport-100 Helmet, Red ...	Helmets ...
10	707	Sport-100 Helmet, Red ...	Helmets ...
11	707	Sport-100 Helmet, Red ...	Helmets ...

Rysunek 3-10 Wymiar produktów

3.5 Wymiar pracownicy

W wymiarze pracowników, zostały zawarte kluczowe informacje z perspektywy analiz na temat osób zatrudnionych na stanowisku sprzedawcy w firmie Adventure Works Cycles. W późniejszych analizach te dane wspomogły ocenić wartość jaką pracownicy wnosili do firmy na podstawie ich wyników finansowych. Dim_Pracownicy_AW został utworzony w sposób analogiczny do poprzednich wymiarów. Ten zestaw danych jest o tyle ciekawy, że w procesie tworzenia go należało wymieszać dane z trzech różnych obszarów bazy Adventure Works.

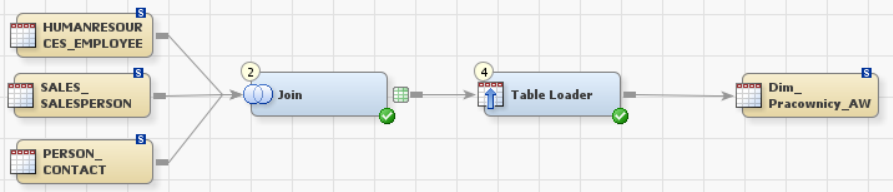


Rysunek 3-11 Węzeł join w schemacie procesu tworzącego wymiar pracowników.

W przypadku tabeli wymiaru pracowników połączenie zostały tabele: HUMANRESOURCES_EMPLOYEE, SALES_SALESPERSON oraz PERSON_CONTACT.

W ograniczeniu *select* powstały kolumny:

1. IDPracownika (Kolumna SalesPersonID z tabeli SALES_SALESPERSON)
2. Imie (Kolumna FirstName z tabeli HUMANRESOURCES_EMPLOYEE) .
3. Nazwisko (Kolumna LastName z tabeli HUMANRESOURCES_EMPLOYEE).



Rysunek 3-12 Schemat procesu tworzącego wymiar pracowników.

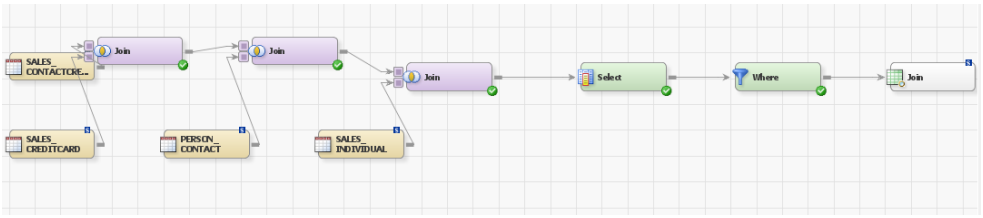
Tak jak w poprzednim procesie węzeł Table Loader został wykorzystany w celu załadowania danych do tabeli wymiaru.

#	IDPracownika	Imie	Nazwisko
1	268	Stephen ...	Jiang ...
2	288	Syed ...	Abbas ...
3	284	Amy ...	Alberts ...
4	280	Pamela ...	Ansman-Wolfe...
5	283	David ...	Campbell ...
6	277	Jillian ...	Carson ...
7	281	Shu ...	Ito ...
8	276	Linda ...	Mitchell ...
9	279	Tsvi ...	Reiter ...
10	282	José ...	Saraiva ...
11	278	Garrett ...	Vargas ...
12	286	Ranjit ...	Varkey Chudu...
13	289	Rachel ...	Valdez ...
14	290	Lynn ...	Tsoflias ...
15	285	Jae ...	Pak ...
16	275	Michael ...	Blythe ...
17	287	Tete ...	Mensa-Annan ...

Rysunek 3-13 Wymiar pracowników

3.6 Wymiar klientów

Ostatni z wymiarów został utworzony poprzez połączenie aż czterech zbiorów danych. Jego głównym celem było zebranie informacji dotyczących klientów. Tabela ta zawiera imię i nazwisko osoby odpowiedzialnej za kontakt w sklepie do którego wysyłane są towary oraz wydawcę karty kredytowej użytej do uiszczenia opłat za zamówione towary.

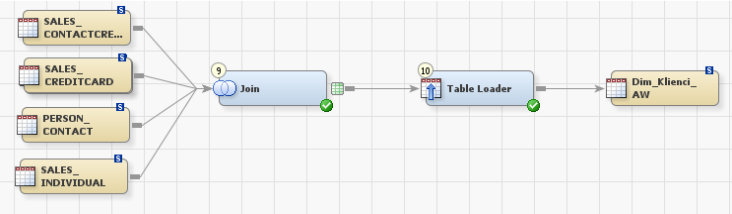


Rysunek 3-14 Węzeł join w schemacie procesu tworzącego wymiar klientów.

Głowna różnica między procesem tworzenia tego wymiaru, a poprzednich jest ilość złączonych tabel. Żeby zebrać wszystkie potrzebne do analiz informacje trzeba było połączyć, aż cztery: PERSON_CONTACT, SALES_INDIVIDUAL, SALES_CONTACTCREDITCARD, SALES_CREDITCARD. Kolumny jakie zostały utworzone stosując *select* to:

1. Wydawca_karty (Kolumna CardType z tabeli SALES_CREDITCARD).
2. Tytul (Kolumna Title z tabeli PERSON_CONTACT).

3. Imie (Kolumna FirstName z tabeli PERSON_CONTACT).
4. Nazwisko (Kolumna LastName z tabeli PERSON_CONTACT).
5. IDKlienci (Kolumna CustomerID z tabeli SALES_INDIVIDUAL).



Rysunek 3-15 Schemat tworzący wymiar klientów.

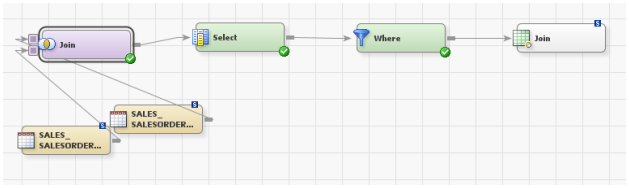
Analogicznie do poprzednich procesów węzeł Table Loader odpowiada za umieszczenie danych w tablicy wymiaru.

#	Wydawca_karty	Tytul	Imie	Nazwisko	IDKlienci
1	Vista	Mr.	David	Robinett	11377
2	Distinguish	Ms.	Rebecca	Robinson	11913
3	SuperiorCard	Ms.	Dorothy	Robinson	11952
4	ColonialVoice	Ms.	Carol Ann	Rockne	20164
5	ColonialVoice	Mr.	Scott	Rodgers	20211
6	SuperiorCard	Mr.	Jim	Rodman	20562
7	Vista	Mr.	Eric	Rothenberg	20668
8	ColonialVoice	Mr.	Michael	Rothkugel	20813
9	SuperiorCard	Mr.	Pablo	Rovira Diez	21190
10	Distinguish	Ms.	Linda	Rousey	21279
11	ColonialVoice	Mr.	Luke	Roy	21286
12	SuperiorCard	Ms.	Lisa	Roy	21403
13	ColonialVoice	Mr.	Michael	Ruggiero	21867

Rysunek 3-16 Wymiar klientów

3.7 Tabela faktów

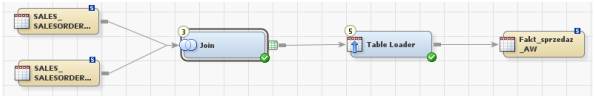
Tabela faktów co do zasady zawiera klucze obce łączące jej struktury z tabelami wymiarów. W przypadku tego projektu zawiera dane sprzedażowe. Poza kluczami zawiera jedynie dane numeryczne.



Rysunek 3-17 Schemat węzła join w procesie tworzącym tabelę faktów.

W ograniczeniu *select* zostały wybrane kolumny odpowiadające kolumnom zawierającym ID z wymiarów (Tabele SALES_SALESORDERHEADER oraz SALES_SALESORDERDETAIL zawierają je wszystkie w sobie) oraz dodatkowe kolumny:

1. Ilosc_Produktu (Kolumna OrderQty z tabeli SALES_SALESORDERDETAIL).
2. Cena_Produktu (Kolumna UnitPrice z tabeli SALES_SALESORDERDETAIL).



Rysunek 3-18 Schemat tworzący tabelę faktów.

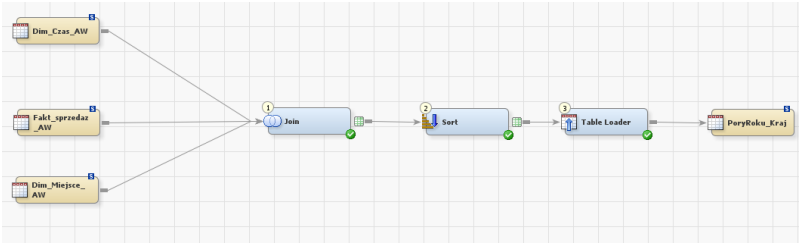
#	IDCzas	IDProdukty	IDKlienci	IDPracownicy	IDKarty	Naleznosc	Ilosc_Produktu	Cena_Produktu	IDZamowienia	IDDostawy	Dostawa_cena	IDMiejsca
1	010701	776	676	279	16281	.	1	$2,024.99	43659	.	.	985
2	010701	777	676	279	16281	.	3	$2,024.99	43659	.	.	985
3	010701	778	676	279	16281	.	1	$2,024.99	43659	.	.	985
4	010701	771	676	279	16281	.	1	$2,039.99	43659	.	.	985
5	010701	772	676	279	16281	.	1	$2,039.99	43659	.	.	985
6	010701	773	676	279	16281	.	2	$2,039.99	43659	.	.	985
7	010701	774	676	279	16281	.	1	$2,039.99	43659	.	.	985
8	010701	714	676	279	16281	.	3	$28.84	43659	.	.	985
9	010701	716	676	279	16281	.	1	$28.84	43659	.	.	985

Rysunek 3-19Tabela faktów

4 Procesy ETL

4.1 Wartość sprzedanych produktów w danym kraju z podziałem na pory roku

Firma Adventure Works Cycles zajmuje się sprzedażową międzynarodową produktów nastawionych głównie na wykorzystanie podczas ciepłych pór roku. Stworzony proces pozwoli sprawdzić czy wartość sprzedanych produktów spada zależnie od pory roku. Może wspomóc to proces produkcyjny i magazynowany ograniczając zużycie przestrzeni na towary, zmniejszając ryzyko nadmiaru zapasów. Dane zostaną dodatkowo podzielone na kraje, żeby sprawdzić czy istnieją między nimi jakieś różnice.



Rysunek 4-1 Proces analizy pierwszej.

W tym celu wymiar czasu, miejsca oraz tabela faktów zostały połączone węzłem *Join.* W ograniczeniu *Select* została utworzona nowa kolumna „Wartość_Sprzedazy" sumująca iloczyn kolumn: Ilosc_Produktu oraz Cena_Produktu.

#		Column	Column Description	Expression
1		Pora_roku		
2		Wartosc_Sprzedazy		sum(Fakt_sprzedaz_AW."Cena_Produktu"n *Fakt_sprzedaz_AW."Ilosc_Produktu"n)
3		Kraj	Name	

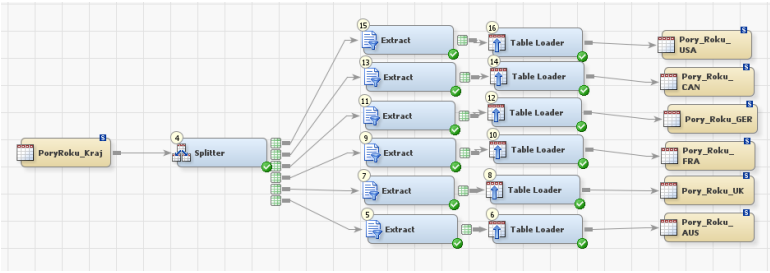
Rysunek 4-2 Kolumny po zastosowaniu select.

Następnie dane zostały pogrupowane stosując instrukcję *Group by* ze względu na kolumny: Pora_roku i Kraj. Tak jak widać na rysunku 5-1 pomiędzy węzłem *Join*, a *Table Loader* został dodany *Sort* odpowiedzialny za posortowanie danych rosnąco ze względu na kraj i porę roku. Uzyskane dane wynikowe zapisano do tabeli PoryRoku_Kraj.

#	Pora_roku	Wartosc_Sprzedazy	Kraj	#	Pora_roku	Wartosc_Sprzedazy	Kraj
1	Jesien	$5,427,477.64	United St...	13	Wiosna	$3,654,039.35	Canada ...
2	Jesien	$420,998.12	France ...	14	Wiosna	$766,907.74	Germany ...
3	Jesien	$178,789.00	Germany ...	15	Wiosna	$1,135,695.41	Australia ...
4	Jesien	$805,645.09	United Ki...	16	Wiosna	$1,276,737.50	France ...
5	Jesien	$1,498,539.92	Canada ...	17	Wiosna	$13,732,678.76	United St...
6	Jesien	$456,079.10	Australia ...	18	Wiosna	$1,148,624.92	United Ki...
7	Lato	$807,478.54	Germany ...	19	Zima	$826,388.87	France ...
8	Lato	$4,212,816.97	Canada ...	20	Zima	$8,013,988.22	United St...
9	Lato	$1,529,537.95	France ...	21	Zima	$595,186.04	United Ki...
10	Lato	$1,433,593.44	United Ki...	22	Zima	$610,703.68	Australia ...
11	Lato	$14,599,792.64	United St...	23	Zima	$1,751,195.52	Canada ...
12	Lato	$1,083,597.13	Australia ...	24	Zima	$385,773.65	Germany ...

Rysunek 4-3 Dane wynikowe analizy pierwszej.

Można zauważyć zróżnicowanie w zyskach, co sugeruje możliwość wpływu pory roku na te wartości. Wiosną oraz latem następuje znaczący wzrost wartości sprzedaży produktów we wszystkich krajach. Ze względu na te obserwacje, podjęto decyzję o przeprowadzeniu analizy mającej na celu dokładne określenie zakresu tego wzrostu. Postanowiłem skoncentrować się na identyfikacji procentowego udziału zysków, który przypada na poszczególne pory roku.



Rysunek 4-4 Schemat procesu do obliczenia odsetka.

Tabela Pory_Roku_Kraj została poddana kolejnym procesom. W pierwszym etapie, wykorzystując węzeł Splitter, dokonano podziału tabeli według kryterium kraju. Później z zastosowaniem *Extract* wymieniono kolumnę Wartosc_Sprzedazy na Odsetek, której wyrażenie znajduje się na rysunku 4-5.

Rysunek 4-5 Wyrażenie kolumny odsetek.

Dane zostały wczytane za pomocą *Table Loader* do odpowiadających im tabel.

#	Pora_roku	Odsetek	Kraj
1	Jesien	13.0%	United St...
2	Lato	34.9%	United St...
3	Wiosna	32.9%	United St...
4	Zima	19.2%	United St...

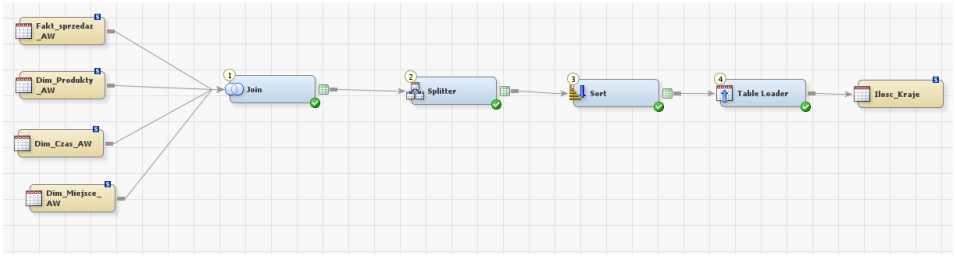
Rysunek 4-6 Wyniki tabeli Pory_Roku_USA

W dalszym etapie analizy danych ujawniło się, że aż około 70% całkowitej wartości sprzedaży w każdym z krajów jest generowanych w okresie wiosennym i letnim. W Niemczech jest to nawet 73,7%. Te wyniki podkreślają istotność sezonowych fluktuacji w kontekście osiąganych zysków, co może stanowić istotny element strategii zarządzania sprzedażą, zwłaszcza

w kontekście kształtowania działań marketingowych i dostosowania oferty do zmieniających się preferencji konsumentów w różnych porach roku.

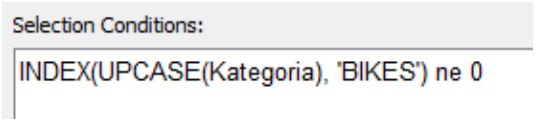
4.2 Ilość sprzedanych modeli rowerów do danego kraju

Drugi proces pozwoli zbadać jakie modele rowerów są najczęściej zamawiane zależnie od kraju. Wyniki mogą pozwolić firmie dostosować lokalną ofertę do specyfiki lokalnych rynków sprzedażową dla swoich klientów oraz zwiększyć atrakcyjność na międzynarodowym rynku rowerowym.



Rysunek 4-7 Proces analizy drugiej.

Tym razem węzłem *Join* trzeba połączyć trzy wymiary z tabela faktow – czasu, produktów i miejsca. Ograniczenie *Select* odpowiada za utworzenie kolumny o nazwie „Ilosc_sprzedanych". Kolumna ta sumuje Ilosc_Produktu z tabeli faktu. Zastosowano również funkcję *Group by* i pogrupowano dane wynikowe uwzględniając kategorię, nazwę produktu oraz kraj. Tak przygotowana tabela zawierała ilość sprzedanych wszystkich produktów. W celu zawężenia zakresu analizy i ograniczenia zbieranych danych do kategorii rowerów, wprowadzono węzeł *Splitter*.



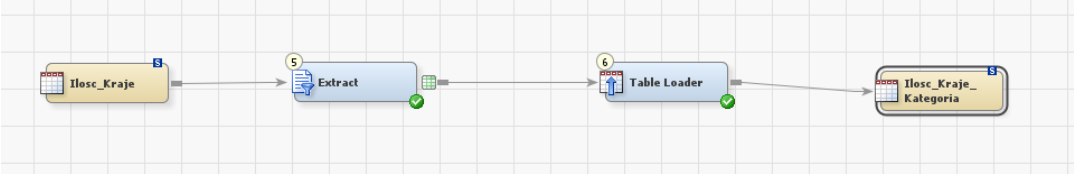
Rysunek 4-8 Węzeł Splitter w analizie drugiej.

Dane wynikowe posortowane ze względu na kolumny Ilosc_sprzedanych oraz Kraj zostały wczytane do tabeli Ilosc_Kraje. Pięć pierwszych wierszów danych wynikowe prezentuje się jak na rysunku 5-6.

#	Ilosc_sprzedanych	Kategoria	Nazwa_produktu	Kraj
1	4596	Mountain Bikes...	Mountain-200 Black, 3...	United St...
2	4572	Road Bikes ...	Road-650 Red, 44 ...	United St...
3	4392	Road Bikes ...	Road-650 Red, 60 ...	United St...
4	4380	Road Bikes ...	Road-650 Black, 52 ...	United St...
5	4068	Mountain Bikes...	Mountain-200 Black, 4...	United St...

Rysunek 4-9 Dane wynikowe.

Tabela, utworzona na podstawie analizy, może stanowić solidną podstawę do przeprowadzenia dalszych badań, takich jak porównania sprzedaży między poszczególnymi krajami, obliczanie średniej sprzedaży oraz identyfikacja ewentualnych trendów. Kolejnym krokiem w procesie analizy będzie szczegółowe zbadanie, jaka ilość przypada na poszczególne typy rowerów. Ten etap pozwoli na lepsze zrozumienie preferencji klientów i dostosowanie strategii sprzedażowej do specyfiki rynku, co może przyczynić się do dalszego zwiększenia efektywności działalności.



Rysunek 4-10 Schemat procesu podziału na kategorie.

Tabela Ilosc_Kraje wykorzystana do stworzenia kolejnego procesu. Usunięto z niej kolumnę kraj oraz nazwa_produktu. Do Ilosc_sprzedanych zostało dodane wyrażenie widoczne na rysunku 4-11.

#		Column	Column Description	Expression
1		Ilosc_sprzedanych		sum("Ilosc_sprzedanych"n)
2		Kategoria	Name	

Rysunek 4-11 Wyrażenie sumujące Ilosc_sprzedanych

Korzystając z węzła *Extract* dane zostały pogrupowane ze względu na kategorię oraz załadowane do tabeli Ilosc_Kraje_Kategoria.

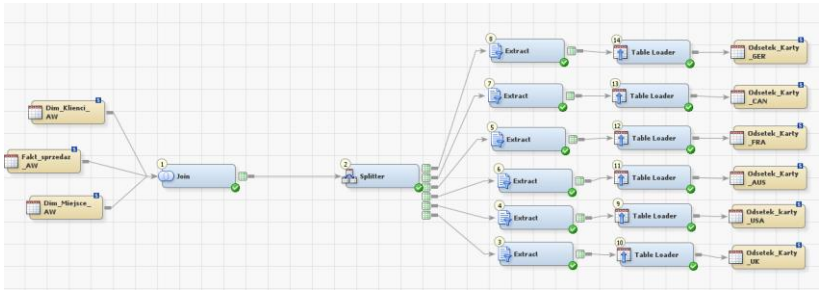
#	Ilosc_sprzedanych	Kategoria
1	73312	Mountain Bikes...
2	123228	Road Bikes ...
3	39644	Touring Bikes ...

Rysunek 4-12 Tabela wynikowa

Można zauważyć, że ponad połowa sprzedawanych w firmie Adventure Works Cycles rowerów stanowią rowery szosowe. Ten fakt sugeruje dominację tej kategorii produktów w ofercie firmy oraz potencjalne zainteresowanie klientów tym konkretnym typem rowerów. Takie spostrzeżenie może być istotne dla strategii marketingowej i planowania asortymentu, umożliwiając lepsze dostosowanie oferty do preferencji klientów oraz zoptymalizowanie działań sprzedażowych.

4.3 Udział procentowy kart kredytowych wykorzystywanych do dokonywania płatności z podziałem na kraje

Celem trzeciego etapu analizy jest zbadanie udziału kart płatniczych w dokonywaniu transakcji zakupu sprzętu rowerowego, uwzględniając zróżnicowanie preferencji płatniczych w różnych krajach. Przeprowadzenie tego procesu pozwoli na lepsze zrozumienie wyborów klientów z poszczególnych regionów.



Rysunek 4-13 Proces analizy trzeciej

Stosując węzeł *Join* udało się połączyć ze sobą trzy tabele – wymiaru klientów, wymiaru miejsca oraz faktu. W instrukcji *Select* kolumny: Kraj oraz Wydawca_karty zostały przeniesione bez zmian z wymiarów. Dalej utworzono nową kolumnę Ile, która odpowiadała za zliczenie ile kolumn zostało zgrupowanych funkcją *Group by* ze względu na wydawcę karty oraz kraj.

#			Column	Column Description	Expression
1			Kraj	Name	
2			Wydawca_karty	CardType	
3			Ile		count(*)

Rysunek 4-14 Funkcja agregująca count w instrukcji select. Analiza trzecia.

Następnie zastosowano węzeł *Splitter* celem podzielenia danych na 6 różnych tabel. Podział ten wynikał z kraju, z którego pochodzą dane karty kredytowe.

Target Tables:	Selection Conditions:
GER (W4PVYE4H)	upcase(Kraj) = "GERMANY"
CAN (W4PVYFU0)	
FRA (W4PWEGLV)	
AUS (W4PWEQ7N)	
USA (W4PWF3F6)	
UK (W4PWFDUT)	

Rysunek 4-15 Węzeł Splitter z przykładowym kodem.

Uzyskanie odsetka wymagało skorzystania z węzła *Extract,* w którym powstała nowa kolumna Odsetek. Do utworzenia jej zastosowano wyrażenie *Ile/sum(Ile)*. Dane wynikowe zostały wczytane do tabeli odpowiadającej krajowi jaki opisywały.

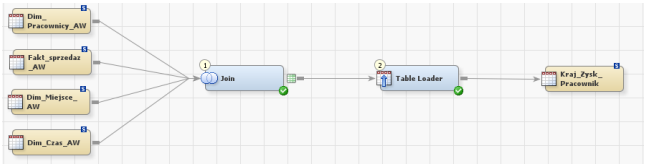
#	Odsetek	Wydawca_karty		Kraj
1	24.6%	ColonialVoice	...	United St...
2	25.9%	Distinguish	...	United St...
3	26.3%	SuperiorCard	...	United St...
4	23.2%	Vista	...	United St...

Rysunek 4-16 Dane wynikowe z tabeli analizy trzeciej dotyczące USA.

Na podstawie dostępnych danych można zauważyć, że w każdym kraju podział posiadaczy kart ze względu na wydawców jest równomierny. Sugeruje to brak dominującego trendu.

4.4 Wartość sprzedaży danego pracownika z podziałem na kraje zamówień

Czwarty proces zbada jak radzą sobie sprzedawcy firmy Adventure Works Cycles z podziałem na kraje. Badanie wartości sprzedaży danego pracownika, uwzględniające podział na kraje zamówień, jest kluczowe dla zrozumienia efektywności działań handlowych w różnych obszarach.



Rysunek 4-17 Proces analizy czwartej.

Zastosowano węzeł *Join* w celu połączenia czterech tabel: wymiaru pracowników, wymiaru miejsca, wymiaru czasu oraz tabeli faktu. W tym przypadku instrukcja *Select* została zastosowana w celu utworzenia pięciu kolumn w tabeli wynikowej: Imie, Nazwisko, Rok, Suma_sprzedazy i Kraj. Wszystkie poza sumą sprzedaży zostały zaczerpnięte z tablic wymiarów bez zmian.

Column	Column Description	Expression
Imie	FirstName	
Nazwisko	LastName	
Rok		
Suma_sprzedazy		sum(Fakt_sprzedaz_AW."Cena_Produktu"n * Fakt_sprzedaz_AW."Ilosc_Produktu"n)
Kraj	Name	

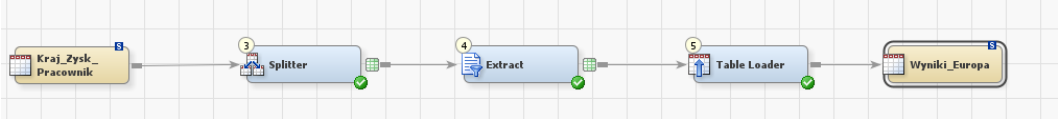
Rysunek 4-18 Kolumny stworzone w funkcji select.

Suma sprzedaży obliczana jest za pomocą wyrażenia na rysunku 5-12. Dane wynikowe grupowane są za pomocą *Group by* ze względu na rok, imię, nazwisko oraz kraj. Ostatni krok to wczytanie ich do tabeli Kraj_Zysk_Pracownik.

#	Imie	Nazwisko	Kraj	Rok	Suma_sprzedazy
1	Pamela ...	Ansman-Wolfe...	United St...	2001	$191,328.49
2	Michael ...	Blythe ...	United St...	2001	$304,755.60
3	David ...	Campbell ...	United St...	2001	$203,348.24
4	Jillian ...	Carson ...	United St...	2001	$455,655.32
5	Shu ...	Ito ...	United St...	2001	$377,432.05

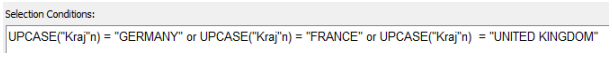
Rysunek 4-19 Pierwsze pięc wierszów z tabeli wynikowej.

Dane wynikowe pozwalają zidentyfikować pracownika operującego na terenie Europy, którego wartość sprzedaży jest najwyższa. Przeprowadzę proces, w którym sprawdzę który z pracowników przyczynia się najbardziej do generowania sprzedaży w tym regionie. Ten proces pozwoli na wskazanie kluczowych graczy w zespole oraz może dostarczyć informacji pomocnych w ocenie skuteczności działań sprzedażowych na obszarze europejskim.

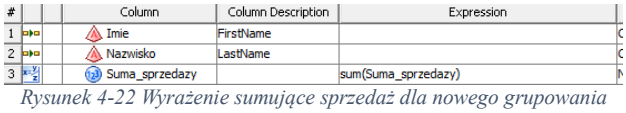


Rysunek 4-20 Schemat procesu sprawdzającego wyniki pracowników w Europie.

Tabela Kraj_Zysk_Pracownik posłużyła jako podstawa kolejnego procesu. W tym kontekście, wykorzystanie wyrażenia zawartego na rysunku 4-20 w węźle Splitter miało na celu ekstrakcję danych dotyczących wyłącznie krajów europejskich.



Rysunek 4-21 Wyrażenie odpowiedzialne za przekazanie danych tylko z krajów europejskich.



#		Column	Column Description	Expression
1		Imie	FirstName	
2		Nazwisko	LastName	
3		Suma_sprzedazy		sum(Suma_sprzedazy)

Rysunek 4-22 Wyrażenie sumujące sprzedaż dla nowego grupowania

W dalszym procesie, w module Extract, ograniczono analizę do kolumn zawierających imiona i nazwiska pracowników oraz sumy sprzedaży, przy użyciu wyrażenia widocznego na rysunku 4-21. Grupowanie nastąpiło ze względu na imię i nazwisko.

#	Imie	Nazwisko	Suma_sprzedazy
1	Amy ...	Alberts ...	$447,370.47
2	José ...	Saraiva ...	$2,939,772.62
3	Rachel ...	Valdez ...	$1,362,326.59
4	Ranjit ...	Varkey Chudu...	$3,413,027.02

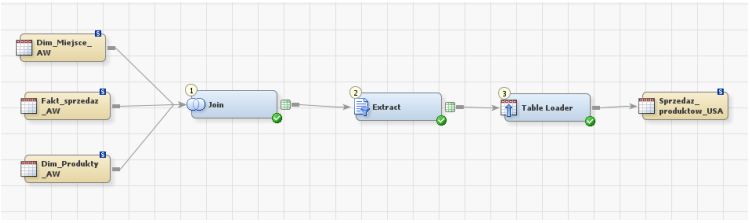
Rysunek 4-23 Tabela wynikowa zawierająca sumę sprzedaży pracowników w Europie

Dane zawarte w tabeli wynikowej ukazują, że najlepszy wynik finansowy w Europie został osiągnięty przez pracownika o imieniu Ranjit. Kolejni na podium są Jose oraz Rachel, którzy również uzyskali znaczące rezultaty finansowe. Natomiast na końcu listy znajduje się Amy Alberts z wynikiem finansowym, który jest ponad 7 razy mniejszy niż wynik Ranjita. To zjawisko mogłoby być przedmiotem dodatkowej analizy w celu zrozumienia przyczyn tak dużych różnic w wynikach finansowych między poszczególnymi pracownikami. Możliwe jest, że istnieją różne

czynniki wpływające na te wyniki, takie jak umiejętności sprzedażowe, obszar odpowiedzialności, relacje z klientami czy strategie działania.

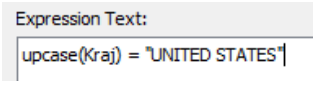
4.5 Najgorzej sprzedające się produkty w Stanach Zjednoczonych

Piąty proces poświęcony jest zbadaniu sprzedaży produktów w Stanach Zjednoczonych. Analiza najgorzej sprzedających się produktów w istotny etap w identyfikowaniu obszarów wymagających uwagi i optymalizacji.



Rysunek 4-24 Schemat procesu analizy piątej.

Żeby zrealizować ten proces należało połączyć dwa wymiary – miejsca oraz produktów z tabelą faktu. Następny za pomocą *Select* utworzono 4 kolumny: Nazwa_produktu, Kategoria, Ilosc_Produktu oraz Kraj. Zgrupowano je ze względu na nazwę produktu oraz kategorię.



Rysunek 4-25 Wyrażenie w Extract.

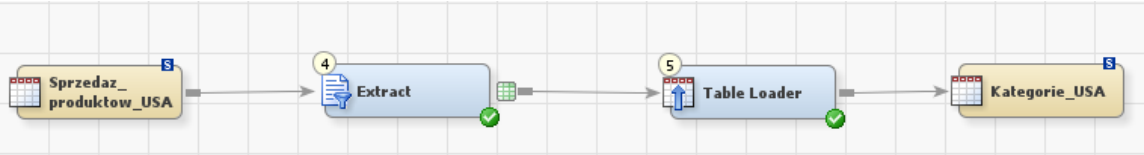
Następnie stosując węzeł *Extract* została wykorzystana funkcja *Where.* Wyrażenie użyte w tej instrukcji znajduje się na rysunku 5-15. Wyniki zostały załadowane do tabeli Sprzedaz_produktow_USA.

#	Nazwa_produktu	Kategoria	Ilosc_Produktu	Kraj
1	LL Road Seat/Saddle ...	Saddles ...	8	United St...
2	ML Mountain Frame-W...	Mountain Fram...	24	United St...
3	LL Mountain Frame - Bl...	Mountain Fram...	28	United St...
4	LL Touring Frame - Blu...	Touring Frame...	28	United St...
5	LL Mountain Frame - Bl...	Mountain Fram...	48	United St...
6	LL Touring Frame - Blu...	Touring Frame...	52	United St...
7	HL Mountain Frame - B...	Mountain Fram...	52	United St...
8	LL Touring Frame - Yell...	Touring Frame...	88	United St...
9	LL Mountain Frame - Si...	Mountain Fram...	92	United St...
10	LL Touring Handlebars...	Handlebars ...	96	United St...
11	HL Touring Frame - Yel...	Touring Frame...	144	United St...
12	ML Mountain Frame - ...	Mountain Fram...	156	United St...
13	ML Touring Seat/Saddl...	Saddles ...	156	United St...
14	ML Crankset ...	Cranksets ...	164	United St...
15	HL Touring Frame - Yel...	Touring Frame...	164	United St...

Rysunek 4-26 Tabela wynikowa procesu piątego

Analiza wykazała, że najgorzej sprzedający się przedmiot w Stanach Zjednoczonych to siodełko LL Road. Przedstawiona tabela wynikowa stanowi solidną bazę do dalszych analiz. W kolejnym etapie procesu planuję szczegółowo zbadać, która kategoria produktów radzi sobie

najgorzej. Ta analiza pomoże zidentyfikować obszary, które wymagają szczególnej uwagi i dostarczy informacji istotnych dla opracowania skutecznych strategii poprawy wyników sprzedaży w tych konkretnych kategoriach produktów.



Rysunek 4-27 Proces mający na celu ograniczenie wyników jedynie do kategorii

W kontynuacji analizy, tabela "Kraj_Zysk_Pracownik" została wykorzystana w kolejnym procesie. Przy użyciu węzła Extract dokonano grupowania danych według kraju i kategorii, jednocześnie usuwając kolumnę zawierającą nazwę produktu z tabeli wynikowej.

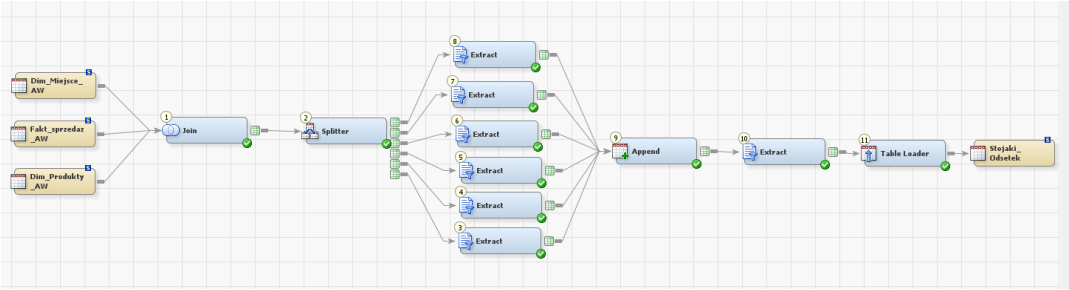
#	Kategoria	Ilosc_Produktu	Kraj
1	Bike Stands ...	340	United St...
2	Chains ...	1640	United St...
3	Forks ...	1660	United St...
4	Bottom Bracket...	1980	United St...
5	Brakes ...	2132	United St...

Rysunek 4-28 Tabela wynikowa procesu ograniczenia jedynie do kategorii.

Obserwując tabelę wynikową, jednoznacznie można stwierdzić, że najgorzej sprzedającym się produktem w Stanach Zjednoczonych są stojaki na rowery. Można przypuszczać, że istnieje pewien problem lub wyzwanie związane z tą kategorią produktów na rynku.

4.6 Odsetek jaki stojaki na rowery stanowią na tle całkowitej sprzedaży w danym kraju

Ostatni proces można nazwać kontynuacją pierwszego. Podjąłem decyzję o przeprowadzeniu analizy dotyczącej sprzedaży stojaków na rowery tym razem uwzględniając inne państwa. Wybór odsetka jako miary porównawczej wynikał z zróżnicowania krajów pod względem całkowitej wielkości sprzedaży.



Rysunek 4-29 Proces ostatniej analizy

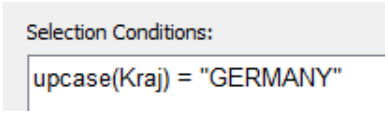
W procesie realizacji, konieczne było połączenie trzech zestawów danych: wymiaru miejsca, wymiaru produktów oraz tabeli faktów. Następnie, w ramach procesu *Select*, dokonano wyboru kolumn, które miały być przekazane do kolejnej fazy analizy. W tym przypadku wybrano

kolumny Kraj, Kategoria oraz Ilosc_Produktu. Z uwagi na planowane grupowanie danych według kraju i kategorii, konieczne było utworzenie kolumny "Ilosc_Produktu" za pomocą wyrażenia sumującego, co zostało zobrazowane na rysunku 4-29.

#			Column	Column Description	Expression
1			Kraj	Name	
2			Kategoria	Name	
3			Ilosc_Produktu	OrderQty	sum(Fakt_sprzedaz_AW."Ilosc_Produktu"n)

Rysunek 4-30 Kolumny stworzone w funkcji select.

Następnie zastosowano węzeł *Splitter*, którego zadaniem było podzielenie wyników funkcji *Join* na 6 oddzielnych tabel, uwzględniając podział ze względu na kraj. Warunek odpowiedzialny za to został zdefiniowany i przedstawiony na rysunku 4-30.



Rysunek 4-31 Przykładowy warunek w węźle Splitter

W drugim etapie analizy, skupiono się na obliczeniu procentowego udziału poszczególnych kategorii produktów w całkowitej sprzedaży dla danego kraju. Proces ten obejmował sześć odrębnych węzłów Extract, z których każdy utworzył kolumny: Kategoria, Kraj oraz Odsetek. Wyrażenie tworzące kolumnę odsetek zostało zdefiniowane zgodnie z rysunkiem 4-31.

#			Column	Column Description	Expression
1			Kategoria	Name	
2			Odsetek	OrderQty	Ilosc_Produktu/sum(Ilosc_Produktu)
3			Kraj	Name	

Rysunek 4-32 Struktura każdej tabeli.

Wyniki zostały zintegrowane do jednej tabeli dzięki zastosowaniu węzła Append, a następnie, przy użyciu operacji Extract, załadowano do tabeli wynikowej jedynie te wartości, które są powiązane z kategorią „Bike stands".

#	Kategoria	Odsetek	Kraj
1	Bike Stands ...	0.2%	Germany ...
2	Bike Stands ...	0.1%	Canada ...
3	Bike Stands ...	0.1%	France ...
4	Bike Stands ...	0.4%	Australia ...
5	Bike Stands ...	0.1%	United St...

Rysunek 4-33 Tabela wynikowa procesu 6

Można zaobserwować, że stojaki na rowery stanowią niewielki procent całkowitej sprzedaży we wszystkich krajach. Ta informacja sugeruje, że istnieje potrzeba optymalizacji i dostosowania procesu produkcji tych elementów. Warto zauważyć, że największy odsetek sprzedaży stojaków na rowery notowany jest w Australii. W związku z tym, rozważenie

zoptymalizowania produkcji stojaków na rowery w kontekście rynku australijskiego może przyczynić się do bardziej efektywnego wykorzystania zasobów.

5 Raport uzyskanych wyników

5.1 Proces 1

Wyniki analizy wyraźnie wskazują, że sprzedaż osiąga najwyższe wartości wiosną i latem, natomiast obroty maleją zimą i jesienią. Ten trend jest zauważalny we wszystkich krajach.

Dodatkowa analiza procentowego udziału zysków w poszczególnych porach roku wskazuje, że aż około 70% całkowitej wartości sprzedaży generowane jest w okresie wiosennym i letnim. W Niemczech ten odsetek sięga nawet 73,7%.

Głównym powodem takiego stanu rzeczy najprawdopodobniej jest fakt, że ciepłe pory roku sprzyjają zakupom produktów związanych z aktywnością na świeżym powietrzu. W okresie, gdy temperatura wzrasta, ludzie częściej angażują się w różnorodne formy rekreacji. Skutkuje to rosnącym zapotrzebowaniem na towary dedykowane tej sferze.

Te wyniki sugerują, że strategie zarządzania sprzedażą powinny uwzględniać zmiany sezonowe, szczególnie w kontekście dostosowania działań marketingowych i oferty do preferencji konsumentów w różnych porach roku.

5.2 Proces 2

Analiza wykazała, że ponad połowa sprzedawanych rowerów to rowery szosowe, co wskazuje na dominującą pozycję tej kategorii w ofercie firmy. To spostrzeżenie ma kluczowe znaczenie dla strategii marketingowej, umożliwiając lepsze dostosowanie oferty do preferencji klientów oraz efektywniejsze prowadzenie działań sprzedażowych.

Niniejsza obserwacja stanowi podstawę do dalszych analiz, obejmujących konkretne modele, rozmiary i kolory rowerów, co może być istotne w kontekście podejmowania decyzji dotyczących produkcji. Wyniki te mogą z kolei stanowić fundament dla ewentualnych zmian produkcyjnych, skierowanych na zwiększenie zyskowności, poprzez dostosowanie oferty do preferencji rynkowych.

5.3 Proces 3

Wyniki analizy dotyczące USA zostały wczytane do odpowiedniej tabeli. Na podstawie dostępnych danych zauważono, że w każdym kraju podział posiadaczy kart ze względu na wydawców jest równomierny, co sugeruje brak dominującego trendu.

5.4 Proces 4

Wyniki analizy wskazały, że Ranjit osiągnął najwyższe wyniki finansowe, natomiast Amy Alberts uzyskała wynik znacznie niższy. Różnice te stanowią istotny punkt wyjścia do przeprowadzenia dalszej analizy, która pozwoli zidentyfikować czynniki wpływające na efektywność sprzedawców. Dodatkowe badania w tych obszarach mogą dostarczyć bardziej szczegółowych informacji, które będą pomocne w opracowaniu strategii zarządczych, mających na celu optymalizację wyników sprzedażowych oraz równomierne rozwijanie potencjału zespołu sprzedażowego.

Identyfikacja mocnych stron Ranjita, które przyczyniły się do osiągnięcia wysokich wyników, oraz słabych stron Amy, które mogły wpłynąć na niższe rezultaty, może pomóc w opracowaniu planów rozwoju dla sprzedawców w firmie Adventure Works Cycles.

5.5 Proces 5

W tabeli wynikowej jednoznacznie wskazano, że stojaki na rowery są najgorzej sprzedającą się kategorią produktów w Stanach Zjednoczonych. To stanowi punkt wyjścia do dalszych działań mających na celu poprawę wyników sprzedaży w tej kategorii.

Aby zrozumieć powody, które doprowadziły do tej sytuacji, wskazane jest przeprowadzenie analizy czynników wpływających na niską sprzedaż stojaków na rowery. Poprzez głębsze zrozumienie tych elementów możliwe będzie sformułowanie skutecznych strategii poprawy wyników sprzedażowych.

5.6 Proces 6

Zauważono, że stojaki na rowery stanowią niewielki procent ogólnej sprzedaży we wszystkich krajach, przy czym najwyższy odsetek sprzedaży tego produktu zanotowano w Australii.

Wysoki na tle innych krajów odsetek sprzedaży w Australii może skłaniać do dalszych badań dotyczących specyficznych czynników, które wpływają na atrakcyjność stojaków na rowery

w tym regionie. Ponadto, identyfikacja różnic w sprzedaży stojaków na rowery między krajami podkreśla potrzebę dostosowania oferty do różnych rynków.