Onderzoeksrapport overzetten brongegevens met behulp van ETL tools

Pentaho Spoon, Enhydra Octopus (TDT) en Jitterbit integration Environment

Auteur:
Wouter van der Meulen Kunee

Versie: 1.0

De Haagse Hogeschool

Voorwoord

Dit rapport is geschreven om de verschillen tussen 3 ETL tools inzichtelijk te maken. Hieruit voortvloeiend zal er een advies komen voor het gebruik en de toegankelijkheid van de verschillende pakketten.

Voor het onderzoek naar de verschillende pakketten heb ik gebruik gemaakt van het internet en heb ik zelf 3 ETL tools getest, hierbij komen ook de mogelijkheden van Open Source oplossingen aan bod.

Den Haag, 14 november 2008

Wouter van der Meulen Kunee

1 Samenvatting

Dit onderzoeksrapport is gericht op de vergelijking tussen 3 ETL tools, hierbij heb ik gekozen voor 3 bekende pakketten: Pentaho Spoon, Enhydra Octopus (TDT) en Jitterbit integration Environment.

Ik heb een analyse gemaakt van de software pakketten en hoe de applicaties werken. Dit heb ik gedaan door middel van het raadplegen van het internet en praktijkervaring opdoen met de verschillende pakketten. Met deze analyse heb ik de verschillende software pakketten op functionaliteit met elkaar vergeleken.

Op een duidelijke manier heb ik de stappen en acties beschreven om een test correct uit te voeren. Ook voor niet technische mensen, moest het duidelijk moeten zijn om de testen uit te voeren.

2 Inhoudsopgave:

3 Inleiding

Ik heb de opdracht gekregen om onderzoek te verrichten naar de verschillen tussen 3 ETL tools.

Ik heb hiervoor de volgende pakketten gekozen

De pakketten:

- Pentaho Spoon 3.1.0
- Enhydra Octopus, TDT (Together Data Transformer 3.6)
- Jitterbit integration Environment 2

Hierbij doe ik een onderzoek naar de basisfunctionaliteiten

- Extract (Selecteren van de benodigde tabellen, velden, informatie)
- Transform (Het bewerken van de data om deze op de gewenste manier te laden)
- Load (Het daadwerkelijk inladen van de gegevens)

4 Opdracht omschrijving

4.1 Inleiding

In dit hoofdstuk wordt beschreven hoe de opdracht luidt, de opdracht is afgebakend, waarom ik voor de te onderzoeken applicaties heb gekozen, welke aanpak ik heb gehanteerd.

4.2 Opdracht omschrijving

Ga op zoek naar tools waarmee een extractie van brondata naar de nieuwe database gedaan kan worden, inclusief de noodzakelijke transformaties, zogenaamde ETL tools. Onderzoek er 3 producten (zorg ook voor open source mogelijkheden). Doe met elk van de producten testen om de kwaliteit vast te stellen. Formuleer een advies voor een te gebruik tool.

4.2.1 Afbakening

Tijdens het onderzoek beperk ik me tot het analyseren van maar 3 pakketten.
Ik heb mij hierbij gehouden aan de basisfunctionaliteiten en houdt hierbij geen rekening gehouden met extra functionaliteiten die geboden worden.
Wegens het feit dat het niet haalbaar is om een commerciële omgeving (Oracle OWB) op te zetten vanwege de complexiteit die dit met zich meebrengt beperk ik mij tot de opensource mogelijkheden

4.3 De pakketten

4.3.1 Inleiding

Dit hoofdstuk zal kort beschrijven welke pakketten ik ga onderzoeken.

4.3.2 Spoon (Opensource)

Spoon is een ETL tool van het bedrijf Pentaho. Zij omschrijven dit als een Data Integration pakket.
Dit pakket is weer onderdeel van het Pentaho BI suite.
Spoon is een uitgebreid grafisch ETL pakket en heette voorheen Kettle

4.3.3 Enhydra Octopus (TDT, Together Data Transformer) (Opensource)

TDT is een pakket van het bedrijf Enhydra. Dit pakket bestaat echter ook onder de naam Enhydra Octopus.
Het pakket is bedoeld voor zeer basic gebruik.

4.3.4 Jitterbit integration Environment (Opensource)

Jitterbit integration Environment is een product van het bedrijf Jitterbit.
Het pakket is net als Spoon een grafisch uitgebreide tool.

4.4 Aanpak

Om de pakketten gelijkwaardig te onderzoeken heb ik gebruik gemaakt van een vast testplan, welke ik hieronder ga beschrijven.

4.4.1 Werkomgeving

Werkmachine
Intel Centrino 2,26Ghz, 2gb ram, 100gb Hd
Microsoft Windows XP PRO
WAMP MySQL server

Brongegevens
Microsoft Access Database (C:\Person.mdb)

Doel
MySQL database (Datawarehouse)
Tabellen gebruikt: DW, Personnell
Of
Microsoft Access Database (C:\Person2.mdb)

4.4.2 Extract

Nadat de pakketten geïnstalleerd en geconfigureerd zijn moet als eerste de brongegevens beschikbaar zijn. De brongegevens staan in een Access Database (Een personeelsbestand) welke lokaal opgeslagen is. Met behulp van de ETL tool moeten hier brongegevens uit gehaald kunnen worden.

4.4.3 Transform

Tijdens het overhalen van de brongegevens, moeten de gegevens getransformeerd kunnen worden met behulp van de ETL tool. Ik heb ervoor gekozen om hierbij een numeriek veld, het jaarsalaris van het personeel, te vermenigvuldigen met het favoriete nummer van het desbetreffende personeelslid.

4.4.4 Load

Nadat de gegevens tijdens het overhalen zijn getransformeerd, moeten deze ingeladen kunnen worden in een andere omgeving, dit zou een datawarehouse kunnen zijn. Ik heb ervoor gekozen om dit in een lokale MySQL omgeving te laden of een Microsoft Access Database.

5 Resultaten

5.1 Spoon

5.1.1 Installatie / configuratie

Voordat het pakket Spoon opgestart kan worden moet er een Java VM met een versie 1.5 of hoger geïnstalleerd zijn. Verder moeten er ook een aantal environment variabelen ingesteld worden. Hoewel dit niet expliciet bij het installeren wordt genoemd, wordt dit met het opstarten van het pakket wel duidelijk. Hoewel het pakket Spoon heet, heet het opstartbestand wel Kettle.exe

Het opstarten laat meteen een mooi grafisch opstartscherm zien (zie bijlage fig 3.1)
Hierbij kan er wel of niet gekozen worden om gebruik te maken van een repository. Aangezien ik deze niet heb ik hier niet voor gekozen.

Na het opstarten wordt er meteen een soort wizard/help opgestart (zie bijlage fig 3.2)

Score: =

5.1.2 Extract

Het scherm om de connecties en transformaties te maken valt ietwat tegen (zie bijlage fig 3.3) , maar spreekt voor zich.

Het aanmaken van de koppeling van de brongegevens is zeer mooi grafisch weergegeven, hierbij kan gebruik gemaakt worden van Drag & drop (zie bijlage fig 3.3)
Een duidelijke menustructuur laat de verschillende typen brongegevens die gebruik kunnen worden zien. Na het slepen van de Access file in het rechterscherm, het werkscherm, wordt een duidelijk scherm weergegeven waarin de brongegevenskoppeling ingevuld kan worden (zie bijlage fig 3.4 en 3.5)

Wat zeer fraai is, is dat de structuur van de brongegevens meteen bekeken kan worden (zie bijlage fig 3.6 en 3.7)

Score: ++

5.1.3 Transform

Spoon heeft een drag & drop functie welke gebruikt kan worden om transformaties te maken.
In bijlage fig 3.19 is te zien dat er tussen 2 Access Databases een transformatie gebeurt. Via drag and drop van de source op de transformatie ontstaat er een flow van de informatiestroom.
Vervolgens kan in de transformatie, de calculatorfunctie, de betreffende transformatie gedaan worden (zie bijlage fig 3.18).
Vervolgens kan de transformatie gesleept worden op de doeldatabase.

Score: ++

5.1.4 Load

Het laden van de gegevens gaat zeer simpel, men drukt de “voer transformatie uit” knop in, deze ziet er uit als een groene play button. Men krijgt de optie om deze lokaal of op afstand (via server) uit te voeren (zie bijlage fig 3.21)
Nadat de gegevens getransformeerd zijn worden ze succesvol ingeladen in de doeldatabase (zie bijlage fig 3.20)

Score: ++

5.1.5 Foutafhandeling

Foutafhandeling binnen Spoon gebeurt zeer duidelijk. Er worden duidelijke meldingen gemaakt wat er fout gaat (zie bijlage fig 3.9) en zodoende kan de fout snel verholpen worden.

Score: ++

5.1.6 Helpfunctie

Spoon heeft geen built-in helpfunctie, maar een online helpfunctie / documentatie (zie bijlage fig 3.15)
Het nadeel hiervan is dat er altijd een internetverbinding beschikbaar moet zijn om hier gebruik van te maken. Het grote voordeel hiervan is wel dat de online documentatie centraal ge-update kan worden en voor iedereen beschikbaar is.

Score: +

5.1.7 Grafisch

Het opstarten laat een mooi grafisch opstartscherm zien (zie bijlage fig 3.1)
Ook zijn de onderdelen goed gescheiden weergegeven. Dit is gedaan met behulp van gegroepeerde functies. In bijlage fig 3.4 is te zien hoe alle verschillende typen bronnen allemaal onder input vallen, dit is net zo bij transformations, output en de overige opties.

Score: ++

5.2 Enhydra Octopus (TDT, Together Data Transformer)

5.2.1 Installatie / configuratie

TDT vereist geen speciale omgeving. Hierdoor is het opzetten van een omgeving niet aan de orde of stelt weinig voor.
Het opstarten van TDT laat meteen een kaal opstartscherm zien (zie bijlage fig 2.1)
Het aanmaken van een nieuw project is niet moeilijk, uit het start menu kan men de optie nieuw project aanmaken kiezen en een naam opgeven voor het nieuwe project (zie bijlage fig 2.2)

Score: ++

5.2.2 Extract Transform en Load

Het extract, transform en load gedeelte van TDT wordt allemaal op 1 scherm weergegeven.
De opties met betrekking tot het opzetten van de brongegevens zijn zeer beperkt. Men heeft de keus uit talloze typen brongegevens zoals MySQL, PostgresSQL, maar geen Microsoft Access Database!
Bij het kiezen van de doeldatabase heeft men echter wel de keus voor een Microsoft Access Database! Om toch een compromis te maken heb ik ervoor gekozen om een MySQL brondatabase te laden in een Microsoft Access Database.
Bij het opzetten van de configuratiegegevens (zie bijlage fig 2.3) krijg ik een aantal vraagtekens, de helpfunctie raadplegen heeft hierbij niet geholpen, omdat deze zeer karig is (zie bijlage fig 2.5)

Score: -

5.2.3 Helpfunctie en foutafhandeling

Na lang experimenteren met parameters en met trail-and-error ben ik niets wijzer geworden van het programma TDT. De foutmelding die ik krijg met het connecteren van de bron- en doeldatabase is dermate niet toereikend (zie bijlage fig 2.4) dat hier niet mee te werken valt.

Score: --

5.2.4 Grafisch

Grafisch stelt TDT weinig voor (zie bijlage fig 2.1). De indeling van het scherm qua input en output is op wel duidelijk gegroepeerd, jammer dat hier weinig info op wordt gegeven.

Score: =

5.3 Jitterbit

5.3.1 Installatie / configuratie

Het opzetten van de Jitterbit omgeving is niet moeilijk. De Jitterbit Integration Environment vereist een Jitterbit Integration Server. De installatie en configuratie van beide pakketten wijst zich vanzelf. Er kan gebruikt gemaakt worden van een standaard configuratie waarbij database accounts vanzelf gecreëerd worden. (zie bijlage fig 1.1)

Nadat de server is geïnstalleerd kan er gewerkt worden met het ETL pakkett.
Het opstartscherm laat een mooi en duidelijk menu zien waarin een nieuw project aangemaakt kan worden of een oud project geladen kan worden.

Score: +

5.3.2 Extract

Het opzetten van de brongegevens doet men in de Project Tree onder het kopje Sources, dat is vrij duidelijk aangegeven. Bij het aanmaken van een nieuwe brondatabase krijgt men een mooi scherm te zien waarbij men de keus krijgt tussen de verschillende typen databases (Access, SQL, etc.) (zie bijlage fig 1.2).
Bij het opzetten van de Access database onstaan er echter problemen, namelijk de connectie met de ODBC driver van Windows XP. De oplossing hiervoor wordt echter niet in de help functie weergegeven (zie bijlage fig 1.11). Google bood hierbij een goede hulp. In een forum was een gebruiker met hetzelfde probleem, de oplossing die gegeven werd was om een handmatige Connectiestring te maken (zie bijlage fig 1.2)

Score: +

5.3.3 Transform

Om de brongegevens te transformeren moet er een nieuwe “Operation” aangemaakt worden. Dit gebeurt ook in het overzichtelijke menu, Project Tree.
Bij het maken van een nieuwe Operation krijgt men grafisch te zien hoe de Operation eruit komt te zien (zie bijlage fig 1.4)
Na het maken van de Operation, krijgt men meteen een soort van Transform wizard te zien. Hierna kunnen er diverse transformaties gemaakt worden (zie bijlage fig 1.10 en 1.12) en de velden van de brondata met de doeldatabase gekoppeld worden (zie bijlage fig 1.5)

Score: ++

5.3.4 Load

Bij het laden van de gegevens, het deployen, krijgt men een scherm te zien waarop men kan aangeven welke gegevens er gedeployed moeten worden (zie bijlage fig 1.6 en fig 1.7)
De gevulde datawarehouse database (zie bijlage fig 1.9) laadt de succesvolle ETL transactie zien.

Score: ++

5.3.5 Helpfunctie

De helpfunctie van het pakket (zie bijlage fig 1.11) is overzichtelijk en uitgebreid.

Score: ++

5.3.6 Foutafhandeling

Bij het deployen kunnen er echter fouten optreden, maar deze fouten worden netjes weergegeven (zie bijlage fig 1.8). Hierdoor kan de fout makkelijk worden opgespoord en gecorrigeerd.

Score: +

5.3.7 Grafisch

Jitterbit is grafisch mooi weergegeven. Duidelijk ingedeelde schermen, knoppen en zonder poespas weergegeven opties laat de nieuwe gebruiker de interface van Jitterbit snel leren.

Score: +

6 Conclusie en aanbevelingen

6.1 Inleiding

Het is heel moeilijk om de kwaliteit van software pakketten te meten met behulp van kwantitatieve meetgegevens, daarom heb ik gebruik gemaakt van een kwalitatieve schaal. Ik heb hierbij de volgende schaalverdeling gemaakt die ik per functionaliteit invul:

Zeer slecht	--
Matig	-
Neutraal	=
Goed	+
Zeer goed	++

Door gebruik te maken van deze schaal is het niet mogelijk om een totaal score te berekenen door het gemiddelde van alle functionaliteiten te nemen. Dit is ook niet mijn bedoeling. Mijn bedoeling is om een afweging te maken op basis van de functionaliteit die gewenst is en deze kan per gebruiker verschllen.

6.2 Aanbevelingen

Pakket	Installatie/ Configuratie	Extract	Transform	Load	Help-functie	Fout-afhandeling	Grafisch
Spoon	=	++	++	++	+	++	++
TDT	++	-	-	-	--	--	=
Jitterbit	+	+	++	++	++	+	+

In het hoofdstuk resultaten heb ik per functionaliteit per pakket een score ingevuld, in bovenstaande tabel is het totale resultaat te zien.

Vanwege de gebrekkige helpfunctie en onduidelijke werkwijze van TDT is dit pakket niet aan te raden voor gebruikers met gelimiteerde kennis van het opzetten van een datawarehouse.

Zowel Jitterbit als Spoon zijn uitstekende ETL tools. Beide hebben een prima functionaliteit en zijn gebruiksvriendelijk.

Spoon heeft betere ondersteuning als het gaat om koppelingen leggen tussen bron- en doeldata. Jitterbit heeft een betere helpfunctie en de omgeving is iets sneller opgezet dan Spoon. Spoon heeft kant-en-klare transformaties in het linkermenu staan waardoor het voor de gebruiker makkelijk is om te vinden wat hij zoekt.
Al met al is op dit moment Spoon de beste oplossing als ETL tool. Qua functionaliteit is deze het beste getest en grafisch is deze beter opgezet zodat gebruikers de interface sneller kunnen leren. Ook het gebruik van Transformation plugins geeft Spoon een voorsprong, hierdoor is Spoon uitbreidbaar en kan daarmee alle kanten op.
Jitterbit is ook een goed pakket, maar is nog niet opgewassen tegen de mogelijkheden van Spoon.

7 Bijlagen

7.1 Bijlage A

Zie “Bijlage A.doc”