Wprowadzenie do SSIS dla MS SQL Server 2014

Wiadomości wstępne

Wprowadzenie do narzędzia SSIS dla MS SQL Server 2014 jest ćwiczeniem „krok po kroku" przygotowanym w celu zapoznania studentów z narzędziami typu BI w MS SQL Server 2014. Ćwiczenie to nie ma na celu pokazania zasad projektowania hurtowni danych – przykład pokazuje bardzo uproszczoną hurtownię danych, która nie zawiera szeregu istotnych elementów hurtowni danych.

Ćwiczenie składa się z dwóch części (A i B). Zakładamy, że księgarnia potrzebuje hurtowni danych. Przyjmujemy za fakt: sprzedaż danego tytułu książki w konkretnym dniu. W księgarni w dwóch plikach tekstowych (*BookSales1.txt* i *BookSales2.txt*) przechowywana jest informacja odpowiednio o książkach sprzedawanych w danej księgarni (*BookSales1.txt*) i o zaistnieniu faktu sprzedaży (*BookSales2.txt*).

Część A (Utworzenie bazy danych dla księgarni) pokazuje w jaki sposób, wykorzystując narzędzie *SQL Server Management Studio*, można utworzyć bazę danych w MS SQL Server 2014. Struktura bazy danych dla hurtowni jest omówiona w ramach prezentacji stanowiącej cześć wykonywanego ćwiczenia.

Część B (Proces ETL) pokazuje w jaki sposób dane zapisane w plikach BookSales1.txt i BookSales2.txt zintegrować i zapisać w bazie danych utworzonej w części A. W części B wykorzystywane jest narzędzie *SQL Server Data Tools*. Kolejne kroki procesu ETL są opisane przy kolejnych punktach zadania.

A. Utworzenie bazy danych dla księgarni

1. Otwórz narzędzie *Microsoft SQL Server 2014 → SQL Server Management Studio*
2. W oknie połączenia wybierz *Server Type: Database Engine, Server Name: localhost*
3. Wybierz *Connect*
4. Otwórz dokument *sprzedaz.sql* (*File → Open → File*)
5. Wykonaj komendę *Execute*

B. Proces ETL

1. Otwórz narzędzie *Microsoft SQL Server 2014 → SQL Server Data Tools*
2. Utwórz nowy projekt dla procesu integracji danych (*File → New → Project*). Typ tworzonego projektu to *Integration Services Project*

W procesie ETL ważne jest zidentyfikowanie „składnic" danych, z których będziemy korzystać. W projektowanym procesie ETL istnieją trzy takie „składnice": dwa pliki *BookSales1.txt* i *BookSales2.txt*, z których będziemy odczytywać dane oraz baza danych *sprzedaz*, do której odczytane i odpowiednio przekształcone dane zostaną zapisane. Tworząc połączenia do plików tekstowych należy zwrócić szczególną uwagę na sposób rozdzielenia danych oraz typ odczytywanych danych.

3. W oknie *Connection Managers* utwórz trzy połączenia wg. następnych punktów: dwa do plików: *BookSales1.txt* i *BookSales2.txt* i jeden do bazy danych *sprzedaz*.
4. Dla pliku BookSales1.txt:
    a. Wybierz *New Flat File Connection.*
    b. W polu *Connaction manager name* wpisz BookSales1
    c. W polu *File name* wybierz plik BookSales1.txt
    d. Klikając na liście elementów po lewej stronie element *Columns* sprawdź czy został wybrany "," (przecinek) jako *Column delimeter*, a następnie skontroluj dokonany podział na trzy kolumny.
    e. W elemencie *Advanced* ustaw kolejno nazwy kolumn: ISBN, Gatunek, Tytuł oraz wprowadź odpowiednie typy (ISBN - *DT_STR*, Gatunek oraz Tytuł - *DT_WSTR*) i sprawdź długości łańcuchów znaków zgodnie ze schematem bazy danych (*sprzedaz.sql*).
    f. Kliknij *OK*
5. Dla pliku *BookSales2.txt* wykonaj analogiczne kroki jak w punkcie 4, podając odpowiednią nazwę pliku oraz jako separator kolumn wybierając ";" (średnik). Nadaj następujące nazwy kolumn: *ISBN, Data, Cena kupna, Ilość*. Sprawdź typy danych - Data musi być typu *DT_DBDATE*, Cena - *DT_DECIMAL* oraz *DataScale* równe 2 natomiast Ilość - *DT_DECIMAL* oraz *DataScale* równe 0.
6. 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: *sprzedaz.*
    e. Kliknij dwukrotnie *OK.*

Proces ETL polega na wykonaniu kolejnych zadań przepływu danych (przepływ danych to odczytanie, przetworzenie i zapisanie danych). W ramach przedstawianego procesu ETL wykonywane są dwa zadania przepływu danych: przepływ danych z plików do tabeli *Ksiazka* i *Data* bazy danych *sprzedaz* oraz przepływ danych z plików do tabeli *Sprzedaz* tej samej bazy danych. Drugie zadanie jest wykonywane po zakończeniu zadania pierwszego. Takie następstwo

wynika z potrzeby odczytania kluczy głównych (generowanych automatycznie) dla tabel *Ksiazka* i *Data* w trakcie wykonywania zadania przepływu danych do tabeli *Sprzedaz*.

Zadania przepływu danych znajdują się w zakładce *Control Flow*. Zielona strzałka wskazuje następstwo czasowe wykonania poszczególnych zadań. Kolejne kroki każdego zadania przepływu danych definiuje się w zakładce *Data Flow*, gdzie zielona strzałka oznacza wejście/wyjście zadania.

Zadanie przepływu danych z plików do tabeli *Ksiazka* i *Data* wymaga odczytania danych z obu plików *BookSales1.txt* i *BookSales2.txt*. W ramach tego przepływu są wykonywane następujące kroki:
- odczytanie danych z pliku *BookSales1.txt*,
- posortowanie pobranych z pliku *BookSales1.txt* danych po ISBN (tylko posortowane dane mogą stanowić wejście do zadania złączającego dane, dlatego dane odczytane z obu plików muszą być posortowane),
- odczytanie danych z pliku *BookSales2.txt*,
- posortowanie pobranych z pliku *BookSales2.txt* danych po ISBN,
- złączenie danych po ISBN (JOIN),
- wywiedzenie na podstawie kolumny *Cena kupna* nowej kolumny *Przedział cenowy* książki (jeżeli cena jest mniejsza niż 30,00 to książka jest tania, w przeciwnym wypadku książka jest droga),
- rozdzielenie danych na dwa identyczne strumienie (dane z pierwszego strumienia zostaną załadowane do tabeli *Ksiazka*, z drugiego do tabeli *Data*)
- załadowanie danych do tabeli *Ksiazka*
- wyeliminowanie duplikatów dat dla drugiego strumienia (daty w tabeli *Data* muszą być unikalne)
- załadowanie danych do tabeli *Data*.

Odczytanie danych z pliku *BookSales1.txt*

7. Na zakładkę *Control Flow* przeciągnij z *Toolbox*'a element *Data Flow Task*.
8. Zmień nazwę tego zadania na "Informacje o książkach i datach" klikając na niego i wybierając F2.
9. Kliknij dwukrotnie na to zadanie aby przejść do zakładki *Data Flow*.
10. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *Flat File Source* (znajduje się w grupie *Other Sources*).
11. Zmień jego nazwę na "BookSales1".
12. Kliknij dwukrotnie na to zadanie - otworzy się okienko z jego właściwościami. Sprawdź, czy pole *Flat file connection manager* zakładki *Connection Manager* wskazuje na połączenie *BookSales1*. Następnie wybierz *OK*.

Posortowanie pobranych z pliku *BookSales1.txt* danych po ISBN

13. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *Sort*
14. Zmień jego nazwę na "Sort BookSales1".
15. Wykonaj połączenie od "BookSales1" do "Sort BookSales1" z użyciem niebieskiej strzałki
16. Kliknij dwukrotnie na to zadanie - otworzy się okienko z jego właściwościami. Zaznacz kolumnę *ISBN* jako kolumnę do sortowania. Następnie wybierz *OK*.

Odczytanie danych z pliku *BookSales2.txt*

17. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *Flat File Source*.
18. Zmień jego nazwę na "BookSales2".
19. Kliknij dwukrotnie na to zadanie - otworzy się okienko z jego właściwościami. Sprawdź, czy pole *Flat file connection manager* zakładki *Connection Manager* wskazuje na połączenie *BookSales2*. Następnie wybierz *OK*.

Posortowanie pobranych z pliku *BookSales2.txt* danych po ISBN

20. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *Sort*
21. Zmień jego nazwę na "Sort BookSales2".
22. Wykonaj połączenie od "BookSales2" do "Sort BookSales2" z użyciem niebieskiej strzałki
23. Kliknij dwukrotnie na to zadanie - otworzy się okienko z jego właściwościami. Zaznacz kolumnę *ISBN* jako kolumnę do sortowania. Następnie wybierz *OK*.

Złączenie danych po ISBN

24. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *Merge Join*
25. Wyjścia z obu zadań sortowania przekaż na wejście elementu *Merge Join* (niebieska strzałka)
26. Kliknij dwukrotnie na to zadanie - otworzy się okienko z jego właściwościami. Zaznacz ISBN jako klucz łączący a pozostałe kolumny przepisz na wyjście. Następnie wybierz *OK*.

Wywiedzenie na podstawie kolumny *Cena kupna* nowej kolumny *Przedział cenowy* książki

27. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *Derived column*.
28. Zmień jego nazwę na "Przedział cenowy"

29. Wyjście poprzedniego elementu (*Merge Join*) połącz z wejściem tego elementu.
30. Kliknij dwukrotnie na to zadanie - otworzy się okienko z jego właściwościami. Dodaj nową wywiedzioną kolumnę o nazwie "Przedział cenowy". Jako wyrażenie wyznaczające jej wartość wpisz: *[Cena kupna] < 30 ? "tania" : "droga"*. Kliknij *OK*.

Rozdzielenie danych na dwa identyczne strumienie

31. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *Multicast*
32. Wyjście poprzedniego elementu (*Derived column*) połącz z wejściem tego elementu.

Załadowanie danych do tabeli *Ksiazka*

33. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *OLE DB Destination*.
34. Zmień nazwę elementu OLE DB Destination na "ksiazka".
35. Wyjście poprzedniego elementu (*Multicast*) połącz z wejściem tego elementu.
36. 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].[Ksiazka]. Upewnij się, że pole *Keep identity* jest odznaczone. W zakładce *Mappings* ustaw odwzorowania względem nazw (*ID_Ksiazki* nie jest mapowane). Następnie wybierz *OK*.

Wyeliminowanie duplikatów dat dla drugiego strumienia

37. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *Sort*
38. Zmień jego nazwę na "Sort Data".
39. Wyjście elementu *Multicast* połącz z wejściem tego elementu.
40. Kliknij dwukrotnie na to zadanie - otworzy się okienko z jego właściwościami. Zaznacz kolumnę Data jako kolumnę do sortowania. Nie przepuszczaj na wyjście żadnych innych kolumn (odznaczone Pass Through). Zaznacz pole *Remove rows with duplicate sort values*. Następnie wybierz *OK*.

PAMIĘTAJ!!! W Twoim procesie ETL dane z tabeli data mają już być wcześniej wygenerowane (wszystkie daty z zadanego zakresu).

Załadowanie danych do tabeli *Data*

41. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *OLE DB Destination*.
42. Zmień jego nazwę na "data".
43. Wyjście elementu *Sort Data* połącz z wejściem tego elementu.
44. 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].[Data].

Upewnij się, że pole *Keep identity* jest odznaczone. W zakładce *Mappings* ustaw odwzorowania względem nazw (*ID_Daty* nie jest mapowane). Następnie wybierz *OK*.

Zadanie przepływu danych z plików do tabeli *Sprzedaz* wymaga odczytania danych z pliku *BookSales2.txt* oraz z tabel *Ksiazka* i *Data* (w celu odczytania wygenerowanych automatycznie wartości kluczy głównych).

45. Na zakładkę *Control Flow* przeciągnij z *Toolbox*'a kolejny element *Data Flow Task.*
46. Zmień nazwę tego zadania na "Sprzedaż".
47. Poprowadź powiązanie od zadania "Informacje o książkach i datach" do zadania "Sprzedaż"
48. Kliknij dwukrotnie zadanie "Sprzedaż" aby przejść do zakładki *Data Flow*.

W ramach tworzonego przepływu są wykonywane następujące kroki:
- odczytanie danych z pliku *BookSales2.txt*,
- odczytanie danych z tabeli *Ksiazka* bazy danych *sprzedaz*,
- posortowanie pobranych z pliku *BookSales2.txt* danych po ISBN,
- posortowanie pobranych z tabeli *Ksiazka* danych po ISBN,
- złączenie danych po ISBN (JOIN),
- posortowanie złączony danych po dacie,
- odczytanie danych z tabeli *Data* bazy danych *sprzedaz*,
- konwersja daty (data pobrana z bazy danych i data pobrana z pliku ma inny typ. Aby możliwe było dokonanie złączenia niezbędne jest ujednolicenie typów)
- posortowanie dat
- złączenie danych po dacie (JOIN),
- wywiedzenie na podstawie kolumn *Cena kupna* i *Ilosc* nowej kolumny *Zysk* (zysk ze sprzedaży jednego egzemplarza to 10% ceny kupna),
- załadowanie danych do tabeli *Sprzedaz*.

Odczytanie danych z pliku *BookSales2.txt*

49. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *Flat File Source*.
50. Zmień nazwę tego zadania na "BookSales2".
51. Kliknij dwukrotnie na to zadanie - otworzy się okienko z jego właściwościami. Sprawdź, czy pole *Flat file connection manager* zakładki *Connection Manager* wskazuje na połączenie *BookSales2*. Następnie wybierz *OK*.

Odczytanie danych z tabeli *Ksiazka* bazy danych *sprzedaz*

52. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *OLE DB Source*.

53. Zmień jego nazwę na “książka”.
54. 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].[Ksiazka]. W zakładce *Columns* pobierz tylko *ID_Ksiazki* oraz *ISBN*. Następnie wybierz *OK*.

Posortowanie pobranych z pliku *BookSales2.txt* danych po ISBN

55. Na zakładkę *Data Flow* przeciągnij z *Toolbox*’a element *Sort*
56. Zmień jego nazwę na “Sort BookSales2”.
57. Wykonaj połączenie od “BookSales2” do “Sort BookSales2” z użyciem niebieskiej strzałki
58. Kliknij dwukrotnie na to zadanie - otworzy się okienko z jego właściwościami. Zaznacz kolumnę *ISBN* jako kolumnę do sortowania. Następnie wybierz *OK*.

Posortowanie pobranych z tabeli *Ksiazka* danych po ISBN

59. Na zakładkę *Data Flow* przeciągnij z *Toolbox*’a element *Sort*
60. Zmień jego nazwę na “Sort książka”.
61. Wykonaj połączenie od “książka” do “Sort książka” z użyciem niebieskiej strzałki
62. Kliknij dwukrotnie na to zadanie - otworzy się okienko z jego właściwościami. Zaznacz kolumnę *ISBN* jako kolumnę do sortowania. Następnie wybierz *OK*.

Złączenie danych po ISBN

63. Na zakładkę *Data Flow* przeciągnij z *Toolbox*’a element *Merge Join*
64. Wyjścia z obu zadań sortowania przekaż na wejście elementu Merge Join (niebieska strzałka)
65. Kliknij dwukrotnie na to zadanie - otworzy się okienko z jego właściwościami. Zaznacz ISBN jako klucz łączący a pozostałe kolumny przepisz na wyjście. Następnie wybierz *OK*.

Posortowanie złączony danych po dacie

66. Na zakładkę *Data Flow* przeciągnij z *Toolbox*’a element *Sort*
67. Zmień jego nazwę na “Sort merge”.
68. Wykonaj połączenie od elementu *Merge Join* do “Sort merge” z użyciem zielonej strzałki
69. Kliknij dwukrotnie na to zadanie - otworzy się okienko z jego właściwościami. Zaznacz kolumnę *Data* jako kolumnę do sortowania. Następnie wybierz *OK*.

Odczytanie danych z tabeli *Ksiazka* bazy danych *sprzedaz*,

70. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *OLE DB Source*.
71. Zmień jego nazwę na "data".
72. 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].[Data]. W zakładce *Columns* pobierz wszystkie kolumny. Następnie wybierz *OK*.

Konwersja daty

73. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *Data Convertion*
74. Zmień jego nazwę na "Convert data".
75. Wykonaj połączenie od "data" do "Convert data" z użyciem zielonej strzałki
76. Kliknij dwukrotnie na to zadanie - otworzy się okienko z jego właściwościami. Zmień typ kolumny Data (jako jedyna ma być zaznaczona) na DT_DBDATE oraz pozostaw niezmienioną nową nazwę *Copy of Data*. Następnie wybierz *OK*.

Posortowanie dat

77. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *Sort*
78. Zmień jego nazwę na "Sort convert".
79. Wykonaj połączenie od "Convert data" do "Sort convert" z użyciem zielonej strzałki
80. Kliknij dwukrotnie na to zadanie - otworzy się okienko z jego właściwościami. Zaznacz kolumnę *Copy of Data* jako kolumnę do sortowania. Nie przepuszczaj kolumny *Data* (nieskonwertowana) na wyjście. Następnie wybierz *OK*.

Złączenie danych po dacie

81. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *Merge Join*
82. Wyjścia z obu zadań sortowania przekaż na wejście elementu Merge Join (zielona strzałka)
83. Kliknij dwukrotnie na to zadanie - otworzy się okienko z jego właściwościami. Zaznacz Data oraz Copy of Data jako klucz łączący a pozostałe kolumny przepisz na wyjście. Następnie wybierz *OK*.

Wywiedzenie na podstawie kolumn *Cena kupna* i *Ilosc* nowej kolumny *Zysk*

84. Na zakładkę *Data Flow* przeciągnij z *Toolbox*'a element *Derived column*.
85. Zmień jego nazwę na "Zysk"
86. Wyjście poprzedniego elementu (*Merge Join*) połącz z wejściem tego elementu.

87. Kliknij dwukrotnie na to zadanie - otworzy się okienko z jego właściwościami. Dodaj nową wywiedzioną kolumnę o nazwie “Zysk”. Jako wyrażenie wyznaczające jej wartość wpisz: *[Cena kupna] * [Ilość] * 0.1*. Kliknij *OK*.

Załadowanie danych do tabeli *Sprzedaz*

88. Na zakładkę *Data Flow* przeciągnij z *Toolbox’*a element *OLE DB Destination*.
89. Zmień jego nazwę na “sprzedaż”.
90. Wyjście elementu *Zysk* połącz z wejściem tego elementu.
91. 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].[Sprzedaz]. Upewnij się, że pole *Keep identity* jest odznaczone. W zakładce *Mappings* ustaw odwzorowania względem nazw. Następnie wybierz *OK*.

92. Wykonaj oba zadania z zakładki *Control Flow*. Po wykonaniu wszystkie zadania muszą mieć kolor zielony. Następnie sprawdź zawartość docelowej bazy danych (wszystkie tabele muszą być poprawnie wypełnione danymi).

POWODZENIA!