

ANALIZA SPRZEDAŻY W FIRMIE ADVENTURE WORKS

PAWEŁ LECHOWCZ
UNIwersytet Ekonomiczny

Spis treści

Analiza wybranego narzędzia do czyszczenia danych dla potrzeb budowy hurtowni danych ..	2
1.Cel realizacji projektu	9
2.Opis źródeł danych.....	10
3.Model logiczny hurtowni danych w oparciu o model gwiazdy	14
3.1 Czas_dim.....	14
3.2 Klienci_dim.....	15
3.3 Miejsce_dim	16
3.4 Pracownicy_dim.....	17
3.5 Produkty_dim	18
3.6 Sprzedaż_fact	19
4.Opis procesów ETL.....	21
4.1 Analiza sprzedaży produktów w zależności od kategorii oraz podkategorii.	21
4.2 Analiza najlepiej/najgorzej sprzedających się produktów w Wielkiej Brytanii.....	22
4.3 Analiza pracownika z najmniejszą liczbą sprzedanych produktów	25
4.4 Analiza średniej wartości zamówienia dla klienta	27
4.5 Analiza dochodów ze sprzedaży w zależności od roku oraz miesiąca.	28
5. Podsumowanie	30

Analiza wybranego narzędzia do czyszczenia danych dla potrzeb budowy hurtowni danych

Zacznijmy od tego czym jest hurtownia danych. Hurtownia danych (ang. data warehouse) to specjalnie zaprojektowana i skonfigurowana baza danych, która służy do gromadzenia, przechowywania i analizy dużych ilości danych z różnych źródeł w celu wspierania procesów podejmowania decyzji w przedsiębiorstwie. Gromadzone dane pochodzą z różnych systemów operacyjnych, transakcyjnych i innych źródeł, a następnie są zintegrowane, przetworzone i przechowywane w hurtowni danych w taki sposób, aby były łatwo dostępne i gotowe do analizy.

Typowa hurtownia danych obejmuje cztery główne elementy:

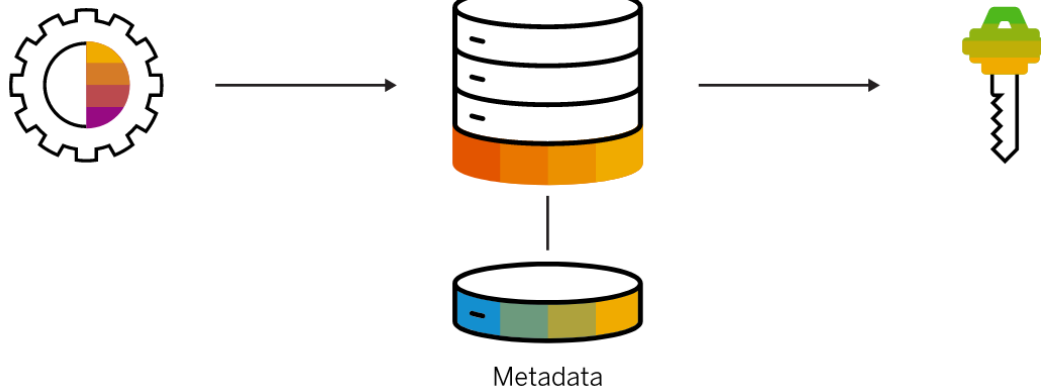
1. **Centralna baza danych:** Jest to specjalnie zaprojektowana baza danych, która służy do gromadzenia, przechowywania i zarządzania dużymi ilościami danych z różnych źródeł. Centralna baza danych umożliwia łatwy dostęp i szybką analizę danych.
2. **Narzędzia ETL (Ekstrakcja, Transformacja, Ładowanie):** Te narzędzia są używane do ekstrakcji danych z różnych źródeł, ich transformacji w odpowiedni format i ładowania do centralnej bazy danych hurtowni danych. Proces ETL pozwala na zintegrowanie danych z różnych systemów.
3. **Metadane:** Metadane to informacje opisujące strukturę i znaczenie danych przechowywanych w hurtowni. Pomagają one zrozumieć, skonfigurować i zarządzać danymi. Metadane mogą zawierać informacje o pochodzeniu danych, ich strukturze, zależnościach i innych istotnych aspektach.
4. **Narzędzia dostępowe:** Narzędzia dostępowe umożliwiają użytkownikom korzystanie z danych przechowywanych w hurtowni. Są to interfejsy graficzne lub zapytania języka SQL, które pozwalają na przeglądanie, analizę i raportowanie danych. Te narzędzia są zoptymalizowane pod kątem szybkiego uzyskiwania wyników i natychmiastowej analizy danych.

Wszystkie te elementy są starannie zaprojektowane i skonfigurowane w hurtowni danych w celu umożliwienia efektywnego gromadzenia, zarządzania i analizy danych, co wspiera procesy podejmowania decyzji w przedsiębiorstwie.

Extract, transform, and load (ETL)

Central database

Access tools



Schemat przedstawiający elementy hurtowni danych

Hurtownie danych mogą przyjmować różne modele w zależności od potrzeb i wymagań organizacji. Oto kilka popularnych modeli hurtowni danych:

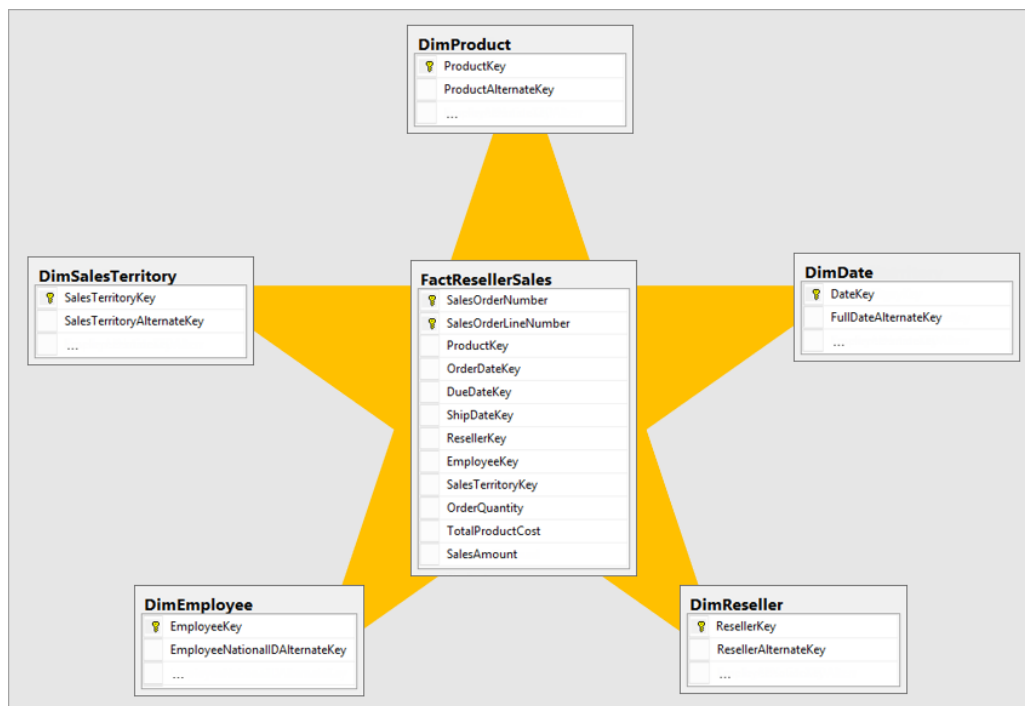
1. **Model Gwiazdy (Star Schema):** Jest to jedno z najczęściej stosowanych podejść. W tym modelu dane są zorganizowane wokół centralnej tabeli faktów, która zawiera liczby i wskaźniki, oraz tabeli wymiarów, które zawierają opisy wymiarów związanych z danymi liczbowymi. Ten model jest prosty, łatwy do zrozumienia i efektywny w analizie danych.
2. **Model Siatki (Snowflake Schema):** To rozwinięcie modelu gwiazdy, w którym tabelki wymiarów są normalizowane, co oznacza, że są podzielone na podtabelki. Choć może to pomóc w oszczędności miejsca na dysku, to jednak może skomplikować zapytania. Niemniej jednak, ten model jest przydatny w przypadku dużych hurtowni danych.
3. **Model Konstelacji (Constellation Schema):** Ten model obejmuje multiple gwiazdy, co oznacza, że istnieje więcej niż jedna tabela faktów, połączona za pomocą wspólnych tabel wymiarów. Jest stosowany, gdy organizacja potrzebuje analizować dane z różnych obszarów biznesowych.
4. **Model Koniunkcyjny (Starflake Schema):** To połączenie modelu gwiazdy i siatki. Tabele wymiarów są częściowo znormalizowane, co oznacza, że niektóre z nich mogą zawierać znormalizowane dane, a inne nie. Ten model łączy prostotę modelu gwiazdy z oszczędnością miejsca modelu siatki.
5. **Model Sztynny (Inmon Data Warehouse Model):** Ten model skupia się na centralnej hurtowni danych, gdzie dane są przechowywane jednokrotnie i mogą być używane do wielu celów. W tym modelu metadane odgrywają kluczową rolę, co ułatwia zarządzanie danymi i zrozumienie ich struktury.

Te modele oferują różne podejścia do projektowania hurtowni danych, a wybór zależy od konkretnych wymagań i celów organizacji. Każdy z nich ma swoje zalety i ograniczenia, a decyzja o wyborze danego modelu zależy od specyfiki danej implementacji hurtowni danych. Na potrzeby naszego projektu będziemy używali model gwiazdy.

Model gwiazdy (ang. Star Schema) to jeden z najpopularniejszych modeli używanych w hurtowniach danych. Charakteryzuje się prostotą, łatwością zrozumienia i efektywnością analizy danych. Głównym elementem tego modelu są dwie rodzaje tabel: tabela faktów (fact table) i tabele wymiarów (dimension tables):

- **Tabela Faktów (Fact Table):** Centralnym elementem modelu gwiazdy jest tabela faktów. Zawiera ona liczby, wskaźniki oraz dane numeryczne, które stanowią podstawę analizy. Tabela faktów jest zwykle duża i zawiera dane zebrane z różnych źródeł. Przechowuje informacje o zdarzeniach, transakcjach lub działaniach, które są istotne dla biznesu.
- **Tabele Wymiarów (Dimension Tables):** Wokół tabeli faktów rozmieszczone są tabele wymiarów, które zawierają opisy wymiarów związanych z danymi liczbowymi. Wymiary to cechy, według których użytkownik chciałby analizować dane. Na przykład, w hurtowni danych dla sprzedaży, wymiarami mogą być data, klient, produkt czy lokalizacja. Tabele wymiarów zawierają opisy tych wymiarów, co ułatwia zrozumienie kontekstu danych numerycznych.

Model gwiazdy jest nazywany tak ze względu na swoje graficzne podobieństwo do gwiazdy, gdzie tabela faktów jest punktem centralnym (centrum gwiazdy), a tabele wymiarów są rozgałęzieniami (promieniami gwiazdy).



Przykładowy model gwiazdy

Gdy już wiemy czym jest Hurtownia Danych jakie ma elementy oraz jaki model HD chcemy stworzyć możemy przejść do etapu czyszczenia danych.

Czyszczenie danych (ang. data cleansing lub data cleaning) to istotny etap w budowie hurtowni danych. Oczyszczone dane są kluczowe dla poprawności i dokładności analizy oraz raportowania w hurtowni danych. Narzędzia do czyszczenia danych pomagają usuwać duplikaty, korygować błędy, standaryzować formaty i eliminować inne nieprawidłowości w danych. Istnieje wiele narzędzi, które wspierają ten proces, a wybór zależy od konkretnych potrzeb organizacji.

Jednak przed przystąpieniem do czyszczenia danych w kontekście budowy hurtowni danych, istnieje kilka kluczowych kroków przygotowawczych, które warto rozważyć. Te kroki pomagają zrozumieć strukturę danych, zidentyfikować potencjalne problemy i dostosować strategię czyszczenia. Oto kilka ważnych kroków poprzedzających proces czyszczenia danych:

- Pierwszym istotnym etapem jest zrozumienie źródeł danych. Niezbędne jest poznanie, skąd pochodzą dane, jak są gromadzone oraz w jakiej formie są przechowywane. To zrozumienie kontekstu danych stanowi fundament skutecznego oczyszczania informacji.
- Następnie, analiza struktury danych staje się kluczowym krokiem. Przegląd struktury obejmuje identyfikację kluczowych kolumn, relacji między nimi oraz potencjalnych duplikatów. Posiadanie pełnego obrazu struktury danych ułatwia późniejsze planowanie procesu czyszczenia.

- Ustalanie celów biznesowych jest równie ważne. Określenie, do jakich konkretnych celów dane zostaną wykorzystane, pozwala dostosować proces czyszczenia do specyficznych potrzeb organizacji.
- Analiza brakujących danych stanowi kolejny kluczowy etap. Zidentyfikowanie czy braki wynikają z błędów, czy też dane są po prostu niedostępne, pozwala na odpowiednie zarządzanie brakującymi danymi, co jest istotne dla utrzymania spójności.
- Przeprowadzenie wstępnej analizy jakości danych obejmuje identyfikację potencjalnych problemów, takich jak błędy ortograficzne, niejednorodności w formatach czy nieprawidłowe wartości. Jest to istotny etap w eliminowaniu nieprawidłowości.
- W przypadku danych pochodzących z różnych źródeł, proces mapowania i standaryzacji staje się kluczowym elementem. Jednolita struktura danych ułatwia późniejsze etapy czyszczenia.
- Przed przystąpieniem do czyszczenia zawsze warto utworzyć kopię zapasową danych. To środek bezpieczeństwa, który chroni przed przypadkową utratą danych podczas procesu oczyszczania.
- Ostatecznym etapem jest ustalanie zasad czyszczenia. Określenie reguł, takich jak eliminacja duplikatów, standaryzacja formatów czy usuwanie błędów, gwarantuje spójność i dokładność danych. Ten proces jest kluczowy dla efektywnego oczyszczania danych przed ich analizą czy wykorzystaniem w budowie hurtowni danych.

W momencie, gdy już przeszliśmy przez wszystkie kroki poprzedzające czyszczenie danych musimy zdecydować się na jakieś konkretne narzędzie do czyszczenia ich których na rynku jest naprawdę wiele. Na potrzeby mojej pracy skoncentrowałem się na jednym z nich a dokładniej na OpenRefine.

OpenRefine, wcześniej znane jako Google Refine, to otwartoźródłowe narzędzie do czyszczenia, przekształcania i przeglądania danych tabularnych. Jest przeznaczone do ułatwienia procesu przygotowywania danych przed analizą lub załadowaniem ich do systemów bazodanowych, w tym hurtowni danych. OpenRefine umożliwia interaktywne przeglądanie, edycję i transformację dużych zbiorów danych.



LOGO PROGRAMU

Zalety OpenRefine:

1) Intuicyjny Interfejs Graficzny:

- OpenRefine oferuje łatwy w użyciu interfejs graficzny, co sprawia, że jest przyjazne dla użytkownika, nawet dla tych bez zaawansowanego doświadczenia w programowaniu.

2) Wszechstronność w Przetwarzaniu Danych:

- Narzędzie obsługuje różne formaty danych, takie jak CSV, TSV, Excel, JSON, co czyni go wszechstronnym w przetwarzaniu różnych typów danych.

3) Wizualizacje Danych:

- OpenRefine dostarcza interaktywnych wizualizacji danych, co ułatwia zrozumienie struktury danych i identyfikację potencjalnych problemów.

4) Możliwość Cofania Zmian:

- Narzędzie automatycznie śledzi historię transformacji, co pozwala na łatwe cofanie i przywracanie zmian oraz analizę, jak dane były przekształcane.

5) Czyszczenie Danych w Trybie Masowym:

- OpenRefine pozwala na stosowanie operacji masowych do czyszczenia danych, co jest efektywne w przypadku dużych zbiorów danych.

6) Dostępność Rozszerzeń i Pluginów:

- Istnieje wiele rozszerzeń i pluginów dostępnych w społeczności OpenRefine, co umożliwia dostosowanie funkcjonalności narzędzia do konkretnych potrzeb użytkowników.

7) Open Source:

- OpenRefine jest projektem open-source, co oznacza, że jest dostępne dla społeczności programistycznej i może być dostosowywane do indywidualnych potrzeb.

Wady OpenRefine:

1) Ograniczenia w Przypadku Dużych Danych:

- W przypadku bardzo dużych zbiorów danych OpenRefine może stać się mniej wydajne, co może wpływać na płynność pracy.

2) Brak Zaawansowanych Analiz Statystycznych:

- OpenRefine skupia się głównie na procesie czyszczenia i transformacji danych, co oznacza, że może brakować zaawansowanych funkcji analizy statystycznej dostępnych w bardziej wyspecjalizowanych narzędziach.

3) Wymaga Świadomości Użytkownika:

- Choć narzędzie jest intuicyjne, pewne operacje mogą wymagać od użytkownika pewnej wiedzy na temat struktury danych i potrzebnych transformacji.

4) Ograniczone Zasoby Pomocy:

- Społeczność OpenRefine jest stosunkowo mniejsza w porównaniu do niektórych innych narzędzi, co może oznaczać ograniczoną dostępność zasobów pomocy i wsparcia.

5) Brak Automatycznego Zarządzania Dużymi Procesami ETL:

- OpenRefine nie jest specjalnie zaprojektowane do obsługi skomplikowanych procesów ETL (Ekstrakcji, Transformacji, Ładowania), co może być istotne w przypadku bardzo rozbudowanych hurtowni danych.

Podsumowując czyszczenie danych to kluczowy etap w przygotowywaniu danych do analizy, obejmujący usuwanie błędów, standaryzację formatów, zarządzanie brakującymi danymi oraz utrzymanie spójności i jakości danych, a narzędzie takie jak OpenRefine może się do tego idealnie nadawać.

1.Cel realizacji projektu

Adventure Works Cycles to fikcyjna, duża, międzynarodowa firma produkcyjna, która produkuje i dystrybuje rowery metalowe i kompozytowe na rynki komercyjne w Ameryce Północnej, Europie i Azji. Siedziba Adventure Works Cycles mieści się w Bothell w stanie Waszyngton, gdzie firma zatrudnia 500 pracowników. Ponadto firma Adventure Works Cycles zatrudnia kilka regionalnych zespołów sprzedaży na całym swoim rynku.



Celem głównym projektu jest opracowanie kompleksowego modelu hurtowni danych dla firmy Adventure Works przy wykorzystaniu zaawansowanego narzędzia SAS Data Integration Studio. Proces ten ma na celu dokładną analizę struktur sprzedażowych, umożliwiając skoncentrowanie się na kluczowych obszarach działalności firmy. Model oparty na schemacie gwiazdy pozwoli na szczegółową ekstrakcję, transformację i ładowanie danych.

Analiza sprzedażowa obejmie różnorodne aspekty, od regionów sprzedaży, poprzez trendy czasowe, wydajność produktów, aż po ocenę działalności pracowników i preferencje klientów. Opracowanie takiej struktury analitycznej pozwoli na głębsze zrozumienie danych oraz lepsze dostosowanie strategii biznesowej firmy Adventure Works.

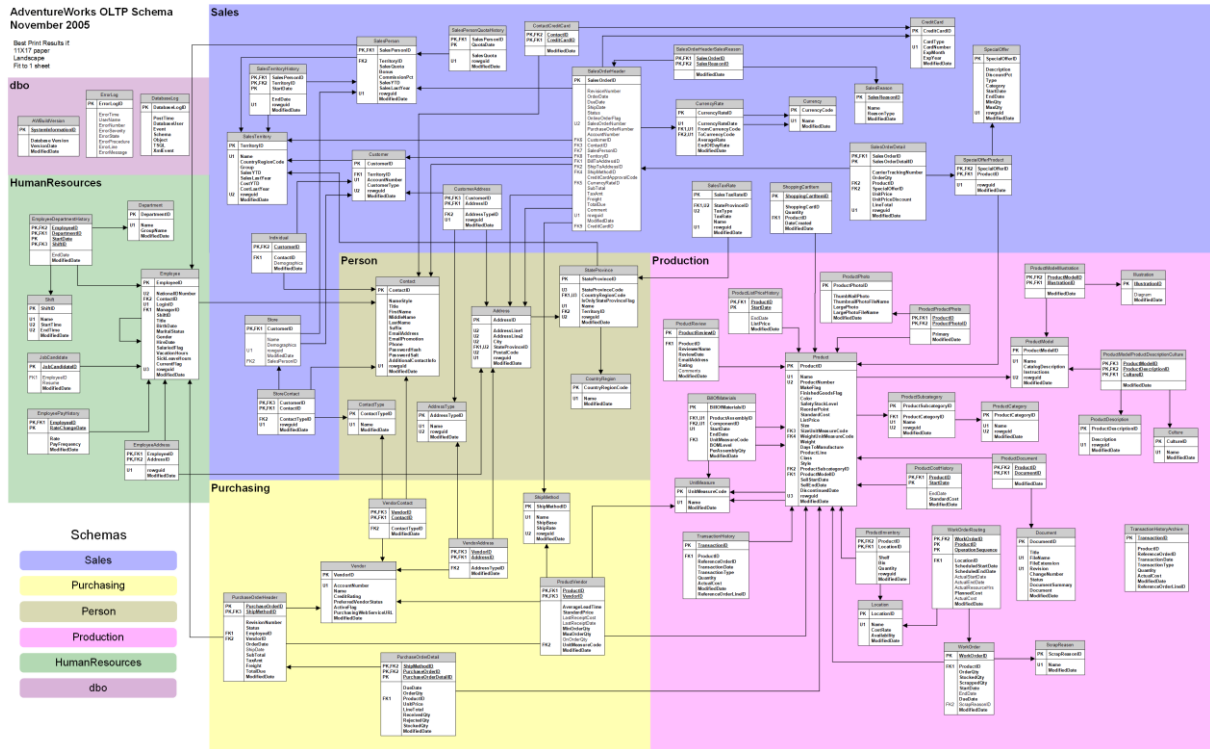
W ramach projektu skoncentruję się na pięciu kluczowych analizach, a każda z nich dostarczy wartościowych informacji na temat efektywności sprzedaży:

- **Analiza dochodów ze sprzedaży produktów w zależności od kategorii.**
- **Analiza najlepiej sprzedających się produktów w UK.**
- **Analiza pracownika z najmniejszą liczbą sprzedanych produktów**
- **Analiza średniej wartości zamówienia dla klienta**
- **Analiza dochodów ze sprzedaży w zależności od roku oraz miesiąca**

2.Opis źródeł danych

W ramach projektu analizy danych dla firmy Adventure Works, kluczowym krokiem jest zrozumienie struktury oraz natury danych, które będą podstawą dla procesów hurtowni danych. Źródłem danych dla naszego projektu jest baza AdventureWorks, która stanowi przykładową bazę danych dla systemów Microsoft SQL Server od wersji 2008 do 2014.

Diagram bazy prezentuje się następująco:



Jak widzimy sama baza danych podzielona jest na 5 głównych obszarów

- Sales
- Purchasing
- Person
- Production
- Human Resources

OBSZAR SALES

W bazie danych bardzo istotnym obszarem jest obszar SALES, gdzie też występują dwie tabele, które posiadają znaczną część kluczy zarówno obcych jak i prywatnych, mowa tutaj o tabeli SALES_SALESORDERHEADER oraz SALES_SALESORDERDETAIL.

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

SALESORDERDETAIL

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

SALESORDERHEADER

Skorzystałem z nich podczas tworzenia jednej z tabel faktów co zostało opisane poniżej.

Z tego obszaru użyłem również tabel takich jak: INDIVIDUAL, SALEPERSON oraz SALESTERRORITY.

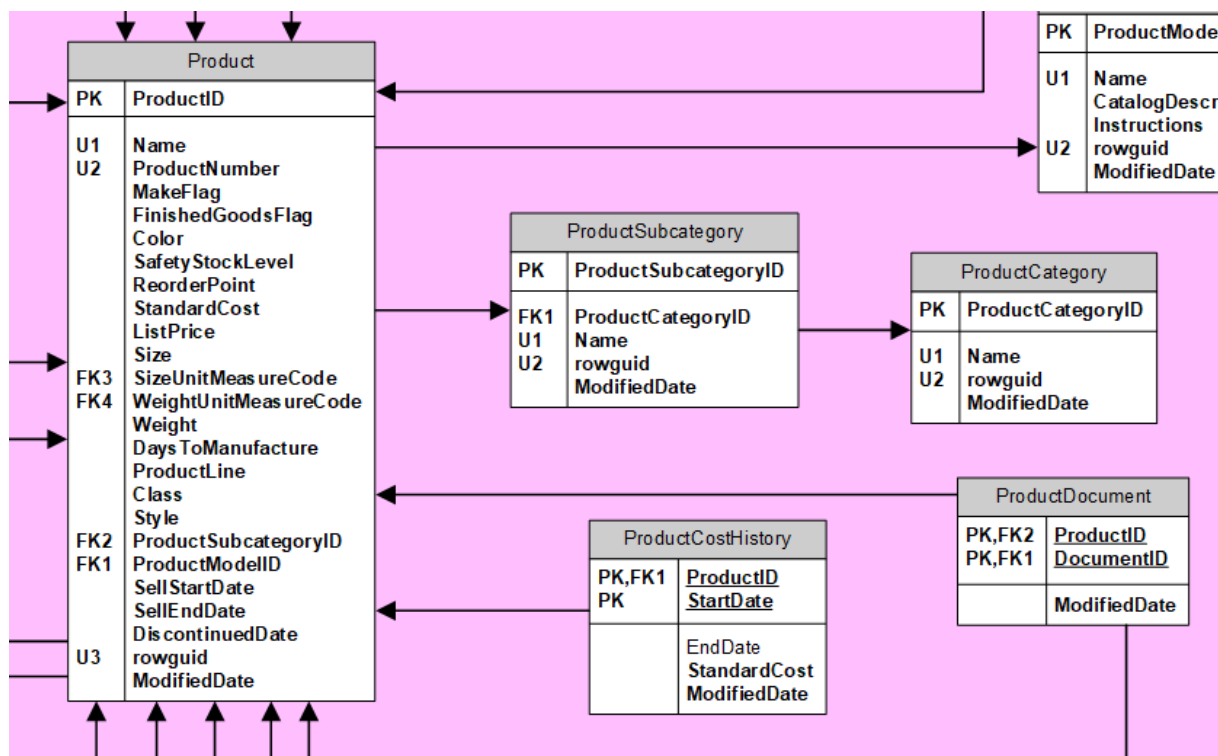
OBSZAR PERSON

Obszar ten zawiera jak sama nazwa wskazuje dane osobowe, w tym Imię, Nazwisko, etc. W mojej pracy użyłem dwukrotnie tabeli PERSON_CONTACT gdzie właśnie uzyskałem dane kontaktowe.

#	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	AdditionalCo...	AdditionalContactInfo
14	rowguid	rowguid
15	ModifiedDate	ModifiedDate

OBSZAR PRODUCTION

W bazie danych Adventure Works, obszar "Production" odnosi się do danych związanych z produkcją i zarządzaniem produktami. Obszar ten był mi potrzebny przy tworzeniu wymiaru produktów. Możemy w nim znaleźć takie dane jak PRODUCTID, kategorie produktów i wiele innych cennych informacji związanych z produkcją.



PRODUCT, PRODUCTSUBCATEGORY oraz PRODUCTCATEGORY to tabele, które użyłem w późniejszej części projektu

OBSZAR HUMAN RESOURCES

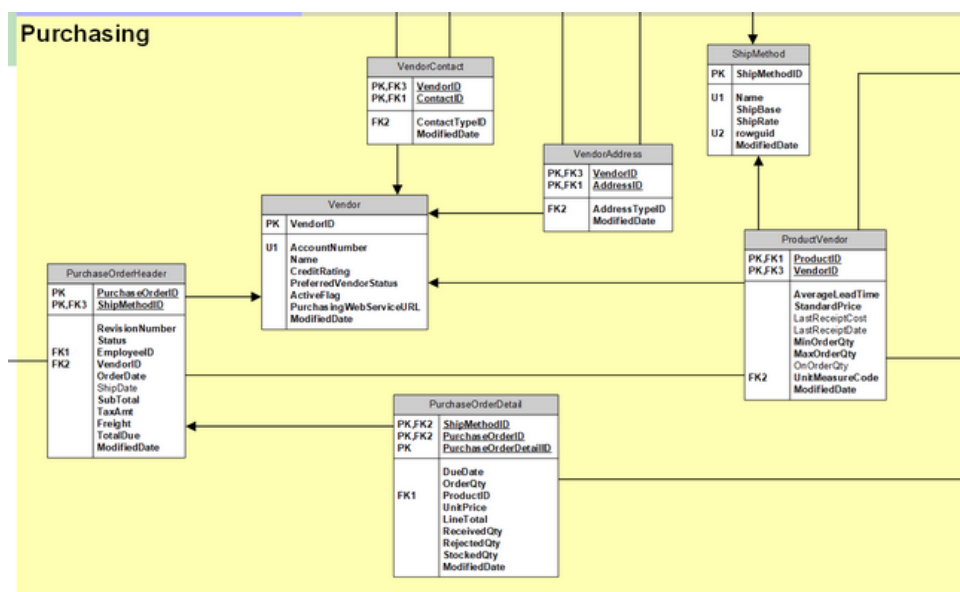
W bazie danych Adventure Works, obszar "Human Resources" obejmuje dane związane z zasobami ludzkimi, czyli informacje dotyczące pracowników, zatrudnienia i zarządzania personelem. Został użyty podczas tworzenia wymiaru z pracownikami. Dokładniej użyłem tabeli HUMANRESOURCES_EMPLOYEE:

#	Name	Description	Type	Length	Informat	Format	Is Nullable
1	EmployeeID	EmployeeID	Numeric	8	11.	11.	Yes
2	NationalIDNu...	NationalIDNumber	Character	1024	\$1024.	\$1024.	Yes
3	ContactID	ContactID	Numeric	8	11.	11.	Yes
4	LoginID	LoginID	Character	1024	\$1024.	\$1024.	Yes
5	ManagerID	ManagerID	Numeric	8	11.	11.	Yes
6	Title	Title	Character	1024	\$1024.	\$1024.	Yes
7	BirthDate	BirthDate	Numeric	8	DATETIME19.	DATETIME19.	Yes
8	MaritalStatus	MaritalStatus	Character	1	\$1.	\$1.	Yes
9	Gender	Gender	Character	1	\$1.	\$1.	Yes
10	HireDate	HireDate	Numeric	8	DATETIME19.	DATETIME19.	Yes
11	SalariedFlag	SalariedFlag	Numeric	8	6.	6.	Yes
12	VacationHours	VacationHours	Numeric	8	6.	6.	Yes
13	SickLeaveHours	SickLeaveHours	Numeric	8	6.	6.	Yes
14	CurrentFlag	CurrentFlag	Numeric	8	6.	6.	Yes
15	rowguid	rowguid	Character	38	\$38.	\$38.	Yes
16	ModifiedDate	ModifiedDate	Numeric	8	DATETIME19.	DATETIME19.	Yes

HUMANRESOURCES_EMPLOYEE

OBSZAR PURCHASING

Jest to obszar związany z procesem zakupu i zarządzaniem dostawami, jednakże nie był potrzebny w moim projekcie.



3. Model logiczny hurtowni danych w oparciu o model gwiazdy

Współczesne organizacje gromadzą ogromne ilości danych w celu lepszego zrozumienia swojej działalności, podejmowania trafnych decyzji biznesowych oraz identyfikowania trendów rynkowych. Wraz z rozwojem technologii informatycznych i rosnącą ilością danych, konieczne staje się efektywne zarządzanie nimi. Jednym z efektywnych podejść jest zastosowanie Hurtowni Danych opartej na modelu gwiazdy.

Model Gwiazdy w kontekście hurtowni danych odnosi się do architektury, w której istnieje centralna tabela faktów, otoczona tabelami wymiarowymi. Tabela faktów zawiera kluczowe informacje biznesowe, natomiast tabele wymiarowe zawierają szczegółowe atrybuty opisujące wymiary analizowane w kontekście biznesowym.

3.1 Czas_dim











Pierwszym tworzonym przeze mnie wymiarem będzie wymiar czasu, który pozwoli nam identyfikować dane względem czasu. Aby utworzyć ten wymiar będziemy potrzebowali tabelę z obszaru SALES a dokładniej tabeli SALES_SALESORDERHEADER. Tabela wymiaru będzie posiadała IDCzas, Dzień, Miesiąc, Rok, Kwartał. IDCzas stworzymy za pomocą kodu:

```
compress(substr(put(YEAR(DATEPART("DataZamówienia "n)),4.),3,2) ||'0' ||
```

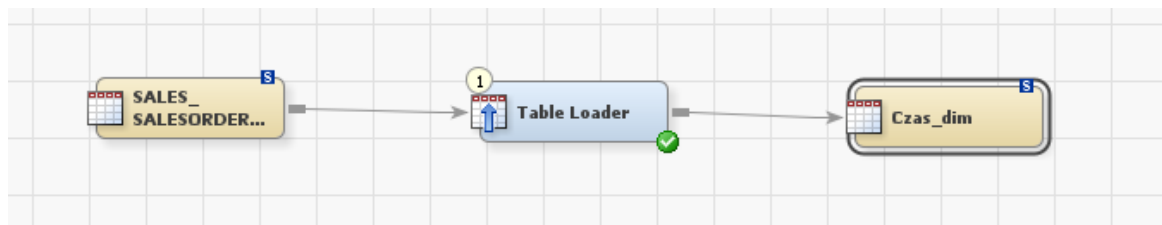
```
put(MONTH(DATEPART(Zamowienia."DataZamówienia"n
```

```
)),2.) ||'0' ||put(DAY(DATEPART(Zamowienia."DataZamówienia"n)),2.))
```

Jest to jedyna kolumna o charakterze tekstowym, ponieważ reszta kolumn jest liczbowa a tworzymy je z pomocą funkcji czasowych: DAY(), MONTH(), YEAR(), QTR()

Target table: Czas_dim (Czas_dim)					
#		Column	Column Description	Expression	Type
1		 IDCzas		compress(substr(put(year(datepart(SALES_SALE...	Character
2		 Dzień		day(datepart(SALES_SALESORDERHEADER."Ord...	Numeric
3		 Miesiąc		month(datepart(SALES_SALESORDERHEADER."...	Numeric
4		 Rok		year(datepart(SALES_SALESORDERHEADER."Or...	Numeric
5		 Kwartał		qtr(datepart(SALES_SALESORDERHEADER."Ord...	Numeric

Proces ETL wygląda następujący sposób:



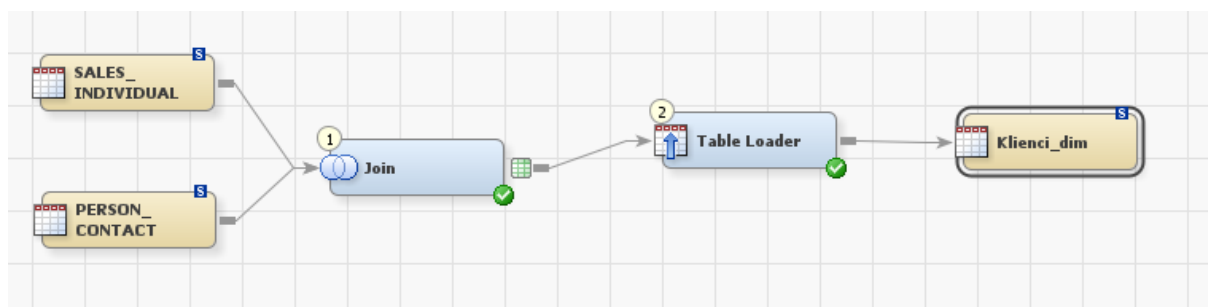
Rysunek 1-tworzenie Czas_dim

Gdy w węźle Table Loader uzupełnimy już nasze kolumny tworzymy tabelę Czas_dim:

#	IDCzas	Dzień	Miesiąc	Rok	Kwartał
1	010701	1	7	2001	3
2	010701	1	7	2001	3
3	010701	1	7	2001	3
4	010701	1	7	2001	3
5	010701	1	7	2001	3
6	010701	1	7	2001	3
7	010701	1	7	2001	3
8	010701	1	7	2001	3
9	010701	1	7	2001	3
10	010701	1	7	2001	3
11	010701	1	7	2001	3
12	010701	1	7	2001	3
13	010701	1	7	2001	3
14	010701	1	7	2001	3
15	010701	1	7	2001	3

3.2 Klienci_dim

Drugim tworzonym wymiarem będzie wymiar odnośnie do klientów. Sam wymiar nie będzie skomplikowany i będzie powstawał z dwóch tabel SALES_INDIVIDUAL oraz PERSON_CONTACT.

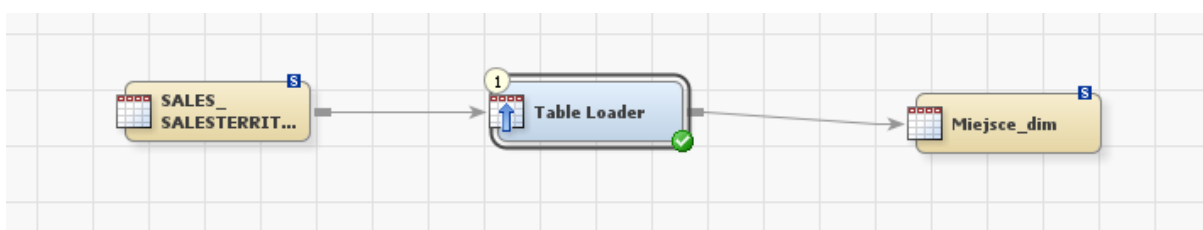


Dzięki temu wymiarowi otrzymujemy informacje na temat IDKlienta, Imienia oraz Nazwiska

#	IDKlienta	Imie	Nazwisko
1	27520	Francis	Navarro
2	21564	Erica	Liu
3	25678	Erica	Yang
4	11220	Erica	Huang
5	25452	Erica	Wu
6	16033	Erica	Lin
7	19062	Erica	Zhou
8	13010	Erica	Ye
9	24963	Erica	Zhao
10	27009	Erica	Lu
11	25245	Francis	Gutierrez
12	20863	Erica	Xu
13	26135	Erica	Sun
14	25221	Erica	Zhu
15	19087	Erica	Gao

3.3 Miejsce_dim

Rozpoczynamy proces tworzenia wymiaru miejsca, który umożliwi nam skategoryzowanie danych ze względu na ich lokalizację. Jednym z pierwszych tworzonych wymiarów będzie wymiar czasu, służący identyfikacji danych z perspektywy czasowej. W tym celu skoncentrujemy się na tabeli SALES_TERRORITY z obszaru SALES.



Rysunek 2- tworzenie wymiar Miejsce_dim

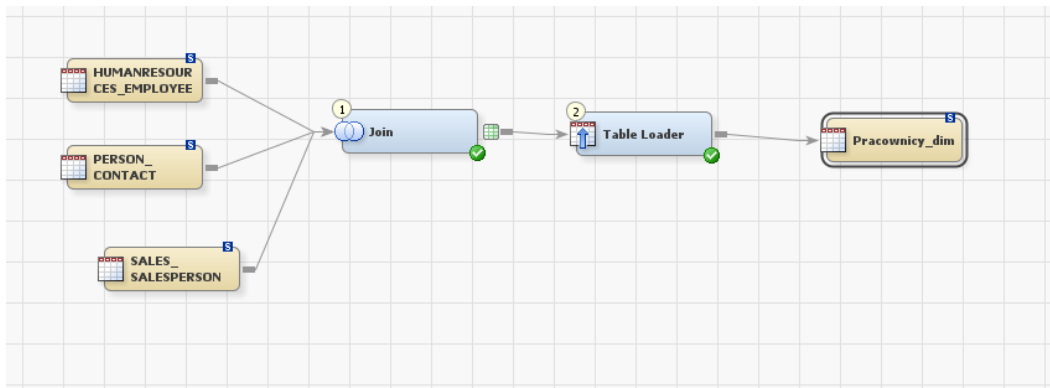
Sama tabela jest dosyć prosta i wygląda następująco:

#	TerritoryID	Nazwa
1	1	Northwest
2	2	Northeast
3	3	Central
4	4	Southwest
5	5	Southeast
6	6	Canada
7	7	France
8	8	Germany
9	9	Australia
10	10	United King...

3.4 Pracownicy_dim

Kolejnym etapem naszego projektu jest tworzenie wymiaru pracowników, który umożliwi nam efektywną kategoryzację danych związanych z personelem firmy. W tym przypadku skorzystamy z dwóch tabel: HUMANRESOURCES_EMPLOYEE oraz PERSON_CONTACT.

Wymiar pracowników będzie zawierał klucz IDPracownika oraz istotne informacje takie jak Imię, Nazwisko, Płeć, Tytuł oraz Stanowisko. Klucz IDPracownika będzie jednoznacznie identyfikować każdego pracownika w kontekście hurtowni danych.



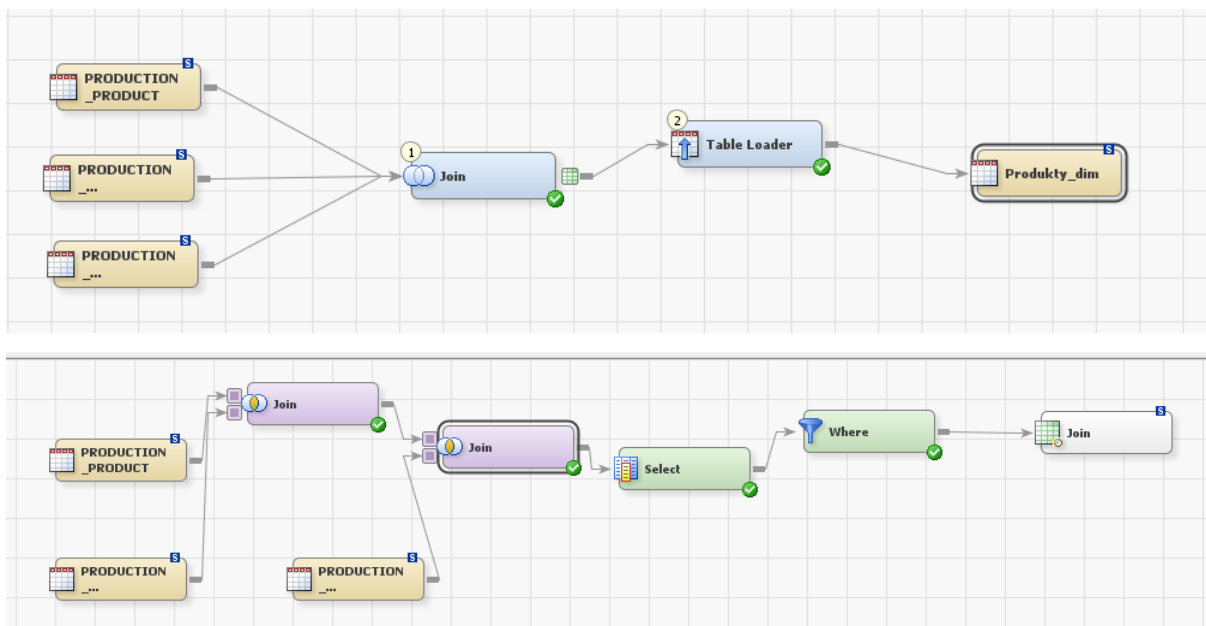
Sama tabela wymiaru pracownicy prezentuje się w taki sposób:

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

3.5 Produkty_dim

W kolejnym etapie naszego projektu skoncentrujemy się na tworzeniu wymiaru produktów, który umożliwi skategoryzowanie danych związanych z produkcją. W tym przypadku wykorzystamy trzy tabele z obszaru PRODUCTION: PRODUCTION_CATEGORY, PRODUCTION_SUBCATEGORY oraz PRODUCTION_PRODUCT.

Wymiar produktów będzie zawierał klucz ProductID, Nazwę Produktu, Nazwę Kategorii oraz Nazwę Podkategorii. Do uzyskania kompleksowych informacji o produktach skorzystamy z węzła JOIN, łącząc dane z różnych tabel w celu utworzenia pełnego obrazu dotyczącego kategorii i podkategorii produktów.



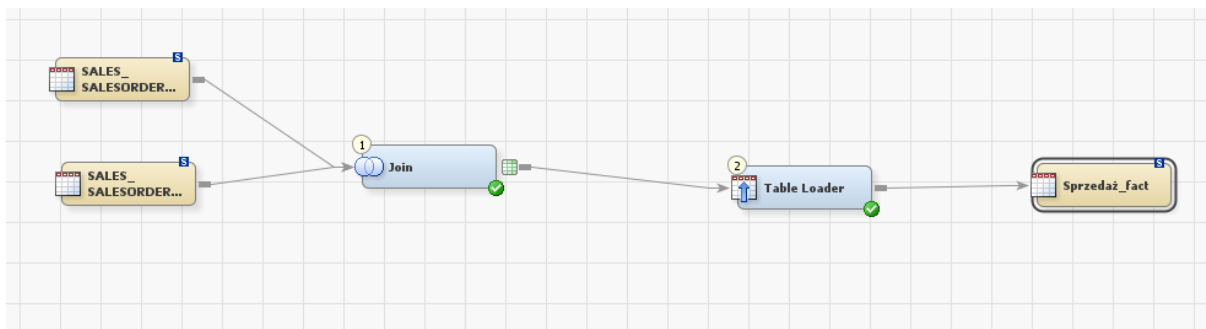
Rysunek 3- węzeł join tworzący wymiar Produkty_dim

Sama tabela wygląda następująco:

#	ProductID	Nazwa produktu	Nazwa_kategorii	Nazwa_podkategorii
1	680	HL Road Frame - Blac...	Road Fra	Componen
2	706	HL Road Frame - Red...	Road Fra	Componen
3	707	Sport-100 Helmet, Re...	Helmets	Accessor
4	708	Sport-100 Helmet, Bla...	Helmets	Accessor
5	709	Mountain Bike Socks, ...	Socks	Clothing
6	710	Mountain Bike Socks, ...	Socks	Clothing
7	711	Sport-100 Helmet, Blu...	Helmets	Accessor
8	712	AWC Logo Cap ...	Caps	Clothing
9	713	Long-Sleeve Logo Jer...	Jerseys	Clothing
10	714	Long-Sleeve Logo Jer...	Jerseys	Clothing
11	715	Long-Sleeve Logo Jer...	Jerseys	Clothing
12	716	Long-Sleeve Logo Jer...	Jerseys	Clothing
13	717	HL Road Frame - Red...	Road Fra	Componen
14	718	HL Road Frame - Red...	Road Fra	Componen
15	719	HL Road Frame - Red...	Road Fra	Componen
16	720	HL Road Frame - Red...	Road Fra	Componen
17	721	HL Road Frame - Red...	Road Fra	Componen
18	722	LL Road Frame - Black...	Road Fra	Componen
19	723	LL Road Frame - Black...	Road Fra	Componen
20	724	LL Road Frame - Black...	Road Fra	Componen
21	725	LL Road Frame - Red,...	Road Fra	Componen
22	726	LL Road Frame - Red,...	Road Fra	Componen
23	727	LL Road Frame - Red,...	Road Fra	Componen
24	728	LL Road Frame - Red,...	Road Fra	Componen
25	729	LL Road Frame - Red,...	Road Fra	Componen
26	730	LL Road Frame - Red...	Road Fra	Componen

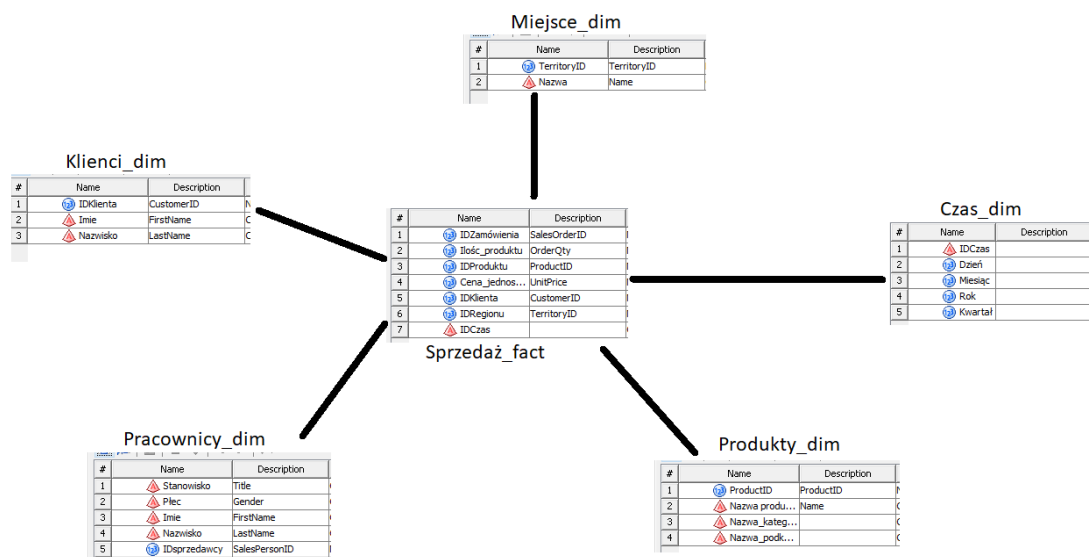
3.6 Sprzedaż_fact

Ostatnią tabelą będzie tabela faktów która zbiera wszystkie potrzebne ID w sobie. Aby ją utworzyć potrzebujemy głównie dwóch tabel: SALES_SALESORDERHEADER oraz SALES_SALESORDERDETAIL. Oprócz ID z wcześniej utworzonych wymiarów interesują nas również kolumny z ilością oraz cena jednostkową produktów.



#	IDZamówienia	Ilość_produktu	IDProduktu	Cena_jednostkowa	IDKlienta	IDRegionu	IDCzas	ID_pracownika
1	43659	1	776	\$2,024.99	676	5	010701	279
2	43659	4	711	\$20.19	676	5	010701	279
3	43659	2	712	\$5.19	676	5	010701	279
4	43659	6	709	\$5.70	676	5	010701	279
5	43659	1	716	\$28.84	676	5	010701	279
6	43659	3	714	\$28.84	676	5	010701	279
7	43659	1	774	\$2,039.99	676	5	010701	279
8	43659	2	773	\$2,039.99	676	5	010701	279
9	43659	1	772	\$2,039.99	676	5	010701	279
10	43659	1	771	\$2,039.99	676	5	010701	279
11	43659	1	778	\$2,024.99	676	5	010701	279
12	43659	3	777	\$2,024.99	676	5	010701	279
13	43660	1	762	\$419.46	117	5	010701	279
14	43660	1	758	\$874.79	117	5	010701	279
15	43661	1	745	\$809.76	442	6	010701	282
16	43661	5	708	\$20.19	442	6	010701	282

Gdy już posiadamy wszystkie potrzebne wymiary oraz tabele faktów możemy stworzyć model gwiazdy naszej hurtowni danych:



Rysunek 4- schemat gwiazdy

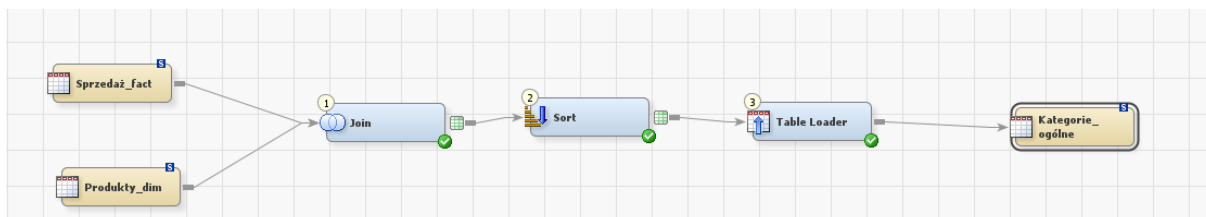
Z gotowym schematem gwiazdy możemy przystąpić do poszczególnych analiz.

4.Opis procesów ETL

4.1 Analiza sprzedaży produktów w zależności od kategorii oraz podkategorii.

Firma Adventure Works posiada różne kategorie produktów, analiza ta pozwoli na stwierdzenie, które z kategorii są najbardziej a które najmniej dochodowe.

Najpierw zaczniemy od stworzenia procesu ETL gdzie skorzystamy z Produkty_dim oraz Sprzedaż_Fact.

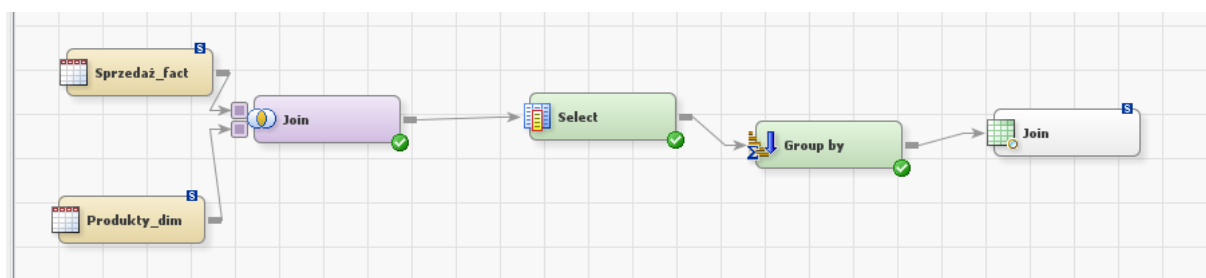


Używamy na starcie węzła join gdzie też w funkcji select wybieramy następujące kolumny.

Target table: JOIN (WORKSOP)						
#		Column	Column Description	Expression	Type	Len
1		Nazwa_kategorii			Character	
2		Wartosc_produktow		sum(Sprzedaż_fact."Cena_jednostkowa"*Sprz...	Numeric	
3		Nazwa_podkategorii			Character	

Tworzymy także własną kolumny Wartosc_produktów, która będzie sumować łączny dochód.

Następnie w węźle Join dodajemy Group by gdzie grupujemy nasze produkty ze względu na kategorie oraz podkategorie.



Kolejnym krokiem jest posortowanie danych malejąco, żeby uzyskać tabele wynikową:

Sort Properties

General | **Sort By Columns** | Mappings | Options | Table Options | Code | Precode and Postcode | Parameters | Notes | Extended Attributes

Available columns:

- Nazwa_kategorii
- Wartosc_produkow
- Nazwa_podkategorii

Sort by columns:

#	Column name	Sort order
1	Wartosc_produkow	Descending

Na koniec naszego procesu ETL używam Table Loader aby załadować pliki do tabeli Kategorie_ogólne:

#	Nazwa_kategorii	Wartosc_produkow	Nazwa_podkategorii
1	Bikes	43978443.249	Road Bik
2	Bikes	36622296.451	Mountain
3	Bikes	14545073.652	Touring
4	Componen	4714725.5206	Mountain
5	Componen	3851978.7188	Road Fra
6	Componen	1644934.0495	Touring
7	Clothing	756665.1574	Jerseys
8	Componen	681426.7804	Wheels
9	Accessor	487663.5202	Helmets
10	Clothing	420414.9764	Shorts
11	Clothing	263544.05	Vests
12	Clothing	246962.4066	Gloves
13	Accessor	246454.8464	Tires an
14	Accessor	239437.2	Bike Rac
15	Clothing	204375.4963	Tights
16	Componen	204064.7632	Crankset
17	Componen	170657.1483	Handleba
18	Clothing	168003.2309	Bib-Shor
19	Componen	147530.884	Pedals
20	Accessor	106329.7638	Hydratio
21	Componen	77968.9588	Forks
22	Componen	70262.56	Deraille
23	Componen	66061.95	Brakes
24	Accessor	64353.5989	Bottles
25	Componen	61135.8845	Headsets
26	Componen	55848.7396	Saddles
27	Componen	51826.374	Bottom B
28	Clothing	51512.2811	Caps
29	Accessor	46619.58	Fenders
30	Accessor	39591	Bike Sta
31	Clothing	30029.4258	Socks
32	Accessor	18518.571	Cleaners
33	Accessor	16264	Locks
34	Accessor	13528.8322	Pumps
35	Componen	9385.6928	Chains

Tabela-kategorie_ogólne

Wnioskiem wyciągniętym z naszej tabeli jest fakt, że podkategoria Road Bik była najbardziej dochodowa a tym samym kategoria, w której się znajdowała, czyli Bikes. Natomiast podkategorią, która przyniosła najmniej pieniędzy była podkategoria Chains czyli Łańcuchy, które stanowiły mniejszy człon kategorii Components.

4.2 Analiza najlepiej/najgorzej sprzedających się produktów w Wielkiej Brytanii.

Kraje oraz regiony są zróżnicowane ze względu na klientów, dlatego też ważne jest, aby wiedzieć do jakiej grupy odbiorców chcemy trafiać w danym regionie. Poniższa analiza będzie miała na celu sprawdzenie jakie produkty sprzedają się najlepiej w Wielkiej Brytanii.

Zaczynamy od zidentyfikowania tabel których będziemy używać a będą to tabele Produkty_dim, Miejsce_dim oraz Sprzedaż_fact. Następnie łączymy je za pomocą węzła Join gdzie też wybieramy interesujące nas kolumny:

#	Column	Column Description	Expression	Type	Length	Info
1	Ilość sprzedanych	OrderQty	sum(Sprzedaż_fact. Ilość produktu * n)	Numeric	8 6.	
2	Nazwa regionu	Name		Character	1024 \$1024	
3	Nazwa produktu	Name		Character	1024 \$1024	

Jedną z kolumn modyfikujemy i sumujemy ilość sprzedanych produktów.

Kolejnym krokiem jest pogrupowanie danych ze względu na nazwę regionu oraz nazwę produktu. Aby uzyskać dane dotyczące tylko Wielkiej Brytanii stosujemy węzeł Extract:

Extract Properties

General Where Group By Order By Mapping

Tekst wyrażenia:

Nazwa_regionu = "United Kingdom"

+ - * / **

Warunek w węźle Extract

Gdy już mamy gotowy węzeł Extract łączymy go z węzłem Sort- pozwoli nam to na posortowanie danych w naszym przypadku będzie to malejąco.

Sort Properties

General Sort By Columns Mappings Options Table Options Code Precode and Postcode Parameters Notes Extended Attributes

Available columns:

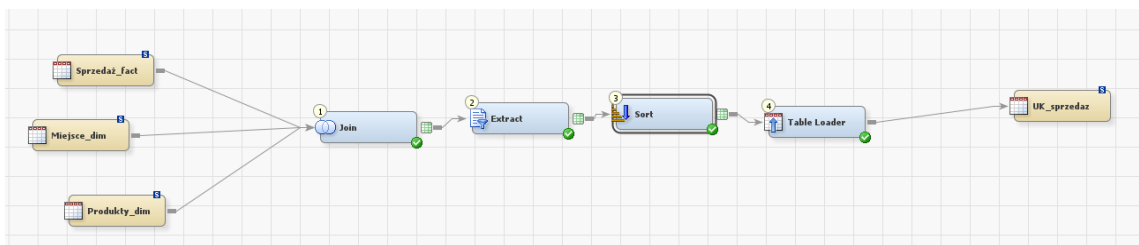
- Ilość produktu
- Nazwa regionu
- Nazwa produktu

Sort by columns:

#	Column name	Sort order
1	Ilość produktu	Descending

Sortowanie w węźle Sort

Ostatnim węzłem jest węzeł Table Loader za pomocą którego przeładujemy dane do tabeli wynikowej. Cały proces wygląda następująco:



Gdy już mamy gotowy proces możemy otworzyć naszą tabelę wynikową:

#	Ilość_produktu	Nazwa_regionu	Nazwa produktu
1	358	United Kingdom	AWC Logo Cap
2	330	United Kingdom	Long-Sleeve Logo Jersey, L
3	311	United Kingdom	Classic Vest, S
4	309	United Kingdom	Sport-100 Helmet, Black
5	307	United Kingdom	Short-Sleeve Classic Jersey, XL
6	283	United Kingdom	Sport-100 Helmet, Blue
7	260	United Kingdom	Half-Finger Gloves, M
8	256	United Kingdom	Sport-100 Helmet, Red
9	242	United Kingdom	Women's Mountain Shorts, S
10	239	United Kingdom	Hitch Rack - 4-Bike
11	222	United Kingdom	Short-Sleeve Classic Jersey, L
12	218	United Kingdom	Classic Vest, M
13	214	United Kingdom	Bike Wash - Dissolver
14	195	United Kingdom	Hydration Pack - 70 oz.
15	192	United Kingdom	Full-Finger Gloves, L
16	189	United Kingdom	Long-Sleeve Logo Jersey, M
17	189	United Kingdom	Water Bottle - 30 oz.
18	174	United Kingdom	Women's Mountain Shorts, L
19	165	United Kingdom	Long-Sleeve Logo Jersey, XL
20	162	United Kingdom	Full-Finger Gloves, M

Wnioskujemy z niej, że najlepiej sprzedającym się produktem było AWC Logo kup które sprzedały się w ilości 358. Analizując dalej jesteśmy w stanie stwierdzić, że na drugim miejscu sprzedał się produkt w ilości 330 i był to Long_Sleeve Logo Jersey o rozmiarze L. Trzecie miejsce w kategorii najczęściej sprzedających się przedmiotów zyskał przedmiot o nazwie Classic Vest.

Teraz aby zobaczyć produkty najgorzej się sprzedające na terenie Wielkiej Brytanii zmienimy wartość na Ascending(Rosnąco) w węźle Sort.

#	Ilość_produktu	Nazwa_regionu	Nazwa produktu
1	1	United Kingdom	LL Mountain Frame - Black, 40
2	1	United Kingdom	LL Mountain Frame - Black, 52
3	1	United Kingdom	LL Touring Frame - Blue, 58
4	1	United Kingdom	LL Touring Frame - Blue, 62
5	1	United Kingdom	LL Touring Frame - Yellow, 58
6	1	United Kingdom	ML Mountain Frame-W - Silver, 38
7	2	United Kingdom	LL Road Frame - Red, 52

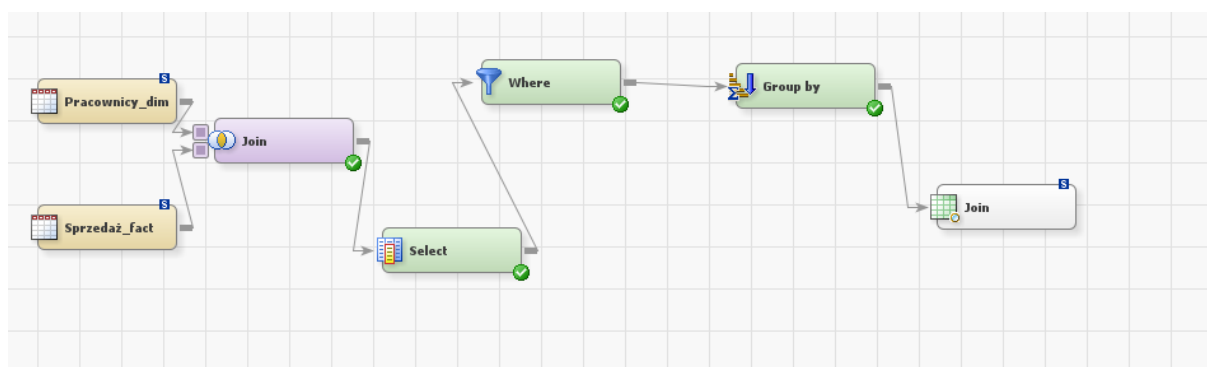
Tabela wynikowa posortowano rosnąco

Jesteśmy w stanie zauważyć, że przedmiotów, które sprzedały się jednorazowo było sześć, A przedmiotem, który widnieje na samej górze jest LL Mountain Frame-Black,40

4.3 Analiza pracownika z najmniejszą liczbą sprzedanych produktów

Efektywna analiza wydajności pracowników jest kluczowym elementem skutecznego zarządzania sprzedażą w każdej firmie. Monitorowanie indywidualnych wyników sprzedażowych pracowników pozwala zidentyfikować obszary do poprawy, dostosować strategię szkoleniową oraz maksymalizować ogólny sukces zespołu. W niniejszej analizie skupimy się na pracowniku, który osiągnął najmniejszą liczbę sprzedanych produktów.

Ponownie pierwszym korkiem jest zidentyfikowanie tabel których będziemy używali podczas tego procesu, a będą to Pracownicy_dim oraz Sprzedaż_fact. Łączymy nasze tabele za pomocą węzła Join gdzie używamy kolejno węzłów Select, Where oraz Group by:

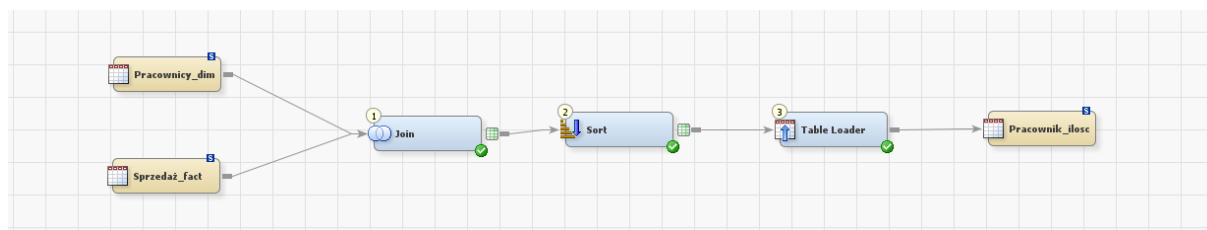


Węzeł Join

W węźle Select wybieramy następujące kolumny: Płeć, Imię, Nazwisko:

#		Column	Column Description	Expression	Type	Length
1		Płeć	Gender		Character	1
2		Imię	FirstName		Character	1024
3		Nazwisko	LastName		Character	1024
4		Ilość produktów	OrderQty	SUM(Sprzedaż_fact."Ilość produktu")	Numeric	8

Tworzymy także własną kolumnę: Ilość produktów w której sumujemy ilość sprzedanych produktów. Kolejnym krokiem będzie posortowanie danych malejąco za pomocą węzła Sort. Sort jest połączony z węzłem Table Loader który przeładowuje dane do tabeli wynikowej



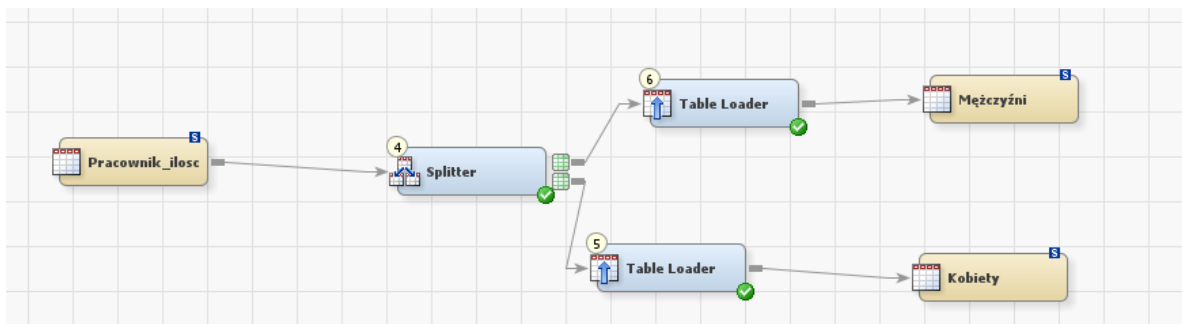
Proces uzyskiwania Pracownicy_ilosc

Dzięki temu procesowi uzyskujemy tabele:

#	Płec	Imie	Nazwisko	Ilość_produktu
1	M	Syed ...	Abbas ...	825
2	F	Amy ...	Alberts ...	2012
3	M	Stephen ...	Jiang ...	3095
4	F	Lynn ...	Tsoflias ...	4123
5	M	Tete ...	Mensa-Annan ...	5650
6	F	Rachel ...	Valdez ...	7033
7	F	Pamela ...	Ansman-Wolfe...	7360
8	M	David ...	Campbell ...	8172
9	M	Garrett ...	Vargas ...	11544
10	M	Ranjit ...	Varkey Chudu...	14085
11	M	José ...	Saraiva ...	15220
12	M	Shu ...	Ito ...	15397
13	M	Tsvi ...	Reiter ...	16431
14	M	Michael ...	Blythe ...	23058
15	F	Jae ...	Pak ...	26231
16	F	Jillian ...	Carson ...	27051
17	F	Linda ...	Mitchell ...	27229

Tabela wynikowa

Nie jest to koniec jednak naszej analizy, ponieważ chcemy uzyskać dwie oddzielne tabele, które będą podzielone ze względu na płeć. Stworzyłem też, dlatego taki proces:



Podzieliłem za pomocą węzła Splitter tabele główna na tabele Mężczyźni oraz Kobiety:

#	Płec	Imie	Nazwisko	Ilość_produktu
1	F	Amy ...	Alberts ...	2012
2	F	Lynn ...	Tsoflias ...	4123
3	F	Rachel ...	Valdez ...	7033
4	F	Pamela ...	Ansman-Wolfe...	7360
5	F	Jae ...	Pak ...	26231
6	F	Jillian ...	Carson ...	27051
7	F	Linda ...	Mitchell ...	27229

#	Płec	Imie	Nazwisko	Ilość_produktu
1	M	Syed ...	Abbas ...	825
2	M	Stephen ...	Jiang ...	3095
3	M	Tete ...	Mensa-Annan ...	5650
4	M	David ...	Campbell ...	8172
5	M	Garrett ...	Vargas ...	11544
6	M	Ranjit ...	Varkey Chudu...	14085
7	M	José ...	Saraiva ...	15220
8	M	Shu ...	Ito ...	15397
9	M	Tsvi ...	Reiter ...	16431
10	M	Michael ...	Blythe ...	23058

Tabela- Kobiety

Tabela- Mężczyźni

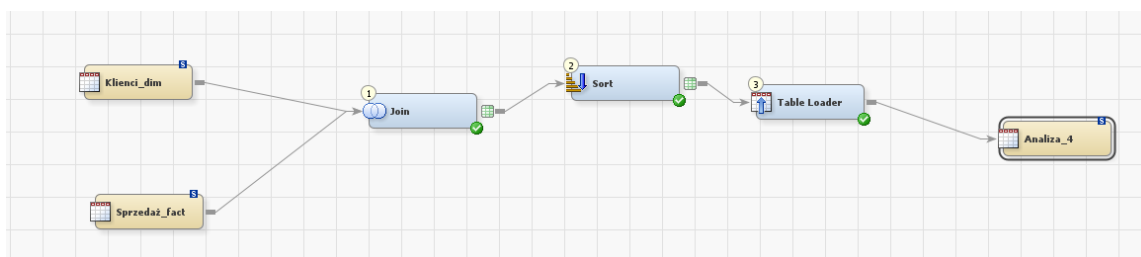
Gdy już mamy dwie ostateczne tabele wynikowe możemy przejść do wyciągania wniosków. Z tabeli Kobiety jesteśmy w stanie odczytać, że Amy Alberts jest pracownikiem, który

sprzedał najmniejszą ilość produktów. Pracownikiem męskim który uzyskał najmniejszy wynik pod względem ilości sprzedanych produktów był Syed Abbas.

4.4 Analiza średniej wartości zamówienia dla klienta

Celem tej analizy jest zidentyfikowanie średniej wartości zamówienia dla każdego klienta, co umożliwi nam wyodrębnienie grup klientów generujących najwyższą wartość transakcji. Zrozumienie, którzy klienci przyczyniają się najbardziej do naszych przychodów, pozwoli na skierowanie działań marketingowych i sprzedażowych w bardziej efektywny sposób.

Jak w każdym z wcześniejszych procesów zaczynamy od połączenia ze sobą tabel źródłowych za pomocą węzła Join:



Proces ETL numer 4

W samym węźle wybieramy kolumny z imieniem oraz nazwiskiem, a także tworzymy dwie własne kolumny

Target table: Join (W 184JCH7)						
#		Column	Column Description	Expression	Type	Length
1		Imie	FirstName		Character	1024
2		Nazwisko	LastName		Character	1024
3		Ilosc_zamowien		COUNT(*)	Numeric	8
4		Srednia_na_zamowienie		AVG(Sprzedaż_fact."Ilość produktu" * Sprzedaż...	Numeric	8

Za pomocą funkcji Count(*) zliczamy ilość zamówień natomiast dzięki funkcji Average uzyskujemy średnią wartość zamówienia. Nasze dane grupujemy według klienta a dokładnie według imienia oraz nazwiska. Kolejnym krokiem jest zastosowanie węzła Sort, w którym posortujemy nasze dane malejąco ze względu na Ilosc_zamówień co pozwoli na zidentyfikowaniu klienta z największą liczbą zamówień. Ostatnim użytym węzłem był Table Loader dzięki któremu przeładujemy dane do tabeli docelowej. Aby uzyskać średnią na zamówienie w dolarach na koniec zmienimy jeszcze format:

Type	Length	Informat	Format
Character	1024	\$1024.	\$1024.
Character	1024	\$1024.	\$1024.
Numeric	8 (None)	(None)	(None)
Numeric	8 (None)	(None)	dollar20.2

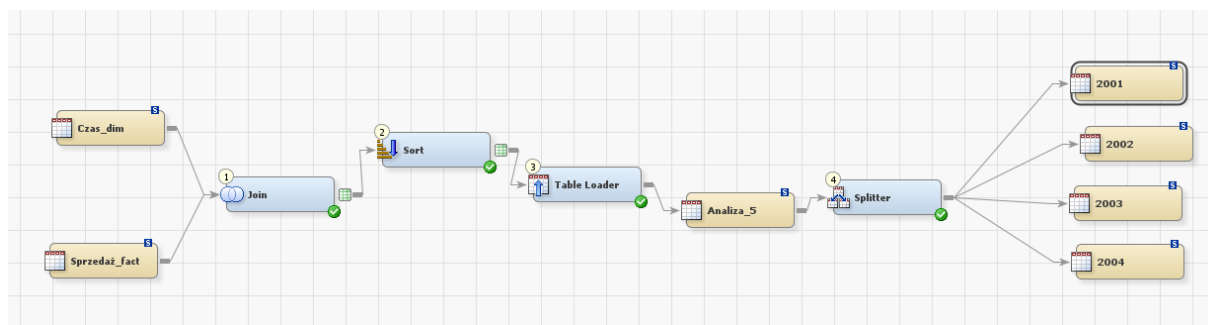
#	Imie	Nazwisko	Ilosc_zamowien	Srednia_na_zamowienie
1	Ashley ...	Henderson ...	68	\$23.77
2	Fernando...	Barnes ...	67	\$22.40
3	Charles ...	Jackson ...	65	\$22.08
4	Jennifer ...	Simmons ...	63	\$17.89
5	Henry ...	Garcia ...	62	\$18.82

Tabela wynikowa

Z tabeli wynikowej odczytujemy, że Ashley Henderson był klientem, który zamówił aż 68 zamówień natomiast jego średnia wartość zamówienia wynosiła 23,76\$.

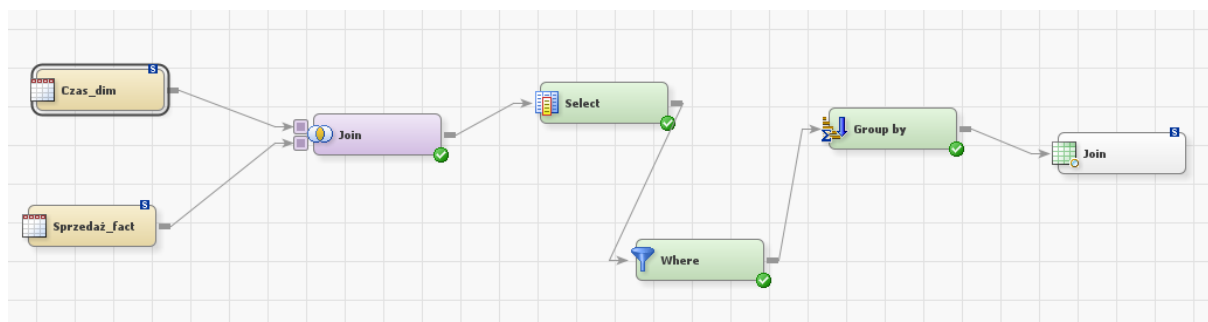
4.5 Analiza dochodów ze sprzedaży w zależności od roku oraz miesiąca.

W ostatnim procesie ETL będę przeprowadzał analizę, która pozwoli na stwierdzenie, który miesiąc w danym roku był najbardziej dochodowym miesiącem. Proces ten będzie używał dwóch tabel Czas_dim oraz Sprzedaż_fact. Pierwszym krokiem jest połączenie ich za pomocą węzła Join:



Proces ETL numer 5

W środku węzła ustawiamy kroki jak: select, group by oraz where:



Węzeł Join

W węźle select ustawiamy następujące kolumny:

#		Column	Column Description	Expression
1		Miesiąc		
2		Suma_na_miesiac		SUM(Sprzedaż_fact."Ilość produktu" * Sprzedaż...
3		Ilosc_na_miesiac		COUNT(*)
4		Srednia_na_miesiac		AVG(Sprzedaż_fact."Ilość produktu" * Sprzedaż...
5		Rok		

Trzy kolumny ustawiamy ręcznie: Suma_na_miesiac- która za pomocą funkcji SUM oblicza sumę dochodów uzyskanych na miesiąc, Ilosc_na_miesiac- która zlicza ilość zamówień w danym miesiącu a także Srednia_na_miesiac- która za pomocą funkcji AVERAGE oblicza średnia wartość zamówienia w danym miesiącu.

W węźle Where łączymy tabele po ID, a później w węźle Group by grupujemy dane

Group by columns		
Table name	Column name	Column ref...
Join	Miesiąc	Name
Join	Rok	Name

Grupowanie

Następnie sortujemy dane malejąco ze względu na suma_na_miesiac, dzięki czemu uzyskujemy najbardziej dochodowy miesiąc na przestrzeni lat 2001-2004.

Kolejnym korkiem jest użycie Table Loader dzięki któremu uzyskujemy naszą tabelę wynikową:

#	Miesiąc	Suma_na_miesiac	Ilosc_na_miesiac	Srednia_na_miesiac	Rok
1	5	1052260178.9	1208067	871.02799675	2004
2	8	1050004107.8	1094311	959.51160847	2003
3	12	981446009.08	1126694	871.08479239	2003
4	9	940602715.11	1027307	915.60041458	2003
5	6	933263333.78	1120566	832.84994706	2004
6	11	878589951.42	934322	940.35027691	2003
7	3	777567342.42	860722	903.38964546	2004

Teraz aby móc rozdzielić tabelę wynikową na tabelę względem roku stosujemy węzeł Splitter. Powoduje on, że powstają 4 tabelę: 2001, 2002, 2003, 2004. Ponownie, aby dane były w formacie pieniężnym zmieniamy ich format w węźle Table loader na dollar20.2

#	Column	Column Description	Type		#	Type	Length	Informat	Format	Is Nullable
1	Miesiąc		Numeric		1	Numeric	8 (None)	(None)	(None)	Yes
2	Suma_na_miesiac		Numeric		2	Numeric	8 (None)	dollar20.2		Yes
3	Ilosc_na_miesiac		Numeric		3	Numeric	8 (None)	(None)	(None)	Yes
4	Srednia_na_miesiac		Numeric		4	Numeric	8 (None)	(None)	(None)	Yes
5	Rok		Numeric		5	Numeric	8 (None)	(None)	(None)	Yes

W efekcie końcowym uzyskujemy tabelę:

#	Miesiąc	Suma_na_miesiac	Ilosc_na_miesiac	Srednia_na_miesiac	Rok
1	11	\$228,526,238.67	107767	\$2,120.56	2001
2	8	\$127,526,891.98	64103	\$1,989.41	2001

Najbardziej dochodowym miesiącem w roku 2001 był listopad, należy również wspomnieć, że w roku tym dane są od lipca do grudnia.

#	Miesiąc	Suma_na_miesiac	Ilosc_na_miesiac	Srednia_na_miesiac	Rok
1	8	\$542,705,584.16	442656	\$1,226.02	2002
2	11	\$428,504,557.98	328190	\$1,305.66	2002

W roku 2002 najbardziej dochodowym miesiącem był sierpień a zaraz po nim listopad, czyli odwrotnie jak rok wcześniej.

#	Miesiąc	Suma_na_miesiac	Ilosc_na_miesiac	Srednia_na_miesiac	Rok
1	8	\$1,050,004,107.77	1094311	\$959.51	2003
2	12	\$981,446,009.08	1126694	\$871.08	2003

W 2003 roku na pierwszym miejscu był sierpień, gdzie wartość przekroczyła miliard dolarów.

#	Miesiąc	Suma_na_miesiac	Ilosc_na_miesiac	Srednia_na_miesiac	Rok
1	5	\$1,052,260,178.95	1208067	\$871.03	2004
2	6	\$933,263,333.78	1120566	\$832.85	2004

W roku 2004 miesiącem, który przyniósł najwięcej pieniędzy był maj (dane w tym roku zawierały miesiące od stycznia do lipca).

5. Podsumowanie

Uzyskane rezultaty w ramach przeprowadzonej pracy stanowią kluczowy punkt wyjścia dla właścicieli firmy, umożliwiając im zgłębienie informacji dotyczących wyników sprzedażowych, klientów oraz asortymentu produktów w różnych obszarach, gdzie firma oferuje swoje produkty.

W projekcie, gdzie analizowaliśmy sprzedaż firmy Adventure Works udało nam wyodrębnić pięć procesów ETL gdzie każdy z nich prowadzi do innych wniosków, na podstawie których można przeprowadzić dalsze raporty lub analizy.

W pierwszej kolejności badaliśmy obszar produkcji a dokładniej przedmiotów, gdzie też podzieliłem je na kategorie oraz podkategorie. Celem analizy były znalezienie najbardziej jak i najmniej dochodowych grup w celu zoptymalizowania późniejszej produkcji czy też zmieniania dofinansowywania danych kategorii. Wnioskiem wyciągniętym z tej analizy był fakt, iż Rowery były kategorią najbardziej dochodową a tym samym podkategorią, która zarobiła najwięcej pieniędzy była podkategoria rowerów szosowych. Dzięki temu zarząd

może kontynuować analizowanie tego zjawisko i stwierdzić skąd ono wynika czy też przeznaczyć większy kapitał pieniędzy na kategorię rowerów.

Drugą przeprowadzoną analizą była Analiza najlepiej/najgorzej sprzedających się produktów w Wielkiej Brytanii. Analiza taka pomaga zrozumieć rynek na danych regionach i dostosować swoją ofertę specjalnie pod dany region geograficzny co też może przełożyć się na większy zysk w przyszłości. Informacje jakie udało się uzyskać to:

- najczęstszym kupowanym przedmiotem na terenie Wielkiej Brytanii była czapka z logiem AWC
- najrzadziej kupowanym przedmiotem na terenie Wielkiej Brytanii była czarna rama górską o rozmiarze 40

Można wnioskować zatem że masowe produkowanie ram nie ma sensu i może lepiej skupić się na indywidualnych dostosowywanych zamówieniach.

Analiza trzecia dotyczyła pracowników, którzy sprzedali najmniejszą liczbę przedmiotów. W analizie tej podzieliliśmy również sprzedających ze względu na płeć a wyniki wyszły następujące:

- kobieta, która sprzedała najmniejszą liczbę przedmiotów to - Amy Alberts
- mężczyzna, który sprzedał najmniejszą liczbę przedmiotów to - Syed Abbas

Wnioski pozwolą na lepsze zarządzanie zasobami ludzkimi a także mogą wpłynąć na rozdysponowywanie płac wśród osób pracujących. Można także na podstawie procesu sporządzić dalsze analizy przyczynowo-skutkowe i zagłębić się w informacje dotyczące pracowników, a także spróbować uzasadnić dane liczby sprzedanych produktów, które mogą zależeć od wielu czynników jak np. sezonowość, miejsce sprzedaży, itp.

Czwarta analiza służyła do sprawdzenia którzy klienci składają najczęściej zamówienia w naszym sklepie a także ile wynosi średnia zamówienia na klienta. Dzięki takiej analizie można rozważyć wprowadzenie kodów rabatowych dla osób częściej kupujących w naszym sklepie, czy też na zrobieniu darmowych dostaw w zależności od przekroczonego progu cenowego. Wyniki z niej prezentują się w taki sposób: Ashley Henderson był klientem, który złożył aż 68 zamówień w naszym sklepie a średnia wartość jego zamówienia wyniosła 23,76\$.

Ostatnią analizą była analiza dotycząca dochodów ze sprzedaży w zależności od roku oraz miesiąca. Gdzie uzyskaliśmy istotne informacje: sierpień oraz końcówka roku zazwyczaj były okresami najbardziej dochodowymi. Dzięki takiej informacji

firma Adventure Works mogłaby zastanowić się nad sporządzaniem kampanii reklamowych czy inwestowaniu więcej pieniędzy właśnie podczas tych okresów. Można także przeprowadzić dalsze analizy w tym kierunku i sprawdzić skąd wynikają dane zależności i dostosować swoją późniejszą ofertę do stwierdzonych wniosków.

Podsumowując, przeprowadzona analiza danych sprzedażowych firmy Adventure Works pozwoliła na identyfikację kluczowych obszarów optymalizacyjnych. Zastosowanie schematu gwiazdy oraz procesów ETL umożliwiło wyodrębnienie istotnych informacji dotyczących asortymentu, rynków regionalnych, pracowników oraz klientów. Wnioski płynące z analizy pozwalają nie tylko na skoncentrowanie się na najbardziej dochodowych produktach czy regionach, ale także na lepsze zarządzanie zasobami ludzkimi oraz dostosowanie strategii marketingowej do zmieniających się potrzeb rynku. Przyszłe decyzje biznesowe oparte na tych wnioskach mają potencjał zwiększenia efektywności operacyjnej i generowania większych zysków dla przedsiębiorstwa Adventure Works.