

ITA-102 Hurtownie Danych

Marcin Gorawski, Michał Jan Gorawski

Moduł 2

Wersja 1.0

Proces ekstrakcji danych (ETL) II

Spis treści

Proces ekstrakcji danych (ETL) II	1
Informacje o module	2
Podstawy teoretyczne zawiera moduł I pt.: Proces ekstrakcji danych (ETL): część 2.	3
Laboratorium podstawowe	3
Problem 1 (czas realizacji 25 min) – kontynuacja modułu 1	3
Problem 2 (czas realizacji 15min) - kontynuacja problemu 1	6
Laboratorium rozszerzone	8
Zadanie 1 (czas realizacji 30 min)	8
Zadanie 2 (czas realizacji 30 min)	8
Zadanie 3 (czas realizacji 30 min)	8

Informacje o module

Opis modułu

W module tym znajdziesz informacje dotyczące zagadnień związanych z procesem ekstrakcji danych. Poznasz podstawową wiedzę na temat ekstrakcji danych. Zobaczysz, czym jest oraz jak zaprojektować prosty proces ETL (Ekstrakcja, Transformacja i Ładowanie danych) w środowisku SQL Server 2008.

Cel modułu

Celem modułu jest przedstawienie możliwości użycia pakietu Integration Services, jednego z komponentów SQL Server 2008, przy projektowaniu i implementowaniu procesu ETL o przeciętnej złożoności. W zadaniach ETL tego modułu nacisk położono na proces przeszukiwania rozmytego.

Uzyskane kompetencje

Po zrealizowaniu modułu będziesz:

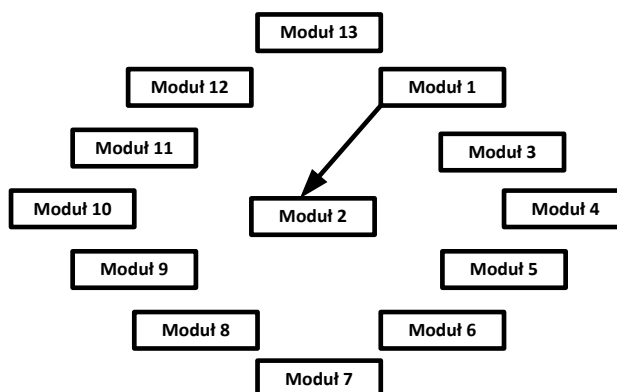
- wiedział, czym jest oraz jak projektować proces ETL
- potrafił zbudować mało złożony proces ETL w SQL Server 2008
- rozumiał mechanikę tworzenia procesów ekstrakcji danych w SQL Server 2008

Wymagania wstępne

Przed przystąpieniem do pracy z tym modulem powinieneś:

- dobrze orientować się w zagadnieniach z zakresu baz danych
- znać zasady pracy w środowisku Visual Studio

Mapa zależności modułu



Rys. 1 Mapa zależności modułu



Podstawy teoretyczne zawiera moduł I pt.: Proces ekstrakcji danych (ETL) I.






Laboratorium podstawowe

Problem 1 (czas realizacji 25 min) – kontynuacja modułu 1

Ćwiczenie pokaże sposób użycia **Sequence Container** (Kontenera Sekwencji) służącego do łączenia ze sobą grup zadań. Zapoznasz się również ze sposobem łączenia zmiennych, ze skryptami oraz z kontrolą **Precedence Constraint** (Wymuszenia Pierwszeństwa). Celem ćwiczenia jest pokazanie możliwości zarządzania kolejnością wykonanie poszczególnych komponentów za pomocą narzędzi **Control Flow** (Kontroli Przepływu). Prezentowany jest również sposób pisania skryptów w środowisku SQL Server 2008 : Integration Services




Zadanie	Tok postępowania
1. Dodanie narzędzia Kontroli Przepływu, Sequence Container (Kontenera Sekwencji) o nazwie Plik Nie Istnieje do przechowania istniejących zadań	<p>Celem tej prezentacji jest sprawdzenie, czy dany plik już istnieje. Jeżeli tak, jedynie część pakietu ma zostać wykonana, jeżeli natomiast plik nie istnieje wykonany musi zostać kompletny pakiet. Za pomocą kontenera sekwencji możliwe jest stworzenie METAprzepływu (większa część przepływu), do którego można wprowadzić sekwencję zadań. Najpierw stworzymy kontener sekwencji dla istniejących zadań, które zostaną wykonane tylko, jeżeli plik nie istnieje</p> <ul style="list-style-type: none"> Na zakładce Control Flow (Kontrola Przepływu) przeciągnij z Toolbox (Panela Narzędzi) komponent Sequence Container (Kontener Sekwencji). <p>Kontener sekwencji pozwala przechować zadanie lub grupę zadań, co pozwala na dodanie bądź usunięcie zadania bez konieczności zmiany całego grafu zadań.</p>
2. Dodanie zmiennej PlikIstnieje	<p>Potrzebny jest sposób na sprawdzenie czy plik wynikowy istnieje i jeżeli tak jest to przejście do stworzonego właśnie kontenera sekwencji. Pierwszym krokiem jest stworzenie zmiennej pakietu.</p> <ul style="list-style-type: none"> Kliknij ppm na zakładkę ControlFlow i wybierz Variables (Zmienne). <p><i>Okno, które się pojawi może zostać umieszczone z boku ekranu</i></p> <ul style="list-style-type: none"> Wybierz przycisk Add Variable z Panelu Variable. <ul style="list-style-type: none"> Name: PlikIstnieje Data Type: Boolean.
3. Dodanie zadania Skrypt Script Task nazwanego SprawdzPlik	<p>Teraz należy nadać wartość zmiennej PlikIstnieje poprzez użycie skryptu.</p> <ul style="list-style-type: none"> Z Toolbox (Panelu Narzędzi) przeciągnij komponent Script Task (Skrypt). Kliknij podwójnie na komponent, aby zmienić jego właściwości.




	<p>Domyślnym językiem skryptu jest Microsoft Visual C# 2008</p> <ul style="list-style-type: none"> • Zmień język skryptu na Visual Basic 2008. • Na zakładce General (Ogólne) zmień nazwę na SprawdzPlik. <ul style="list-style-type: none"> • Na zakładce Script (Skrypt) wybierz Edit Script (Edytuj Skrypt) i dodaj następujący kod do funkcji Main(): <pre> Try Dim myVariable As Variables Dts.VariableDispenser.LockOneForWrite("PlikIstnieje", myVariable) myVariable("PlikIstnieje").Value = _ System.IO.File.Exists(Dts.Connections("Eksportuj Studentów").ConnectionString) Dts.TaskResult = ScriptResults.Success Catch Dts.TaskResult =ScriptResults.Failure End Try </pre> <p>VariableDispenser należy do zadania, nie do pakietu. Jeżeli obsługujemy wiele zmiennych, każda dodawana jest niezależnie do VariableDispenser przy użyciu LockForRead i LockForWrite oraz jedynie nazwy zmiennej jako argumentu. Nie powoduje to zablokowania zmiennej, lecz dodanie jej do listy zablokowanych. Kiedy zmienne znajdują się w VariableDispenser wywołujemy funkcję GetVariables, aby umieścić szereg zmiennych w lokalnej zmiennej (myVariable w tym przypadku). Jeżeli używamy jednej zmiennej funkcja LockOne pozwala na skrócenie kodu. Do testowania istnienia pliku używamy funkcji System.IO. Zamiast wpisywania ścieżki dostępu do pliku używamy funkcji connection (odnosi się ona nie bezpośrednio do pliku, ale do menadżera połączeń, który obsługuje dany plik). Zmienna FileFound zawiera wyniki testu. Niezależnie od istnienia pliku zadanie zwróci Success</p> <p> Nie trzeba zwalniać zmiennej na końcu funkcji. Jest to robione automatycznie na końcu pakietu.</p>
<p>4. Dodanie Precedence Constraint (Wymuszenia Pierwszeństwa), aby uruchomić zadanie SprawdzPlik przed zadaniem Plik Nie Istnieje</p>	<ul style="list-style-type: none"> • Przeciągnij zieloną strzałkę wyjścia z komponentu SprawdzPlik do Plik Nie Istnieje. • Kliknij podwójnie na połączenie (zielona strzałka), aby je edytować : <ul style="list-style-type: none"> • Evaluation Operation: Expression and Constraint • Expression: @PlikIstnieje == False <p> Użycie symbolu @ przed nazwą zmiennej jest niezbędne aby z niej korzystać.</p> <ul style="list-style-type: none"> • Wybierz przycisk Test, aby sprawdzić poprawność wyrażenia.
<p>5. Dodanie Sequence</p>	<p>Dodanie kontenera na zadania wykonane gdy plik istnieje.</p>

Container (Kontenera Sekwencji) o nazwie Plik Istnieje	 <ul style="list-style-type: none"> Na zakładce Control Flow (Kontrola Przepływu) przeciągnij z Toolbox (Panelu Narzędzi) komponent Sequence Container. Zmień jego nazwę na: Plik Istnieje
6. Dodanie Precedence Constraint (Wymuszenia Pierwszeństwa), aby uruchomić zadanie SprawdzPlik przed zadaniem Plik Istnieje	<ul style="list-style-type: none"> Przeciągnij zieloną strzałkę wyjścia z komponentu SprawdzPlik do Plik Istnieje. Kliknij podwójnie na połączenie (zielona strzałka), aby je edytować: <ul style="list-style-type: none"> Evaluation Operation: Expression and Constraint. Expression: @PlikIstnieje == True.
7. Usunięcie pliku wynikowego i dwukrotne uruchomienie pakietu	<ul style="list-style-type: none"> Usuń plik StudenciDestination.txt. Uruchom pakiet. Jedynie gałąź Plik Nie Istnieje powinna zostać wykonana. Przerwij debugowanie i wykonaj pakiet ponownie.  <p>Jedynie gałąź Plik Istnieje powinna zostać wykonana.</p> <ul style="list-style-type: none"> Przerwij debugowanie.
8. Wymuszenie wykonania gałęzi Plik Istnieje	 <p>Teraz jedynie jedna gałąź jest wykonywana, jednakże nawet, jeśli gałąź Plik Nie Istnieje wykona się, gałąź Plik Istnieje również musi się wykonać. Można dodać Wymuszenie Pierwszeństwa pomiędzy kontenerami, aby wymusić wykonanie gałęzi Plik Istnieje.</p> <ul style="list-style-type: none"> Przeciągnij zieloną strzałkę wyjścia z komponentu File Plik Nie Istnieje do Plik Istnieje i kliknij podwójnie na połączenie (zielona strzałka) aby je edytować. Dla nowego wymuszenia wybierz wymuszenie logical OR.  <p>Ponieważ jest więcej niż jedno wymuszenie na kontenerze Plik Istnieje, trzeba określić czy obydwa wymuszenia (logical AND) czy tylko jedno z nich (logical OR) musi być spełnione, aby wykonać zadania w kontenerze. W tym przypadku tylko jedno wymuszenie musi być wykonane, nigdy obydwa, więc używamy opcji wymuszenia OR.</p>
9. Usunięcie pliku wynikowego i uruchomienie pakietu	<ul style="list-style-type: none"> Usuń plik wynikowy StudenciDestination.txt. Wykonaj pakiet.  <p>Obydwie gałęzie wykonają się.</p> <ul style="list-style-type: none"> Przerwij debugowanie.

Problem 2 (czas realizacji 15min) - kontynuacja problemu 1

Ćwiczenie ilustruje użycie komponentu **Fuzzy Lookup** (Rozmytego Przeszukiwania) do porównywania wartości ze źródła wejściowego z przeszukiwaną tablicą. Omówione zostanie również użycie komponentu **Data Viewer**, który umożliwia podgląd danych przetwarzanych podczas debugowania.

Zadanie	Tok postępowania
1. Przeglądanie pliku InputA.txt	<ul style="list-style-type: none"> Otwórz plik z lokalizacji wskazanej przez prowadzącego np.: C:\HD\UID\InputA.txt. <p> Niektóre wiersze odpowiadają liście studentów, podczas gdy inne są zbliżone lub błędne</p>
2. Dodanie Data Flow Task (Zadania Przepływu Danych) nazwane Znajdź Nazwiska do kontenera Plik Istnieje	<ul style="list-style-type: none"> W SSBIDS, na zakładce Control Flow (Kontrola Przepływu) przeciągnij Data Flow Task (Zadanie Przepływu Danych) z Toolbox (Panelu Narzędzi) do kontenera Plik Istnieje. Nazwij zadanie: Znajdź Nazwiska.
3. Dodanie Źródła Flat File Source (Danych jako Plik Płaski).	<ul style="list-style-type: none"> Kliknij podwójnie na Znajdź Nazwiska, aby otworzyć zakładkę Data Flow (Przepływ Danych). Przeciągnij Flat File Source (Źródło Danych jako Plik Płaski) z Toolbox (Panelu Narzędzi).
4. Dodanie Connection Manager (Menadżera Połączeń) Wprowadź Nazwiska .	<ul style="list-style-type: none"> Kliknij podwójnie na Flat File Source (Źródło Danych jako Plik Płaski) a następnie wybierz przycisk New, aby stworzyć nowego Connection Manager (Menadżera Połączeń) Wprowadź nazwę menadżera: Wprowadź Nazwiska. Jako źródła wskaż z lokalizacji podanej przez prowadzącego plik: np.: C:\HD\UID\InputA.txt Wybierz Advanced (Zaawansowane): <ul style="list-style-type: none"> Name: Nazwisko. OutputColumnWidth: 250. <p> Przeszukiwanie Rozmyte (Fuzzy Lookup) ostrzega, gdy wielkość kolumny źródła jest różna od kolumny, do której się odnosi.</p> <ul style="list-style-type: none"> W Edytorze Flat File Source (Danych jako Plik Płaski), wybierz zakładkę Columns (kolumny) aby ustawić mapowanie między kolumną zewnętrzną a kolumną wynikową.
5. Dodanie transformacji Fuzzy Lookup (Przeszukiwanie rozmyte)	<p> Teraz jesteś gotowy, aby porównać listę nazwisk w pliku wejściowym z tabelą ImieNazwisko.</p> <ul style="list-style-type: none"> Przeciągnij komponent Fuzzy Lookup (Przeszukiwanie Rozmyte) z Toolbox 'a (Panelu Narzędzi). Połącz wyjście Flat File Source (Źródło Danych jako Plik Płaski) z Fuzzy

	<p>Lookup (Przeszukiwanie Rozmyte).</p> <ul style="list-style-type: none"> Kliknij podwójnie na Fuzzy Lookup (Przeszukiwanie Rozmyte), aby otworzyć edytor: <ul style="list-style-type: none"> Connection manager: localhost.Studenci. Wygeneruj nowy index (new index) przy użyciu referencji do tabeli zawierającej nazwisko i nr indeksu np. NrAlbumu. Przeszukiwanie Rozmyte tworzy skomplikowany indeks na przeszukiwanej tabeli. Może tworzyć indeks za każdym razem bądź też przechowywać go w bazie danych. Dla dużych tablic, które nie zmieniają się zbyt często indeks powinien być przechowywany. Wybierz zakładkę Columns (kolumny) i stwórz połączenie pomiędzy Nazwisko a Student. Zaznacz Pass Through w Available Input Columns, aby nazwa pliku źródłowego była załączona w pliku wynikowym. Wybierz kolumnę Student jako kolumną przeszukiwaną. Wybierz OK.
<p>6. Dodanie SQL Server Destination (Źródła Wynikowe Jako SQL Server) z nowym menadżerem połączeń, aby połączyć bazę danych Studenci oraz nową tabelę ZnalezioneNazwiska</p>	<p> Umieść rezultaty wyszukiwania w nowej tabeli.</p> <ul style="list-style-type: none"> Przeciągnij SQL Server Destination (Źródło Wynikowe Jako SQL Server) z Toolbox 'a (Panelu Narzędzi). Przeciągnij zieloną strzałkę wyjścia Fuzzy Lookup (Przeszukiwania Rozmytego) do SQL Server Destination (Źródło Wynikowe Jako SQL Server). Kliknij podwójnie na SQL Server Destination (Źródło Wynikowe Jako SQL Server) i wybierz przycisk New, aby stworzyć nowego Menadżera Połączeń. Wybierz menadżera: localhost.Studenci. Wybierz przycisk New aby stworzyć nową tabelę <ul style="list-style-type: none"> Zmień nazwę tabeli z SQL Server Destination na ZnalezioneNazwiska <p> Zwróć uwagę na dodatkową kolumnę stworzoną do przechowywania statystyk.</p> <ul style="list-style-type: none"> Wybierz Mapowania (Mappings) w Edytorze SQL Server Destination (Źródło Wynikowe Jako SQL Server), aby poprawnie zmapować wejściową i wyjściową kolumnę Zamknij edytor.
<p>7. Dodanie Data Viewer (podglądu danych) dla Przeszukiwania Rozmytego (Fuzzy Lookup)</p>	<p> Podgląd Danych pozwala na monitorowanie przepływu danych. Umożliwia on na sprawdzenie wyników bez odnoszenia się do SQL Server Management Studio po zakończeniu zadania.</p> <ul style="list-style-type: none"> Kliknij ppm na połączenie pomiędzy Fuzzy Lookup (Przeszukiwaniem Rozmytym) a SQL Server Destination (Źródło Wynikowe Jako SQL Server) i wybierz Data Viewer (Podglądu Danych). W zakładce Data Viewers wybierz przycisk Add, aby dodać nowy Podgląd Danych typu Grid.
<p>8. Przetestowanie zadania Znajdź Nazwiska</p>	<p>Można przetestować część pakietu poprzez komendę Execute Task (Wykonaj Zadanie).</p>

	<ul style="list-style-type: none"> Na zakładce Control Flow (Kontrola Przepływu) kliknij ppm na zadaniu Znajdź Nazwiska. <p>Okno Podglądu zawiesza wykonanie zadania. Zwróć uwagę na dokładne trafianie (Adam Radomski, Aleksander Wukara, Artur Trulik, Arkadiusz Dorzyński), dla których statystyki „similarity” oraz „confidence” równe są 1.</p> <p>Aby kontynuować, zamknij okno podglądu, lub kliknij przycisk kontynuuj (zielona strzałka w lewym górnym rogu okna podglądu).</p> <ul style="list-style-type: none"> Wybierz przycisk Kontynuuj, zamknij podgląd oraz zatrzymaj debugger.
9. Dodanie Execute SQL Task (wykonanie zadania SQL)	<p>Jeśli uruchomisz zadanie po raz drugi te same wiersze zostaną dodane ponownie do tabeli ZnalezioneNazwiska. Trzeba więc czyścić tabelę przed każdym przeszukaniem.</p> <ul style="list-style-type: none"> Na zakładce Control Flow (Kontrola Przepływu) dodaj Execute SQL Task (Wykonanie Zadania SQL) do kontenera Plik Istnieje: <ul style="list-style-type: none"> Name: Wyczyść ZnalezioneNazwiska Connection: localhost.Studenci SQL Statement: Delete from ZnalezioneNazwiska.
10. Dodanie Wymuszenia Pierwszeństwa, aby uruchomić Wyczyść ZnalezioneNazwiska przed wykonaniem Znajdź Nazwiska	<ul style="list-style-type: none"> Przecignij strzałkę wyjścia z Wyczyść ZnalezioneNazwiska do Znajdź Nazwiska.

Laboratorium rozszerzone

Zadanie 1 (czas realizacji 30 min)

W Laboratorium podstawowym pisaliśmy prosty skrypt w języku Visual Basic 2008. Jednakże skrypt, który napisałeś nie jest satysfakcjonujący dla twojego przełożonego który nie akceptuje użycia funkcji Connection. Zmodyfikuj skrypt tak, aby nie było konieczności użycia tej funkcji. W razie problemów polecono ci skorzystać ze strony <http://technet.microsoft.com/en-us/library/ms345171.aspx>.

Zadanie 2 (czas realizacji 30 min)

W Laboratorium podstawowym pisaliśmy prosty skrypt w języku Visual Basic 2008, otrzymałeś polecenie przepisania skryptu w języku Microsoft Visual C# 2008 . W razie problemów polecono ci skorzystać ze strony <http://technet.microsoft.com/en-us/library/ms345171.aspx>

Zadanie 3 (czas realizacji 30 min)

Zlecono ci dokonać przeszukiwania rozmytego na tabeli NrAlbumu (Laboratorium rozszerzone moduł 1) przy użyciu pliku z numerami albumów Input1.txt oraz wyświetlenie rezultatów (oraz przechowanie ich w nowej tabeli systemu SQL Server) – numerów albumów oraz nazwisk studentów.

Wskazówka: określenia kolumn przeszukiwanych i wyświetlanych dokonujemy w komponencie **Fuzzy Lookup**. Pamiętaj o czyszczeniu wynikowej tablicy