Filip Strózik 260377

Hurtownie Danych - laboratorium
Lista 7

Procesy ETL, Kostka danych

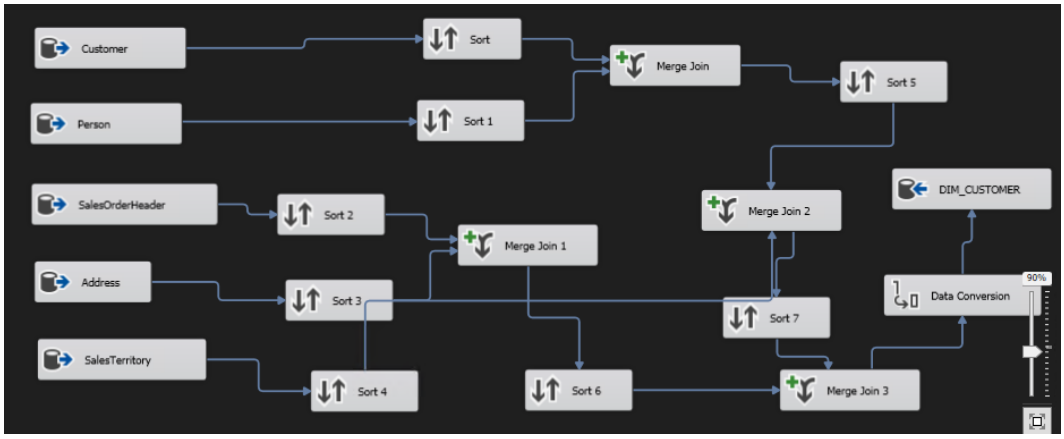
Zad. 1. ETL (prawie) bez SQLa

Przygotować proces ETL analogiczny do przygotowanego na poprzedniej liście. Zastąpić instrukcje SQL związane z wstawianiem danych do tabel wymiarów:
DIM_CUSTOMER;
DIM_PRODUCT;
DIM_SALESPERSON;
Przepływami danych budowanych z narzędzi dostępnych w zakładce Data Flow takich jak np.: OLE DB Source/Destination, Merge Join, Sort, Derived Column, Fuzzy Lookup, Fuzzy Grouping, itp.

Uruchomić pakiet i sprawdzić i udokumentować poprawność jego działania. Porównać wyniki z poprzednią wersją procesu.

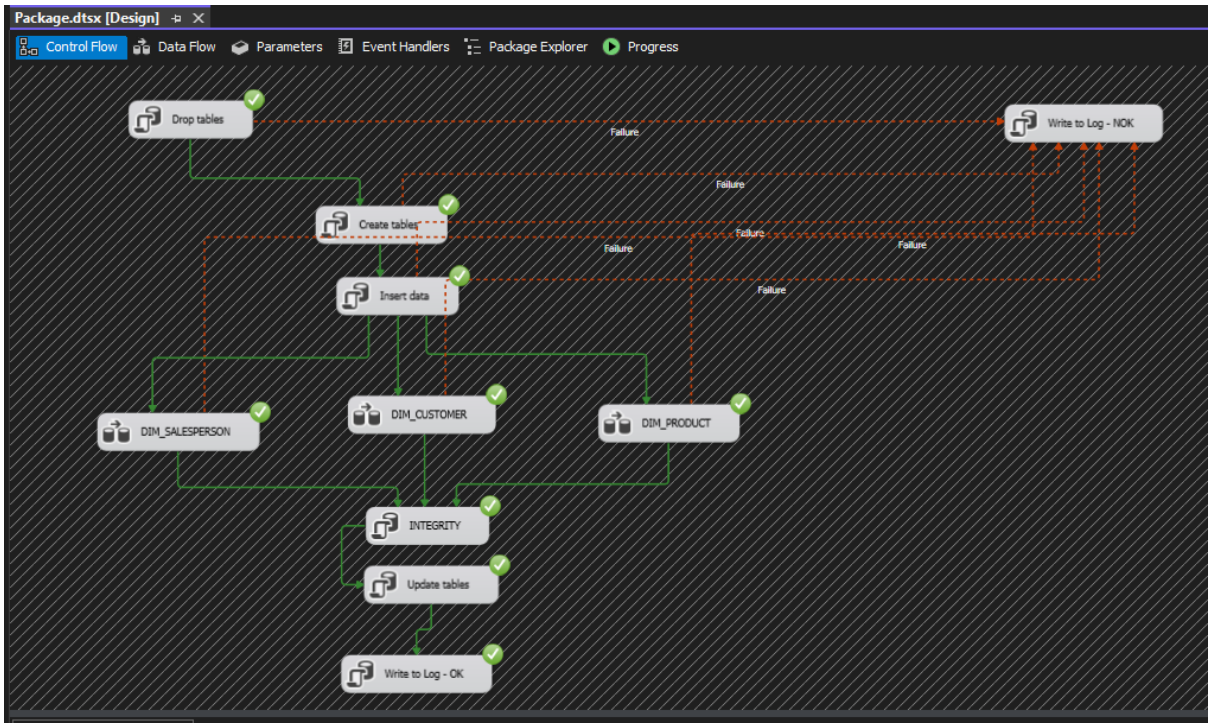
Uwagi:
Przykładowy proces:

Przykładowy przepływ danych:

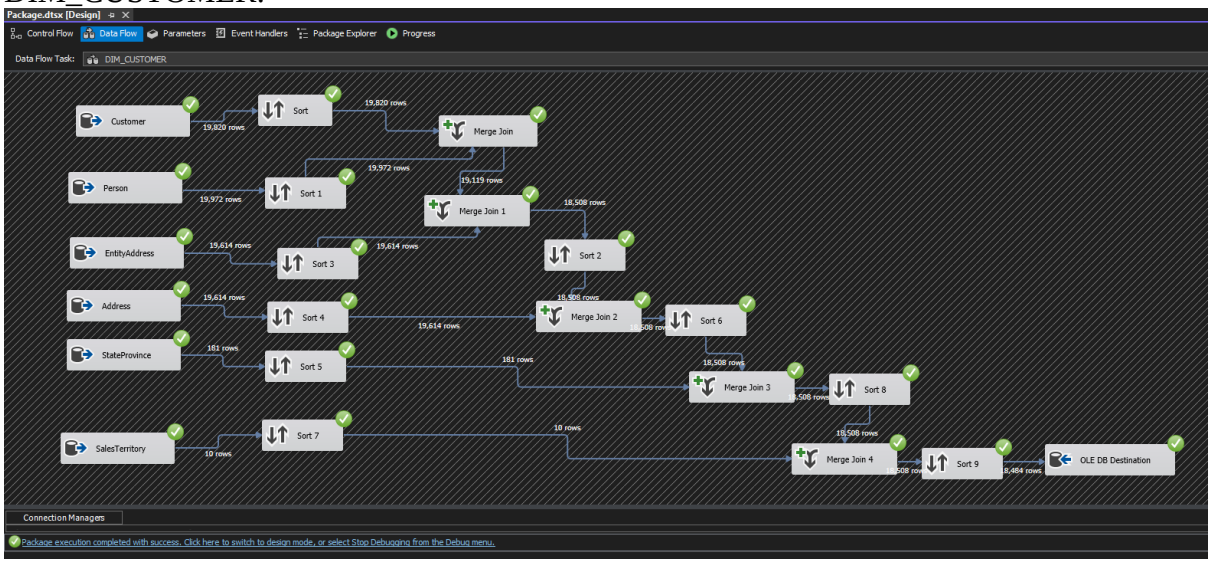


Rozwiązania:

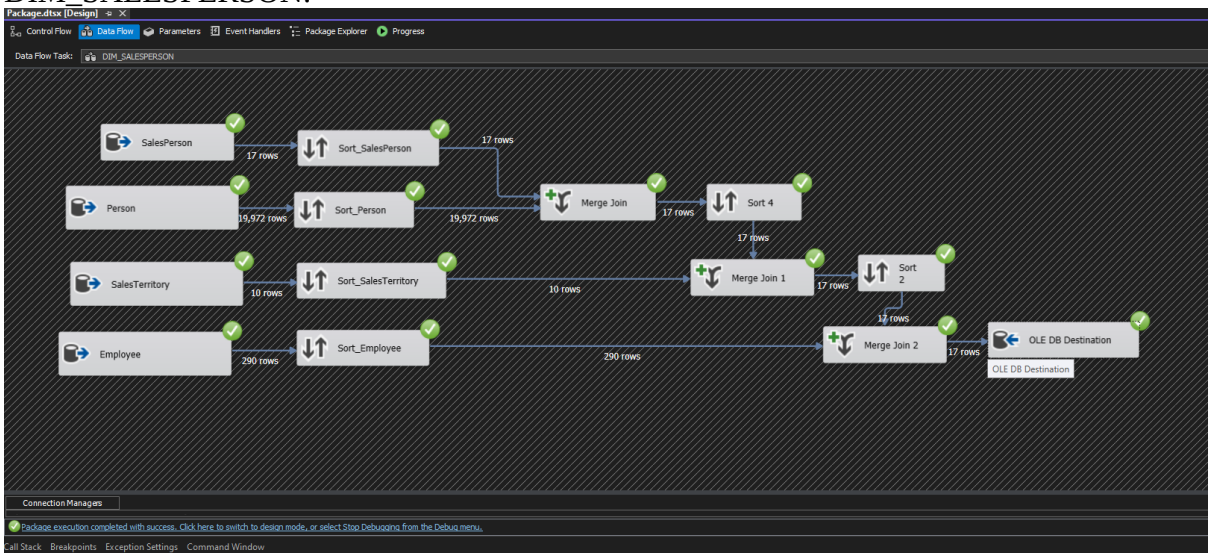
Control Flow:



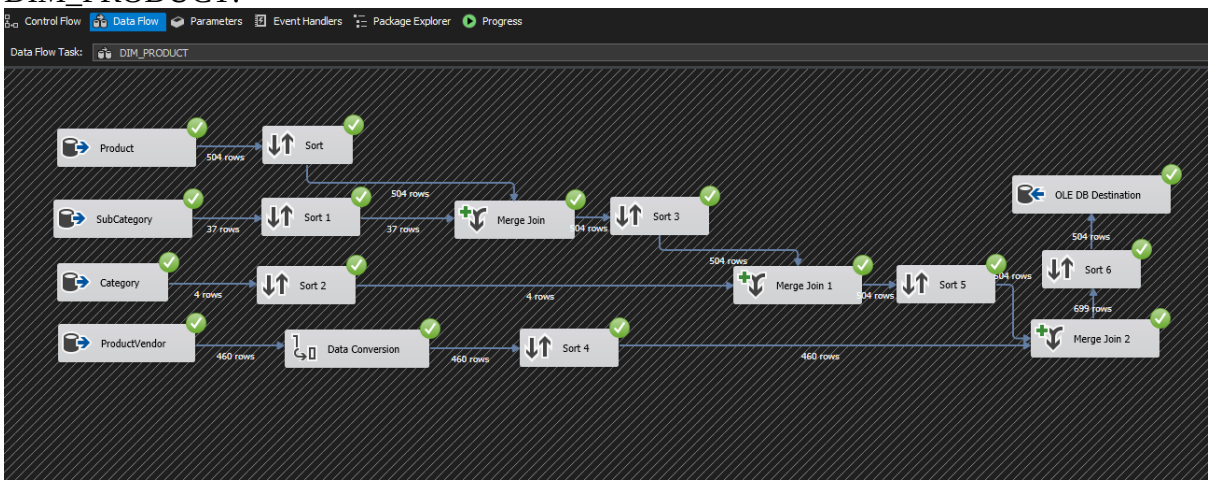
DIM_CUSTOMER:

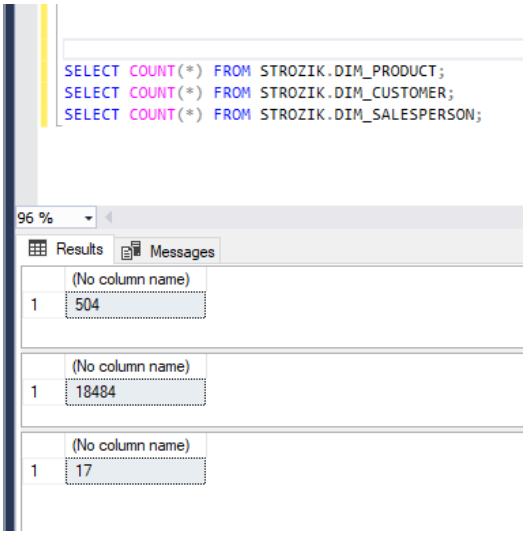


DIM_SALESPERSON:



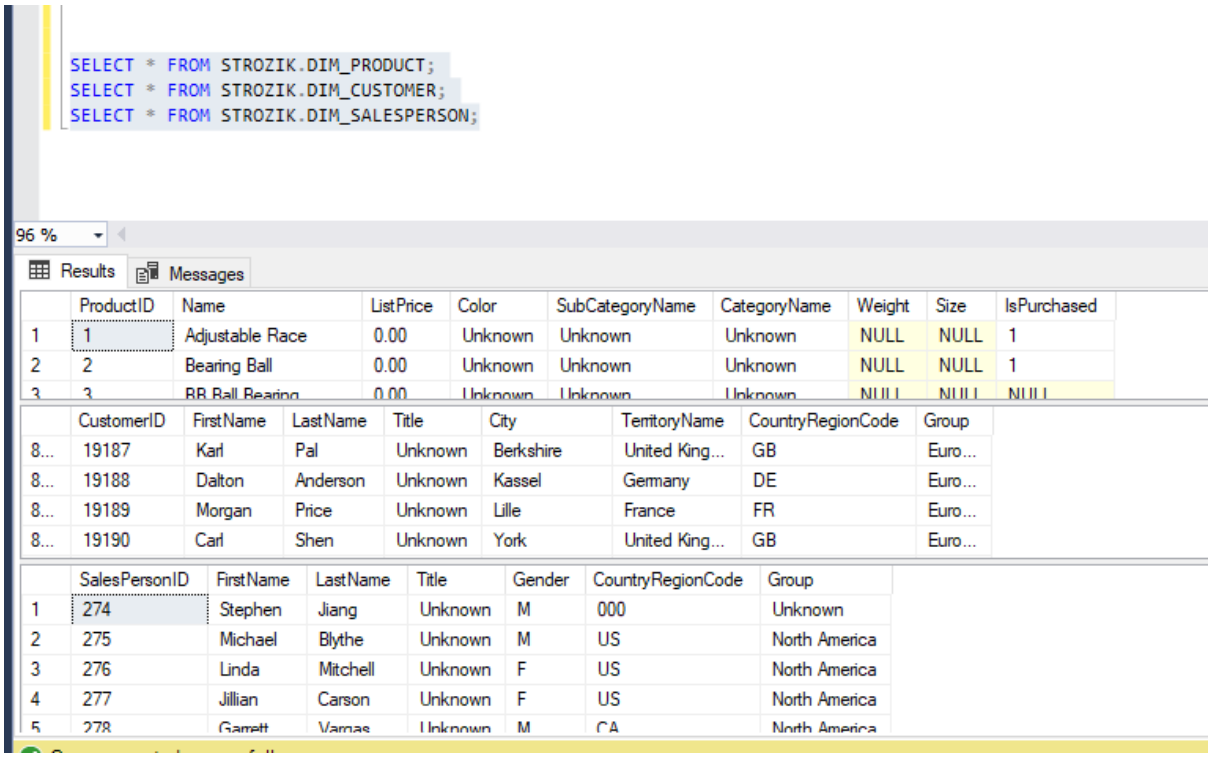
DIM_PRODUCT:





Porównujący wyniki ilości wierszy są one identyczne do wcześniejszego rozwiązania. W ramach rozwiązania dodano kolejny węzeł „Update tables“, który zmieniał dane tak jak było to wymagane w liście 6 z NULL na np „Unknown“. Gdyby proces ETL nie miał węzła „Drop tables“ to procedura ładowania danych byłaby bardziej skomplikowana a to ze względu na to

że trzeba byłoby zamieniać wszystkie kolumny o wartościach NULL, na odpowiednio „Unknown” czy „000” i porównywać rekordy z aktualnymi rekordami w tabelach przed wstawieniem nowych aby uniknąć powtórzeń.



```
SELECT * FROM STROZIK.DIM_PRODUCT;
SELECT * FROM STROZIK.DIM_CUSTOMER;
SELECT * FROM STROZIK.DIM_SALESPERSON;
```

96 %

Results | Messages

	ProductID	Name	ListPrice	Color	SubCategoryName	CategoryName	Weight	Size	IsPurchased
1	1	Adjustable Race	0.00	Unknown	Unknown	Unknown	NULL	NULL	1
2	2	Bearing Ball	0.00	Unknown	Unknown	Unknown	NULL	NULL	1
3	3	BB Ball Bearing	0.00	Unknown	Unknown	Unknown	NULL	NULL	NULL

	CustomerID	FirstName	LastName	Title	City	TerritoryName	CountryRegionCode	Group
8...	19187	Karl	Pal	Unknown	Berkshire	United King...	GB	Euro...
8...	19188	Dalton	Anderson	Unknown	Kassel	Germany	DE	Euro...
8...	19189	Morgan	Price	Unknown	Lille	France	FR	Euro...
8...	19190	Carl	Shen	Unknown	York	United King...	GB	Euro...

	SalesPersonID	FirstName	LastName	Title	Gender	CountryRegionCode	Group
1	274	Stephen	Jiang	Unknown	M	000	Unknown
2	275	Michael	Blythe	Unknown	M	US	North America
3	276	Linda	Mitchell	Unknown	F	US	North America
4	277	Jillian	Carson	Unknown	F	US	North America
5	278	Garrett	Vargas	Unknown	M	CA	North America

Na powyższym zrzucie ekranu widać, po wykonanym procesie ETL, zaktualizowane wartości wskazany kolum na odpowiednio „Unknown” i „000”.

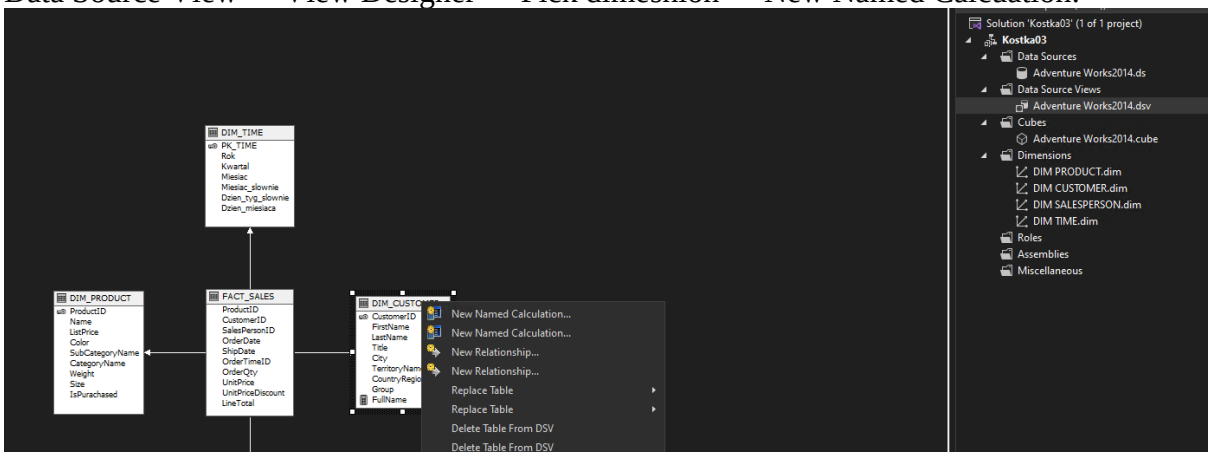
Zad. 2. Modyfikacja wymiarów i tabeli faktów

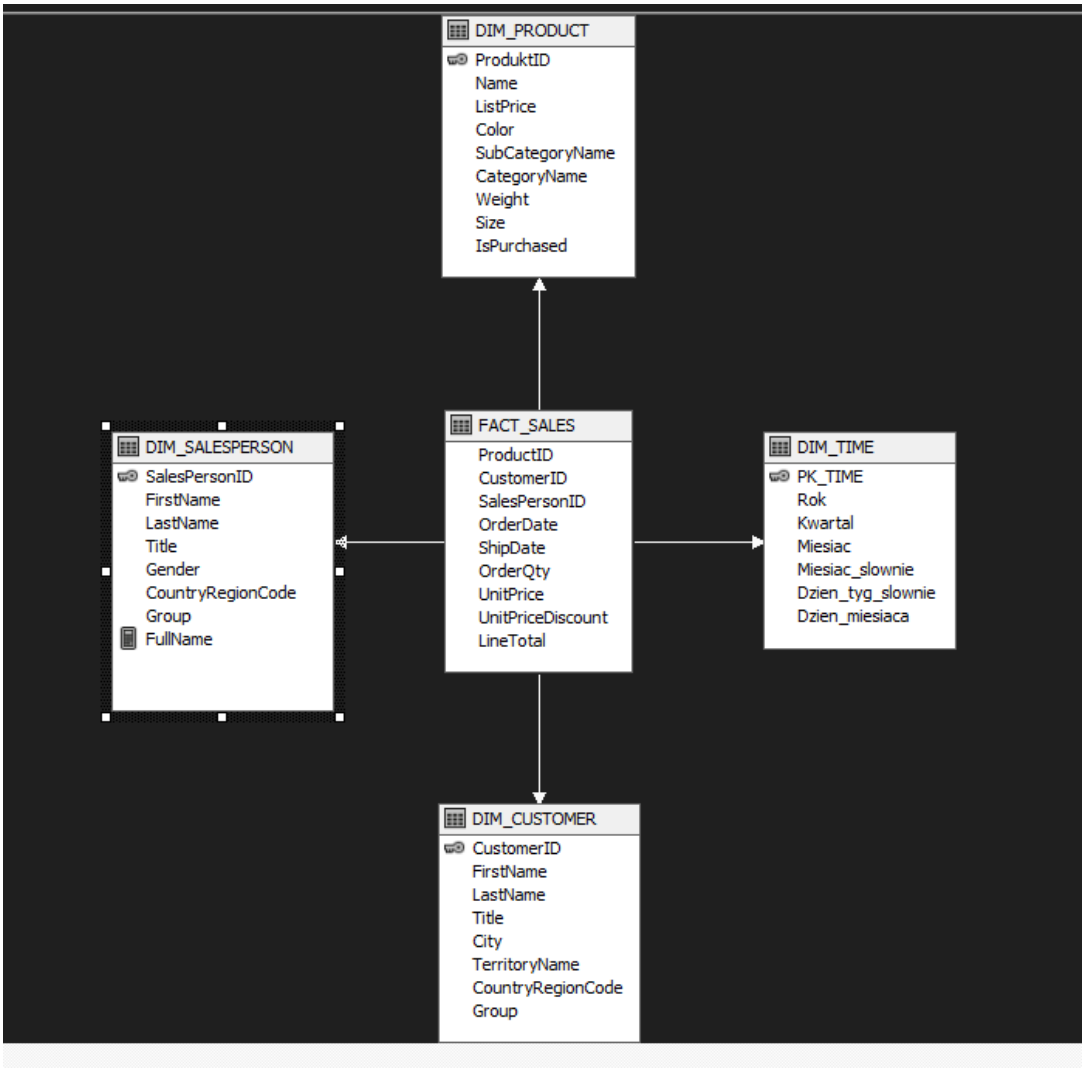
Bazując na kostce utworzonej przy realizacji poprzednich list, należy:

a) zmodyfikować definicję wymiarów tak, aby:

- w wymiarach CUSTOMER i SALESPERSON dodać wyliczany atrybut FullName (FirstName+’ ‘+LastName).

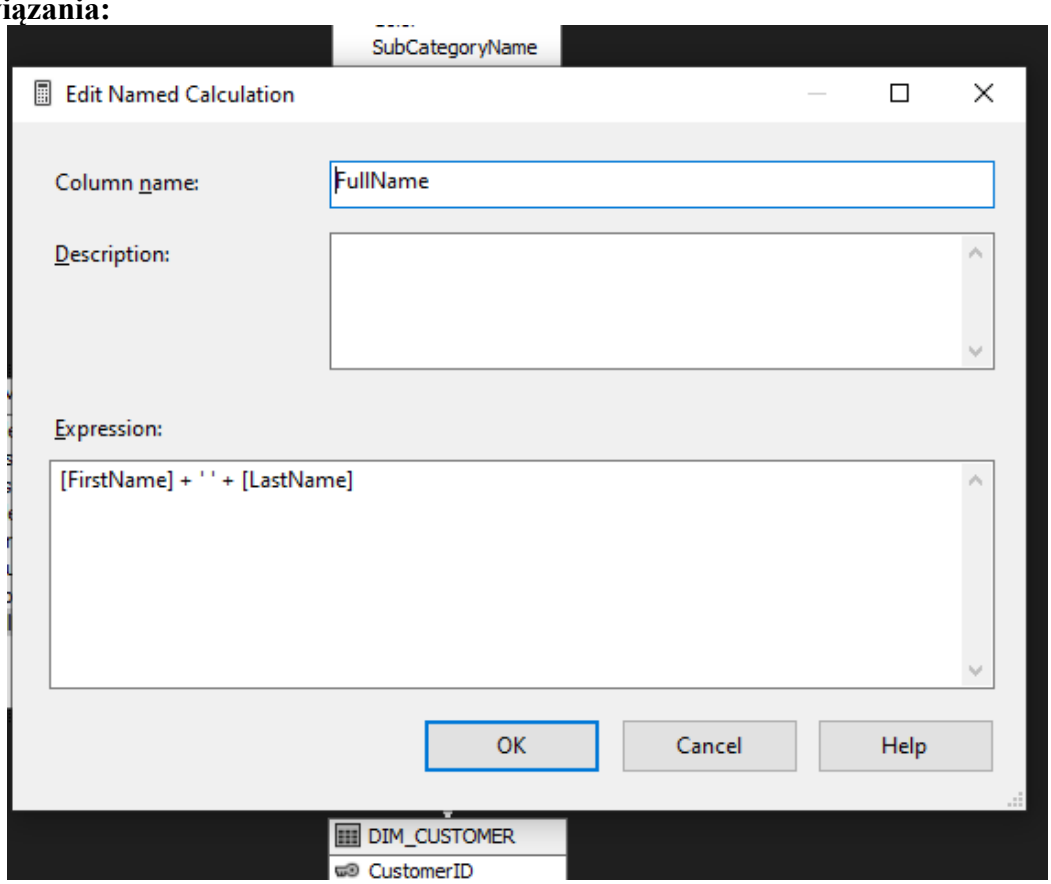
Data Source View -> View Designer -> Pick dimesnion -> New Named Calcuation:

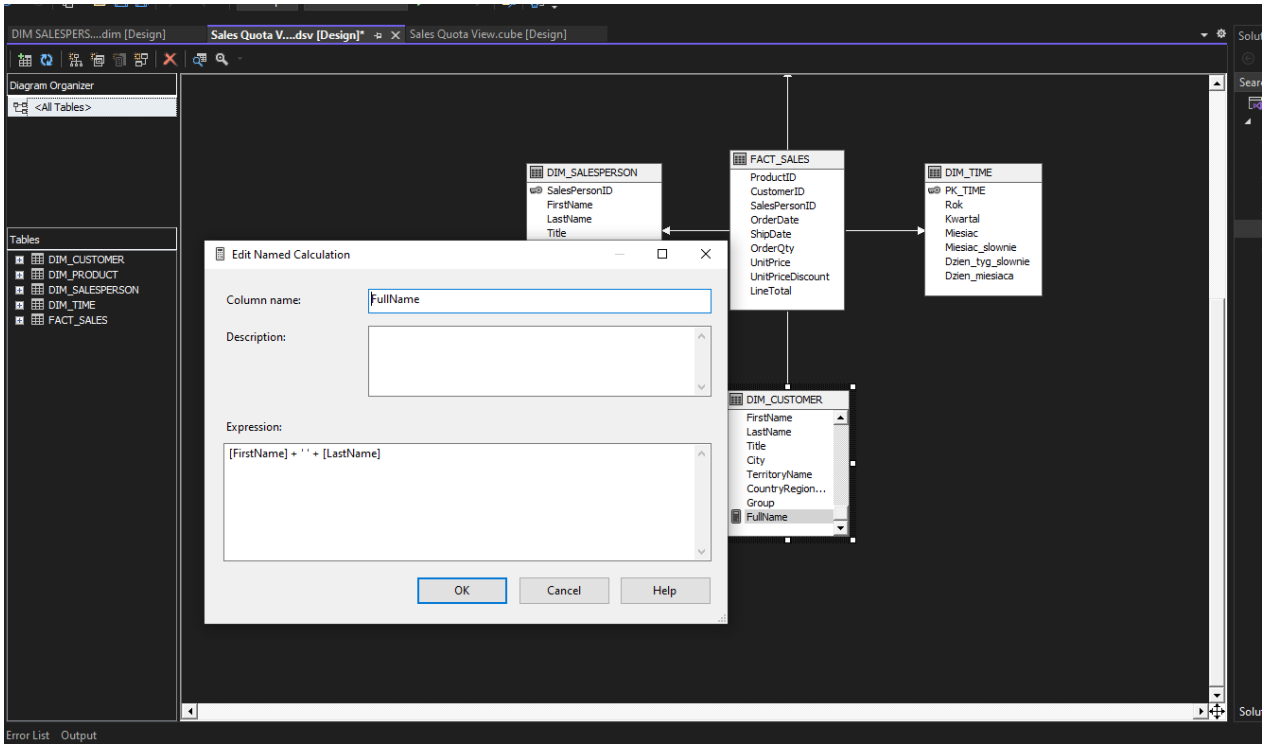




Dimension -> View Designer -> Hierchies

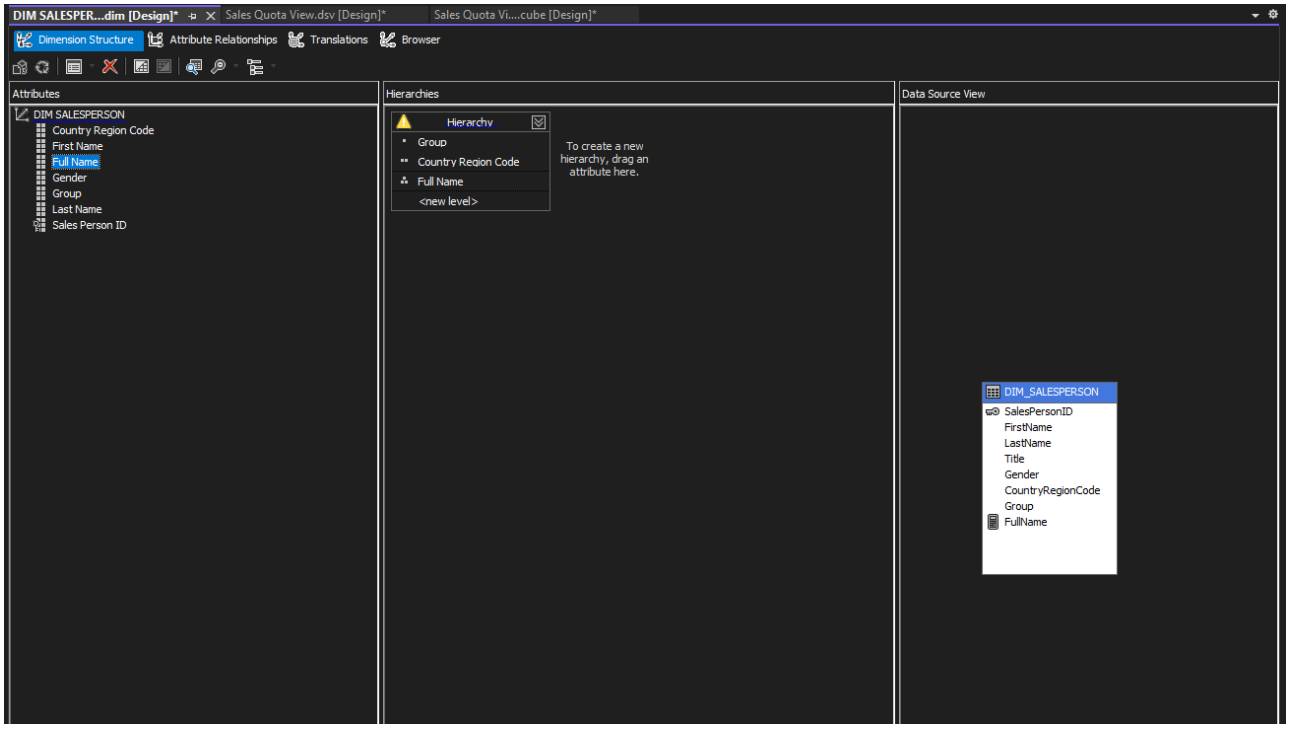
Rozwiązania:



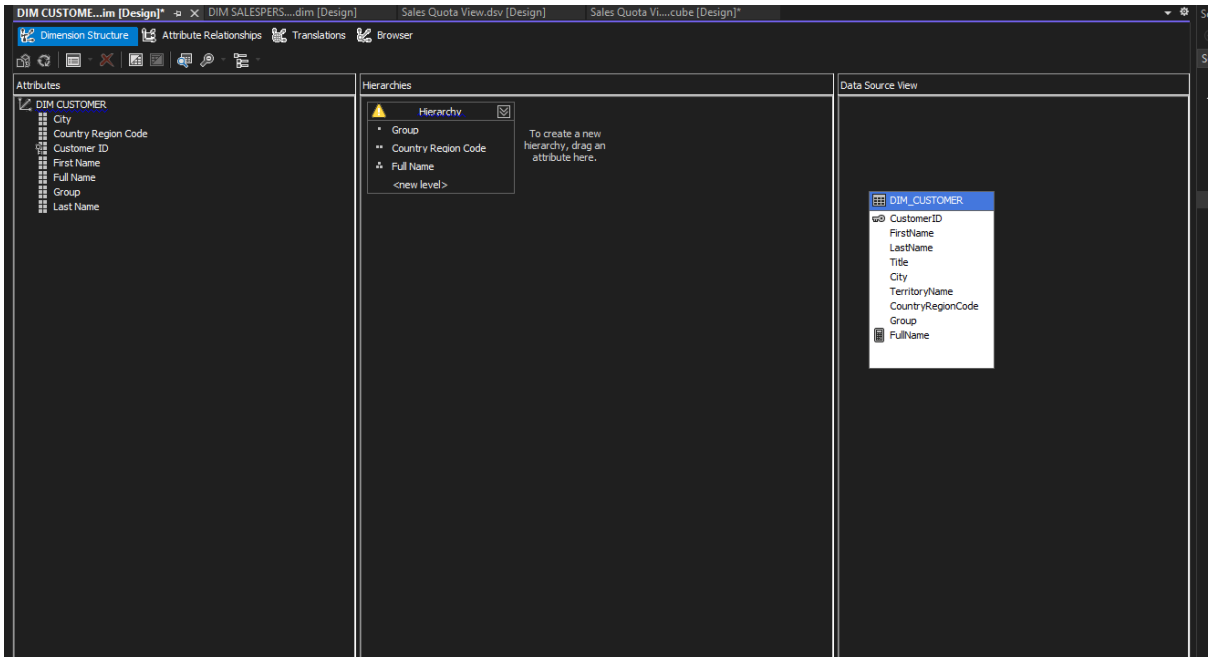


Aby utworzyć wyliczany atrybut FullName, należało odpowiednio użyć wyrażenia który konkatenował wartości dwóch kolumn FirstName i LastName, oraz dodawał jeden pusty znak między nimi. [FirstName] + ‘ ‘ + [LastName],

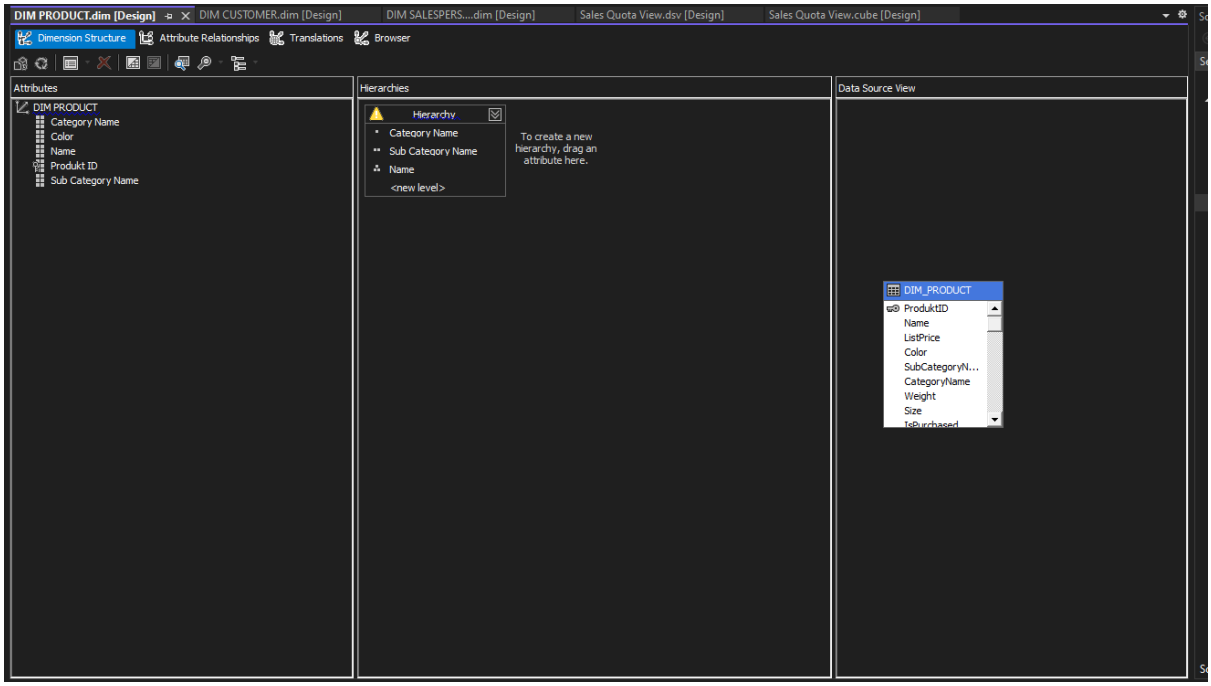
- w wymiarze SALESPERSON pojawiła się hierarchia Group – CountryRegion-Code – FullName



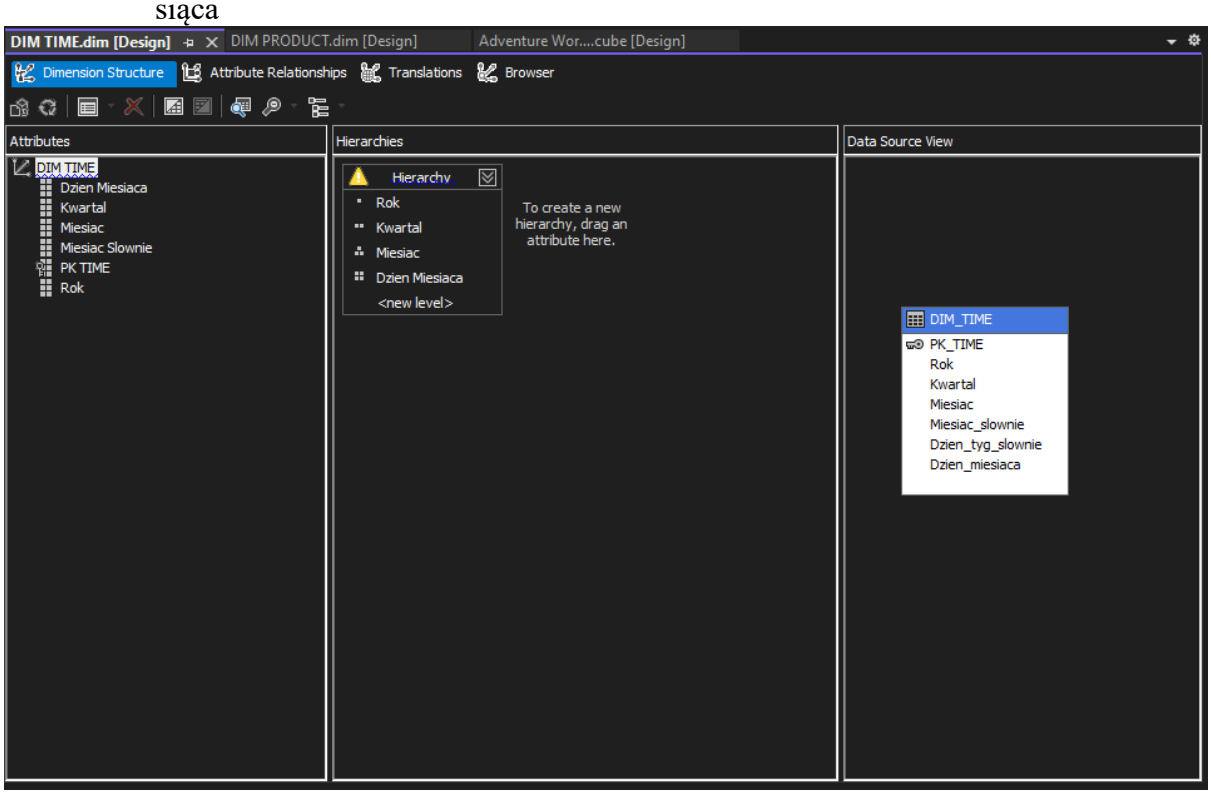
- w wymiarze CUSTOMER pojawiła się hierarchia Group – CountryRegionCode – FullName



- w wymiarze PRODUCT pojawiła się hierarchia CategoryName – SubCategory-Name – Name



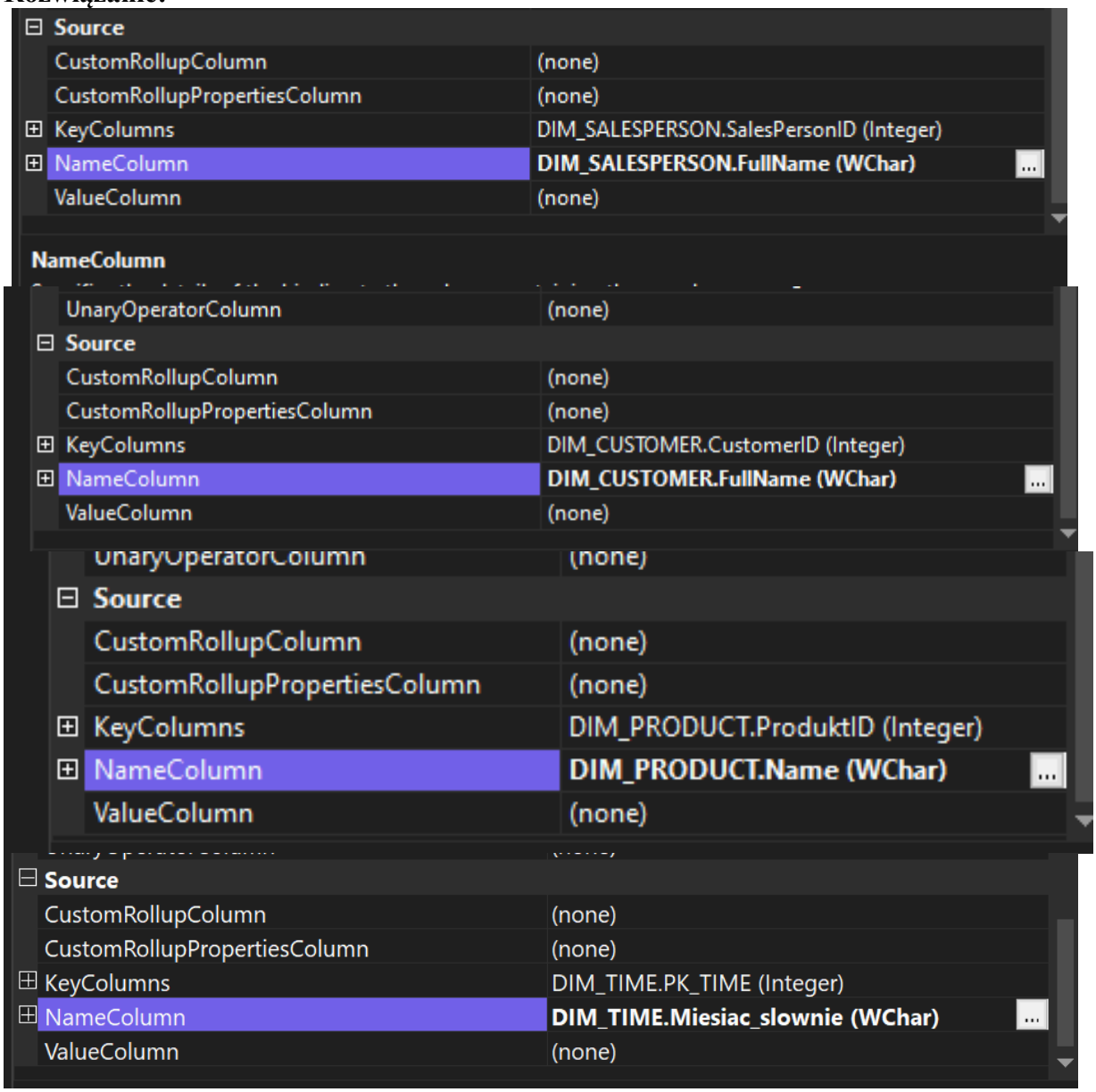
- w wymiarze TIME pojawiła się hierarchia Rok – Kwartał – Miesiąc – Dzień miesiąca



b) dla każdego atrybutu kluczowego wymiaru, którego wartościami są liczby całkowite, zmodyfikować właściwości (Properties). Zmodyfikować parametr NameColumn, tak aby nazwy kolejnych elementów wymiaru nie były liczbami. (Przykładowo dla wymiaru dotyczącego Produktu można wykorzystać atrybut Name).

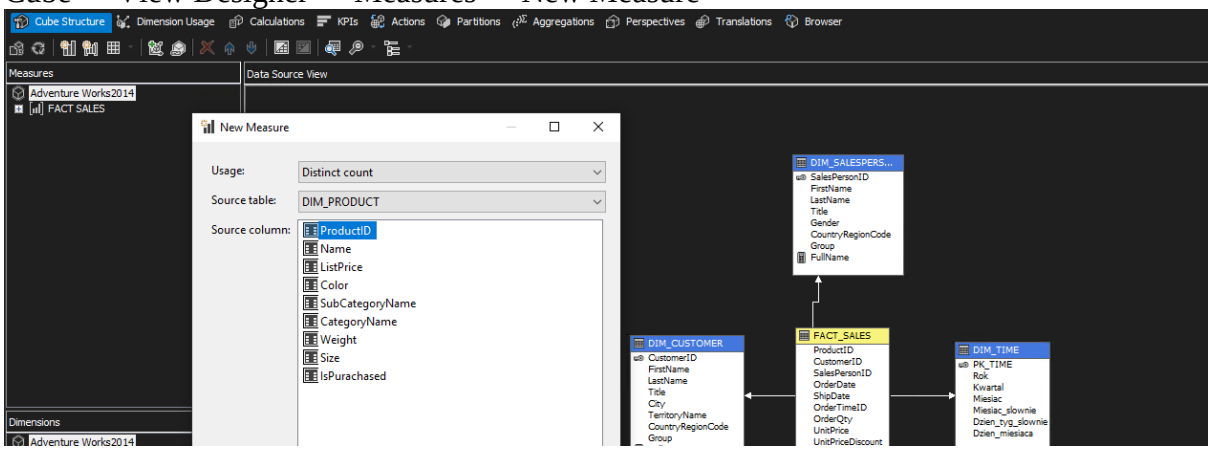
Przykładowe rozwiązanie.

Rozwiązanie:

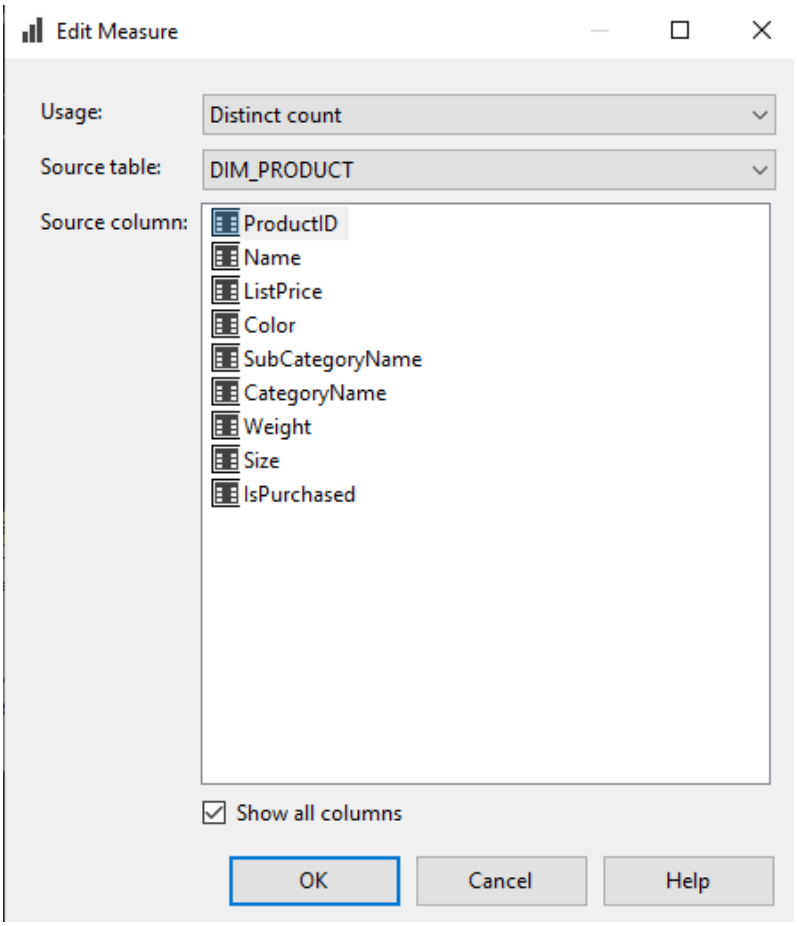


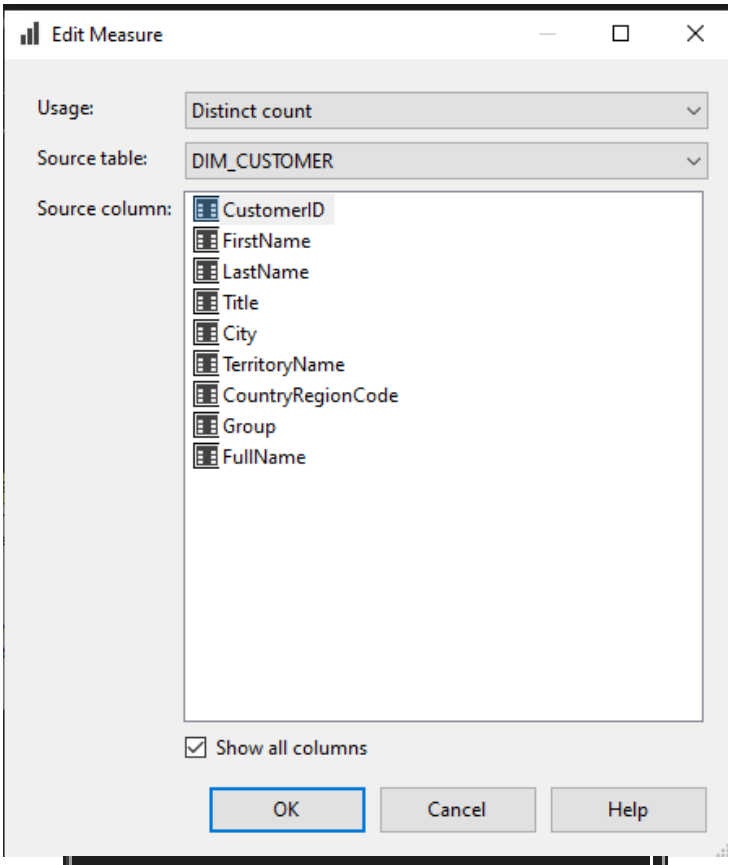
c) utworzyć nowe miary, które będą odzwierciedlać:
- Liczbę różnych klientów (aggregatedFunction: distinct count)
- Liczbę różnych produktów

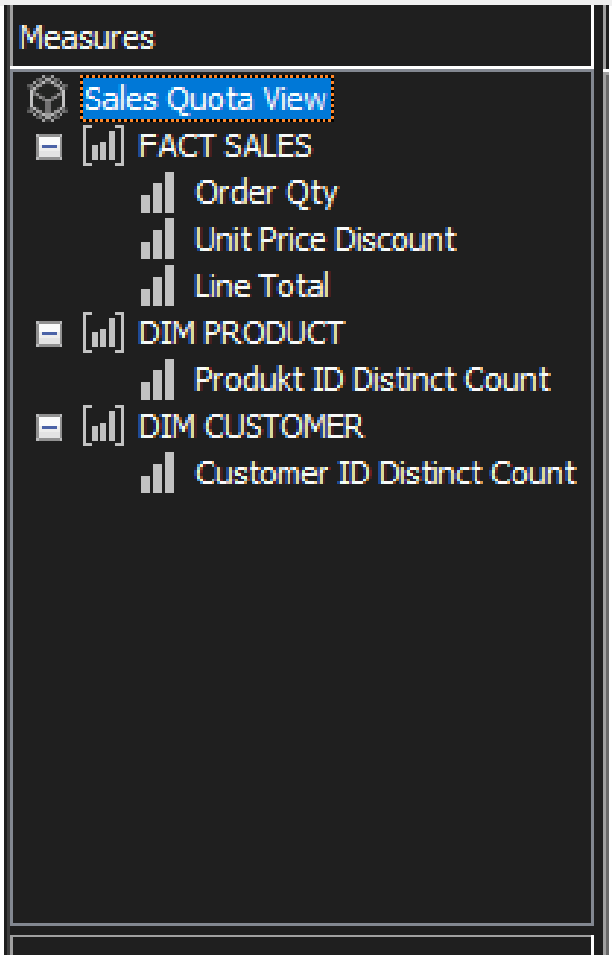
Cube -> View Designer -> Measures -> New Measure



Rozwiązanie:

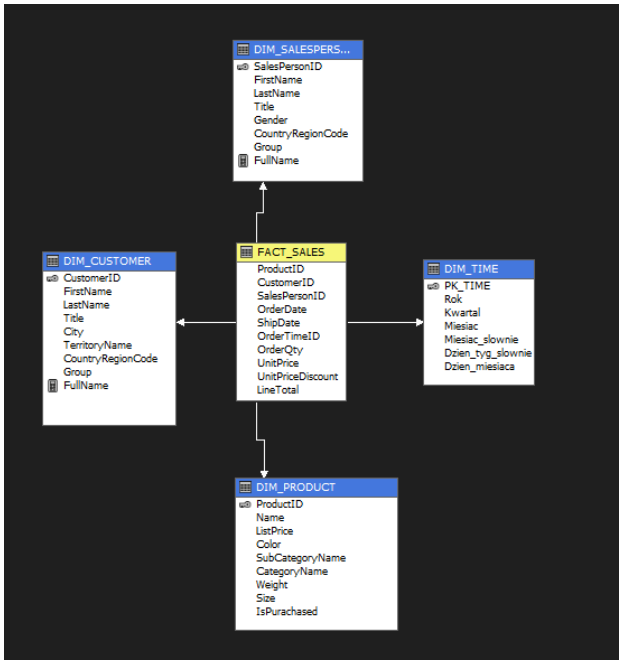






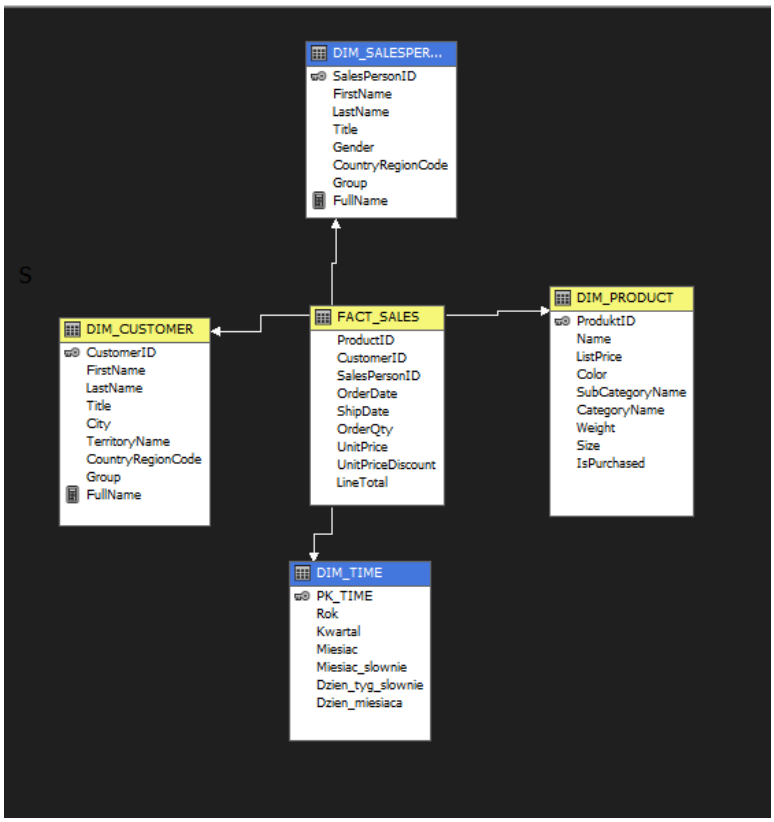
Już stworzone miary zliczajace distinct count liczby produktów i klientów.

d) wdrożyć i przeprocesować kostkę.

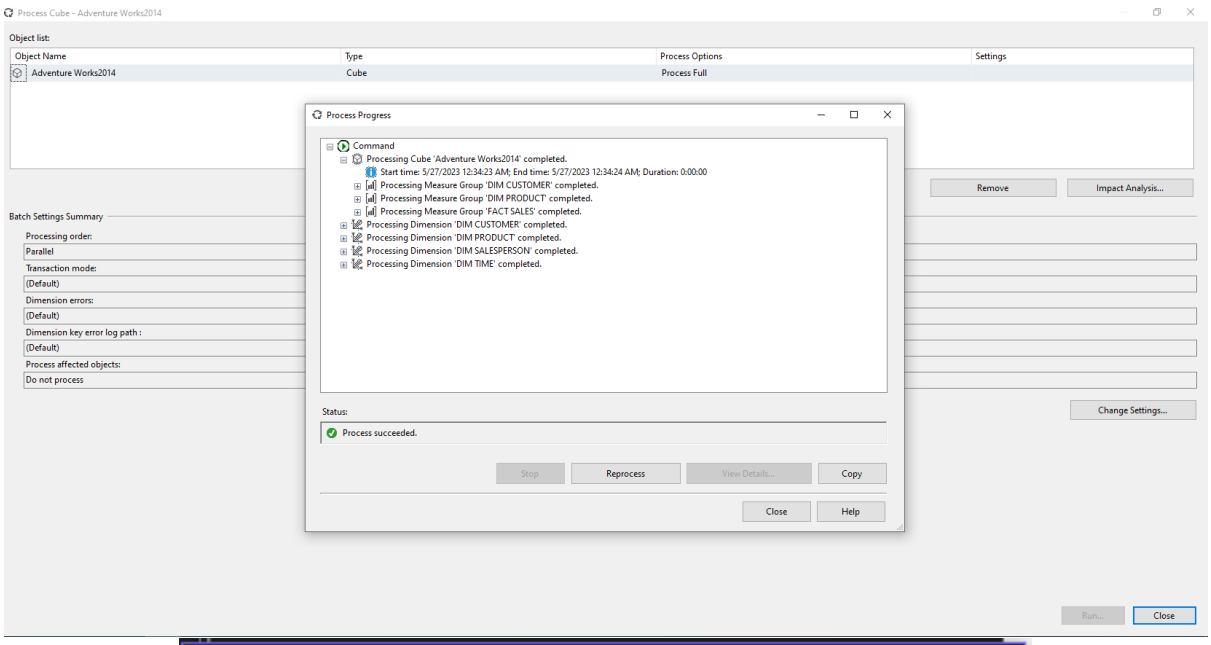


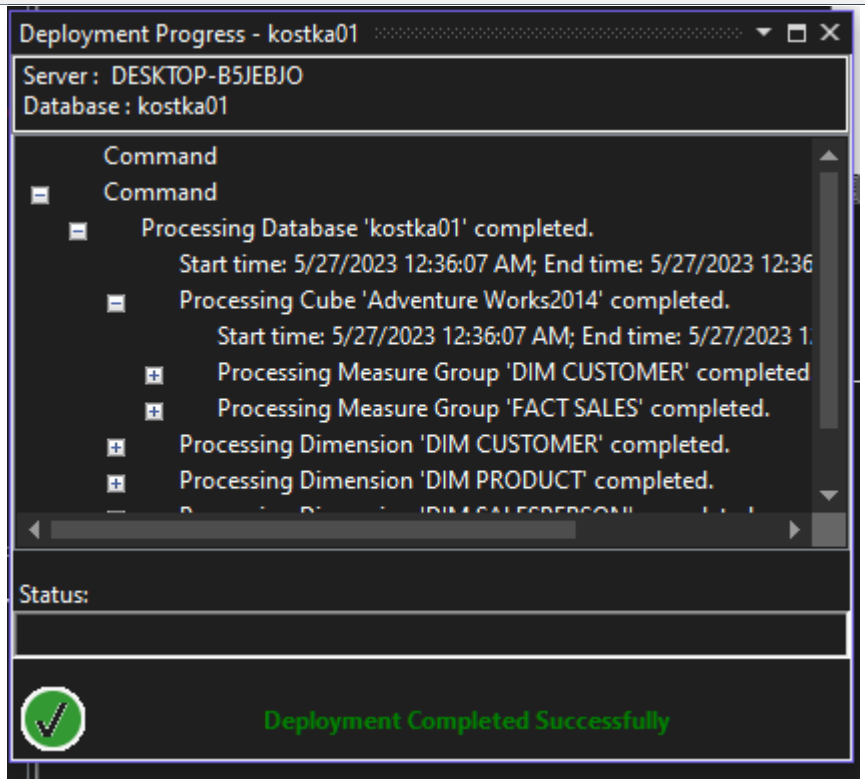
Przykładowe rozwiązanie – schemat wynikowej kostki.

Rozwiązanie:



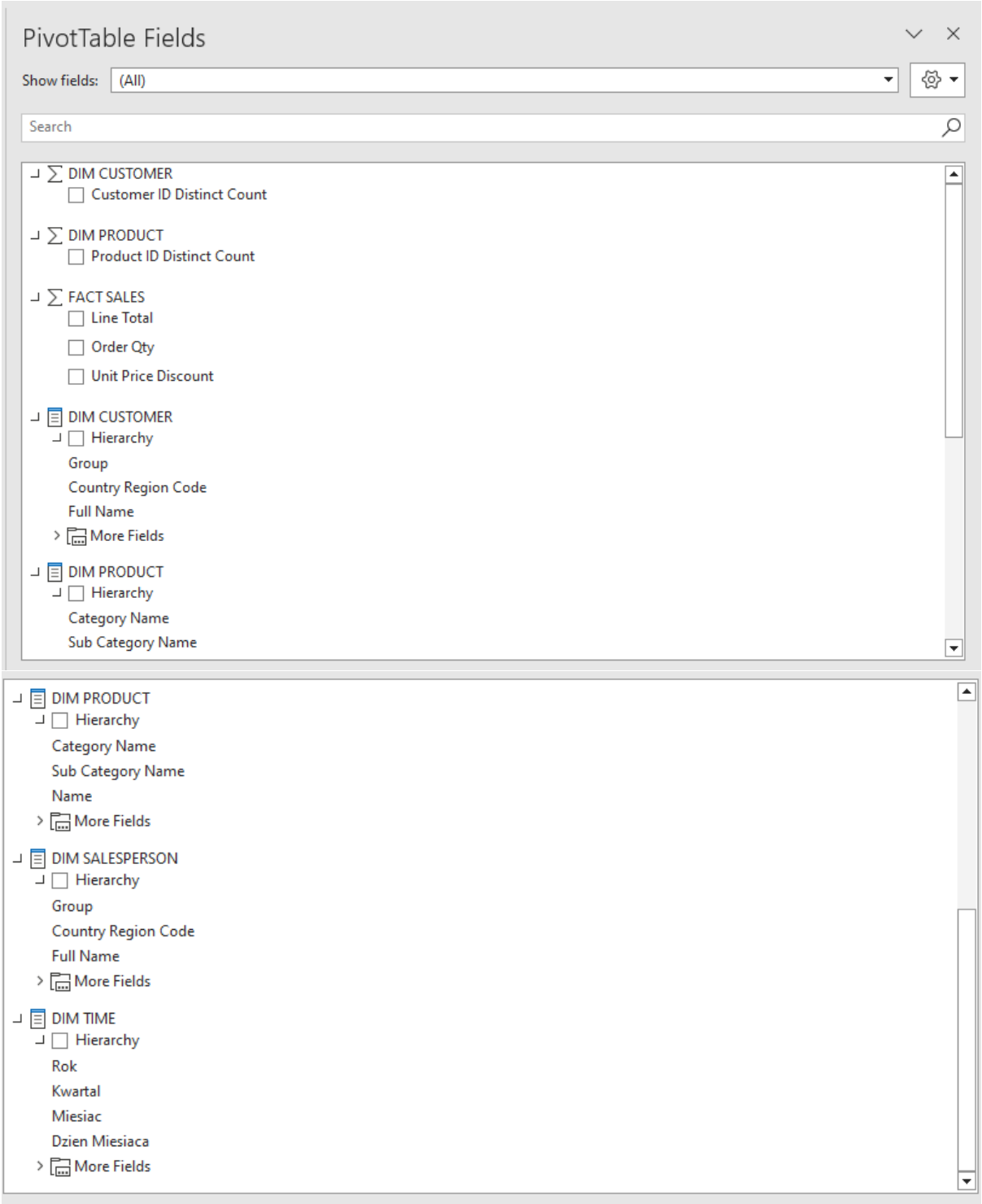
Filip Strózik 260377



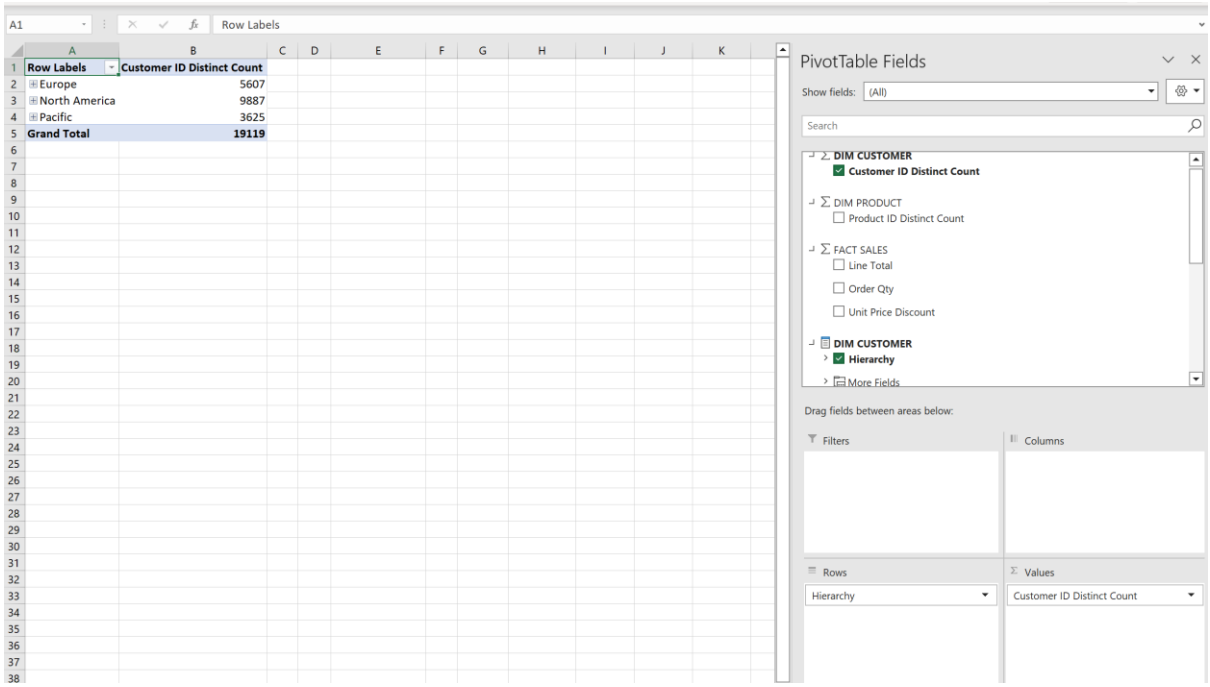


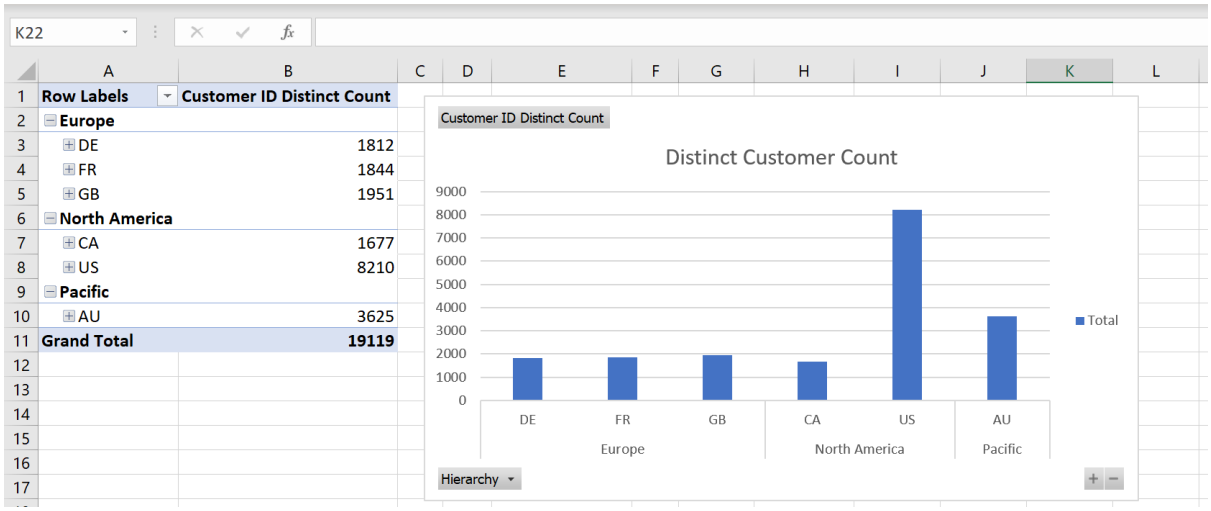
Przeprocesowanie i wdrożenie kostki na serwerze DESKTOP-B5JEBJO

Zad. 3. Przykładowe raporty
Połączyć się z programem MS Excel i przygotować 2 przykładowe raporty (tabele i wykresy przestawne), w których zostaną ujęte ciekawe zależności pomiędzy danymi.

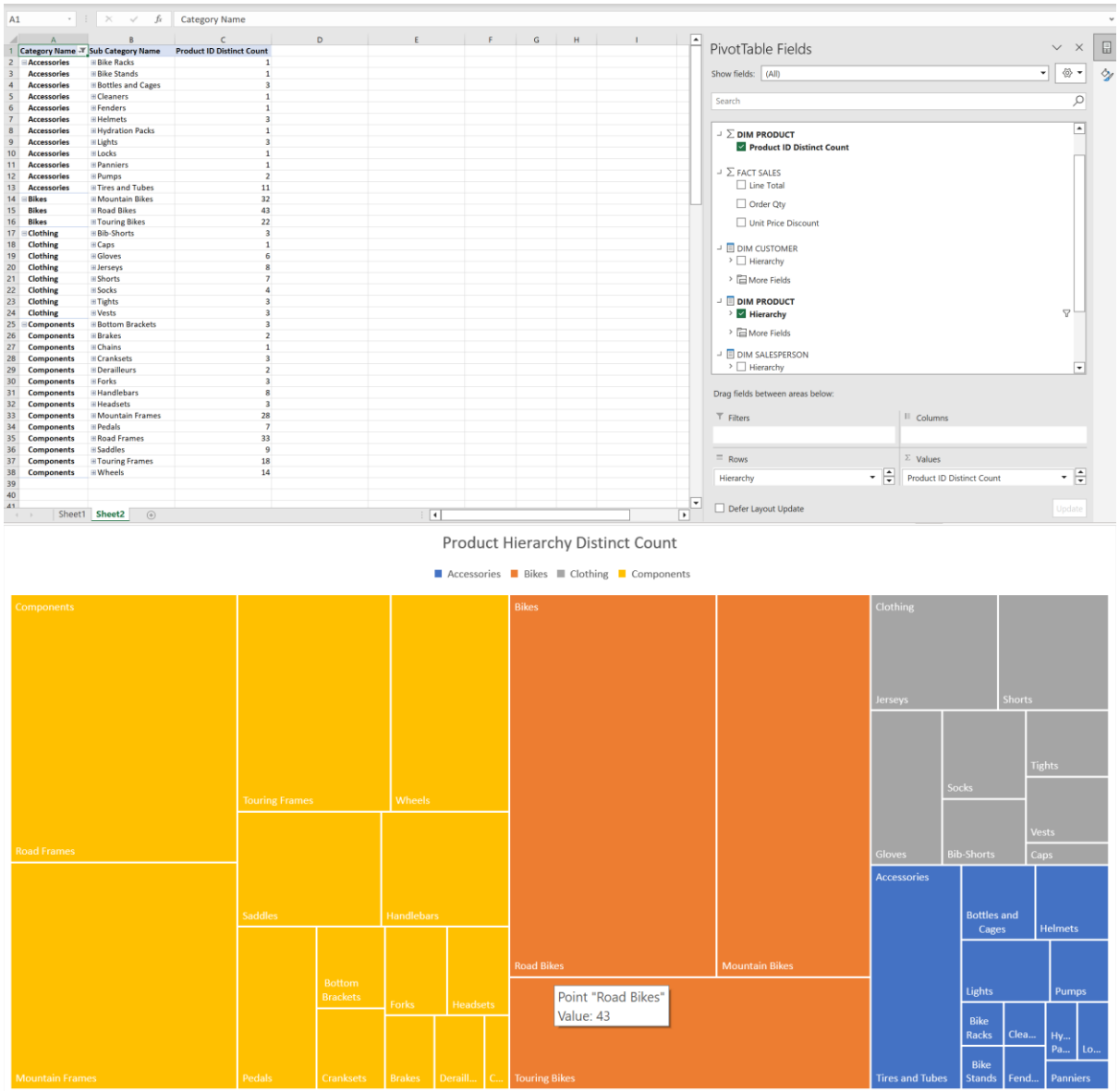


Udane połączenie z programu MS Excel do Analysys Services – wdrożonej kostki.





Pokazanie nowo wprowadzonych właściwości do kostki wydawałoby się czymś zupełnie naturalnym. W ramach tej listy wprowadziliśmy rozszerzone hierarchie dla w pewnych wymiarów oraz dodatkowe kalkulowane miary w projekcie kostki. Aby pokazać nowo wprowadzoną hierarchię dla wymiaru DIM_CUSTOMER oraz dodatkową miarę dotyczącą klientów zdecydowano pokazać ciekawą zależność na powyższym wykresie. Dzięki nowo wprowadzonej wymiarze oraz hierarchii możemy łatwo Zauważyć do jakiej grupy Części świata należą badane regiony, widzimy te które znajdują się w Europie, Ameryce Północnej oraz w regionie Oceanu Spokojnego. Daje to jakąś referencje oraz dodatkową informację o tym jak interpretujemy zgromadzone dane. Przydając samą wiedzę o hierarchii zauważy połączenia faktów które stawiają wyniki w innym świetle.



W tym przypadku decyzja a o wykorzystaniu akurat tego typu wykresu i nowo wprowadzonych właściwości do kostki nie jest przypadkowa. Wykres typu „TreeMap” może świetnie sprawdzić się do przedstawiania danych, które można hierarchizować. Mamy tu do czynienia również z wykorzystaniem miary kalkulowanej w kostce oraz z hierarchią wymiaru DIM_PRODUCT, z którego akurat użyliśmy kategorii i podkategorii produktu. Można zaznaczyć, że TreeMap jest szczególnie dobrym rozwiązaniem, ponieważ pozwala na efektywne wykorzystanie przestrzeni wykresu, na zauważenie wyraźne hierarchii oraz na możliwość porównywania, ponieważ kafelki są proporcjonalne co do wartości rozmiarem. Niestety z PivotTable nie da się od razu stworzyć wykresu TreeMap ale po kilku prostych krokach jest to już możliwe. Na wykresie możemy dobrze się zaznajomić z ilością różnych produktów w danej podkategorii w danej kategorii a użycie akurat tej reprezentacji pozwala na lepszą i bardziej „porównywalną” interpretację wyników.

Wnioski:

Zauważyłem, że rozbicie poleceń SQL na odpowiednie węzły Data Flow w ramach procesu ETL może być skomplikowane, ponieważ aby uzyskać ten sam efekt, trzeba wykonać o wiele więcej kroków niż w przypadku wykreowania tego samego wyniku za pomocą zapytań SQL.

Kolejnym wnioskiem może być to, że mimo trudności to można odczuć, że się ma większą kontrolę podczas tworzenia całego procesu. Opakowanie poleceń SQL w węzły oraz wszelka walidacja z ciągłym połączeniem do bazy może prowadzić do rzadszych błędów podczas np. wypełniania tabel danymi. Użytkownik podejmuje tylko konieczne decyzje, większość dzieje się automatycznie np mapping pól. W ramach walidacji są sprawdzane typy danych co pozwala na wykrycie błędu i uniemożliwienie na częściową realizację procesu ETL. Dodatkowo podczas uruchomienia procesu ETL w ramach sesji debugowej, widać ile wierszy zostało przetworzonych na każdym węźle co daje lepsze pojęcie o stanie danych podczas uruchomienia procesu.

Problematycznym było wybieranie tabel z bazy, które jeszcze nie istniały w bazie, ponieważ miały być utworzone w ramach jednego z pierwszych kroków procesu, nie wiadomo co trzeba było zmienić, aby Visual Studio pozwoliło na wybranie takiej tabeli. Problem został rozwiązany poprzez uruchomienie tylko i wyłącznie jednego kroku aby tabele rzeczywiście stworzyć, po czym już walidacja Visual Studio pozwalała na uruchomienie całego procesu. Daje to wrażenie, że ETL wcale nie jest tak zautomatyzowany jak jest opisywany. Być może jest to kwestia dobrej konfiguracji projektu.

Duży wpływ według mnie miały zmiany wprowadzone do kostki danych. Dodatkowe hierarchie, pola wymiarów oraz miary sprawiły, że kostka oferowała możliwości bardziej szczegółowej analizy danych lub nadawały pewnym zestawieniom inny sens lub dodatkowe podłoże informacji. Odpowiednio wykorzystując hierarchie można zawrzeć na wykresach informacje, które wcześniej nie były widoczne gołym okiem, pozwala to na pewną granulację i grupowanie danych, która ma świetnie zastosowanie w przypadku porównywania danych ze sobą.

Projekt typu Analysis Services umożliwia wygodne wybieranie, przetwarzanie i prezentowanie danych. Daje możliwość modyfikacji przedstawionych danych bez konieczności ingerencji w bazę danych, dzięki funkcji Named Calculations oraz tworzeniu hierarchii, co ułatwia dalsze precyzowanie danych, na przykład na wykresach. Nie wymaga pisania zapytań do samej bazy w celu utworzenia tabel przestawnych, które prezentują potrzebne dane.

Uwaga!

- Sprawozdanie, bez wniosków podsumowujących aspekt zagadnień analizowanych na zajęciach laboratoryjnych i zawartych w sprawozdaniu, jest automatycznie oceniane negatywnie!