

HURTOWNIE I EKSPLOACJA DANYCH

POLITECHNIKA WARSZAWSKA - OKNO

Studia Magisterskie

Laboratorium Hurtowni i Eksploracji Danych

Sprawozdanie Hurtowni Danych

Wykonał: Jan Bajena

Data wykonania: 13.03.2017

Wersja: 1

MODELOWANIE PROCESÓW BIZNESOWYCH

MODELOWANIE PROCESÓW BIZNESOWYCH

Spis treści

<u>1. URUCHOMIENIE BAZY DANYCH PRACE DYPLOMOWE</u>	<u>4</u>
<u>2. PROCES ETL DO BAZY DANYCH</u>	<u>4</u>
<u>3. PROJEKT HURTOWNI DANYCH PRACEDYPLOMOWEDW</u>	<u>15</u>
<u>4. BUDOWA KOSTKI WIELOWYMIAROWEJ OLAP</u>	<u>18</u>
<u>5. BUDOWA RAPORTÓW NA KOSTCE WIELOWYMIAROWEJ</u>	<u>22</u>
<u>6. PODSUMOWANIE I WNIOSKI</u>	<u>24</u>

MODELOWANIE PROCESÓW BIZNESOWYCH

1. Uruchomienie bazy danych Prace Dyplomowe

Aby uruchomić bazę danych należało wykonać następujące kroki:

- a) W programie Microsoft SQL Server Management Studio wybrać Databases -> New Database i podać poprawną nazwę bazy danych: “PraceDyplomowe”**
- b) Otworzyć skrypt “TabeleBazyDanychPraceDyplomowe.sql” w SSMS i kliknąć “Execute”**

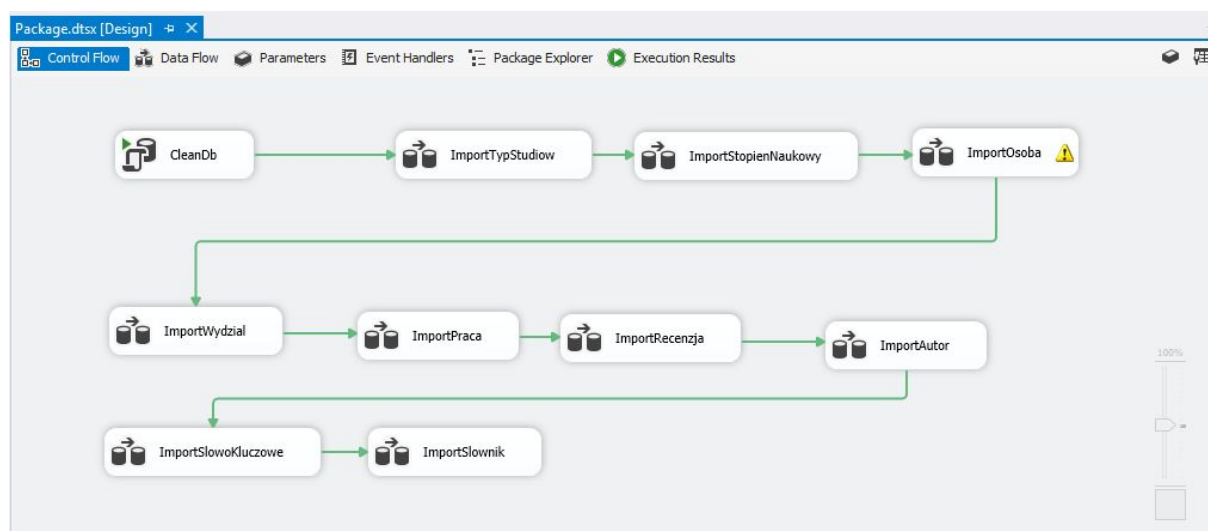
Skrypt stworzył wszystkie potrzebne tabele oraz klucze główne/obce.

2. Proces ETL do bazy danych

Proces ETL do tego kroku znajduje się w folderze o nazwie “CreatePraceDyplomowe”

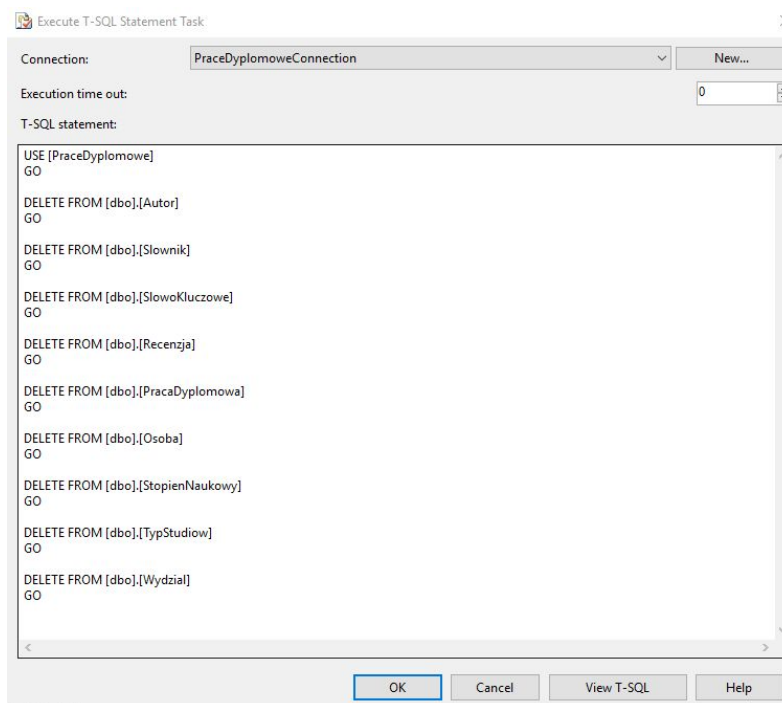
Schemat blokowy procesu ETL importującego dane do bazy danych wygląda następująco:

MODELOWANIE PROCESÓW BIZNESOWYCH



MODELOWANIE PROCESÓW BIZNESOWYCH

a) CleanDb - zadanie wykonujące prosty skrypt T-SQL czyszczący dane z tabel bazy “PraceDyplomowe”:



b) Zadania importujące dane do poszczególnych tabel z plików tekstowych

Każde z zadań wykorzystuje dwa komponenty:

- Flat File Source - komponent wczytujący dane z pliku płaskiego.

MODELOWANIE PROCESÓW BIZNESOWYCH

- **SQL Server destination - komponent zapisujący wczytane wcześniej dane do odpowiednich kolumn w bazie danych.**

Niektóre z zadań wymagały dodatkowych czynności zanim dane mogły zostać zapisane w bazie danych:

- **Plik “Osoba.txt” zawiera nieprawidłowe rekordy. Niektóre rekordy w kolumnie “Stopien” nie mają żadnej wartości, a w niektórych brakuje wartości w kolumnie “Imie”. Dane zostały naprawione za pomocą komponentu “Script component”. W przypadku braku wartości w kolumnie “Stopien” wpisana została wartość “1” odpowiadająca oznaczająca brak stopnia naukowego.**

Skrypt poprawiający dane wygląda następująco:

MODELOWANIE PROCESÓW BIZNESOWYCH

```
public override void OsobaInput_ProcessInputRow(OsobaInputBuffer Row)
{
    if (Row.IDStopien == 0)
    {
        int stopien;
        if (int.TryParse(Row.Nazwisko, out stopien))
        {
            Row.IDStopien = (sbyte)stopien;
            Row.Nazwisko = Row.Imie;
            Row.Imie = "";
        }
        else
        {
            Row.IDStopien = 1;
        }
    }
}
```

- Plik “Wydzial.txt” zawiera nazwy wydziałów, które poprzedzone są znakiem spacji. W celu poprawy wartości użyty został komponent “Derived column”, który pozwala na prostą modyfikację wartości. W tym wypadku użyta została funkcja LTRIM, która usuwa białe znaki na początku słowa.

MODELOWANIE PROCESÓW BIZNESOWYCH

Konfiguracja komponentu “Derived Column” wygląda następująco:

Specify the expressions used to create new column values, and indicate whether the values update existing columns or populate new columns.

Variables and Parameters
Columns

Mathematical Functions
String Functions
Date/Time Functions
NULL Functions
Type Casts
Operators

Description:

Derived Column Name	Derived Column	Expression	Data Type
NazwaWydzialu	Replace 'NazwaWydzialu'	LTRIM(NazwaWydzialu)	string [DT_STR]

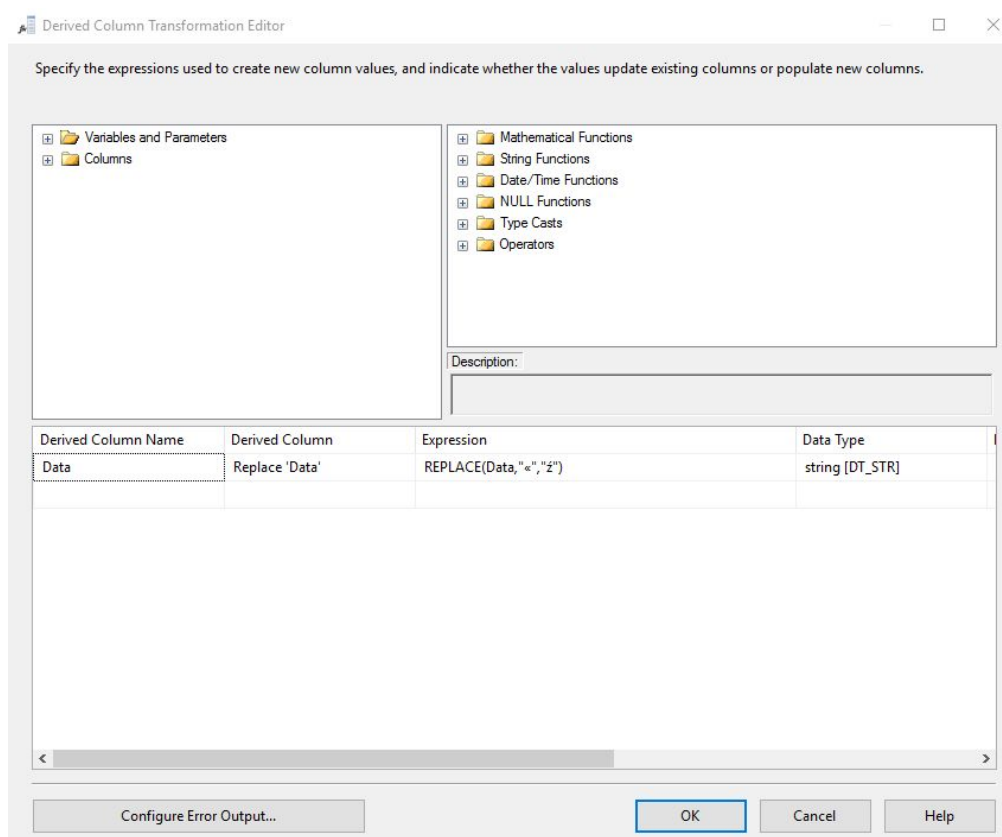
Configure Error Output... OK Cancel Help

MODELOWANIE PROCESÓW BIZNESOWYCH

- **Plik “PracaDyplomowa.txt” zawiera nieprawidłowe znaki w kolumnie data - zamiast litery “Ż” znajduje się w niej znak “«”. Poprawienie tej kolumny wymaga użycia komponentów “Derived Column” oraz “Data Conversion”. Pierwszy zamienia znaki w błędnych wierszach, a drugi konwertuje kolumnę z typu string do typu date.**

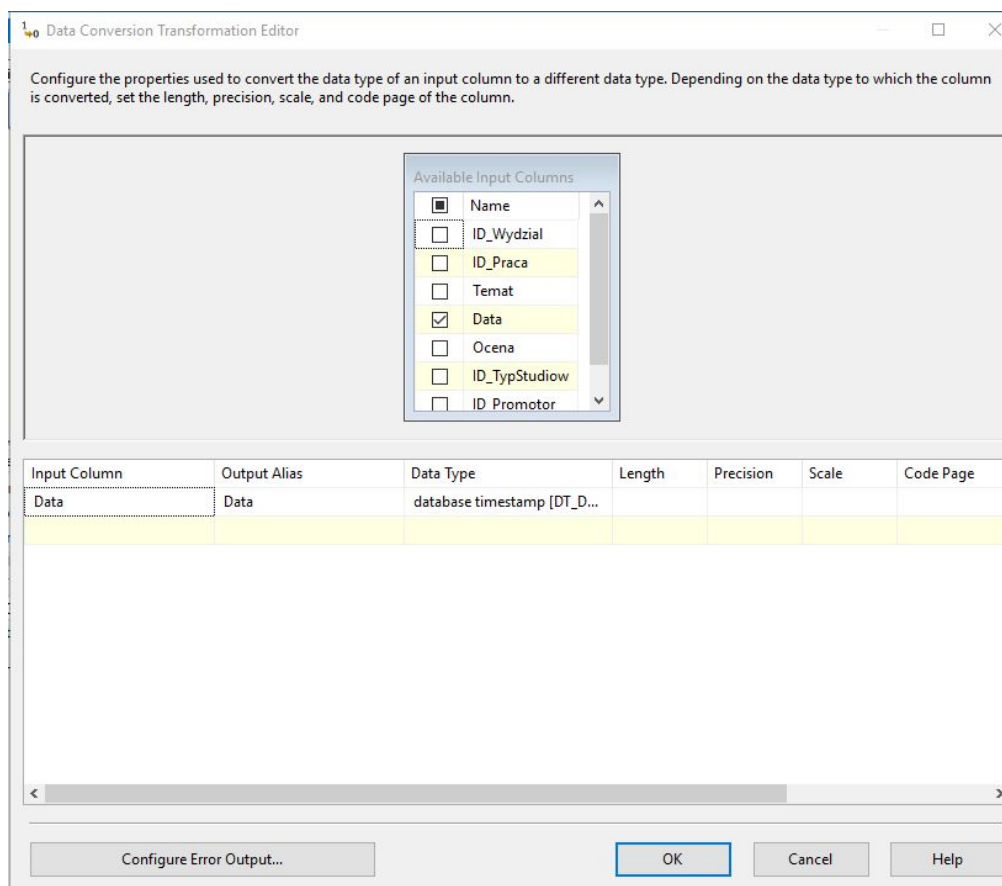
Konfiguracja komponentu “Derived Column”:

MODELOWANIE PROCESÓW BIZNESOWYCH



Konfiguracja komponentu “Data conversion”:

MODELOWANIE PROCESÓW BIZNESOWYCH



- Plik “Autor.txt” zawiera ID osób, których nie ma w pliku “Osoba.txt”, co powodowało błąd wynikający z próby zapisu nieprawidłowej wartości klucza obcego. Aby import przebiegł pomyślnie należy odznaczyć opcję “check constraints” w komponencie “Sql Server Destination”

MODELOWANIE PROCESÓW BIZNESOWYCH

Configure the properties used to bulk copy data into a local instance of the Database Engine.

Connection Manager
Mappings
Advanced

Specify the options for a bulk insert.

☐ Keep identity ☒ Table lock ☐ Fire triggers
☐ Keep nulls ☐ Check constraints

First row:

Last row:

Maximum number of errors:

Timeout:

- Pliku “Słownik.txt” zawiera wiersz odwołujący się do rekordu SłowoKluczowe o ID = 1, którego nie ma w pliku “SłowoKluczowe.txt”. Podczas importu należy pominąć wiersze z ID_SłowoKluczowe==1, wykorzystując komponent “ConditionalSplit”.

Konfiguracja komponentu “Conditional Split” wygląda następująco:

MODELOWANIE PROCESÓW BIZNESOWYCH

Conditional Split Transformation Editor

Specify the conditions used to direct input rows to specific outputs. If an input row matches no condition, the row is directed to a default output.

+ Variables and Parameters

+ Columns

+ Mathematical Functions

+ String Functions

+ Date/Time Functions

+ NULL Functions

+ Type Casts

+ Operators

Description:

Order	Output Name	Condition
1	InvalidSłowoKluczowe	ID_SłowoKluczowe == 1

Default output name: Conditional Split Default Output

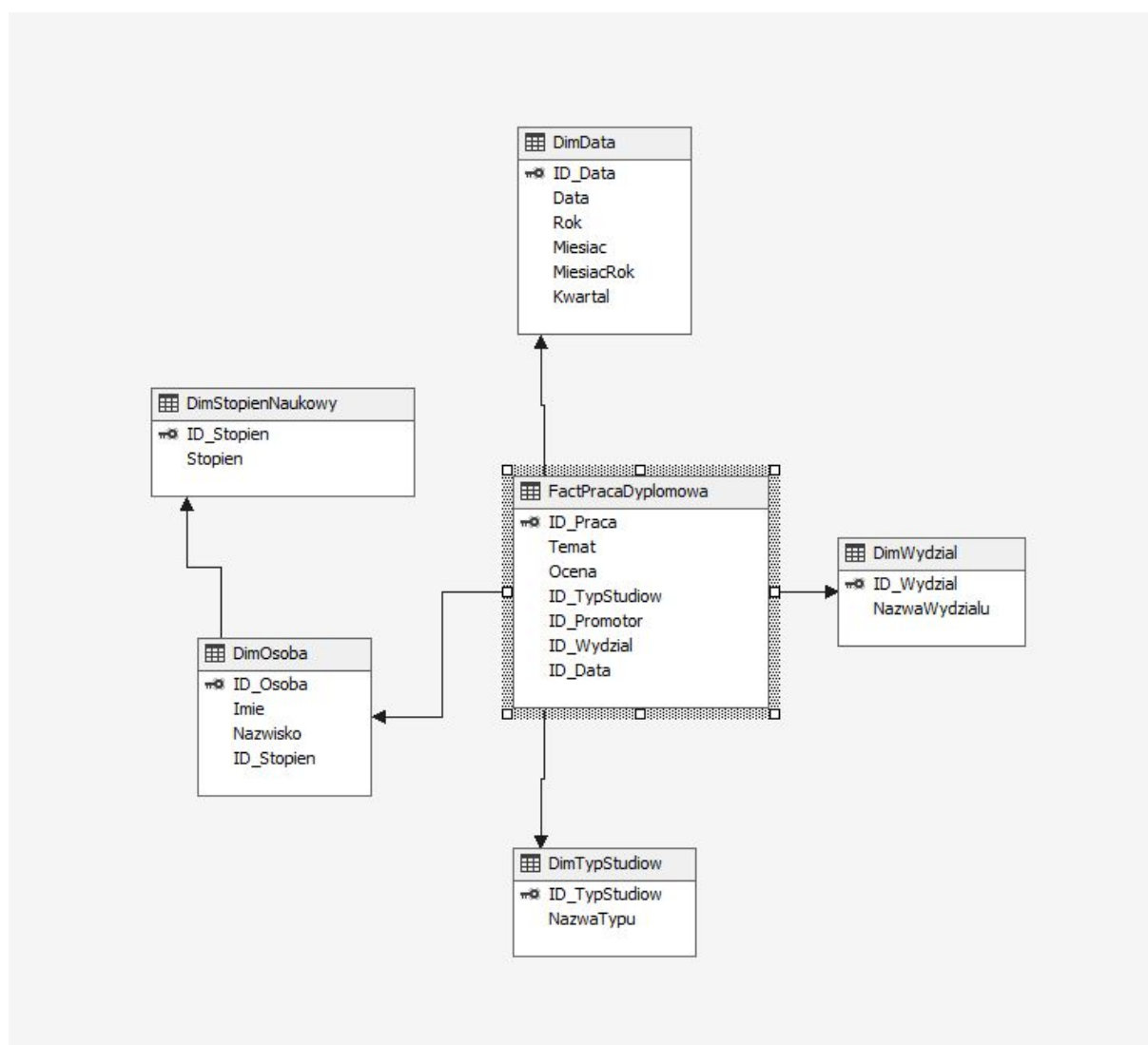
Configure Error Output... OK Cancel Help

3. Projekt hurtowni danych PraceDyplomoweDW

MODELOWANIE PROCESÓW BIZNESOWYCH

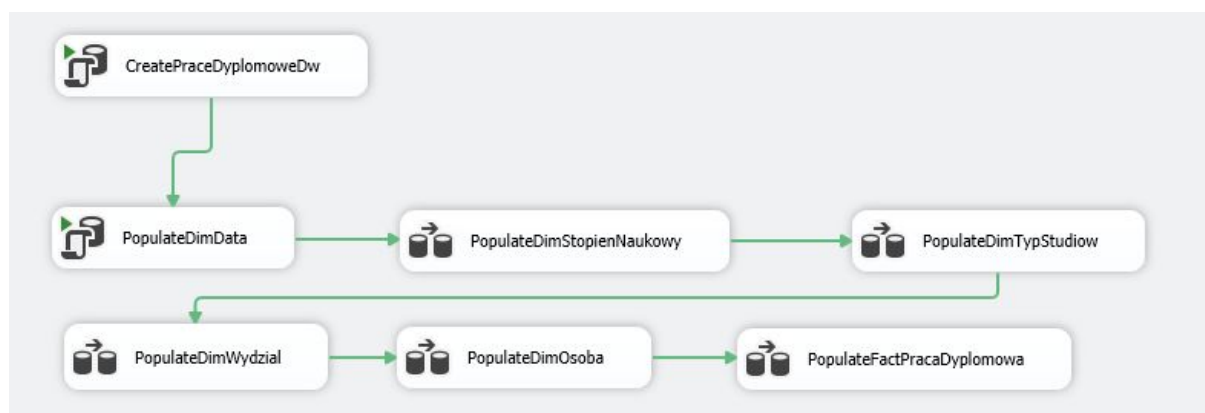
W projektowanej hurtowni danych faktami są prace dyplomowe. Dla tej tabeli faktów zdefiniowane zostały następujące wymiary:

- **Data** - określa termin złożenia pracy dyplomowej. Zawiera informacje takie, jak rok, miesiąc czy kwartał.
- **Wydział** - określa wydział, na którym złożona została praca.
- **Osoba** - zawiera informacje o promotorze pracy
- **Stopień naukowy** - podwymiar osoby - zawiera informacje o stopniu naukowym promotora pracy.

MODELOWANIE PROCESÓW BIZNESOWYCH

MODELOWANIE PROCESÓW BIZNESOWYCH

Dane z bazy danych “PraceDyplomowe” zostały zaimportowane do bazy “PraceDyplomoweDw” za pomocą dodatkowego procesu ETL o następującej strukturze:



- **CreatePraceDyplomoweDw** - tworzy bazę danych “PraceDyplomoweDw” o strukturze zgodnej z opisanym wcześniej schematem
- **PopulateDimData** - wypełnia wymiar “DimData” danymi o czasie
- Pozostałe elementy - kopiują dane z tabel bazy “PraceDyplomowe” do odpowiadających im tabel bazy “PraceDyplomoweDw”

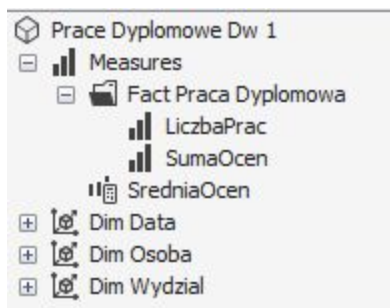
Proces ETL z tego kroku znajduje się w folderze “CreatePraceDyplomoweDw”.

MODELOWANIE PROCESÓW BIZNESOWYCH

4. Budowa kostki wielowymiarowej OLAP

Projekt kostki OLAP znajduje się w folderze “DataWarehouseProject”

Struktura projektu kostki w Visual Studio wygląda następująco:



Projekt zawiera następujące miary:

- **LiczbaPrac** - Informuje o liczby prac w danej części kostki
- **SumaOcen** - Zawiera zsumowaną liczbę ocen w danej części kostki.
Ta miara jest potrzebna była jedynie do wyliczenia miary **SredniaOcen**.


MODELOWANIE PROCESÓW BIZNESOWYCH

- **SredniaOcen** - miara obliczana jako iloraz dwóch poprzednich miar. Definicja tej miary w Visual Studio wygląda następująco:

The screenshot shows the 'Name' property window in Visual Studio. The 'Name' field contains '[SredniaOcen]'. Under the 'Parent Properties' section, 'Parent hierarchy' is set to 'Measures' and 'Parent member' is empty. Under the 'Expression' section, the formula '[Measures].[SumaOcen] / [Measures].[LiczbaPrac]' is entered.

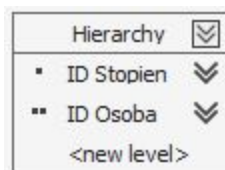
Kostka posiada następujące wymiary:

- **Data** - zawiera 2 hierarchie pozwalające w łatwy sposób przeglądać prace z podziałem na miesiące/kwartały:

 Hierarchv 	 Hierarchv.1 
▪ Rok	▪ Rok
▪▪ Kwartal	▪▪ Miesiac
<new level>	<new level>

MODELOWANIE PROCESÓW BIZNESOWYCH

- **Osoba** - zawiera hierarchię, która umożliwia granulację promotorów ze względu na stopień naukowy:



- **Wydział** - zawiera jedynie nazwę wydziału

Po zakończeniu deploymentu na serwer SSAS można za pomocą zakładki “Browse” przeglądać zawartość kostki.

Poniższy przykład przedstawia kostkę przy użyciu wymiarów “Wydział” i “Rok” oraz miar “LiczbaPrac” i “SredniaOcen”:

MODELOWANIE PROCESÓW BIZNESOWYCH

Edit as Text

Import...

Prace Dyplomowe Dw 1

..

Metadata

Measure Group:

<All>

Prace Dyplomowe Dw 1

Measures

Fact Praca Dyplomowa

SredniaOcen

KPIs

Dim Data

Dim Osoba

Dim Wydzial

ID Wydzial

Calculated Members

Dimension

Hierarchy

<Select dimension>

Rok	ID Wydzial	LiczbaPrac	SredniaOcen
1973	Administr...	10	3,5
1973	Architektury	14	4,07142857...
1973	Chemiczny	12	3,91666666...
1973	Elektroniki...	20	3,35
1973	Elektryczny	11	3,63636363...
1973	Fizyki	11	3,27272727...
1973	Geodezji i...	12	4
1973	Inzynierii ...	18	3,77777777...
1973	Inzynierii ...	16	3,75
1973	Inzynierii ...	14	3,71428571...
1973	Inzynierii ...	17	3,64705882...
1973	Inzynierii ...	19	3,78947368...
1973	Matematy...	19	3,52631578...
1974	Administr...	17	3,35294117...
1974	Architektury	28	3,60714285...
1974	Chemiczny	32	3,34375
1974	Elektroniki...	31	3,48387096...
1974	Elektryczny	32	3,34375
1974	Fizyki	29	3,41379310...
1974	Geodezji i...	36	3,80555555...
1974	Inzynierii ...	29	3,62068965...
1974	Inzynierii ...	31	3,74193548...

MODELOWANIE PROCESÓW BIZNESOWYCH

5. Budowa raportów na kostce wielowymiarowej

Projekt zawierający definicje raportów znajduje się w folderze “PraceDyplomoweReports”. Wygenerowane raporty można obejrzeć w folderze “GeneratedReports”.

Za pomocą usługi SSRS zdefiniowane zostały dwa raporty wykorzystujące wcześniej stworzoną kostkę OLAP jako źródło danych:

MODELOWANIE PROCESÓW BIZNESOWYCH

- **“Liczba_Srednia_Rok_Kwartal.rdl”** - raport zawierający informacje o liczbie prac dyplomowych w poszczególnych kwartałach kolejnych lat.

Liczba prac dyplomowych

	Q1		Q2		Q3		Q4	
	Średnia Ocen	Liczba Prac	Średnia Ocen	Liczba Prac	Średnia Ocen	Liczba Prac	Średnia Ocen	Liczba Prac
1973	3.68	193	Brak	0	Brak	0	Brak	0
1974	3.43	192	Brak	0	Brak	0	3.61	192
1975	Brak	0	3.50	385	Brak	0	Brak	0
1977	Brak	0	3.39	192	Brak	0	Brak	0
1978	Brak	0	Brak	0	Brak	0	3.51	193
1979	Brak	0	Brak	0	Brak	0	3.55	193
1980	Brak	0	Brak	0	3.52	193	Brak	0
1983	3.55	192	Brak	0	3.49	192	Brak	0
1984	Brak	0	Brak	0	3.47	384	3.37	193
1987	3.56	192	Brak	0	Brak	0	Brak	0
1988	Brak	0	Brak	0	3.53	192	Brak	0
1989	Brak	0	3.50	192	Brak	0	Brak	0
1990	3.57	192	Brak	0	Brak	0	Brak	0
1991	Brak	0	3.56	192	Brak	0	Brak	0
1992	Brak	0	Brak	0	Brak	0	3.55	193
2003	Brak	0	Brak	0	Brak	0	3.30	193
2007	3.36	192	3.46	192	Brak	0	Brak	0
2009	3.41	384	Brak	0	Brak	0	Brak	0
2010	Brak	0	3.42	192	Brak	0	Brak	0

Dla niektórych kwartałów brak było danych. W takiej sytuacji komórki należało sformatować wykorzystując następujące wyrażenia:

- Dla średniej: `=IIF(IsNothing(Fields!SredniaOcen.Value), "Brak", Fields!SredniaOcen.Value)`
- Dla liczby prac: `=IIF(IsNothing(Fields!LiczbaPrac.Value), "0", Fields!LiczbaPrac.Value)`

MODELOWANIE PROCESÓW BIZNESOWYCH

- “Stopien_Srednia_Wydzial.rdl” - raport zawierający informacje o średniej ocen prac dyplomowych w zależności od stopnia naukowego promotora i wydziału.

Przykładowo dla wydziału Architektury raport wygląda następująco:

Wydział: Architektury

Stopień Naukowy Promotora	Srednia Ocen	Liczba Prac
	3.47	62
dr	3.68	34
dr hab	3.49	51
dr hab inz.	3.71	28
dr. inz	3.72	39
inz.	3.73	30
mgr	3.41	44
mgr inz.	3.62	47
prof dr hab	3.76	33
prof dr hab inz.	3.19	26

6. Podsumowanie i wnioski

MODELOWANIE PROCESÓW BIZNESOWYCH

Wykonanie zadania pozwoliło na zapoznanie się z podstawami budowania hurtowni danych, analizy danych za pomocą kostki OLAP oraz generowania przyjaznych dla użytkownika raportów w SSRS.

W praktycznych zastosowaniach dane i wymagania użytkowników końcowych z pewnością są dużo bardziej skomplikowane, jednak wiedza wyniesiona z zadania daje solidną podstawę do dalszego rozwoju.