Spis treści

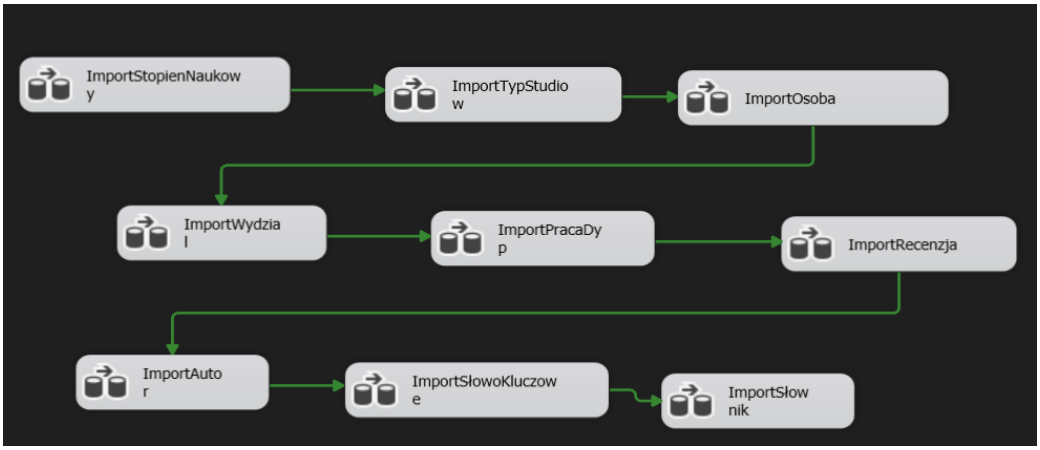
1. Uruchomienie bazy danych Prace Dyplomowe

Aby utworzyć bazę danych należało uruchomić program SQL Server management studio, następnie połączyć się z odpowiednim serwerem, w moim przypadku był to DESKTOP-H7RIFKS. Tworzenie Bazy danych polegało na zastosowaniu SQL, mianowicie użycie wyrażenia CREATE DATABASE, a następnie użycie dostarczonego pliku który stworzył tabele oraz klucze główne jak i obce.

2. Proces ETL do bazy danych

Proces ETL to skrót od słów Extract, Transform, Load, co oznacza proces pozyskiwania danych z różnych źródeł, transformacji ich na odpowiedni format oraz ładowania ich do bazy danych. W kontekście bazy danych Prace Dyplomowe, proces ETL został użyty do zaimportowania danych z plików flat file do bazy danych. Proces ETL zaczynał się od etapu ekstrakcji danych, czyli pobierania danych z plików. Następnie dane były przetwarzane, czyli transformowane do formatu, który można łatwiej przetworzyć i umieścić w bazie danych. W końcowym etapie, dane były ładowane do bazy danych, gdzie można było je wykorzystać w dalszych analizach. Proces ETL jest kluczowym etapem w tworzeniu i zarządzaniu danymi w bazie danych, ponieważ umożliwia zintegrowanie danych z różnych źródeł i umieszczenie ich w spójnej strukturze, co ułatwia dalsze przetwarzanie i analizowanie danych.

Schemat jakiego użyłem aby zaimportować odpowiednie dane do bazy danych przedstawiony jest poniżej:



Rysunek 1 Diagram ETL do bazy danych

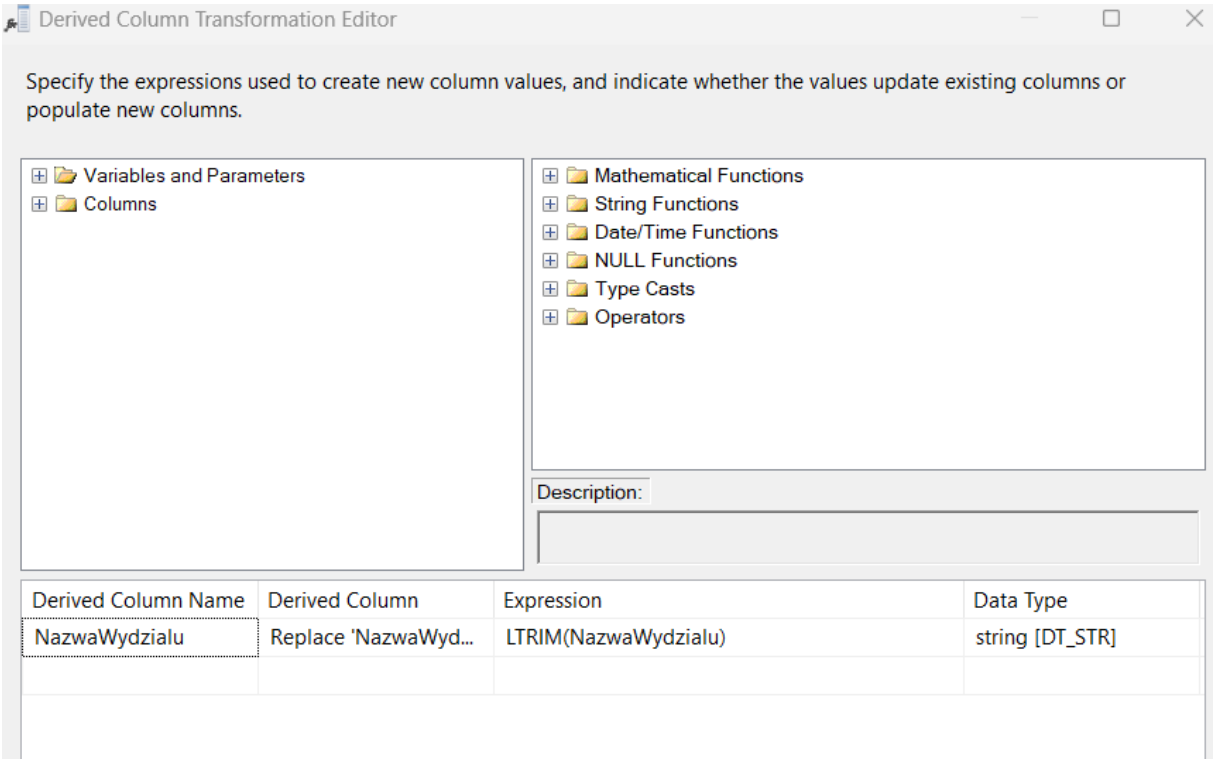
Najprostrze zadania zawierają w sobie dwa komponenty:

- Flat file source służy on do wczytania danych z plików płaskich, czyli w moim przypadku dostarczonego pliku .txt.

MODELOWANIE PROCESÓW BIZNESOWYCH

- SQL Server Destination służy on do zapisania wczytanych wcześniej danych w odpowiednich kolumnach w bazie danych.

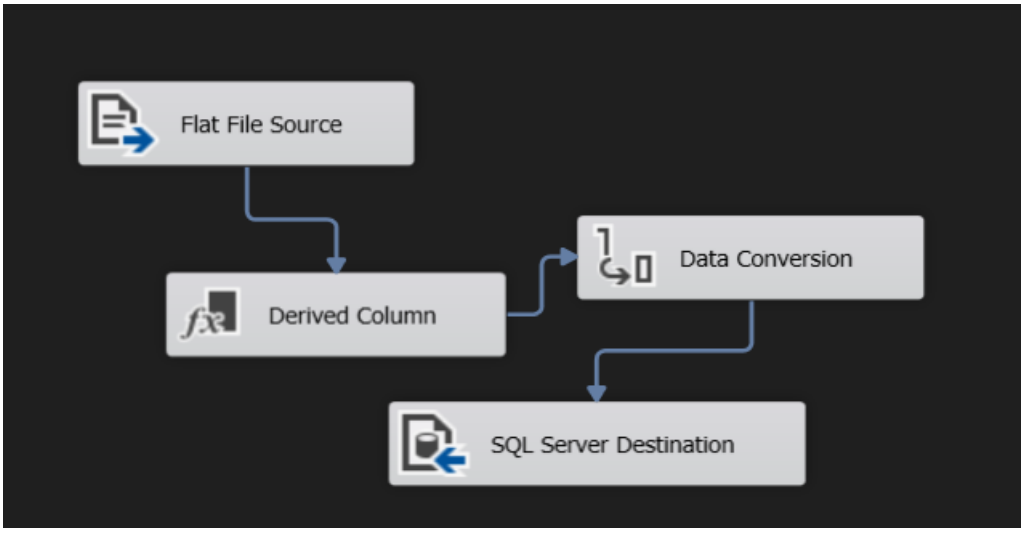
W niektórych dostarczonych plikach tekstowych znajdowały się błędy, mianowicie podczas importu danych z pliku Wydzial.txt, trzeba było zastosować komponent derived column transformation a w nim wyrażenie LTRIM() aby pozbyć się spacji które były przed nazwą wydziałów.



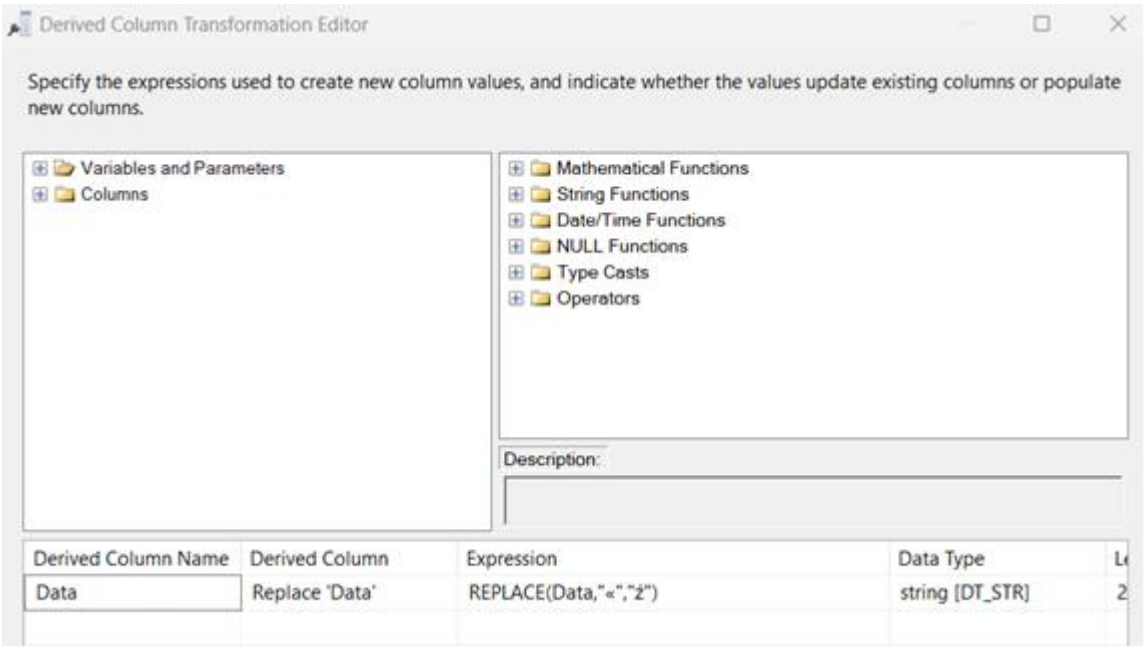
Rysunek 2 Rozwiązanie błędu w Nazwie wydziału

Podczas importu pliku Praca dyplomowa należało użyć, aż dwóch komponentów pomiędzy pobraniem danych, a ich zapisaniem w bazie danych. Derivied Column służył do zmiany nieprawidłowego znaku w kolumnie Data, zamiast „ź” w orginalnym dokumencie był znak “«”. Zaraz po Derivied Column użyłem Data Conversion do zmiany typu kolumny z orginalnego typu String na typ date, a dokładnie timestamp.

MODELOWANIE PROCESÓW BIZNESOWYCH

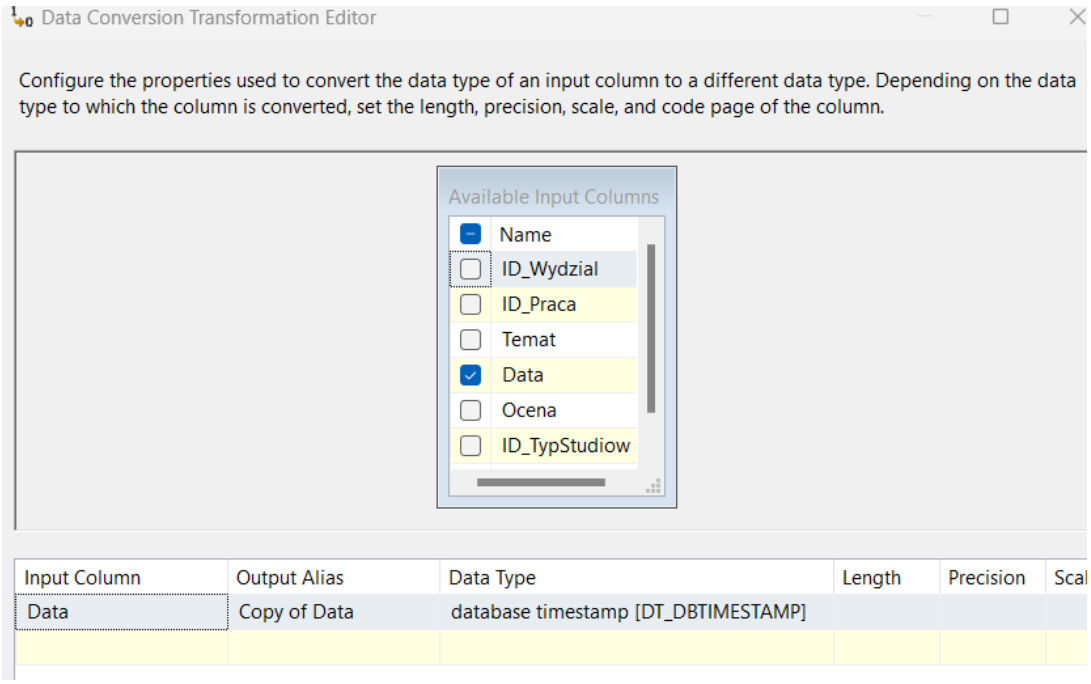


Rysunek 3 Użyte komponenty w imporcie Prace Dyplomowe



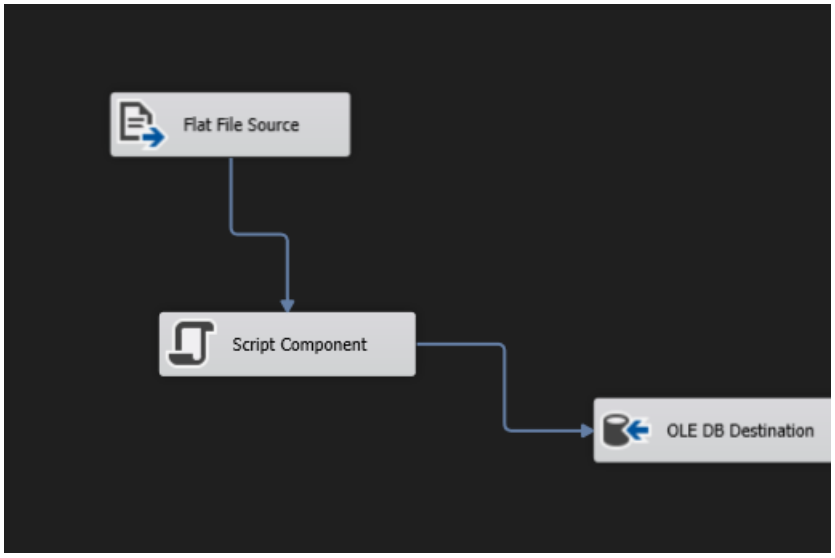
Rysunek 4 Zamiana znaków w imporcie Prace Dyplomowe

MODELOWANIE PROCESÓW BIZNESOWYCH



Rysunek 5 Zmiana typu danych w imporcie Prace Dyplomowe

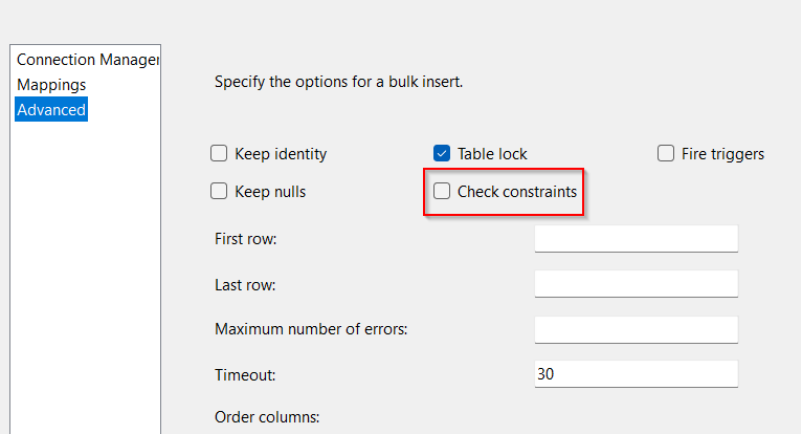
Następną zmianą jaką trzeba było zastosować to dodanie Script component przy imporcie danych z pliku Osoba.txt. W tym pliku nie wszystkie wartości w kolumnie Stopien były wpisane oraz brakowło wartości w kolumnie imie, aby uniknąć błędów zastosowałem skrypt. Nowy komponent wpisywał wartość 1 do kolumny stopien, a ta 1 oznaczała brak stopnia naukowego.



Rysunek 6 Użyte komponenty w imporcie Osoba

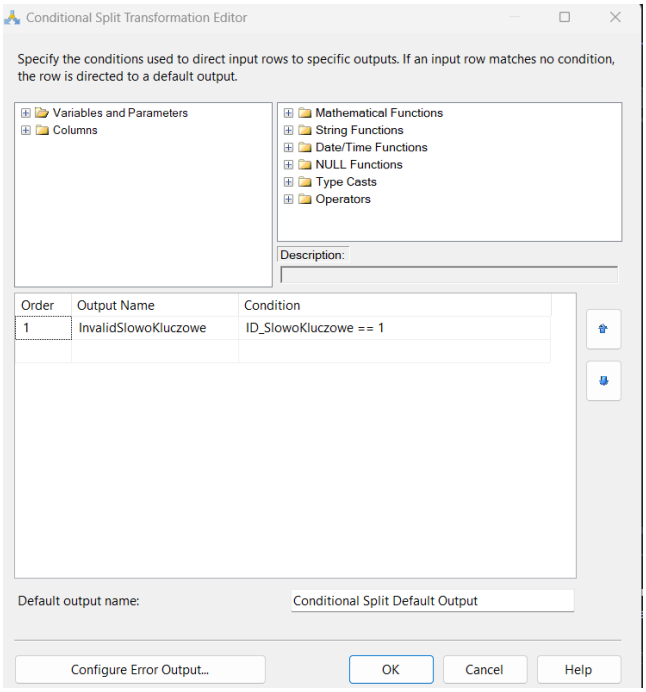
MODELOWANIE PROCESÓW BIZNESOWYCH

Aby uniknąć problemu z brakującymi ID osób którzy są w pliku Autor.txt ale nie ma w Osoba.txt, należało odznaczyć opcję check constraints w komponencie Sql Server Destination



Rysunek 7 Odznaczenie check constraints

Ostatni błąd znajował się w pliku Slownik.txt, który zawierał wiersz odwołujący się do rekordu SlowoKluczowe o identyfikatorze ID = 1, a tego rekordu nie było w pliku SlowoKluczowe.txt, dlatego podczas import należało pominąć wiersze z ID_SlowoKluczowe==1. Wykorzystałem do tego komponent ConditionalSplit.



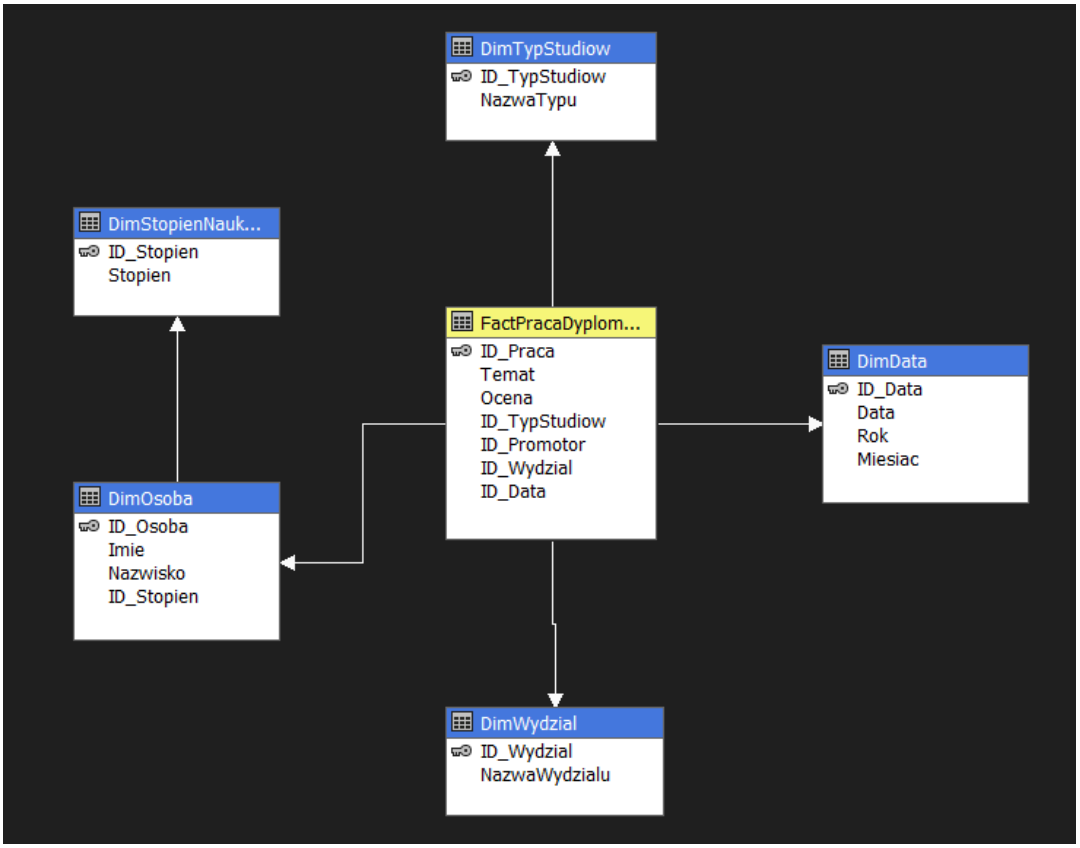
Rysunek 8 Conditional Split w imporcie Slownik

MODELOWANIE PROCESÓW BIZNESOWYCH

3. Projekt hurtowni danych PraceDyplomoweDW

Początkowym krokiem podczas projektowania hurtowni danych było wybranie typu modelowania. W moim projekcie użyłem typu hybrydowego, czyli połączenia cech zarówno modelu gwiazdy jak i płatka śniegu. Wybrałem ten typ ze względu na prostą strukturę oraz szybki czas odpytywania wymiarów które są zamodelowane jako typ gwiazdy, ale też typ ten pozwala na bardziej szczegółową analizę danych oraz redukuje nadmiarowość danych przy stosowaniu płatka śniegu. Moim zdaniem elastyczna struktura danych jaka jest dostarczana poprzez typ hybrydowy jest kluczowa w mojej hurtowni danych.

Tabela faktów zawiera informacje na temat prac dyplomowych, w tym ich tytuł i ocenę. Wokół tabeli faktów znajdują się tabele wymiarów, które zawierają dodatkowe informacje na temat tych prac. Na przykład tabela TypStudiow zawiera informacje o typie studiów, na których powstała praca dyplomowa, tabela Data określa termin złożenia pracy dyplomowej, tabela Wydział zawiera informacje o nazwie wydziału, a tabela Osoba zawiera informacje o promotorze pracy dyplomowej, w tym zawiera podwymiar StopienNaukowy który określa jego stopień naukowy.

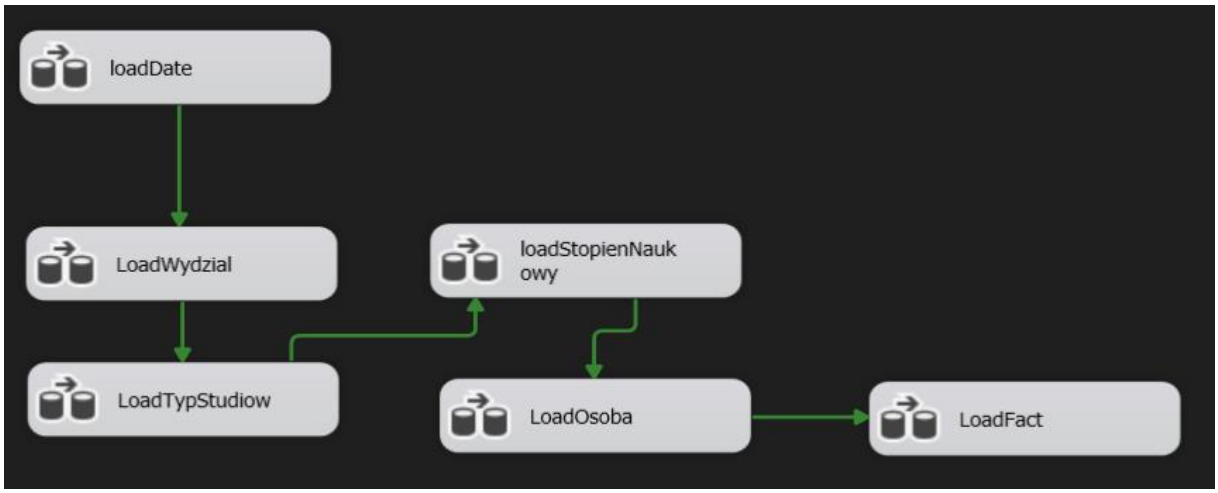


Rysunek 9 Diagram Hurtowni Danych

MODELOWANIE PROCESÓW BIZNESOWYCH

Po ustaleniu struktury należało stworzyć bazę danych którą nazwałem PraceDyplomoweDW. Do załadowania danych do nowo powstałej hurtowni danych znowy potrzebny był proces ETL, poniżej przedstawiona jest struktura w jaki sposób dane były ładowane do tabel. Kolejność ładowania danych jest ważna, ponieważ tabela faktów zależy od tabel wymiarów, a ich klucze obce są używane w celu połączenia danych między tabelami. Kiedy dane są ładowane do tabel wymiarów, tworzone są klucze unikalne, które są później wykorzystywane jako klucze obce w tabeli faktów. Kolejność ładowania danych jest ważna, ponieważ jeśli dane faktów zostaną załadowane przed danymi wymiarów, klucze obce w tabeli faktów mogą nie zostać poprawnie przypisane do kluczy unikalnych w tabelach wymiarów, co może prowadzić do błędów lub utraty danych.

Dlatego, należy pamiętać o właściwej kolejności ładowania danych do bazy danych - najpierw do tabel wymiarów, a następnie do tabeli faktów.



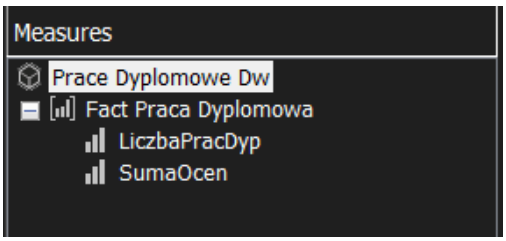
Rysunek 10 Struktura Hurtowni Danych

MODELOWANIE PROCESÓW BIZNESOWYCH

4. Budowa kostki wielowymiarowej OLAP

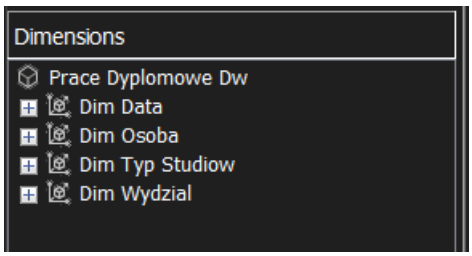
Jak wiadomo kostka OLAP składa się z trzech podstawowych elementów: miar, wymiarów i hierarchii.

Miary: są to wartości, które chcemy analizować w kontekście wymiarów. Miary to zazwyczaj liczby, takie jak suma, średnia, minimalna i maksymalna wartość. Projekt zawiera miary takie jak LiczaPracDyp oraz SumaOcen.



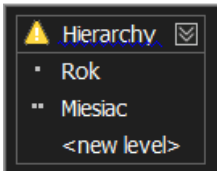
Rysunek 11 Miary Kostki OLAP

Wymiary: są to elementy, według których dane są grupowane i filtrowane. Wymiary na kostce to data, osoba, typ studiów oraz wydział.



Rysunek 12 Wymiary kostki OLAP

Hierarchie: są to struktury organizujące dane w ramach wymiarów. Hierarchie składają się z poziomów, które pozwalają na analizę danych na różnych poziomach szczegółowości. Projekt posiada hierarchie czasu.



Rysunek 13 Hierarchia kostki OLAP

MODELOWANIE PROCESÓW BIZNESOWYCH

5. Budowa raportów na kostce wielowymiarowej

Budowa raportów na kostce wielowymiarowej OLAP polega na wykorzystaniu jej struktury, czyli wymiarów, miar i hierarchii do generowania wykresów, tabel i innych wizualizacji, które pozwalają na analizę danych. Budowa raportów na kostce wielowymiarowej umożliwia szybkie i dokładne analizowanie dużych ilości danych, co może być przydatne w biznesie, badaniach naukowych czy innych dziedzinach, gdzie wymagana jest analiza danych. W projekcie skorzystałem z programu Report Builder, który pozwolił mi stworzyć dwa raporty. Pierwszy raport pokazuje średnią ocenę dla każdego wydziału, natomiast drugi raport pokazuje średnią ocenę oraz liczbę prac dyplomowych oddanych w danym roku.

Raport 1

Nazwa Wydzialu	Srednia Ocen
Administracji i Nauk Spolecznych	3.56
Architektury	3.57
Chemiczny	3.50
Elektroniki i Technik Informacyjnych	3.48
Elektryczny	3.38
Fizyki	3.39
Geodezji i Kartografii	3.54
Inzynierii Chemicznej i Procesowej	3.45
Inzynierii Ladowej	3.47
Inzynierii Materialowej	3.49
Inzynierii Produkcji	3.46
Inzynierii Srodowiska	3.49
Matematyki i Nauk Informacyjnych	3.53

Rysunek 14 Raport 1

MODELOWANIE PROCESÓW BIZNESOWYCH

Raport 2

Rok	Liczba Prac Dyp	Średnia Ocen
1973	193	3.68
1974	384	3.43
1975	385	3.40
1977	192	3.39
1978	193	3.51
1979	193	3.55
1980	193	3.52
1983	384	3.55
1984	577	3.47
1987	192	3.56
1988	192	3.53
1989	192	3.50
1990	192	3.57
1991	192	3.56
1992	193	3.55
2003	193	3.30
2007	384	3.36
2009	384	3.34
2010	192	3.42
Total	**5000**	**3.68**

Rysunek 15 Raport 2

MODELOWANIE PROCESÓW BIZNESOWYCH

6. Podsumowanie i wnioski

To sprawozdanie opisuje proces tworzenia hurtowni danych dla projektów dyplomowych. Rozpocząłem od utworzenia nowej bazy danych i użyłem języka SQL do tworzenia tabel oraz kluczy głównych i obcych. Następnie przeprowadzona została procedura ETL, aby zaimportować dane z plików płaskich do bazy danych. W tym celu wykorzystałem komponenty takie jak Flat file source oraz SQL Server Destination. W trakcie procesu napotkano kilka błędów, takich jak spacje przed nazwami wydziałów i nieprawidłowe znaki w kolumnie Data, które zostały rozwiązane dzięki zastosowaniu komponentów takich jak Derived Column i Data Conversion. Dodatkowo, w tabeli Osoba został dodany komponent Skryptu, który uzupełnia domyślną wartość dla brakujących stopni naukowych. Aby uniknąć problemów z brakującymi rekordami, odznaczyłem opcję "Check Constraints" w komponencie SQL Server Destination. Następnie zaprojektowałem hybrydowy model wykorzystujący zarówno schemat gwiazdy, jak i schemat płatka śniegu. Na koniec stworzyłem wielowymiarową kostkę i zbudowałem raporty.

Podsumowując, projekt ten pokazuje moją umiejętność tworzenia hurtowni danych i wykorzystywania procesów ETL do importowania danych. Wykazałem również wiedzę na temat różnych komponentów i technik manipulacji danych, takich jak Derived Column, Data Conversion i komponent Skryptu. Zaprojektowałem również hybrydowy model i stworzyłem na jego podstawie raporty, co świadczy o mojej wprawie w modelowaniu danych i tworzeniu raportów.

Co nie zmienia faktu, że w rzeczywistych zastosowaniach, dane oraz wymagania użytkowników końcowych zwykle są bardziej złożone. Jednakże, wykonanie tego zadania dostarczyło solidnej podstawy wiedzy.