Odwzorowanie hierarchii rekurencyjnej w procesie ETL

Poniższy przykład prezentuje sposób prawidłowego przetworzenia danych w procesie ETL z zachowaniem poprawności hierarchii rekurencyjnej. Wykorzystane zostaną w nim dane dotyczące personelu kliniki medycznej, zawierające hierarchię rekurencyjną typu pracownik-przełożony.

Utworzenie bazy danych dla kliniki

1. Otwórz narzędzie *Microsoft SQL Server 2012 → SQL Server Management Studio*.
2. W oknie połączenia wybierz *Server Type: Database Engine*, *Server Name: localhost*.
3. Wybierz *Connect*.
4. Otwórz dokument *create_klinika.sql* (*File→ Open→ File*)
5. Wykonaj komendę *Execute*.

W wyniku wykonania powyższych operacji zostanie utworzona baza danych zawierająca tabelę przechowującą informacje o pracownikach. Do tej tabeli zostaną załadowane dane w procesie ETL.

UWAGA: W celach przejrzystości w utworzonej bazie znajduje się wyłącznie jedna tabela, która mogłaby reprezentować tablicę wymiarów w rzeczywistej hurtowni. Pominięte zostały tablica faktów i tablice pozostałych wymiarów.

Proces ETL

Dane źródłowe pracowników są zawarte w pliku tekstowym *personel.txt*. W każdym wierszu znajdują się kolejno: pesel pracownika, imię i nazwisko, stanowisko zajmowane w klinice, pesel bezpośredniego przełożonego. W przeprowadzanym procesie ETL należy zadbać o poprawne odwzorowanie hierarchii pracowników. W tym celu należy:

A. Załadować do tabeli *pracownicy* bazy danych *klinika* informacje o personelu, ustawiając wartości atrybutu przełożony na *null* dla każdego pracownika.
B. Zaktualizować tabelę pracowników poprzez dodanie informacji o przełożonych (po wykonaniu punktu A).

Rysunek 1: plik źródłowy z danymi o personelu

Czynności wstępne

1. Otwórz narzędzie *Microsoft SQL Server 2012 → SQL Server Data Tools*.
2. Utwórz nowy projekt dla procesu integracji danych (*File→ New→ Project*). Typ tworzonego projektu to *Integration Services Project*.
3. W oknie *Connection Managers* utwórz trzy połączenia wg następnych punktów.
4. Dla pliku *Personel.txt*:
    a. Wybierz *New Flat File Connection*,
    b. W polu *Connection Manager Name* wpisz Personel.txt,
    c. Ustaw code page na 65001 (UTF-8),
    d. Klikając na liście elementów po lewej stronie element *Columns* wybierz ";" (średnik) jako *Column Delimeter*, kliknij *Refresh*, a następnie sprawdź, czy został dokonany właściwy podział na cztery kolumny,
    e. W elemencie *Advanced* ustaw kolejno nazwy kolumn: Pesel, ImieNazwisko, Stanowisko, Przełożony oraz wprowadź odpowiednie typy (*DT_WSTR* dla wszystkich kolumn) i sprawdź długości łańcuchów znaków zgodnie ze schematem bazy danych (*create_klinika.sql*),
    f. Kliknij *OK*.
5. Dla bazy danych:
    a. Wybierz new *OLE DB Connection*,
    b. Wybierz *New*,
    c. Jako *Server* name wpisz *localhost*
    d. Z listy rozwijanej przy polu *Select or enter database name* wybierz nazwę bazy danych: *klinika*.
    e. Kliknij dwukrotnie *OK*.

Załadowanie danych o pracownikach

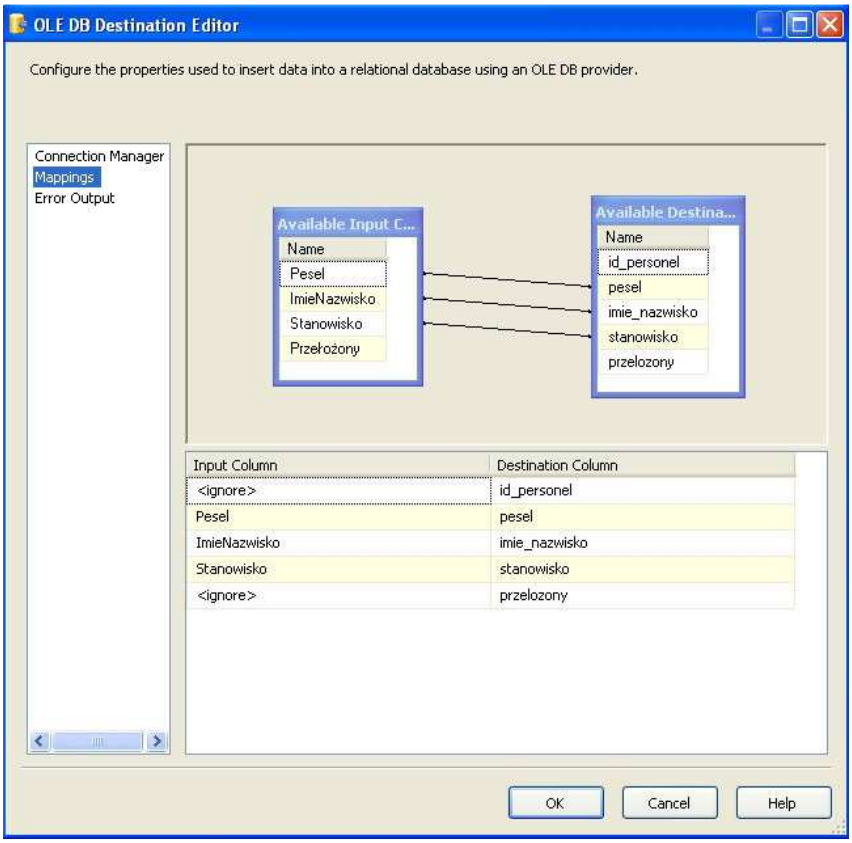
Odczytanie danych z pliku *personel.txt*:

6. Na zakładkę *Control Flow* przeciągnij z *Toolbox*'a element *Data Flow Task*.
7. Zmień nazwę tego zadania na "Informacje o personelu", klikając na niego i wybierając F2.
8. Kliknij dwukrotnie na to zadanie, aby przejść do zakładki *Data Flow*.

9. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *Flat File Source*.
10. Zmień jego nazwę na "Pracownicy".
    a. Kliknij dwukrotnie na to zadanie - otworzy się okienko z jego właściwościami. Sprawdź, czy pole *Flat File Connection Manager* zakładki *Connection Manager* wskazuje na połączenie *Personel*. Następnie wybierz *OK*.

Załadowanie danych do tabeli *Personel:*

11. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *OLE DB Destination*.
12. Zmień nazwę elementu *OLE DB Destination* na "personel".
13. Wyjście poprzedniego elementu połącz z wejściem tego elementu.
14. Kliknij dwukrotnie na to zadanie - otworzy się okienko z jego właściwościami. W polu *Name of the table or the view* zakładki *Connection Manager* wybierz [dbo].[Personel].
15. Upewnij się, że pole *Keep identity* jest odznaczone. W zakładce *Mappings* ustaw odwzorowania względem nazw (pole przełożony nie jest mapowane – rysunek 2). Następnie wybierz *OK*.
16. Wróć na zakładkę *Control Flow*. Kliknij prawym przyciskiem myszy na obiekt *Informacje o pracownikach* i wybierz polecenie *Execute task.*



Rysunek 2: odwzorowania w zakładce mappings

Wykonanie powyższych kroków spowoduje dodanie do tabeli *Personel* bazy danych *klinika* danych pracowników zawartych w źródłowym pliku tekstowym. Póki co atrybut *przełożony* ma wartość *null* dla każdego pracownika.

Results | Messages

	id_personel	pesel	imie_nazwisko	stanowisko	przelozony
1	1	45102309416	Layne Fib	dyrektor	NULL
2	2	15070802789	Ema Midway	ordynator	NULL
3	3	62090202737	Marni Sill	ordynator	NULL
4	4	56030100650	Cooper Shack	pielęgniarka	NULL
5	5	90040106009	Denys Bitten	pediatra	NULL
6	6	92021209001	Massimiliano Balsam	dermatolog	NULL
7	7	04021605937	Rosalynd Chromatography	urolog	NULL
8	8	87080907790	Dev Therefrom	neurolog	NULL
9	9	89032801767	Dee Us	anestezjol...	NULL
10	10	96062005078	Godart Activism	stomatolog	NULL
11	11	51031702957	Malissia Snail	onkolog	NULL
12	12	06071105621	Yardley Country	kardiolog	NULL
13	13	22041000204	Merry Cremate	psycholog	NULL
14	14	73081908512	Antoine Pulp	laryngolog	NULL
15	15	06020600142	Taylor Dwarf	dermatolog	NULL

Rysunek 3: tabela personel po dodaniu informacji o pracownikach

Dodanie informacji o przełożonych

Dodanie informacji o przełożonych sprowadza się do następujących czynności:

A. Załadowanie danych o pracownikach z pliku *personel.txt*.
B. Załadowanie danych z tabeli *personel* bazy danych *klinika*.
C. Złączenie obu zestawów danych po peselu – dzięki temu wiemy, jakiemu numerowi pesel (kluczowi biznesowemu) odpowiada jaki identyfikator id_personel z tabeli personelu.
D. Ponowne wczytanie danych o personelu z tabeli bazy danych i złączenie ich z zestawem wygenerowanym w punkcie C – dzięki temu możliwe jest poprawne przypisanie wartości *id_personel* do atrybutu *przełożony*.
E. Zaktualizowanie tabeli *personel* o informacje o przełożonych.

Załadowanie danych o pracownikach z pliku *personel.txt* i posortowanie wg numeru pesel:

17. Na zakładkę *Control Flow* przeciągnij z *Toolbox*'a element *Data Flow Task.*
18. Zmień nazwę tego zadania na "Informacje o przełożonych", klikając na niego i wybierając F2.
19. Kliknij dwukrotnie na to zadanie aby przejść do zakładki *Data Flow*.
20. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *Flat File Source*.
21. Zmień jego nazwę na "Pracownicy 1".
    a. Kliknij dwukrotnie na to zadanie - otworzy się okienko z jego właściwościami. Sprawdź, czy pole *Flat File Connection Manager* zakładki *Connection Manager* wskazuje na połączenie *pracownicy*. W zakładce Columns pozostaw zaznaczone jedynie pola *Pesel* oraz *Przełożony.* Następnie wybierz *OK.*
22. Przeciągnij z *Toolbox*'a element *Sort* i zmień jego nazwę na „Sort Pracownicy 1". Wyjście elementu „Pracownicy 1" połącz z wejściem tego elementu. Kliknij dwukrotnie na to zadanie i wybierz pole *Pesel* do sortowania.

Załadowanie danych z tabeli *personel* bazy danych *klinika* i posortowanie wg numeru pesel*:*

23. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *OLE DB Source*.
24. Zmień jego nazwę na "Pracownicy 2".
25. Kliknij dwukrotnie na to zadanie - otworzy się okienko z jego właściwościami. W polu *Name of the table or the view* zakładki *Connection Manager* wybierz [dbo].[Personel].
26. W zakładce *Columns* pozostaw zaznaczone kolumny: id_personel, pesel, przełożony (pozostałe odznacz).
27. Przeciągnij z *Toolbox*'a element *Sort* i zmień jego nazwę na „Sort Pracownicy 2". Wyjście elementu „Pracownicy 2" połącz z wejściem tego elementu. Kliknij dwukrotnie na to zadanie i wybierz pole p*esel* do sortowania.

Złączenie obu zestawów danych po peselu:

28. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *Merge Join.*
29. Wyjścia z obu zadań sortowania przekaż na wejście elementu Merge Join.
30. Kliknij dwukrotnie na to zadanie - otworzy się okienko z jego właściwościami. Zaznacz pesel jako klucz łączący. Przepisz na wyjście kolumny *Przełożony* (z elementu Sort Pracownicy 1) oraz *pesel* i id*_personel* (z Sort Pracownicy 2). Następnie wybierz *OK*.

Posortowanie złączonych danych wg peselu przełożonego:

31. Przeciągnij z *Toolbox*'a element *Sort* i zmień jego nazwę na „Sort Przelozony 1". Wyjście elementu „Merge Join" połącz z wejściem tego elementu. Kliknij dwukrotnie na to zadanie i wybierz pole Przełożony do sortowania. Przepisz na wyjście kolumny id_personel i Przełożony.

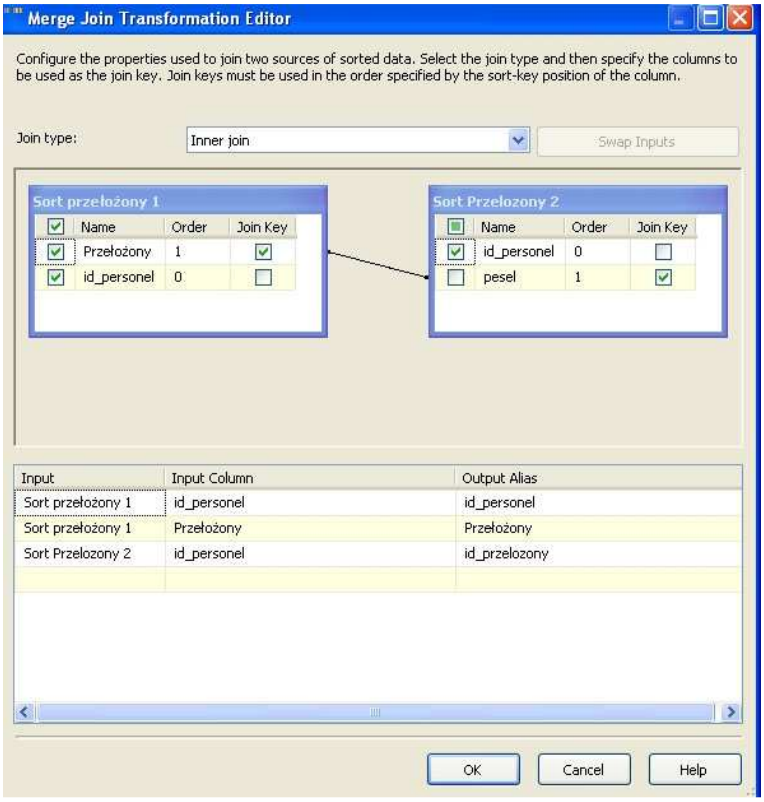
Załadowanie danych z tabeli *personel* bazy danych *klinika* i posortowanie wg numeru pesel (dane te będą służyły do identyfikacji przełożonych)*:*

32. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *OLE DB Source*.
33. Zmień jego nazwę na "Przelozeni".
34. Kliknij dwukrotnie na to zadanie - otworzy się okienko z jego właściwościami. W polu *Name of the table or the view* zakładki *Connection Manager* wybierz [dbo].[Personel].
35. W zakładce *Columns* pozostaw zaznaczone kolumny: *id_personel, pesel* (pozostałe odznacz).
36. Przeciągnij z *Toolbox*'a element *Sort* i zmień jego nazwę na „Sort Przelozony 2". Wyjście elementu „Pracownicy 2" połącz z wejściem tego elementu. Kliknij dwukrotnie na to zadanie i wybierz pole p*esel* do sortowania.

Złączenie obu zestawów danych:

37. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *Merge Join.*
38. Wyjścia z obu zadań sortowania (Sort przełożony 1 i Sort przełożony 2) przekaż na wejście elementu Merge Join.
39. Kliknij dwukrotnie na to zadanie - otworzy się okienko z jego właściwościami. Zaznacz pola *pesel* oraz *Przełożony* jako klucz łączący. Przepisz na wyjście kolumny *Przełożony* i *id_personel* (z elementu Sort Przelozony 1) oraz *id_personel* (z Sort

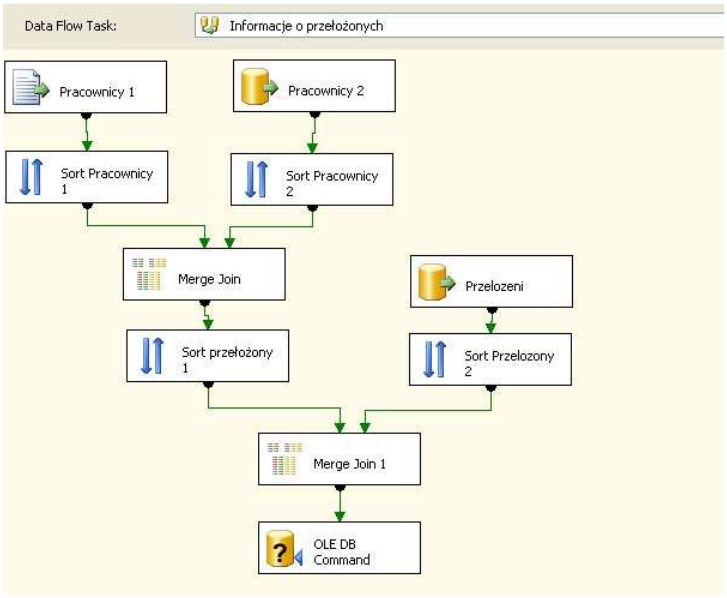
Przelozony 2). Zmień alias pola *id_personel* pochodzącego z Sort Przelozony 2 na *id_przelozony* (patrz rysunek 4). Następnie wybierz *OK*.



Rysunek 4: Złączenie danych

Wykonanie polecenia update na tabeli personel bazy danych klinika w celu wstawienie informacji o przełożonych:

40. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *OLE DB Command.* Wyjście poprzedniego elementu Marge Join połącz z wejściem tego elementu.
41. Kliknij dwukrotnie na to zadanie – w pierwszej zakładce w polu *Connection Manager* wybierz localhost.klinika.
42. Przejdź do zakładki *Component Properties.* W polu SqlCommand wpisz polecenie:
    UPDATE personel SET przelozony = ? WHERE id_personel = ?
43. Przejdź do zakładki *Column Mappings*. Przypisz id_przelozony do Param_0 i id_personel do Param_1. Kliknij OK.
44. Wróć na zakładkę *Control Flow*. Kliknij prawym przyciskiem myszy na obiekt *Informacje o przełożonych* i wybierz polecenie *Execute task.*



Rysunek 5: przepływ "informacje o przełożonych"

Po wykonaniu zadania w tabeli personel pojawią się poprawne klucze obce w polu przełożony, odpowiadające hierarchii zawartej w pliku źródłowym.

Results | Messages

	id_personel	pesel	imie_nazwisko	stanowisko	przelozony
1	1	45102309416	Layne Fib	dyrektor	NULL
2	2	15070802789	Ema Midway	ordynator	1
3	3	62090202737	Marni Sill	ordynator	1
4	4	56030100650	Cooper Shack	pielęgniarka	3
5	5	90040106009	Denys Bitten	pediatra	2
6	6	92021209001	Massimiliano Balsam	dermatolog	3
7	7	04021605937	Rosalynd Chromatography	urolog	3
8	8	87080907790	Dev Therefrom	neurolog	2
9	9	89032801767	Dee Us	anestezjol...	6
10	10	96062005078	Godart Activism	stomatolog	5
11	11	51031702957	Malissia Snail	onkolog	4
12	12	06071105621	Yardley Country	kardiolog	7
13	13	22041000204	Merry Cremate	psycholog	7
14	14	73081908512	Antoine Pulp	laryngolog	6
15	15	06020600142	Taylor Dwarf	dermatolog	8
16	16	27111104364	Augusto Scotia	stomatolog	8
17	17	59101203680	Carlin Rheostat	położna	8
18	18	95030204306	Rand Brittany	onkolog	7

Rysunek 6: tabela personel po wstawieniu danych o przełożonych