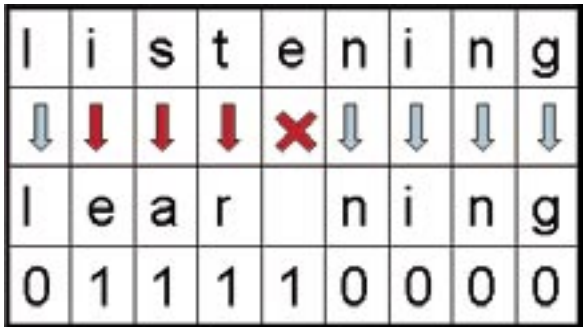
Damir Varga is senior consultant bij Avanade Netherlands BV

Fuzzy zoeken in SQL Server 2005

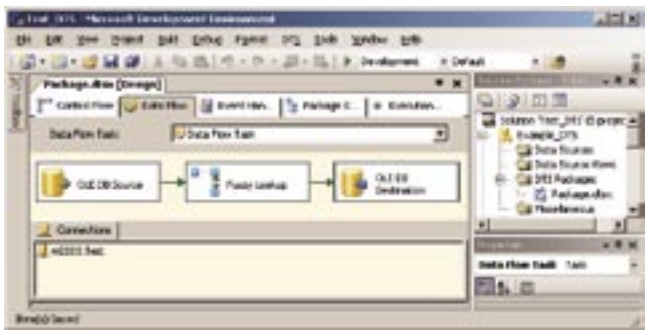
GEGEVENS SCHONEN MET FUZZY LOOKUP

Gegevens, die wij dagelijks als invoer voor onze informatiesystemen ontvangen, zijn vaak met spelfouten en ontbrekende of dubbele woorden afgeleverd. Daarom moeten wij zelf handmatig de juiste gegevens opzoeken, of dure systemen bouwen die dat voor ons gaan doen. Het bewerken van verkoopgegevens of aankoopbestellingen die je van een klant in een bestand hebt gekregen, kan behoorlijk wat tijd in beslag nemen. Klanten maken namelijk gebruik van hun eigen artikelomschrijvingen en namen die net niet overeenkomen met jouw artikelnamen. Met Data Transformation Services (DTS) voor SQL Server 2005 en Fuzzy Lookup- en Fuzzy Grouping-transformaties is het nu mogelijk om op een gemakkelijke manier deze onjuiste gegevens op te schonen, te veranderen of verder te bewerken.

Tot nu toe was het mogelijk met Data Transformation Services invoergegevens met referentiegegevens binnen een 'normale' lookup-transformatie te vergelijken en te matchen. Met fuzzy lookup en grouping DTS-transformaties biedt SQL Server 2005 DTS nu de mogelijkheid om gegevens te matchen die niet 100% gelijk zijn aan de referentiegegevens. De vraag is hoe de fuzzy match eigenlijk werkt. Om te beginnen moeten wij een referentietabel hebben. Deze referentietabel wordt aan het begin van het matching-proces verwerkt. Op basis van de inhoud wordt een nieuwe tabel gemaakt, de Error Tolerant Index-tabel. Deze wordt vervolgens gevuld met woorden en tokens (delen van woorden) uit de referentietabel. Bijvoorbeeld, een string "Microsoft Company" levert de volgende tokens: 'micr', 'roso', 'soft', 'comp' 'ompa' en 'mpan'. Deze tokens zijn gemaakt op basis van algoritmes die rekening houden met de token-volgorde en hun relatieve frequentie. De volgende stap is om van de invoergegevens op dezelfde manier tokens te genereren en die met tokens in de Error Tolerant Index-tabel te vergelijken. Het vergelijken is gebaseerd op een algoritme in SQL Server 2005 die gebruik maakt van een eigen, domeinonafhankelijke afstandsfunctie. Deze functie houdt rekening met de bewerkingsafstand en het aantal tokens.



Afbeelding 1. Voorbeeld van transformatie



Afbeelding 2. Data Flow-tab

Met de tokens en de zojuist genoemde functies berekent SQL Server 2005 de gelijkenis en waarschijnlijkheid van een mogelijk matchresultaat. De bewerkingsafstand is berekend op basis van de prijs om een invoerwaarde naar een referentiewaarde te transformeren. Elke transformatie (toevoeging, verwijdering, verandering) heeft zijn eigen prijs. In ons voorbeeld (afbeelding 1) is de transformatie van 'listening' naar 'learning' berekend als $4/9 \approx 0{,}44$ waar 4 staat voor het aantal transformaties en 9 het maximum aantal tekens in de invoerwaarde of de referentiewaarde weergeeft.

Gelijkenis en waarschijnlijkheid zijn waarden tussen 0 en 1, waarbij 1 een exacte match betekent. Gelijkenis vertelt ons hoe invoer- en referentiegegevens op elkaar lijken en waarschijnlijkheid vertelt ons hoe zeker we zijn van het matchresultaat. Stel, we hebben een referentietabel met twee referentiewaarden: 'SQL 2000' en 'SQL 2005'. Wanneer onze invoerwaarde 'SQL 2005' is, dan worden resultaten verkregen voor een match op 'SQL 2005' met gelijkenis 1 en waarschijnlijkheid 1 en een tweede match op 'SQL 2000' met gelijkenis 0,876 en waarschijnlijkheid 0,29. Het tweede resultaat toont hoge gelijkenis en lage waarschijnlijkheid hetgeen betekent dat een andere referentiewaarde meer gelijkenis vertoont met de invoerwaarde.

Fuzzy Lookup

Na dit korte overzicht hoe fuzzy matching werkt, gaan we nu zelf een matching-transformatie bouwen. Allereerst moet je een nieuw Data Transformation-project in de Business Intelligence Development Studio maken. Vervolgens moet je op de Control Flow-tab een Data Flow-taak toevoegen. Op de Data Flow-tab moet een source-, destination- en fuzzy lookup-component toegevoegd worden; zie afbeelding 2.

_key_in	_key_out	_score	Company	City	Country	Company_clean	City_clean	Country_clean
1	1	1	Avanade B.V.	Almere	Nederland	Avanade B.V.	Almere	Nederland
2	1	0.8687	Avnade B.V.	Almere	Nederland	Avanade B.V.	Almere	Nederland
3	3	1	Avanade	Seattle	USA	Avanade	Seattle	USA
5	5	1	Microsoft B.V.	Schiphol Rijk	Nederland	Microsoft B.V.	Schiphol Rijk	Nederland
4	5	0.7434	Microsoft	Schiphol Rijk	Nederland	Microsoft B.V.	Schiphol Rijk	Nederland
6	6	1	Micro soft	Schiphol Rijk	Nederland	Micro soft	Schiphol Rijk	Nederland

Afbeelding 4. Resultaten fuzzy grouping

Alle componenten zijn in de toolbox aan de linkerkant beschikbaar. Met een dubbelklik op de component open je een eigenschappenscherm. Begin bij de source-component. In de source-component kun je het bronbestand aangeven. Wanneer je klaar bent met alle overige source-componenteigenschappen verbind je de source-component met de fuzzy lookup-component. Dubbelklik op de fuzzy lookup-component. In het geopende eigenschappenscherm stel je de referentietabel in, koppel je de invoervelden en de referentievelden en stel je de parameters in voor de onderwaarde voor de gelijkenis, de token-scheidingstekens en het maximum aantal resultaten per matchopdracht; zie afbeelding 3.

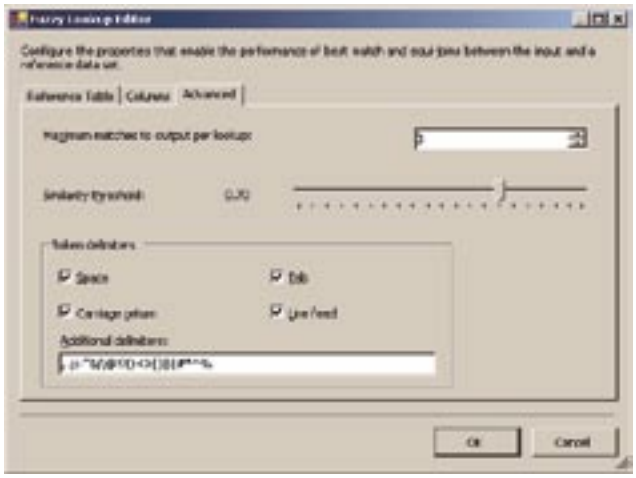
Ook heel interessant zijn de additionele opties onder de Reference-tab. Daar kun je instellen dat je een nieuwe Error Tolerant Index-tabel wilt creëren en opslaan voor later gebruik. Je kunt ook instellen dat je een bestaande ETI-tabel wilt gebruiken als een referentietabel. Deze mogelijkheden zijn echt handig als je met een grote referentietabel moet werken waarbij de opbouw van de ETI-tabel lang duurt. In dat geval kun je de ETI-tabel tijdens een eerste matching opbouwen. Bij volgende transacties kun je de opties zo instellen dat je gebruik wilt maken van deze reeds opgebouwde ETI-tabel.

De volgende stap in ons matching-project is om de uitvoer van de fuzzy lookup-component met de destination-component te verbinden. Voordat we met de laatste stap in ons matching-project verder kunnen gaan, moeten we een tabel in SQL Server 2005 hebben waarin we de resultaten van de matching kunnen opslaan. Het is mogelijk om in deze tabel de volgende velden op te slaan: referentievelden, invoervelden en extra uitvoervelden uit de fuzzy lookup-component. De extra uitvoervelden zijn bijvoorbeeld _Similarity en _Confidence (respectievelijk gelijkenis en waarschijnlijkheid).

Ten slotte moet je als laatste stap in de destination-component een verbinding maken tussen de uitvoervelden uit de fuzzy component en de velden uit de resultaattabel.

Geavanceerde opties

Met een druk op de rechtermuisknop is een Advanced Editor-optie te selecteren. Je kunt in die Advanced Editor additionele settings voor je matchproject instellen. Zo is het mogelijk om een hiërarchie en match voor de invoervelden aan te geven. Met een hiërarchie kun je weegfactoren toekennen aan bijvoorbeeld 1. Bedrijfsnaam, 2. Stad, 3. Provincie en 4. Land. Met deze instelling telt het eerste veld, de bedrijfsnaam, het zwaarst in de matching. Een andere optie die wij hier kunnen kiezen is de wijze waarop we de invoervelden willen vergelijken. Zo kan fuzzy match worden uitgeschakeld en alleen een exacte match worden ingesteld.



Afbeelding 3. De Fuzzy Lookup editor

Fuzzy Grouping

Een andere nieuwe transformatie in SQL Server 2005 DTS is de fuzzy grouping-transformatie. Met deze transformatie is het mogelijk om invoergegevens die op elkaar lijken te groeperen waardoor dubbele invoergegevens worden geëlimineerd. Tijdens het groeperen wordt de meest representatieve invoer als match voor de gehele groep gekozen.
Fuzzy grouping heeft minder opties dan de fuzzy lookup-transformatie, maar de gelijkenis, token-scheidingstekens, weegfactoren, match-type, enzovoort zijn aanwezig en te gebruiken. Het maken van een fuzzy grouping is gelijk aan het maken van een fuzzy lookup.

Afbeelding 4 toont de resultaten van een fuzzy grouping die als drempelwaarde voor de gelijkenis 0.6 heeft meegekregen en een hiërarchie van Company, City en Country. _key_in is een sleutel voor de invoergegevens, _key_out is een sleutel voor gegroepeerde gegevens en het _score-veld is de gelijkenis tussen invoergegevens en de groep-match.
Fuzzy grouping creëert een tijdelijke ETI-tabel die proportioneel groeit met de grootte van de invoergegevens. Hiermee dient dus rekening te worden gehouden bij de calculatie voor de grootte van de database.

Experimenteer!

Met de nieuwe SQL Server 2005 is het gemakkelijk geworden om de gegevens die we van onze handelspartners, klanten en zelfs van andere afdelingen binnen ons bedrijf op dagelijkse basis ontvangen, op te schonen en verder te verwerken. De getoonde fuzzy lookup- en fuzzy grouping-transacties kunnen alleen een klein stapje in een groter proces zijn waar additionele verwerkingen, verrijkingen, enzovoort voor of na fuzzy transformaties plaatsvinden. We kunnen bijvoorbeeld eerst een exacte match in onze referentietabel proberen te vinden en als dat niet lukt de fuzzy lookup inschakelen. De drempelwaarde voor de gelijkenis en hiërarchie met weegfactoren zijn krachtige opties ondanks dat ze er simpel eruitzien. Experimenteer met deze opties en bekijk wat voor jou het beste werkt.

Damir Varga is senior consultant bij Avanade Netherlands BV (www.avanade.nl). Dit artikel is tot stand gekomen mede dankzij Sebastian Hek

Nuttige internetadressen

- http://msdn.microsoft.com/sql
- http://msdn.microsoft.com/SQL/2005/
- http://msdn.microsoft.com/library/en-us/dnsql90/html/fzdtssql05.asp
- http://research.microsoft.com
- http://www.microsoft.com/sql