Analiza sprzedaży w firmie Adventure Works

BARTŁOMIEJ MACHURA
INŻYNIERIA HURTOWNI DANYCH

Uniwersytet Ekonomiczny w Katowicach

Spis treści

Bezpieczeństwo danych w hurtowniach danych: strategie ochrony przed zagrożeniami i naruszeniami

Hurtownie danych cieszą się w dzisiejszych czasach wielkim zainteresowaniem. Ilość danych posiadanych przez przedsiębiorstwa ciągle rośnie a umiejętność ich analizy jest kluczowa do podejmowania trafnych decyzji. Z najprostszej definicji hurtowania danych to system służący do zarządzania danymi. Dzięki niej przedsiębiorstwa mogą przechowywać ogromne ilości danych i analizować je pod kątem, którego obecnie potrzebują. Są one wykorzystane w praktycznie wszystkich dziedzinach np. lotnictwie, bankowości, handlu, medycynie. Jednak wraz ze wzrostem popularności tego rozwiązania pojawiły się zagrożenia związane z bezpieczeństwem danych.

W hurtowniach znajdziemy wiele informacji, duża część z nich ma wartość tylko dla naszego przedsiębiorstwa, ale znajdą się też takie, które są interesujące dla potencjalnych atakujących. Przykładem mogą być dane osobowe pracowników lub klientów, dane finansowe przedsiębiorstw, czy też dane użyteczne dla konkurencji. Powoduje to, iż naturalna staje się potrzeba zadbania o bezpieczeństwo tych danych i ochrona ich przed nieautoryzowanym dostępem.

Aby skutecznie chronić nasze dane nie polegamy na pojedynczym zabezpieczeniu a na kilku zabezpieczeniach, które często nazywamy warstwami. Wektorów ataku jest na tyle dużo, że jedna metoda to za mało, aby system był bezpieczny. Ataków można dokonać zarówno na system sam w sobie, czyli na oprogramowanie jak i na użytkowników. Statystyki pokazują, że najczęściej do wycieku danych przyczynia się błąd człowieka.

1. FIREWALL ORAZ SYSTEMY IPS

Jest to pierwsza warstwa naszej ochrony przed zagrożeniami. Firewall odpowiada za filtrowanie ruchu sieciowego oraz w razie potrzeby blokowanie go. Zwykle działa on z systemem IPS, czyli Intrusion Prevention System. Jest to sprzętowe lub programowe rozwiązanie, które ma za zadanie wykryć potencjalną próbę ataku i poinformować o niej użytkownika. Sam jednak nie może go zablokować, do tego potrzebny jest firewall, który razem z IPS tworzy tzw. Intrusion Detection System. Dzięki temu połączeniu oprogramowanie będzie zdolne do wykrycia i zablokowania ataku. Poglądowy schemat

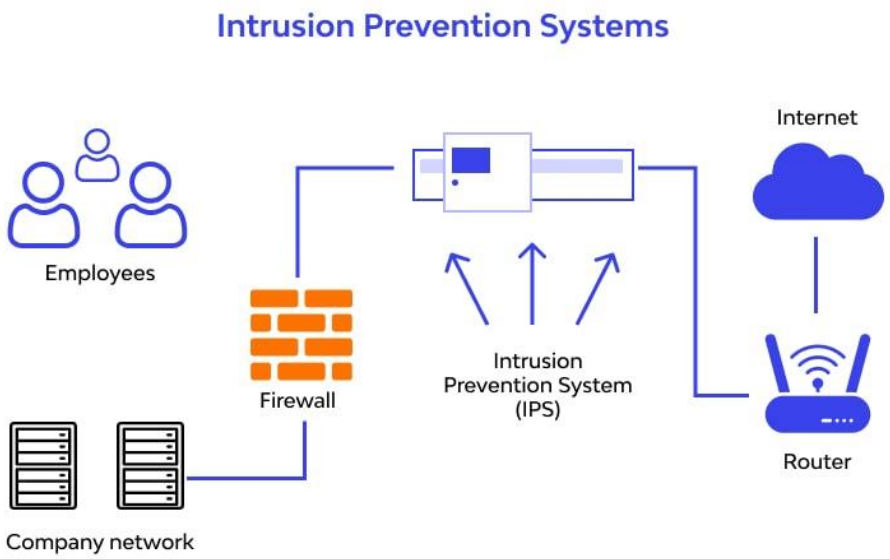
działania systemu pokazuje rysunek nr 1. Istnieją 3 podstawowe warianty działania systemu IDS:

- **Host-Based IDS (HIDS)**
- **Network-Based IDS**
- **Network node IDS (NNIDS)**

Host-Based IDS (HIDS) – jest to jedno z pierwszych wymyślonych zabezpieczeń typu IDS. Zbiera informacje z całej sieci na tzw. komputer gospodarza (najczęściej serwer sieciowy) i na innej jednostce analizuje je. Zbiera informacje z logów systemowych i aplikacji, szuka anomalii w systemie np. logowania w nocy i raportuje je. Wykrywa on także nieautoryzowaną modyfikację pliku. Poprzez sprawdzanie sum kontrolnych ważnych plików systemowych i rejestru szuka Rootkitów i Trojanów, czyli oprogramowania, którego celem jest ukrywanie procesów i innych programów w komputerze, które pomaga atakującemu pozostanie niezauważonym w systemie. Jest także odpowiedzialny za wykrywanie sygnatur, które mogą wskazywać na próbę ataku.

Network-Based IDS - w odróżnieniu od HIDS chroni on infrastrukturę przez analizę ruchu sieciowego tj. pakietów. Jeżeli wykryje, że pakiety wychodzą do nietypowych odbiorców, innych niż zwykle czy niezdefiniowanych wyśle on ostrzeżenie o takim ruchu do administratora. Stosuje on również analizę kontekstową poprzez długotrwałą analizę pakietów oraz wysyła fałszywe odpowiedzi do potencjalnego atakującego, aby myślał, że nie został wykryty. Jest on w stanie również dekodować pakiety HTTP czy FTP co pozwala na detekcję ataków za ich pomocą. Podobnie jak system HIDS wykrywa anomalie jednak on robi to w ruchu sieciowym. Szuka ruchu, który jest nietypowy i odbiega od normy np. skanowanie portów.

Network node IDS (NNIDS) – ma w sobie funkcje zarówno Network-Based IDS jak i Host-Based IDS oraz jest w stanie analizować ruch sieciowy skierowany do chronionego hosta.



Rysunek 1: Schemat działania Systemu IPS Źródło: https://www.wallarm.com/what/intrusion-prevention-system

2. ***Szyfrowanie danych***

Szyfrowanie podczas przesyłania danych

Zabezpieczenie danych podczas przesyłania i przechowywania jest kluczową sprawą, aby uchronić dane przed przejęciem. Obecne metody szyfrowania są na tyle skuteczne, że potencjalni atakujący zdają sobie sprawę, iż obecny sprzęt nie umożliwia łamania tak zaawansowanej kryptografii. Podczas przesyłu danych są one zabezpieczone za **pomocą TLS (Transport Layer Security)**. Jest on następcą protokołu SSL (Secure Socket Layer), nazywa się go często SSL/TSL ze względu na przyzwyczajenie użytkowników do nazwy SSL. TLS korzysta zarówno z szyfrowania symetrycznego jak i asymetrycznego. Szyfrowanie to opiera się na dwóch kluczach: publicznym i prywatnym. Publiczny jest udostępniany wszystkim a prywatny zna tylko użytkownik. Wykorzystywany jest tu algorytm RSA, który opiera się na tym, iż bardzo trudno znaleźć pierwiastek pierwotny z liczby pierwszej.

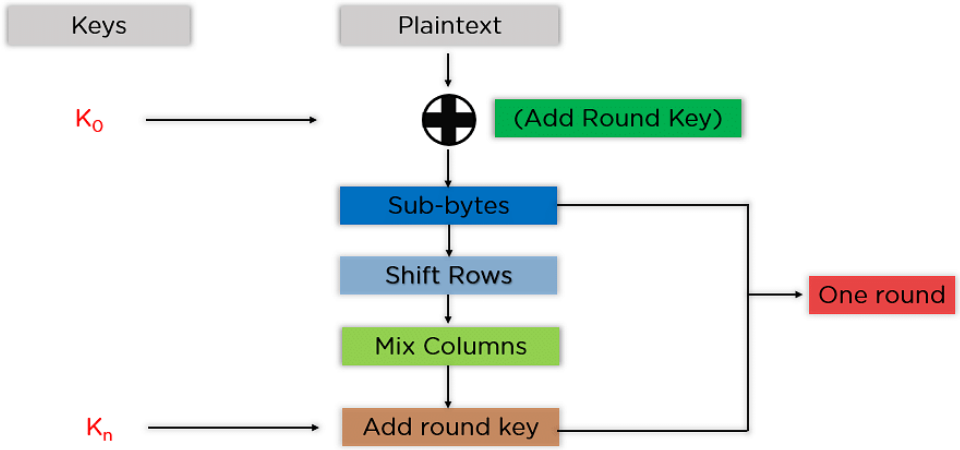
Szyfrowanie danych podczas przechowywania

Dla bezpieczeństwa danych przechowywanych lokalnie stosuje się połączenie OTP (one-time pad) oraz AES (Advanced Encryption Standard). Jest to metoda, która szyfruje dane przed załadowaniem ich do hurtowni danych.

OTP – jest to technika szyfrowania, gdzie każdy utworzony klucz jest wykorzystywany tylko jeden raz. Dodatkowo jest on tak samo długi jak treść, którą ma zaszyfrować, co czyni tę metodę jedną z najtrudniejszych do złamania. Jest on generowany w sposób **całkowicie**

losowy co jest jego wielką zaletą jak i wadą. Powoduje to, iż jest bardzo bezpieczny, jednak przez swoją długość wygenerowanie tak długiego losowego ciągu znaków jest bardzo problematyczne. Dodatkowo przekazanie odbiorcy takiego klucza też jest wyzwaniem.

AES – symetryczny algorytm szyfrowania (klucz do szyfrowania i deszyfrowania jest taki sam). Posiada on kilka rozmiarów klucza (128-bitowy, 196-bitowy, 256-bitowy). Rozmiar ten określa, ile transformacji zostanie dokonanych na tekście wejściowym. Dla 256-bitowego jest to 14 cykli. Poglądowy schemat został ukazany na rysunku 2.

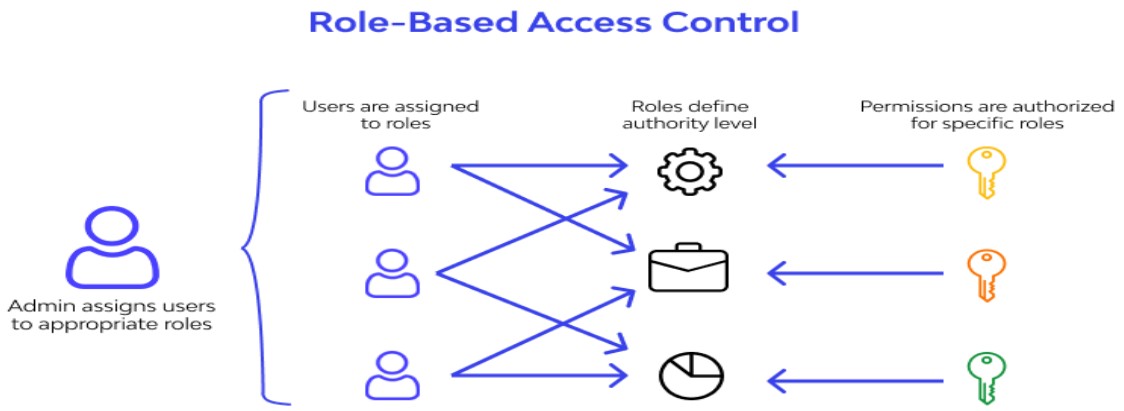


Rysunek 2. Schemat działania algorytmu AES Źródło: https://www.simplilearn.com/tutorials/cryptography-tutorial/aes-encryption

Sposób szyfrowania - najpierw generujemy jednorazowy klucz OTP, a następnie szyfrujemy dane algorytmem AES, za pomocą wygenerowanego klucza.

3. ***Kontrola dostępu do danych:***

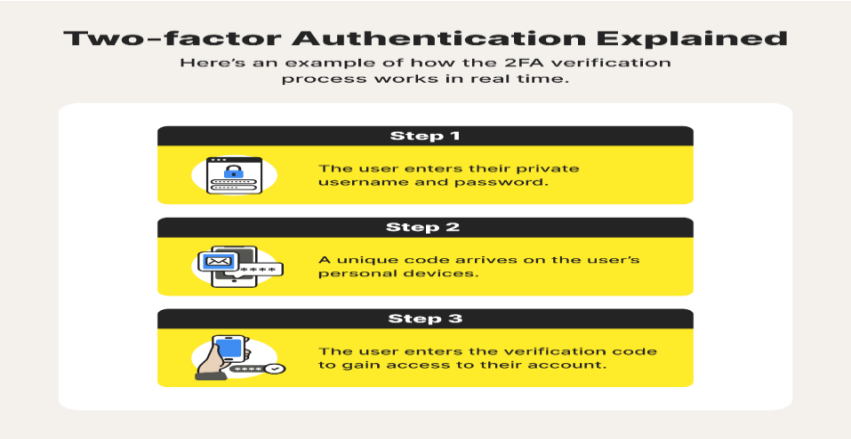
Jest ona kluczowa do zabezpieczenia naszych danych przed nieautoryzowanym dostępem. Jedną z najpopularniejszych i najskuteczniejszych metod jest RBAC (Role-Based Access Control). Polega na dokładnym określeniu ról, jakie pełnią osoby mające dostęp do hurtowni danych oraz udzielenie im dostępu tylko do tych danych, które są dla nich konieczne do pracy. Dodatkowo określamy jakie czynności w hurtowni będą oni mogli wykonywać. Schemat poglądowy na rysunku nr 3.



Rysunek nr 3. Schemat działania metody RBAC Źródło: https://www.wallarm.com/what/what-exactly-is-role-based-access-control-rbac

4. ***Uwierzytelnianie wieloskładnikowe***

Dobrą praktyką stosowaną w organizacjach jest wieloskładnikowe logowanie. Najpopularniejszym wariantem jest 2FA, czyli dwustopniowe logowanie. Najczęstszym wariantem takiego rozwiązania jest stałe hasło użytkownika oraz jednorazowe hasło generowane przez aplikacje (np. Google Authenticator). Schemat na rysunku nr 4. Inne metody to weryfikacja przez SMS na wskazany numer lub klucz U2F. Klucz ten swoim wyglądem przypomina zwykły pendrive, a sposób jego działania jest bardzo prosty. Na odpowiednim etapie logowania podłączamy go do urządzenia lub jeżeli obsługuje funkcję NFC zbliżamy go. Rozwiązania takiego możemy użyć w przypadku logowania do urządzenia, jeżeli nasi pracownicy mają urządzenia służbowe. Dodatkowo coraz więcej systemów hurtowni danych umożliwia wdrożenie uwierzytelniania wieloskładnikowego przy logowaniu. Zaletą tego systemu logowania jest to, że nawet w przypadku wycieku loginu i hasła atakujący nie będzie w stanie zalogować się na konto użytkownika.



Rysunek nr 4. Schemat działania logowania z zastosowaniem 2FA Źródło: https://us.norton.com/blog/privacy/what-is-2fa

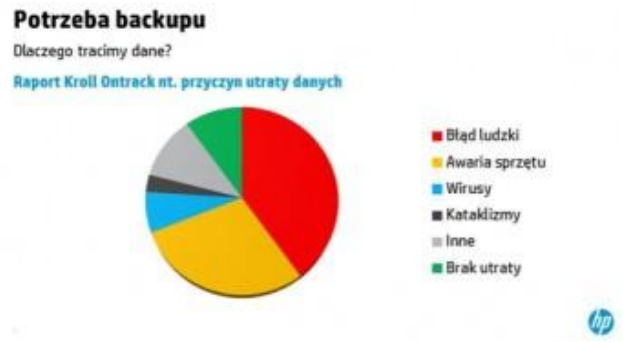
5. ***Systemy DLP***

DLP – Data Loss Prevention to rozwiązanie informatyczne, którego celem jest ochrona przed wyciekiem danych. System ten wykorzystuje uczenie maszynowe do wyszukiwania określonych wzorców w danych. Mogą to być np. dane kart kredytowych, numery PESEL itp. System monitoruje w jaki sposób te dane są udostępniane i kto ma do nich dostęp. Jest on w stanie zablokować np. wysłanie maila, w którym są wrażliwe dane przesyłane w nieodpowiedni sposób. Przykładowe działanie przedstawiono na rysunku nr 5. Jest on zgodny z obecnymi wymogami RODO. Dodatkowo może pomóc przy identyfikacji rodzaju danych w dużych zbiorach. Posiada również funkcję uniemożliwiającą zapis i odczyt danych na urządzeniach przenośnych, co chroni organizacje przed możliwością wynoszenia danych przez pracowników lub zainstalowania złośliwego oprogramowania przez urządzenie przenośne.

Rysunek nr 5. Przykładowy schemat działania systemu DLP. Źródło: https://blog.sendsafely.com/sendsafely-server-less-email-gateway-for-aws

6. ***Kopie zapasowe i plany odzyskiwania.***

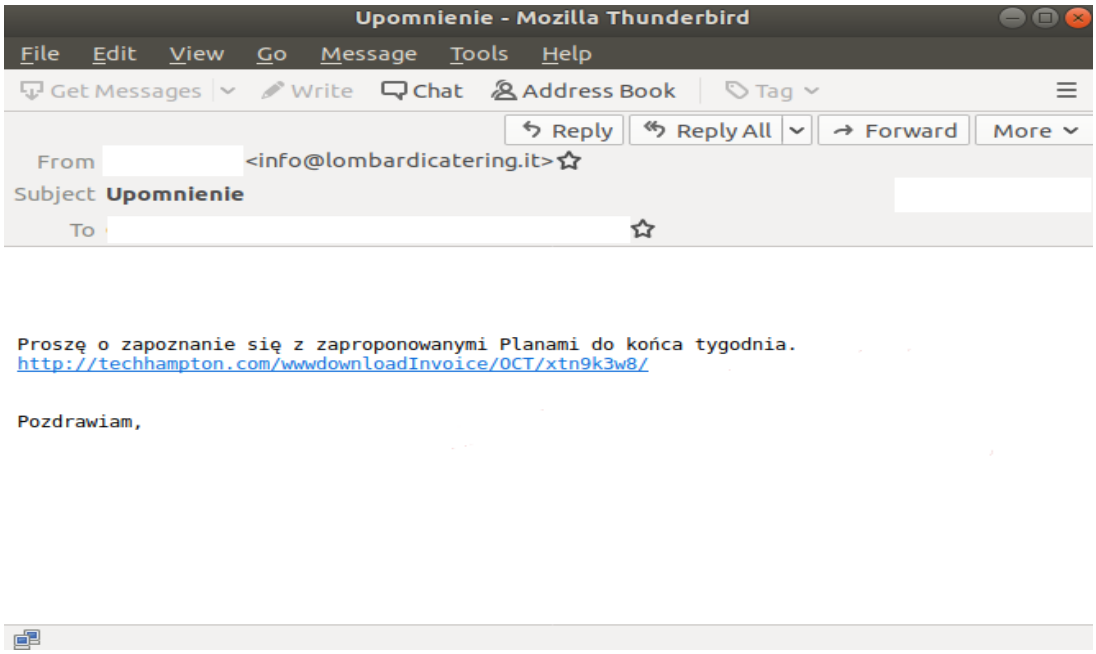
Regularne tworzenie kopii zapasowej jest kluczowe do ochrony danych. W przypadku, w którym nie jesteśmy w stanie wykonać kopii wszystkich danych, powinniśmy wybrać te które są kluczowe. Incydenty które mogą się zdarzyć to np. przypadkowe usunięcie, awaria serwerowni, wirusy itp. Jak pokazują statystyki najczęstszą przyczyną utraty danych jest błąd ludzki oraz awaria sprzętu rysunek nr 6. Dobrą praktyką jest trzymać taką kopię zaszyfrowaną, na zewnętrznych serwerach. Najlepiej, żeby taka kopia była odcięta od naszej sieci, gdyż w przypadku ataku na naszą infrastrukturę wirusa typu ransomware moglibyśmy stracić również kopię danych. Warto ułożyć również plan odzyskiwania danych, czyli schemat działania w razie utraty danych. Powinno się również testować przywracanie danych, aby upewnić się czy jest ono skuteczne.



Rysunek nr 6. Powody utraty danych przez użytkowników Źródło: https://www.ontrack.com/pl-pl/blog/utraty-danych-systemach-wirtualnych

7. ***Szkolenia użytkowników***

Warto pamiętać, że ważnym filarem bezpieczeństwa jest czynnik ludzki. Znajomość tematu cyberbezpieczeństwa przez osoby używające systemu może uniemożliwić atakującym najpopularniejszą metodę ataku, czyli phishing. Jest to atak, podczas którego celem hackera jest uzyskanie danych logowania do systemu lub wysłanie linku ze szkodliwym oprogramowaniem. Najpopularniejszą metodą jest wysyłanie fałszywych emalii do użytkowników z prośbą o zalogowanie się na fałszywej stronie w celu przejęcia danych logowania lub z linkiem nakazującym wejście na złośliwą stronę. Nawet pomimo stosowania wielu filtrów część takich wiadomości dotrze do użytkowników. Poprzez szkolenia możemy nauczyć ich jak rozpoznawać takie wiadomości. Przykładowy email pokazano na rysunku nr 7.



Rysunek nr 7. Przykład wiadomości z linkiem do złośliwej strony Źródło: https://kwestiabezpieczenstwa.pl/phishing/

8. *Analiza zachowania użytkownika*

UBA, czyli User Behavior Analytics to rozbudowany system wykorzystujący machine learning w celu wykrycia nietypowego zachowania użytkownika. W razie wykrycia podejrzanej aktywności blokuje on dostęp i wysyła powiadomienie do administratora. UBA zbiera i analizuje logi systemowe, aplikacji w celu utworzenia wzorca zachowań użytkownika.

9. *Aktualizacja oprogramowania i sprzętu*

Aktualizacja oprogramowania i sprzętu jest kluczowa, aby zachować bezpieczeństwo. W kontekście sprzętu warto być na bieżąco i sprawdzać czy zakup nowego sprzętu, mógłby podnieść bezpieczeństwo. Przykładem takiej sytuacji jest moment wprowadzenia na rynek systemu Windows 11. Wraz z wprowadzeniem systemu Microsoft ulepszył funkcję szyfrowania, jednak wymagało to, aby sprzęt posiadał mikroukład TPM 2.0. Warto być na bieżąco z nowinkami cyberbezpieczeństwa oraz śledzić panujące trendy. Aktualizacja oprogramowania jest równie ważna. Wydawcy często udostępniają aktualizacje związane z funkcjonalnością jak i z poprawkami bezpieczeństwa. Większość ataków odbywa się za pomocą luk w programach, które są znane, więc takie działanie uchroni nas przed nimi. Powinniśmy pamiętać o aktualizacji systemu operacyjnego, naszych systemów bezpieczeństwa (Firewalli, IDS itd.) jak i również aplikacji z których korzystamy.

Podsumowanie

Bezpieczeństwo danych jest kluczowym aspektem funkcjonowania hurtowni danych. Ochrona składa się z wielu warstw, które razem świadczą o sile naszej organizacji. Proces ten jest dość skomplikowany i wymaga zaangażowania zarówno administratora jak i wszystkich użytkowników. Niestety nie da się całkowicie wyeliminować ryzyka naruszeń, jednak możemy zdecydowanie je ograniczyć poprzez ograniczanie potencjalnych wektorów ataku.

Cel projektu

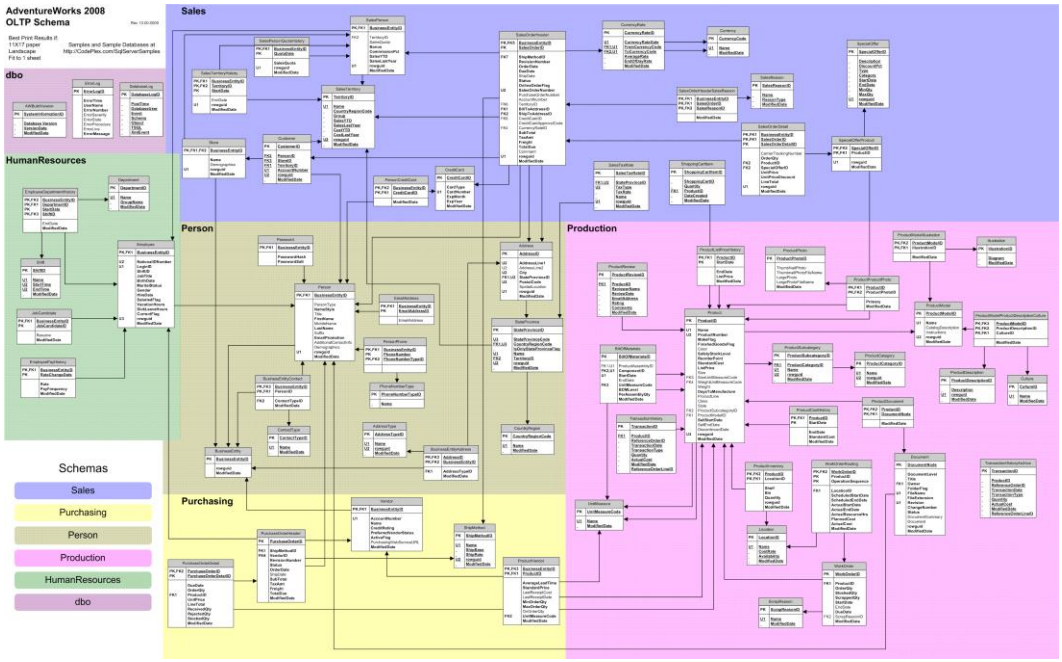
Celem mojego projektu jest ukazanie możliwości hurtowni danych oraz zbudowanie schematu gwiazdy w programie Sas Data Integration Studio. Moja analiza będzie dotyczyć sprzedaży produktów w firmie AdventureWorks. Zajmował się będę przeprowadzeniem kompleksowej analizy sprzedaży w firmie produkcyjno-handlowej specjalizującej się w sprzedaży rowerów na światowym rynku. Analiza skupi się na kluczowych grupach docelowych, opłacalności danego segmentu, lokalizacji magazynów, sezonowości sprzedaży, wydajności sprzedażowej produktów oraz wynikach pracowników sprzedaży. Wnioski z analizy obejmą rekomendacje dotyczące strategii biznesowej, lokalizacji magazynów, planowania produkcji, promocji produktów oraz zarządzania zespołem sprzedażowym.

Jest to fikcyjna firma udostępniona przez Microsoft w celach szkoleniowych i demonstracyjnych. Zajmuje się ona sprzedażą rowerów oraz akcesoriów. W mojej analizie skupie się na aspekcie sprzedaży. Skategoryzuje ją z uwzględnieniem miejsca, pracowników, produktów, klientów oraz czasu. Dokładne tematy mojej analizy są następujące:

- **Analiza wartości sprzedaży z podziałem na rodzaj klienta**
- **Analiza sprzedaży ze względu na kraj.**
- **Analiza sprzedaży w czasie**
- **Analiza sprzedaży produktów ze względu na rodzaj produktu**
- **Analiza wyników sprzedażowych pracowników**

Opis źródeł danych

Do realizacji projektu użyłem treningowej bazy, udostępnionej przez Microsoft o nazwie AdventureWorks. Opisuje ona działanie firmy, która zajmuje się sprzedażą produktów. Schemat bazy na znajduje się na rysunku numer 8.



Rysunek nr 8. Schemat bazy AdventureWorks. Źródło: https://moidulhassan.files.wordpress.com/2014/07/adventureworks2008_schema.gif

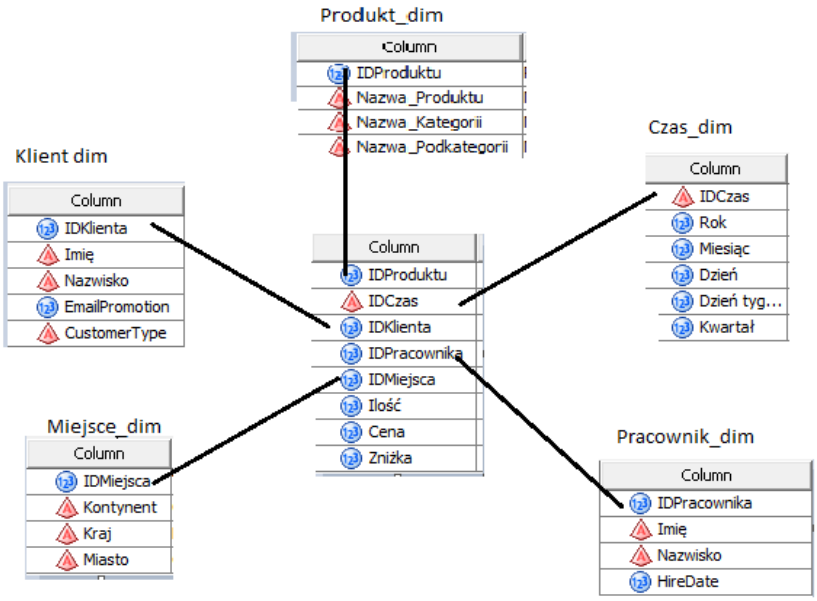
Została ona podzielona na 5 obszarów:

- Human Resources (Zasoby ludzkie)
- Sales (Sprzedaż)
- Person (Dane osobowe)
- Purchasing (Zakupy)
- Production (Produkcja)

W mojej pracy będę korzystał głownie z obszaru związanego ze sprzedażą oraz z danymi osobowymi. Jest to firma, którą zajmuje się sprzedażą rowerów, akcesoriów komponentów i ubrań. Swoje produkty dystrybuuje zarówno do odbiorców indywidualnych jak i resellerów. Baza danych dostarcza nam kompleksowych danych o tym procesie. Opisane są wszystkie najważniejsze obszary działalności firmy, co umożliwia dokonanie wielu analiz.

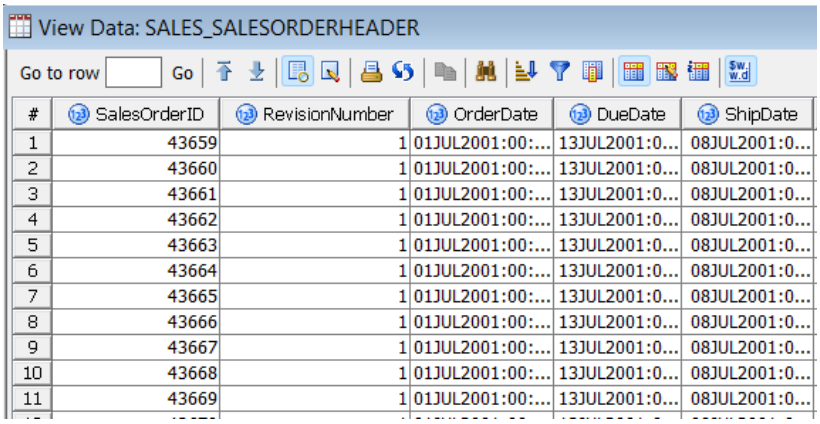
Model schematu gwiazdy

W celu przeprowadzenia analizy użyłem modelu schematu gwiazdy. Jest to jeden z najprostszych modeli projektowania w hurtowni danych. Jego cechą charakterystyczną jest posiadanie tzw. centralnej tabeli, która jest nazywana tabelą faktów. Jest ona tworzona przez odpowiednie klucze połączone z tabelami wymiarów. Dzięki takiemu rozwiązaniu jesteśmy w stanie przeglądać poszczególne kategorie, prowadzić agregacje jak i filtrować dane. Zaletą tego modelu jest łatwość w zrozumieniu i zarządzaniu. Przykładowy schemat na rysunku numer 9.



Rysunek nr 9. Schemat gwiazdy.

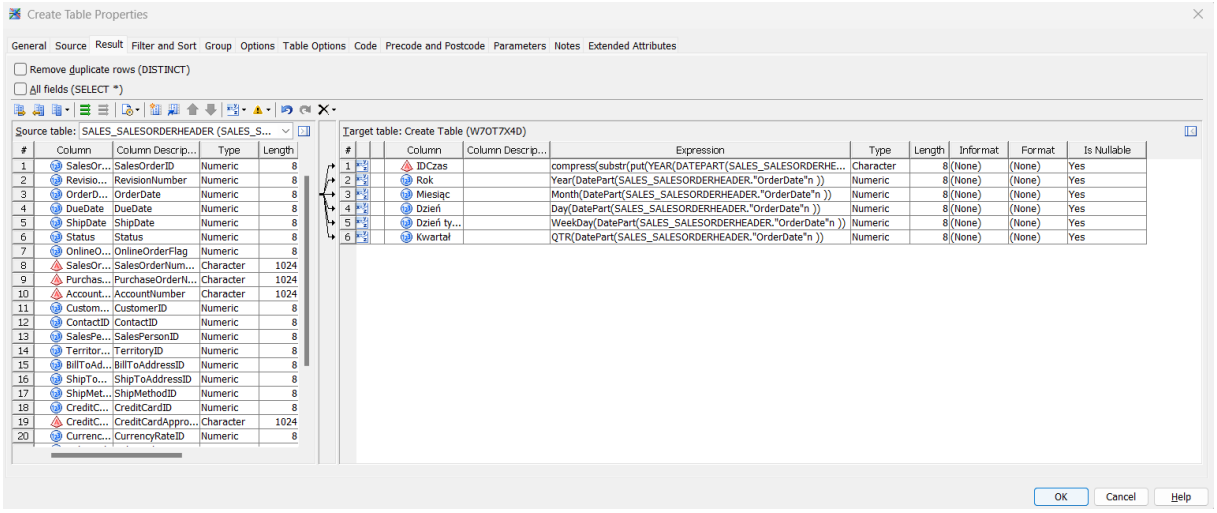
Wymiar czas – tabela Czas_dim będzie określała czasowy aspekt zamówienia. Do jej utworzenia posłuży nam do tego tabela SALES_SALESORDERHEADER.



#	SalesOrderID	RevisionNumber	OrderDate	DueDate	ShipDate
1	43659	1	01JUL2001:00:...	13JUL2001:0...	08JUL2001:0...
2	43660	1	01JUL2001:00:...	13JUL2001:0...	08JUL2001:0...
3	43661	1	01JUL2001:00:...	13JUL2001:0...	08JUL2001:0...
4	43662	1	01JUL2001:00:...	13JUL2001:0...	08JUL2001:0...
5	43663	1	01JUL2001:00:...	13JUL2001:0...	08JUL2001:0...
6	43664	1	01JUL2001:00:...	13JUL2001:0...	08JUL2001:0...
7	43665	1	01JUL2001:00:...	13JUL2001:0...	08JUL2001:0...
8	43666	1	01JUL2001:00:...	13JUL2001:0...	08JUL2001:0...
9	43667	1	01JUL2001:00:...	13JUL2001:0...	08JUL2001:0...
10	43668	1	01JUL2001:00:...	13JUL2001:0...	08JUL2001:0...
11	43669	1	01JUL2001:00:...	13JUL2001:0...	08JUL2001:0...

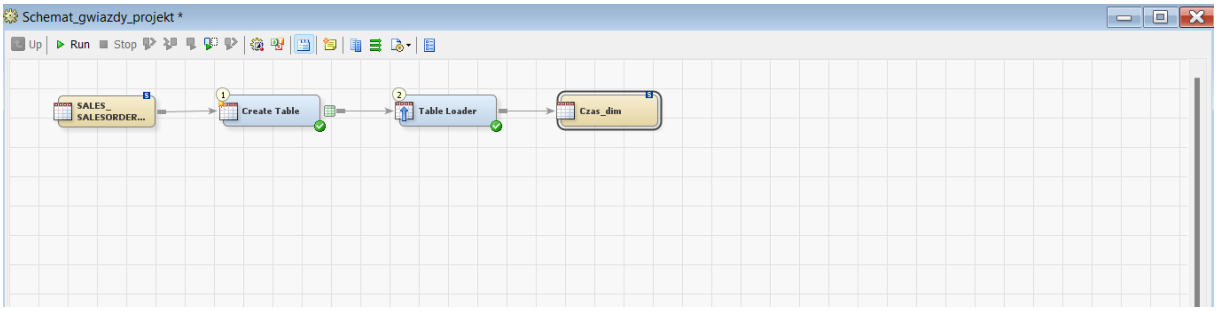
Rysunek 9. Tabela SALES_SALESORDERHEADER.

Zawiera ona kilka kolumn związanych z czasem oraz innymi danymi, jednak dla nas najważniejsza będzie kolumna OrderDate, która zawiera datę zamówienia. Tworzymy nowy proces ETL, gdzie dodajemy interesującą nas tabele. Następnie dodajemy węzeł Create Table. W zakładce result w tabeli docelowej tworzymy kolumny IDCzas, Rok, Miesiąc, Dzień, Dzień tygodnia, Kwartał. Ustawiamy typ IDCzas na character, a resztę jako wartości numeric. **IDCzas** będzie kluczem głównym naszej tabeli i tworzymy go za pomocą wyrażenia: compress(substr(put(YEAR(DATEPART(SALES_SALESORDERHEADER."OrderDate"n))),4.),3,2) ||'0'||put(MONTH(DATEPART(SALES_SALESORDERHEADER."OrderDate"n))),2.)||'0'||put(DAY(DATEPART(SALES_SALESORDERHEADER."OrderDate"n))),2.)). Pozostałe pola tworzymy za pomocą, funkcji: YEAR(), MONTH(), DAY(), QTR() WEEKDAY(). Używamy również funkcji DATEPART(), aby dokonać konwersji daty na obsługiwaną przez SAS.



Rysunek 10. Przygotowany Węzeł Create Table.

Teraz dodajemy węzeł Table Loader oraz tworzymy tabele o nazwie Czas_dim, która będzie zawierać takie pola jakie stworzyliśmy w węźle Create Table. Następnie łączymy ze sobą kolejne węzły i uruchamiamy.



Rysunek nr 11. Proces ETL tworzący wymiar czasu.

Dzięki temu uzyskujemy gotową tabelę Czas_dim.

#	IDCzas	Rok	Miesiąc	Dzień	Dzień tygodnia	Kwartał
1	0101001	2001	10	1	2	4
2	01010010	2001	10	10	4	4
3	01010011	2001	10	11	5	4
4	01010012	2001	10	12	6	4
5	01010013	2001	10	13	7	4
6	01010014	2001	10	14	1	4
7	01010015	2001	10	15	2	4
8	01010016	2001	10	16	3	4
9	01010017	2001	10	17	4	4
10	01010018	2001	10	18	5	4
11	01010019	2001	10	19	6	4
12	0101002	2001	10	2	3	4
13	01010020	2001	10	20	7	4
14	01010021	2001	10	21	1	4
15	01010022	2001	10	22	2	4
16	01010023	2001	10	23	3	4
17	01010025	2001	10	25	5	4
18	01010026	2001	10	26	6	4
19	01010027	2001	10	27	7	4
20	01010028	2001	10	28	1	4

q

Rysunek nr 12. Gotowa tabela Czas_dim.

Wymiar Produkt – tabela Produkt_dim będzie zawierała podstawowe informacje o produkcie, takie jak ID, nazwa, kategoria, podkategoria. Aby ją utworzyć użyjemy 3 tabel: PRODUCTION_PRODUCT, PRODUCTION_PRODUCTCATEGORY, PRODUCTION_PRODUCTSUBCATEGORY.

Zaczynamy od dodania tabel do procesu ETL, a następnie transformacji join. Tabele łączymy w sposób pokazany na rysunku numer 13.

#	Boolean	(	Operand	O...	Operand	)
1			PRODUCTION_PRODUCT."ProductSubcategoryID"n	=	PRODUCTION_PRODUCTSUBCATEGORY."ProductSubcategoryID"n	
2	AND		PRODUCTION_PRODUCTCATEGORY."ProductCategoryID"n	=	PRODUCTION_PRODUCTSUBCATEGORY."ProductCategoryID"n	

Rysunek nr 13. Łączenie tabel dotyczących produktu.

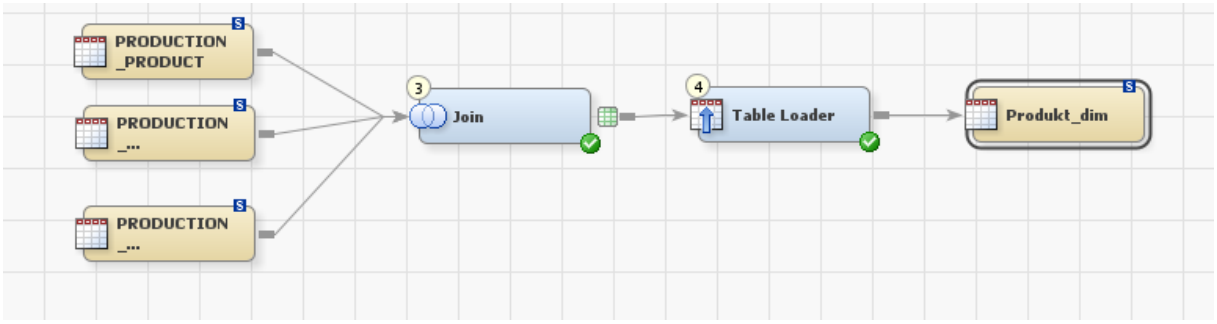
Następnie wybieramy interesujące nas kolumny. W moim przypadku wybrałem ProductID oraz kolumny Name z różnych tabel. Pozmieniałem nazwy w sposób pokazany na rysunku 14, aby ułatwić zrozumienie danych.

#	Column	Colum...	Table	Table Description	Type	Ler
1	ProductID	ProductID	PRODUCTION_...		Numeric	
2	Name	Name	PRODUCTION_...		Character	1
3	ProductNumber	Product...	PRODUCTION_...		Character	1
4	MakeFlag	MakeFlag	PRODUCTION_...		Numeric	
5	FinishedGoodsFlag	Finished	PRODUCTION		Numeric	

#	Column	Column Descrip...	Expressio
1	IDProduktu	ProductID	
2	Nazwa_Produktu	Name	
3	Nazwa_Kategorii	Name	
4	Nazwa_Podkat...	Name	

Rysunek nr 14. Wybrane kolumny do załadowania do tabeli.

Następnie tworzymy tabele Produkt_Dim zawierającą odpowiednie kolumny oraz dodajemy węzeł Table Loader.



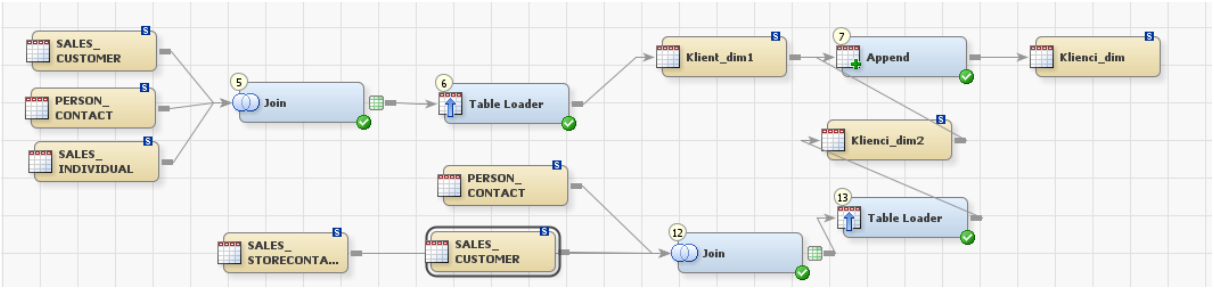
Rysunek nr 15. Proces ETL tworzący wymiar produkt.

Po uruchomienie dane zostały załadowane to tabeli Produkt_Dim.

#	IDProduktu	Nazwa_Produktu	Nazwa_Kategorii	Nazwa_Podkategorii
1	842	Touring-Panniers, La...	Accessories	Panniers
2	843	Cable Lock	Accessories	Locks
3	844	Minipump	Accessories	Pumps
4	845	Mountain Pump	Accessories	Pumps
5	846	Taillights - Battery-P...	Accessories	Lights
6	847	Headlights - Dual-Be...	Accessories	Lights
7	848	Headlights - Weather...	Accessories	Lights
8	873	Patch Kit/8 Patches	Accessories	Tires and Tubes
9	880	Hydration Pack - 70 o...	Accessories	Hydration Packs
10	921	Mountain Tire Tube	Accessories	Tires and Tubes
11	922	Road Tire Tube	Accessories	Tires and Tubes
12	923	Touring Tire Tube	Accessories	Tires and Tubes
13	928	LL Mountain Tire	Accessories	Tires and Tubes
14	929	ML Mountain Tire	Accessories	Tires and Tubes
15	930	HL Mountain Tire	Accessories	Tires and Tubes
16	931	LL Road Tire	Accessories	Tires and Tubes
17	932	ML Road Tire	Accessories	Tires and Tubes
18	933	HL Road Tire	Accessories	Tires and Tubes
19	934	Touring Tire	Accessories	Tires and Tubes
20	680	HL Road Frame - Bla...	Components	Road Frames

Rysunek nr 16. Tabela Produkt_dim wypełniona danymi.

Wymiar Klient – tabela Klienci_dim będzie zawierała informacje na temat klientów. Skupimy się na tym abyśmy znali IDKlienta, imię, nazwisko oraz typ (Klient indywidualny lub reseller). Została ona zbudowana z następujących tabel: SALES_CUSTOMER, PERSON_CONTACT, SALES_STORECONTACT, SALES_INDIVIDUAL.



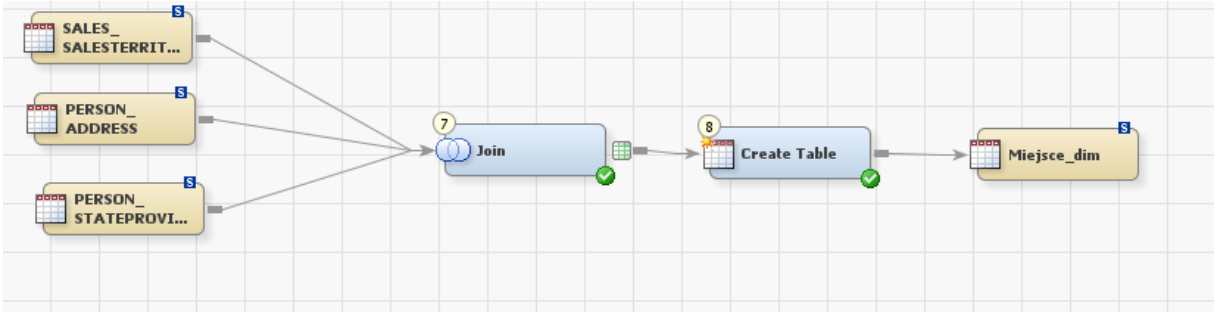
Rysunek nr 17. Schemat ETL tworzący wymiar klientów.

Przy pomocy 1 transformacji od góry utworzyłem tabelę Klient_dim1, która zawiera klientów indywidualnych. Następnie przy pomocy transformacji pokazanych na dole utworzyłem tabelę Klient_dim2, która zawiera dane resellerów. Potem połączyłem te tabele za pomocą węzła Append i utworzyłem tabele Klienci_dim zawierającą kompletne dane klientów.

#	IDKlienta	Imię	Nazwisko	EmailPromotion	CustomerType
1	11000	Jon ...	Yang ...	1	I
2	11001	Eugene ...	Huang ...	0	I
3	11002	Ruben ...	Torres ...	2	I
4	11003	Christy ...	Zhu ...	0	I
5	11004	Elizabeth...	Johnson ...	1	I
6	11005	Julio ...	Ruiz ...	0	I
7	11006	Janet ...	Alvarez ...	0	I
8	11007	Marco ...	Mehta ...	0	I
9	11008	Rob ...	Verhoff ...	0	I
10	11009	Shannon...	Carlson ...	1	I
11	11010	Jacquely...	Suarez ...	0	I
12	11011	Curtis ...	Lu ...	0	I
13	11012	Lauren ...	Walker ...	2	I
14	11013	Ian ...	Jenkins ...	1	I
15	11014	Sydney ...	Bennett ...	0	I
16	11015	Chloe ...	Young ...	1	I
17	11016	Wyatt ...	Hill ...	1	I
18	11017	Shannon...	Wang ...	2	I
19	11018	Clarence...	Rai ...	0	I
20	11019	Luke ...	Lal ...	0	I

Rysunek nr 18. Tabela Klienci_dim

Wymiar miejsce – będzie on odpowiedzialny za określenie miejsca. Będzie zawierał takie dane jak IDMiejsca, Kontynent, Kraj, Miasto. Aby go stworzyć użyłem tabel: SALES_SALESTERRITORY, PERSON_ADDRESS, PERSON_STATEPROVINCE.



Rysunek nr 19. Proces ETL

Łączę tabele w sposób ukazany na rysunku 20.

#	Boolean	(	Operand	Operator	Operand	)
1			"SALES_SALESTERRITORY"n."TerritoryI...	=	"PERSON_STATEPROVINCE"n."TerritoryID"n	
2	AND		PERSON_ADDRESS."StateProvinceID"n	=	"PERSON_STATEPROVINCE"n."StateProvinceID"n	

Rysunek nr 20. Łączenie tabel służących do utworzenia wymiaru miejsca

Wybieram w węźle Join interesujące mnie kolumny i zmieniam ich nazwę zgodnie z moimi założeniami.

#	Column	Colum...	Table	Table Description	Type	Ler	#	Column	Column Descrip...	Expression
1	TerritoryID	Territor...	SALES_SALES...		Numeric		1	IDMiejsca	TerritoryID	
2	Name	Name	SALES_SALES...		Character	1	2	Kontynent	Group	
3	CountryRegionCode	Country...	SALES_SALES...		Character	1	3	Kraj	Name	
4	Group	Group	SALES_SALES...		Character	1	4	Miasto	City	

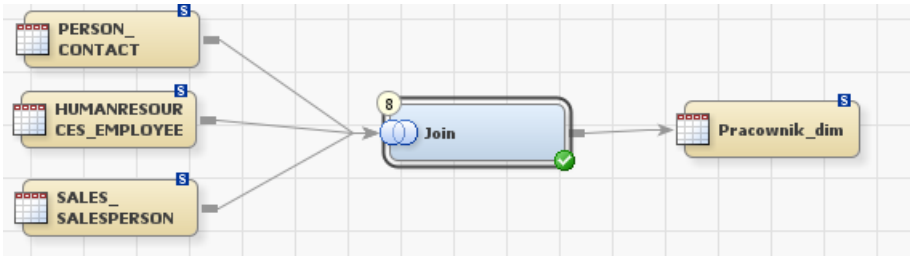
Rysunek nr 21. Wybór kolumn

Następnie uruchamiam proces ETL i otrzymuję tabelę Miejsce_dim.

#	IDMiejsca	Kontynent	Kraj	Miasto
1	1	North America...	Northwe...	Albany ...
2	1	North America...	Northwe...	Anacortes ...
3	1	North America...	Northwe...	Ballard ...
4	1	North America...	Northwe...	Beaverton ...
5	1	North America...	Northwe...	Bellevue ...
6	1	North America...	Northwe...	Bellingham...
7	1	North America...	Northwe...	Billings ...
8	1	North America...	Northwe...	Boise ...
9	1	North America...	Northwe...	Bothell ...
10	1	North America...	Northwe...	Bountiful ...
11	1	North America...	Northwe...	Bremerton ...
12	1	North America...	Northwe...	Burien ...
13	1	North America...	Northwe...	Carnation ...
14	1	North America...	Northwe...	Casper ...
15	1	North America...	Northwe...	Cedar City ...
16	1	North America...	Northwe...	Chehalis ...
17	1	North America...	Northwe...	Cheyenne ...
18	1	North America...	Northwe...	Clackamas...
19	1	North America...	Northwe...	Corvallis ...
20	1	North America...	Northwe...	Duvall ...

Rysunek nr 22. Tabela Miejsce_dim

Wymiar pracownik – jest to tabela, która będzie zawierała podstawowe dane o pracownikach odpowiedzialnych za sprzedaż, takie jak ID, imię, nazwisko, datę zatrudnienia. Do jej utworzenia użyję tabel PERSON_CONTACT, HUMANRESOURCES_EMPLOYEE, SALES_SALESPERSON.



Rysunek nr 23. Proces ETL tworzący tabele Pracownik_dim

Łączę interesujące mnie tabele w sposób pokazany na rysunku 24, a następnie wybieram interesujące mnie kolumny w węźle Join.

#	Boolean	(	Operand	Operator	Operand	)
1			PERSON_CONTACT."ContactID"n	=	HUMANRESOURCES_EMPLOYEE."ContactID"n	
2	AND		SALES_SALESPERSON."SalesPersonID"n	=	HUMANRESOURCES_EMPLOYEE."EmployeeID"n	

Rysunek nr 24. Łączenie tabel związanych z pracownikami

#	Column	Colum...	Table	Table Description	Type
1	ContactID	ContactID	PERSON_CON...		Numeric
2	NameStyle	NameSt...	PERSON_CON...		Numeric
3	Title	Title	PERSON_CON...		Character
4	FirstName	FirstName	PERSON_CON...		Character
5	MiddleName	MiddleN...	PERSON_CON...		Character

#			Column	Column Descrip...
1			IDPracownika	EmployeeID
2			Imię	FirstName
3			Nazwisko	LastName
4			HireDate	HireDate

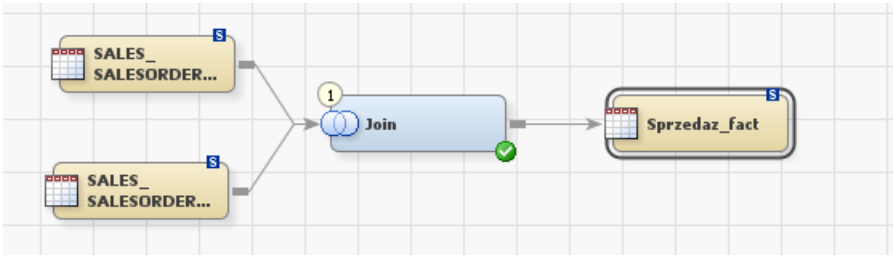
Rysunek nr 25. Węzeł Join

Po uruchomieniu otrzymałem tabelę Pracownik_dim wypełnioną danymi.

#	IDPracownika	Imię	Nazwisko	HireDate
1	268	Stephen ...	Jiang ...	04FEB2001:0...
2	288	Syed ...	Abbas ...	15APR2003:...
3	284	Amy ...	Alberts ...	18MAY2002:...
4	280	Pamela ...	Ansman-Wolf...	01JUL2001:0...
5	283	David ...	Campbell ...	01JUL2001:0...
6	277	Jillian ...	Carson ...	01JUL2001:0...
7	281	Shu ...	Ito ...	01JUL2001:0...
8	276	Linda ...	Mitchell ...	01JUL2001:0...
9	279	Tsvi ...	Reiter ...	01JUL2001:0...
10	282	José ...	Saraiva ...	01JUL2001:0...
11	278	Garrett ...	Vargas ...	01JUL2001:0...
12	286	Ranjit ...	Varkey Chudu...	01JUL2002:0...
13	289	Rachel ...	Valdez ...	01JUL2003:0...
14	290	Lynn ...	Tsoflias ...	01JUL2003:0...
15	285	Jae ...	Pak ...	01JUL2002:0...
16	275	Michael ...	Blythe ...	01JUL2001:0...
17	287	Tete ...	Mensa-Annan...	01NOV2002:...

Rysunek nr 26. Tabela Pracownik_dim

Wymiar sprzedaży – jest to centralna tabela modelu gwiazdy. Odnosi się ona do dokonanych zamówień i zawiera klucze obce odnoszące się do pozostałych tabel symbolizujących wymiary. Nazwałem ją Sprzedaz_fact. Aby ją utworzyć użyję tabel SALES_SALESORDERDETAIL oraz SALES_SALESORDERHEADER. Proces ETL tworzę w sposób ukazany na rysunku nr 27.



Rysunek nr 27. Tworzenie tabeli Sprzedaz_fact

W węźle join wybieram interesujące nas kolumny pamiętając o potrzebnych kluczach obcych. Dodatkowo tworzę klucz IDCzas w taki sam sposób jak przy tworzeniu tabeli Czas_dim.

#	Column	Colum...	Table	Table Description	Type	Ler
1	SalesOrderID	SalesOr...	SALES_SALES...		Numeric	
2	SalesOrderDetailID	SalesOr...	SALES_SALES...		Numeric	
3	CarrierTrackingNumber	Carrier...	SALES_SALES...		Character	1
4	OrderQty	OrderQty	SALES_SALES...		Numeric	
5	ProductID	ProductID	SALES_SALES...		Numeric	
6	SpecialOfferID	Special...	SALES_SALES...		Numeric	
7	UnitPrice	UnitPrice	SALES_SALES...		Numeric	
8	UnitPriceDiscount	UnitPric...	SALES_SALES...		Numeric	

#	Column	Column Descrip...	Expr
1	IDProduktu	ProductID	
2	IDCzas		compress(substr(put(YEAR(DATE
3	IDKlienta	CustomerID	
4	IDPracownika	SalesPersonID	
5	IDMiejsca	TerritoryID	
6	Ilość	OrderQty	
7	Cena	UnitPrice	
8	Zniżka	UnitPriceDiscount	

Rysunek nr 28. Wybór kolumn w węźle Join

Następnie uruchamiam proces i otrzymuję gotową tabelę Sprzedaz_fact.

#	IDProduktu	IDCzas	IDKlienta	IDPracownika	IDMiejsca	Ilość	Cena	Zniżka
1	776	010701	676	279	5	1	$2,024.99	0
2	777	010701	676	279	5	3	$2,024.99	0
3	778	010701	676	279	5	1	$2,024.99	0
4	771	010701	676	279	5	1	$2,039.99	0
5	772	010701	676	279	5	1	$2,039.99	0
6	773	010701	676	279	5	2	$2,039.99	0
7	774	010701	676	279	5	1	$2,039.99	0
8	714	010701	676	279	5	3	$28.84	0
9	716	010701	676	279	5	1	$28.84	0
10	709	010701	676	279	5	6	$5.70	0
11	712	010701	676	279	5	2	$5.19	0
12	711	010701	676	279	5	4	$20.19	0
13	762	010701	117	279	5	1	$419.46	0
14	758	010701	117	279	5	1	$874.79	0
15	745	010701	442	282	6	1	$809.76	0
16	743	010701	442	282	6	1	$714.70	0
17	747	010701	442	282	6	2	$714.70	0
18	712	010701	442	282	6	4	$5.19	0
19	715	010701	442	282	6	4	$28.84	0
20	742	010701	442	282	6	2	$722.59	0

Rysunek nr 29. Tabela Sprzedaz_fact.

Zostały stworzone już wszystkie potrzebne tabele, co oznacza, że mogę zająć się analizą danych.

Analiza procesów ETL

Proces 1. Analiza wartości sprzedaży z podziałem na rodzaj klienta

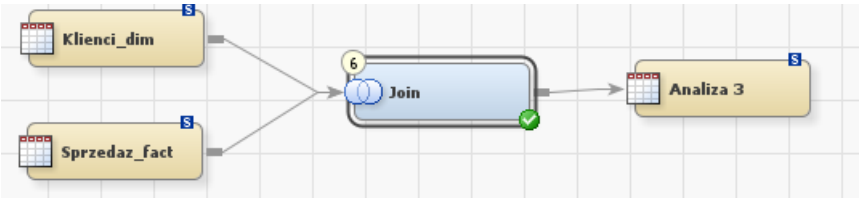
W 1 analizie chciałem sprawdzić, jak wygląda sprzedaż firmy w kontekście podziału klientów na indywidualnych i resellerów. W tym celu użyłem tabeli Klienci_dim oraz Sprzedaz_fact. W węźle Join utworzyłem 3 kolumny:

#		Column	Column Descrip...	E
1		Typ_Klienta	CustomerType	
2		Suma_ilości		Sum(Sprzedaz_fact."Ilość"n)
3		Suma sprzedaży		SUM((Sprzedaz_fact."Cena"n

Rysunek nr 30. Kolumny w węźle join

Typ klienta będzie określał czy jest to klient indywidualny czy reseller. Kolumna Suma_ilości informuje o łącznej ilości zakupionych produktów, a Suma_sprzedaży o ich łącznej wartości. Dane zostały pogrupowane ze względu na Typ Klienta. Użyłem tutaj formuł

Sum(Sprzedaz_fact."Ilość"n) oraz SUM((Sprzedaz_fact."Cena"n -(Sprzedaz_fact."Cena"n *Sprzedaz_fact."Zniżka"n))*Sprzedaz_fact."Ilość"n) do określenia ilości i wartości sprzedaży.



Rysunek nr 31. Proces ETL

Utworzyłem tabelę i po uruchomieniu procesu załadowane zostały do niej dane.

#	Typ_Klienta	Suma_ilości	Suma sprzedaży
1	I	60398	$29,358,677.22
2	S	234176	$86,782,433.61

Rysunek nr 32. Tabela Analiza 1.

Jak widać resellerzy kupili prawie 4 razy więcej produktów i wygenerowali prawie 3 razy więcej przychodu. Uznałem, że warto będzie sprawdzić, którzy resellerzy są dla nas kluczowymi klientami. W tym celu dokonałem analizy ilości i wartości zakupionych produktów przez pojedynczych klientów. Użyłem w tym celu tych samych tabel tym razem grupując ze względu na IDKlienta.

Rysunek 33. Proces ETL

#	IDKlienta	CustomerType	Suma_sprzedazy	Suma_ilosci
1	72	S	$1,492,635.06	3376
2	24	S	$1,272,452.94	5108
3	54	S	$1,075,056.19	4152
4	697	S	$877,107.19	1558
5	170	S	$853,849.18	1322
6	678	S	$841,908.77	2737
7	78	S	$826,109.63	2406
8	328	S	$816,755.58	1736
9	514	S	$799,277.90	1931
10	75	S	$791,734.55	2406

Rysunek nr 34. Tabela Analiza 1.1

Jak widać 5 najważniejszych naszych klientów to ci z ID 72, 24, 54, 697, 170. Jednak tutaj dane zostały pogrupowane ze względu na wartość sprzedaży. Postanowiłem sprawdzić dalej jak to będzie wyglądało w przypadku ilości produktów.

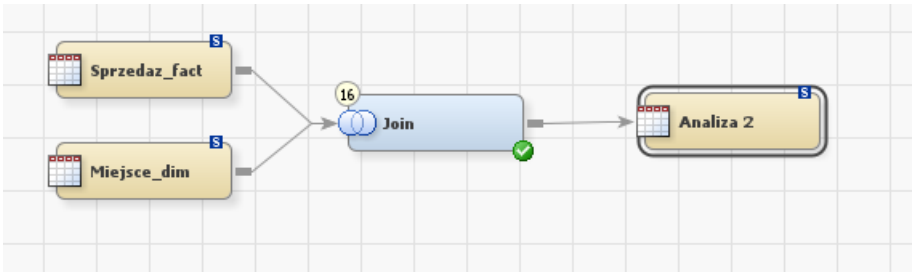
#	IDKlienta	CustomerType	Suma_ilosci
1	24	S	5108
2	54	S	4152
3	3	S	3390
4	72	S	3376
5	12	S	3156
6	678	S	2737
7	75	S	2406
8	78	S	2406
9	496	S	2350
10	175	S	2313

Rysunek nr 35. Tabela Analiza 1.2

Analiza ilości pokazuje, że najważniejsi klienci to ci z ID 24, 54, 3, 72, 12. Pokazuje to, iż 3 klientów się pokrywa, jeśli chodzi o ilość i wartość sprzedaży. Zatem są to dla nas kluczowi klienci, koncentracja na nich jest wysoka i utrata takiego klienta będzie miała istotny wpływ na przychody firmy. W związku z powyższym warto mieć to na uwadze i zadbać w ich przypadku o dobre relacje handlowe, wysoką jakość obsługi, a także rozważenie dążenia do zwiększenia sprzedaży do innych klientów, aby dokonać rozproszenia portfela.

Proces 2. Analiza sprzedaży ze względu na kraj.

W przypadku tej analizy chciałem sprawdzić, jak rozkłada się sprzedaż w kontekście geograficznym. W tym celu użyłem tabel Sprzedaz_fact i Miejsce_dim.



Rysunek nr 36. Proces ETL

W węźle join wybrałem kolumnę kraj oraz utworzyłem kolumny Suma_ilosci oraz Suma_sprzedaży analogicznie jak w przypadku 1 analizy.

#			Column	Column Descrip...	
1			Kraj	Name	
2			Suma_ilości		Sum(Sprzedaz_fact."Ilość"
3			Suma_sprzedaży		SUM((Sprzedaz_fact."Cen

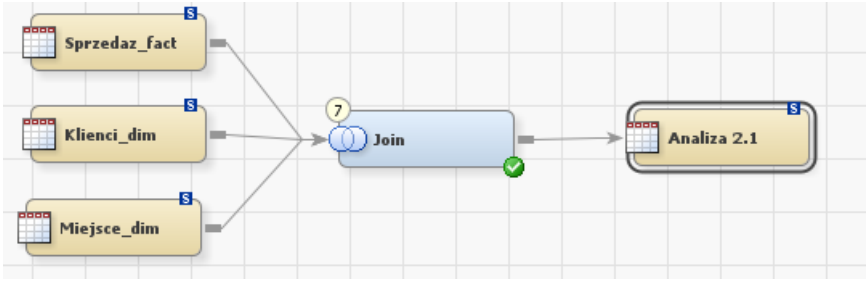
Rysunek nr 37. Kolumny docelowe.

Dane pogrupowałem według kraju, a następnie posortowałem malejąco.

#	Kraj	Suma_ilości	Suma_sprzedaży
1	Southwest ...	7624545	$3,119,814,638.29
2	Northwest ...	3420168	$1,495,899,656.93
3	Canada ...	2419669	$801,432,752.29
4	Southeast ...	1208000	$504,297,924.62
5	Central ...	1052622	$427,086,486.32
6	Australia ...	731720	$426,213,438.38
7	Northeast ...	1091365	$381,665,596.46
8	United Kingdom...	703465	$268,475,236.24
9	France ...	696710	$253,804,447.64
10	Germany ...	499434	$186,785,488.64

Rysunek nr 38. Tabela Analiza 2

Tabela pokazuje jednoznacznie, że kluczowym obszarem są Stany Zjednoczone. Uznałem, że warto będzie także sprawdzić, jak rozkłada się sprzedaż indywidualna i do resellerów w poszczególnych rejonach świata. W tym celu dołożyłem do Joina tabele Klienci_dim oraz dodałem grupowanie po typie klienta.



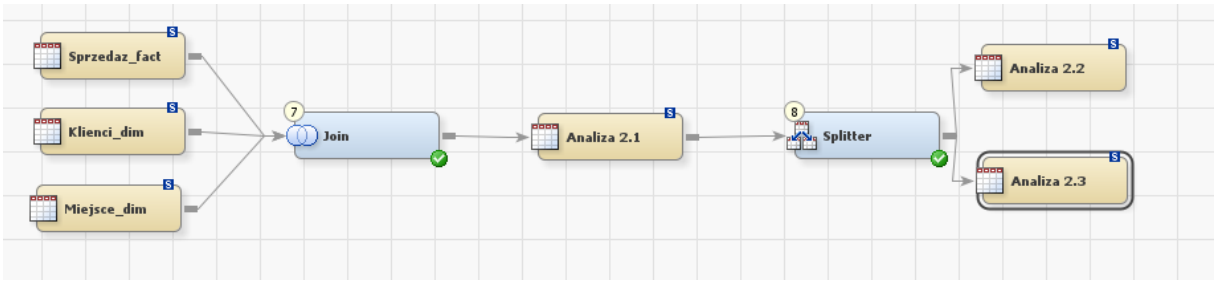
Rysunek nr 39. Proces ETL.

#	Kraj	Typ_Klienta	Suma_ilości	Suma_sprzedaży
1	Australia ...	S	214160	$70,140,069.04
2	Australia ...	I	533800	$362,440,023.38
3	Canada ...	I	373380	$96,914,398.24
4	Canada ...	S	2195690	$742,462,710.00
5	Central ...	I	1080	$162,044.80
6	Central ...	S	1109538	$455,570,932.19
7	France ...	I	194530	$92,540,620.00
8	France ...	S	502180	$161,263,827.64
9	Germany ...	S	301872	$78,803,397.85
10	Germany ...	I	213750	$109,983,868.85
11	Northeast ...	I	1485	$359,285.75
12	Northeast ...	S	1305535	$457,155,817.63
13	Northwest ...	I	836349	$339,437,589.27
14	Northwest ...	S	2628738	$1,171,442,781.29
15	Southeast ...	I	2496	$783,286.37
16	Southeast ...	S	1304704	$555,257,008.32
17	Southwest ...	I	1582185	$737,641,454.79
18	Southwest ...	S	7171239	$2,695,563,217.87
19	United King...	I	241710	$118,709,927.38
20	United King...	S	461755	$149,765,308.86

Rysunek nr 40. Tabela Analiza 2.1

Uzyskane wyniki pokazują, że w prawie wszystkich obszarach sprzedaż do resellerów przewyższa sprzedaż indywidualną. Wyjątkiem od tej reguły jest Australia. Biorąc pod uwagę

warunki klimatyczne tego kontynentu można zdecydowanie założyć, że tam produkty firmy wykorzystywane są przez cały rok w podobnym stopniu, bo umożliwiają to warunki pogodowe. Dla czytelniejszej analizy podzieliłem wyniki ze względu na rodzaj sprzedaży, aby zidentyfikować jaki rodzaj w danym obszarze jest największa. W tym celu zastosowałem węzeł Splitter.



Rysunek nr 41. Proces ETL.

Dzięki temu otrzymałem 2 tabele Analiza 2.2 oraz Analiza 2.3

W przypadku sprzedaży indywidualnej

#	Kraj	Typ_Klienta	Suma_ilości	Suma_sprzedaży
1	Southwest ...	I	1582185	$737,641,454.79
2	Australia ...	I	533800	$362,440,023.38
3	Northwest ...	I	836349	$339,437,589.27
4	United Kingdo...	I	241710	$118,709,927.38
5	Germany ...	I	213750	$109,983,868.85
6	Canada ...	I	373380	$96,914,398.24
7	France ...	I	194530	$92,540,620.00
8	Southeast ...	I	2496	$783,286.37
9	Northeast ...	I	1485	$359,285.75
10	Central ...	I	1080	$162,044.80

Rysunek nr 42. Tabela Analiza 2.2

W przypadku sprzedaży do resellerów.

#	Kraj	Typ_Klienta	Suma_ilości	Suma_sprzedaży
1	Southwe...	S	7171239	$2,695,563,217.87
2	Northwe...	S	2628738	$1,171,442,781.29
3	Canada ...	S	2195690	$742,462,710.00
4	Southea...	S	1304704	$555,257,008.32
5	Northea...	S	1305535	$457,155,817.63
6	Central ...	S	1109538	$455,570,932.19
7	France ...	S	502180	$161,263,827.64
8	United Ki...	S	461755	$149,765,308.86
9	German...	S	301872	$78,803,397.85
10	Australia...	S	214160	$70,140,069.04

Rysunek nr 43. Analiza 2.3

Na podstawie tej analizy można dojść do wniosku, że największych magazynów logistycznych potrzebujemy na terenie Ameryki ze względu na dużą sprzedaż indywidualną i

do resellerów na tym obszarze. Warto rozważyć też mniejszy na terenie Europy. Biorąc pod uwagę znaczną przewagę sprzedaży detalicznej nad sprzedażą do resellerów na terenie Australii nieekonomiczne byłoby tworzenie tam dużego magazynu.

Proces 3. Analiza sprzedaży w czasie

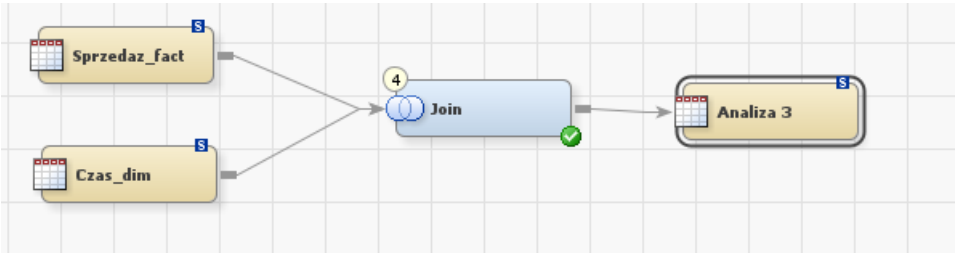
W przypadku tej analizy chciałem dowiedzieć się jak wyglądała ilość oraz wartość sprzedanych produktów w czasie, aby zaplanować proces zarządzania produkcją i dostawami.

W tym celu użyłem tabel Sprzedaz_fact oraz Czas_dim.

#			Column	Column Descrip...	Expression
1			Rok		
2			Suma _ilości		Sum(Sprzedaz_fact."Ilość"n)
3			Suma_sprzedaży		SUM((Sprzedaz_fact."Cena"n -(Sprzedaz_fa

Rysunek nr 44. Węzeł Join

Ustawiłem kolumny tak jak na rysunku powyżej, pogrupowałem dane według roku i uruchomiłem proces.

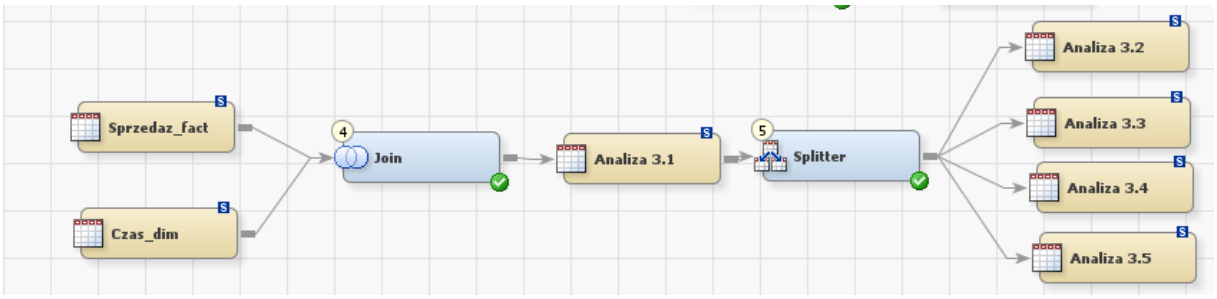


Rysunek nr 45. Proces ETL

#	Rok	Suma _ilości	Suma_sprzedaży
1	2001	11848	$11,331,808.96
2	2002	60918	$30,674,773.17
3	2003	124699	$42,011,037.16
4	2004	77449	$25,828,762.10

Rysunek nr 46. Tabela Analiza 3

Dane, które uzyskałem pokazują, że sprzedaż rosła od 2001 do 2003 roku i spadła w 2004. Są to jednak niezbyt precyzyjne dane. Postanowiłem, że dodam do mojej analizy jeszcze, miesiące, aby sprawdzić, jak sprzedaż rozkładała się w poszczególnych miesiącach roku. W tym celu dodałem kolumnę miesiąc w węźle Join oraz grupowanie według miesiąca. Następnie użyłem węzła Splitter, aby porozdzielać konkretna lata.



Rysunek nr 47. Proces ETL

#	Rok	Miesiąc	Suma _ilości	Suma_sprzedaży
1	2001	7	966	$962,716.74
2	2001	8	2209	$2,044,600.00
3	2001	9	1658	$1,639,840.11
4	2001	10	1403	$1,358,050.47
5	2001	11	3132	$2,868,129.20
6	2001	12	2480	$2,458,472.43

Rysunek nr 48. Tabela Analiza 3.2

Dane z 2001 roku pokazują tylko miesiące od lipca do grudnia. Widać jednak 3 miesiące, które się wyróżniają największą ilością i wartością sprzedaży w ciągu roku i są to sierpień, listopad i grudzień.

#	Rok	Miesiąc	Suma _ilości	Suma_sprzedaży
1	2002	1	1040	$1,309,863.25
2	2002	2	2303	$2,451,605.62
3	2002	3	1841	$2,099,415.62
4	2002	4	1467	$1,546,592.23
5	2002	5	3179	$2,942,672.91
6	2002	6	2418	$1,678,567.42
7	2002	7	7755	$2,894,054.68
8	2002	8	11325	$4,147,192.18
9	2002	9	9066	$3,235,826.19
10	2002	10	5584	$2,217,544.45
11	2002	11	8268	$3,388,911.41
12	2002	12	6672	$2,762,527.22

Rysunek nr 49. Tabela Analiza 3.3

Rok 2002 pokazuje, że sprzedaż firmy na początku roku jest stosunkowo niewielka. Skok jest widoczny w 2 miesiącach – w sierpniu i wrześniu. Następnie sprzedaż maleje i ponownie zwiększa swój poziom w listopadzie.

#	Rok	Miesiąc	Suma _ilości	Suma_sprzedaży
1	2003	1	3532	$1,756,407.01
2	2003	2	5431	$2,873,936.93
3	2003	3	4132	$2,049,529.87
4	2003	4	5694	$2,371,677.70
5	2003	5	8278	$3,443,525.24
6	2003	6	6444	$2,542,671.93
7	2003	7	11288	$3,554,092.32
8	2003	8	18986	$5,068,341.51
9	2003	9	18681	$5,059,473.22
10	2003	10	11607	$3,364,506.26
11	2003	11	14771	$4,683,867.05
12	2003	12	15855	$5,243,008.13

Rysunek nr 50. Analiza 3.4

W 2003 roku skok sprzedaży występuje podobnie jak w 2002 roku w sierpniu, przy czym utrzymuje wysoki poziom do końca roku.

#	Rok	Miesiąc	Suma _ilości	Suma_sprzedaży
1	2004	1	9227	$3,009,197.42
2	2004	2	10999	$4,167,855.43
3	2004	3	11314	$4,221,323.43
4	2004	4	12239	$3,820,583.49
5	2004	5	15656	$5,194,121.52
6	2004	6	15805	$5,364,840.18
7	2004	7	2209	$50,840.63

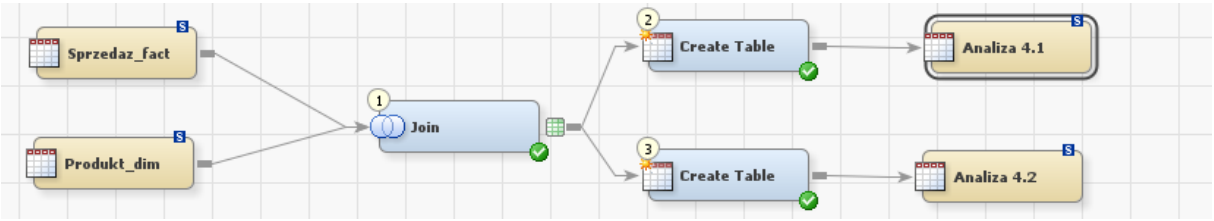
Rysunek nr 51. Analiza 3.5

W roku 2004 sprzedaż jest na podobnym poziomie przez cały rok. Szczyt osiąga w czerwcu, w lipcu nastąpił gwałtowny spadek, ale prawdopodobnie może to być związane z tym, że dane nie obejmują całego miesiąca.

Dane sprzedażowe z lat 2002-2004 pokazują, że sprzedaż firmy jest większa w 2 połowie roku, a swój szczyt osiąga w okolicach sierpnia. Należy zaplanować w związku z tym z odpowiednim wyprzedzeniem proces produkcji i zaopatrzenia magazynów oraz zadbać o infrastrukturę związaną z dostawami towarów do poszczególnych odbiorców, aby przed szczytem zwiększonej sprzedaży byli odpowiednio zaopatrzeni w produkty.

Proces 4. Analiza sprzedaży produktów ze względu na rodzaj produktu

W tej analizie chciałem sprawdzić, jak kształtuje się sprzedaż poszczególnych produktów firmy. W tym celu użyłem tabel Sprzedaz_fact oraz Produkt_dim.



Rysunek nr 52 Proces ETL

W węźle Join określiłem interesujące mnie kolumny takie jak: Suma_ilości, Suma_sprzedaży, Nazwa_produktu, Nazwa_kategorii. Dane zostały pogrupowane według nazwy produktu. Następnie za pomocą węzłów Create Table utworzyłem 2 tabele, gdzie w jednej z nich posortowałem Sumę sprzedaży rosnąco, a w drugiej malejąco.

#	Nazwa_Produktu	Suma	Suma_ilości	Nazwa_Kategorii
1	Mountain-200 Black, ...	$4,400,59...	2977	Bikes ...
2	Mountain-200 Black, ...	$4,009,49...	2664	Bikes ...
3	Mountain-200 Silver, ...	$3,693,67...	2394	Bikes ...
4	Mountain-200 Silver, ...	$3,438,47...	2234	Bikes ...
5	Mountain-200 Silver, ...	$3,434,25...	2216	Bikes ...
6	Mountain-200 Black, ...	$3,309,67...	2111	Bikes ...
7	Road-250 Black, 44 ...	$2,516,85...	1642	Bikes ...
8	Road-250 Black, 48 ...	$2,347,65...	1498	Bikes ...
9	Road-250 Black, 52 ...	$2,012,44...	1245	Bikes ...
10	Road-150 Red, 56 ...	$1,847,81...	664	Bikes ...
11	Road-350-W Yellow, ...	$1,774,88...	1622	Bikes ...
12	Road-150 Red, 62 ...	$1,769,09...	600	Bikes ...
13	Touring-1000 Blue, 6...	$1,721,24...	1120	Bikes ...
14	Road-350-W Yellow, ...	$1,657,19...	1477	Bikes ...
15	Road-250 Red, 58 ...	$1,587,00...	946	Bikes ...
16	Touring-1000 Blue, 4...	$1,586,95...	1002	Bikes ...
17	Road-150 Red, 48 ...	$1,540,80...	493	Bikes ...
18	Touring-1000 Yellow,...	$1,518,13...	1114	Bikes ...
19	Road-250 Black, 58 ...	$1,506,37...	910	Bikes ...
20	Road-250 Red, 44	$1,448,12	895	Bikes

Rysunek nr 53. Tabela Analiza 4.1

Jak pokazuje analiza najbardziej dochodowe produkty to rowery, a szczególnie rower Mountain-200 w kilku wersjach rozmiarowych oraz kolorystycznych. Oprócz niego wysoki przychód zapewniają również rowery marek Road oraz Touring.

#	Nazwa_Produktu	Suma	Suma_ilości	Nazwa_Kategorii
1	LL Road Seat/Saddle...	$162.72	10	Components ...
2	Mountain Bike Socks,...	$513.00	90	Clothing ...
3	LL Touring Frame - B...	$800.21	4	Components ...
4	LL Mountain Frame - ...	$1,198.99	8	Components ...
5	LL Touring Seat/Sad...	$1,480.75	91	Components ...
6	ML Mountain Frame-...	$1,529.18	7	Components ...
7	LL Touring Handlebar...	$1,548.62	56	Components ...
8	LL Headset ...	$1,949.40	95	Components ...
9	ML Touring Seat/Sad...	$1,972.66	84	Components ...
10	LL Mountain Frame - ...	$2,248.11	15	Components ...
11	LL Touring Frame - B...	$3,000.78	15	Components ...
12	HL Road Seat/Saddle...	$4,232.26	134	Components ...
13	LL Touring Frame - B...	$5,001.30	25	Components ...
14	LL Road Handlebars ...	$5,422.54	213	Components ...
15	LL Mountain Seat/Sa...	$5,636.96	347	Components ...
16	Mountain Bike Socks,...	$6,060.39	1107	Clothing ...
17	LL Mountain Frame - ...	$6,970.92	44	Components ...
18	Touring Pedal ...	$7,143.32	147	Components ...
19	LL Touring Frame - Y...	$7,201.87	36	Components ...
20	Touring Tire Tube ...	$7,425.12	1488	Accessories ...

Rysunek nr 54. Tabela Analiza 4.2

Jak można było się spodziewać, najniższą wartość sprzedaży osiągają komponenty i ubrania, bo produkty te są znacznie tańsze niż rowery. Jednak istotne jest posiadanie ich w ofercie, bo wówczas jest ona kompleksowa. Ponadto są towarami szybciej zużywającymi się i z założenia powinny być częściej kupowane. Postanowiłem to sprawdzić i przeprowadziłem dalszą analizę używając tych samych tabel i pogrupowałem dane według kategorii.

Rysunek nr 55. Proces ETL.

#	Nazwa_Kategorii	Suma_ilości	Suma_sprzedaży
1	Bikes ...	90268	$94,651,172.70
2	Components ...	49044	$11,802,593.29
3	Clothing ...	73670	$2,120,542.52
4	Accessories ...	61932	$1,272,072.88

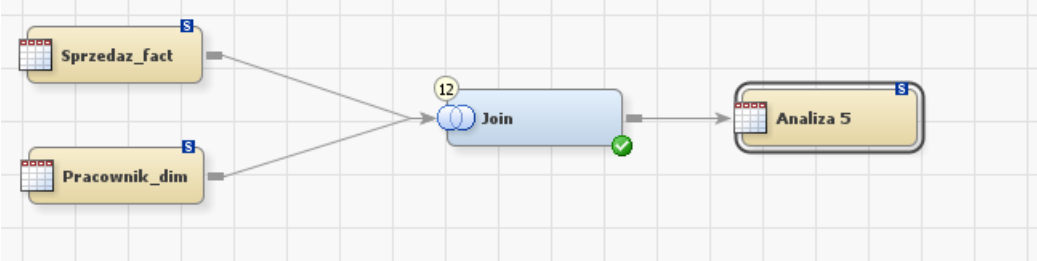
Rysunek nr 56. Tabela Analiza 4.3

Wyniki pokazują, że rowery zapewniają nam największy przychód. Jednak, jeśli chodzi o ilość to łącznie suma pozostałych trzech kategorii znacznie je przewyższa. Klienci kupując nasze rowery, często dobierają także akcesoria, odzież i komponenty. Powinniśmy zatem w celu zwiększenia ich sprzedaży pomyśleć o systemie sprzedaży łączonej. Jako zachętę można byłoby wprowadzić promocję na te akcesoria, komponenty i ubrania przy zakupie nowego

roweru. Należy również dbać o jakość pozostałych produktów i nieustannie poszerzać ich ofertę.

Proces 5. Analiza wyników sprzedażowych pracowników

W tej analizie sprawdzę, jak wygląda przychód ze sprzedaży wygenerowany przez poszczególnych pracowników. W tym celu użyłem tabel Sprzedaz_fact oraz Pracownik_dim. Dodałem węzeł Join oraz pogrupowałem dane według imienia i nazwiska pracownika.



Rysunek nr 57. Proces ETL

#	IDPracownika	Imię	Nazwisko	Suma_sprzedazy	HireDate
1	276	Linda ...	Mitchell ...	$10,367,007.43	01JUL2001:0...
2	277	Jillian ...	Carson ...	$10,065,803.54	01JUL2001:0...
3	275	Michael ...	Blythe ...	$9,293,903.00	01JUL2001:0...
4	285	Jae ...	Pak ...	$8,503,338.65	01JUL2002:0...
5	279	Tsvi ...	Reiter ...	$7,171,012.75	01JUL2001:0...
6	281	Shu ...	Ito ...	$6,427,005.55	01JUL2001:0...
7	282	José ...	Saraiva ...	$5,926,418.36	01JUL2001:0...
8	286	Ranjit ...	Varkey Chudu...	$4,509,888.93	01JUL2002:0...
9	283	David ...	Campbell ...	$3,729,945.35	01JUL2001:0...
10	278	Garrett ...	Vargas ...	$3,609,447.21	01JUL2001:0...
11	280	Pamela ...	Ansman-Wolf...	$3,325,102.59	01JUL2001:0...
12	287	Tete ...	Mensa-Annan...	$2,312,545.69	01NOV2002:...
13	289	Rachel ...	Valdez ...	$1,827,066.71	01JUL2003:0...
14	290	Lynn ...	Tsoflias ...	$1,421,810.92	01JUL2003:0...
15	268	Stephen ...	Jiang ...	$1,092,123.86	04FEB2001:...
16	284	Amy ...	Alberts ...	$732,759.18	18MAY2002:...
17	288	Syed ...	Abbas ...	$172,524.45	15APR2003:...

Rysunek nr 58. Tabela Analiza 5

Jak pokazuje analiza najlepsze wyniki generują osoby z najdłuższym stażem. Wyjątkiem jest Stephan Jiang, który zajmuje 15 miejsce, a jest najdłużej pracującym pracownikiem. Te informacje są przydatne w procesie zarządzania zasobami ludzkimi. Wyniki przeprowadzonej analizy wskazują którzy handlowcy są istotni dla firmy, ze względu na osiągane wyniki. Należy mieć to na uwadze i ich odpowiednio wynagradzać wprowadzając np. dodatkowe premie za najwyższe wyniki w sprzedaży. Dbałość o takiego pracownika wymiernie przekłada się na przychody firmy.

Podsumowanie

Analizę przeprowadziłem w pięciu obszarach sprzedaży tj. wartość sprzedaży z podziałem na rodzaj klienta, ze względu na kraj sprzedaży, czas i rodzaj produktu oraz wyników sprzedażowych pracowników. Przeprowadzone analizy pozwoliły firmie uzyskać pełniejsze zrozumienie swojej pozycji na rynku, a także dostarczyły konkretne rekomendacje mające na celu zwiększenie efektywności operacyjnej i sprzedażowej.

Wyniki pierwszej analizy pomogły mi określić kluczowe grupy docelowe wskazując, że sprzedaż w segmencie resellerów jest dla firmy najbardziej opłacalna zarówno pod względem przychodowym jak i ilościowym. Ponadto zwraca uwagę fakt, że w tym segmencie TOP 3 najlepszych klientów generują największe przychody ze sprzedaży. Te dane są istotne do określenia poziomu koncentracji na odbiorcach, gdyż zarządzając portfelem klientów należy to uwzględnić i zastosować odpowiednią strategię współpracy, aby ją rozwijać lub przynajmniej utrzymać na tym samym poziomie. Drugim aspektem do rozważenia jest też podjęcie działań zmierzających do zwiększenia rozproszenia wśród odbiorców, aby ograniczyć ryzyko spadku obrotów firmy w przypadku utraty któregoś z klientów z TOP 3.

W drugiej analizie zająłem się segmentacją klientów pod kątem położenia geograficznego oraz podzieliłem sprzedaż na rodzaj klienta. Wykonując tę analizę pozyskałem informację, która jest przydatna do określenia lokalizacji magazynów z produktami firmy. Wyniki pokazały, że największe przychody osiągane są w Stanach Zjednoczonych, a biorąc pod uwagę rodzaj klienta to dominująca jest sprzedaż hurtowa we wszystkich rejonach świata z wyjątkiem Australii, gdzie dominuje sprzedaż detaliczna. Z analizy wynika zatem, że umiejscowienie największych magazynów logistycznych firmy będzie opłacalne na terenie Ameryki oraz Europy. Nie ma uzasadnienia ekonomicznego dla inwestowania w otwarcie magazynu w Australii, gdzie dominuje sprzedaż detaliczna.

Prowadząc analizę trzecią zweryfikowałem, jak rozkłada się sprzedaż w czasie, aby zidentyfikować, czy występują sezonowe zmiany w sprzedaży, co umożliwia lepsze zarządzanie zapasami i kampaniami promocyjnymi. Uzyskałem wynik, że sprzedaż produktów firmy podlega sezonowości. Dane z lat 2002-2004 wskazują, że druga połowa roku to okres wzrostu sprzedaży. Planując produkcję i zaopatrzenie klientów należy uwzględnić te wyniki, aby przed szczytem sezonu produkty były dostępne, a dostawy do klientów zaplanowane z wyprzedzeniem.

Przeprowadzając czwartą analizę sprzedaży produktów ze względu na rodzaj oczekiwałem uzyskania odpowiedzi o wydajności sprzedażowej i zidentyfikowania najlepiej sprzedających się produktów. Przeprowadzenie regularnych analiz tego rodzaju umożliwia monitorowanie zmian w preferencjach klientów, dostosowywanie strategii biznesowej do zmieniającego się rynku i maksymalizowanie potencjału sprzedażowego. Wyniki analizy wskazały, jak się można było spodziewać, że najwyższe przychody firma osiąga ze sprzedaży rowerów, jednak przychód ze sprzedaży komponentów, ubrań i akcesoriów to prawie 15% wartości ogólnej sprzedaży. Przychody ze sprzedaży tej części asortymentu mają potencjał wzrostu, bo ich cechą jest to, że szybciej się zużywają i częściej klienci powtarzają ich zakup. Dodatkowo, mając w ofercie poza rowerami również komponenty, ubrania i akcesoria firma dostarcza klientom pełen zakres produktów i może łatwiej konkurować z firmami z tej samej branży, które skupiają się wyłączenie na sprzedaży rowerów. W celu aktywacji wzrostu przychodów ze sprzedaży należy rozważyć większe wypromowanie ich sprzedaży np. przez sprzedaż powiązaną.

Ostatnim punktem analizy było zweryfikowanie wyników pracowników sprzedaży. Analiza tych aspektów pozwala na lepsze zrozumienie środowiska sprzedażowego i podejmowanie bardziej świadomych decyzji w zakresie zarządzania zespołem sprzedażowym. Analiza pokazała, że najlepsze wyniki osiągają pracownicy z najdłuższym stażem, poza jednym wyjątkiem. Wyniki sprzedażowe pracowników mają wymierne przełożenie na wartość przychodów firmy ze sprzedaży i należy wprowadzić zmiany w sposobie zarządzania zasobami ludzkimi. Wieloletni pracownicy znający dobrze rynek i produkty firmy powinni być odpowiednio wynagradzani, a ponadto ich doświadczenie można wykorzystać przez przygotowanie systemu szkoleń wewnętrznych np. w formie dzielenia się dobrymi praktykami w zespole. Dla pracowników z gorszymi wynikami dobrze będzie zidentyfikować obszary, gdzie potrzebują wsparcia i rozwinąć system szkoleń, aby podnieść ich umiejętności sprzedażowe.