Aantrekkelijke aanvullingen op standaard BI-implementaties

# Open Source Analytics

Jos van Dongen

**In een BI-special mag een artikel over Open Source BI-tools natuurlijk niet ontbreken. In eerdere nummers van DB/M zijn al Open Source ETL-tools, databases en rapportage-pakketten besproken. Deze hulpmiddelen vormen weliswaar een goede basis voor een complete OS BI-oplossing, maar zijn meer gericht op het bij (eind)gebruikers krijgen van informatie en niet op het bieden van complexe, interactieve analyses.**

We hebben het dan over OLAP, datamining en statistische analyse. In hoeverre de Open Source wereld in staat is ook hiervoor bruikbare software aan te bieden wordt in dit artikel besproken. De meest bekende OLAP speler in de OS wereld is Mondrian, al geruime tijd onderdeel van de Pentaho Suite en hoofdrolspeler in dit artikel waar het gaat om OLAP. Ook het open source data-mining-project Weka is onderdeel van Pentaho en wordt hier onder de loep genomen. Wanneer gekeken wordt naar statistische analyse is er nog geen BI-suite die hiervoor componenten aan-biedt. Waarschijnlijk heeft dit te maken met het feit dat statistiek een wat taaier onderwerp is voor de gemiddelde business user. Dat weerhoudt ons er echter niet van om wat dieper in te gaan op het 'R' project om het trio toepassingen compleet te maken.

## Mondrian

Een OLAP front end, soms zelfs in combinatie met een eigen OLAP server, maakt meestal deel uit van het aanbod van elke zichzelf respecterende BI-leverancier. Het zal daarom geen verbazing wekken dat ook Pentaho, één van de grotere BI-spelers in de Open Source markt, een OLAP onderdeel in zijn suite heeft opgenomen. Het betreft hier Mondrian, een bijzonder product omdat het geen echte database is (de data staan in een SQL database naar keuze), maar ook geen OLAP front-end. Wat is het dan wel? Simpel gezegd vormt Mondrian de middleware tussen OLAP clients en SQL databases, waardoor multi-dimensionele analyse mogelijk wordt op een relationele database. Eigenlijk is het dus een pure ROLAP-tool. Met Mondrian kunnen MDX query's worden afgevuurd op een SQL database, waarbij Mondrian zorgt voor parsing van de query en cachen van informatie.

Er is echter één groot verschil met de commerciële producten: de MDX query dient wél zelf (correct) geformuleerd te worden. Inzage in de beschikbare dimensies en members is er niet via de front-end, waardoor het toch wel omslachtig is om de goede selectie te maken. Gelukkig is er wel de Mondrian Workbench waarmee de cubes gedefinieerd kunnen worden en die tevens dient om de gewenste MDX te testen. Het resultaat van de MDX query vormt de dataset waarmee het eigenlijke analysewerk kan beginnen. In de meeste gevallen zal hiervoor het OS product jPivot gebruikt worden als analyse front-end, maar het staat eenieder vrij om een tool naar keuze te gebruiken, aangezien Mondrian ook over een XML/A API beschikt.

## Architectuur Mondrian

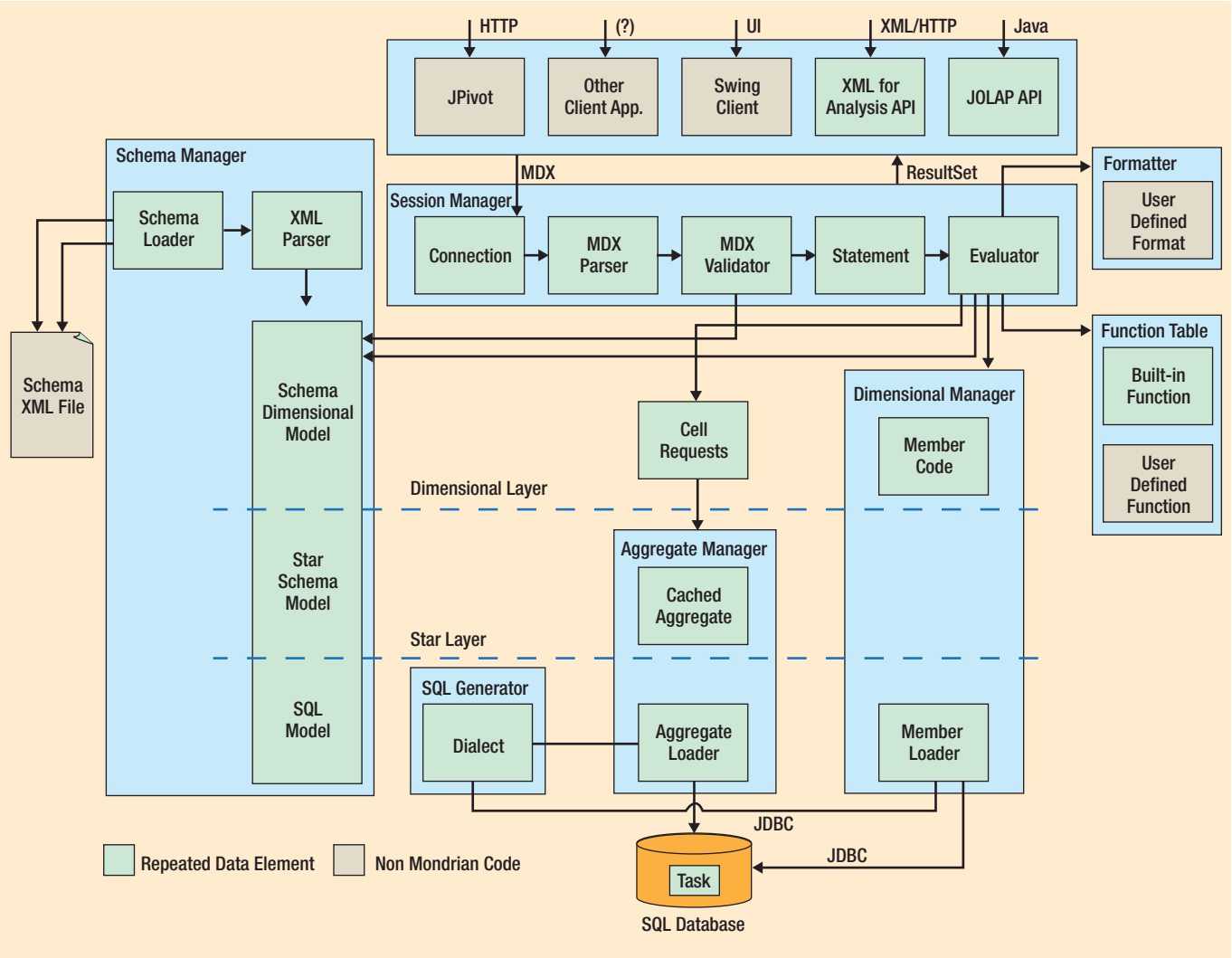
De Mondrian software bestaat uit vier lagen, zie afbeelding 1:

*Presentatielaag.* Deze zorgt voor het weergeven van datasets in de vorm van pivot tables of grafieken. De presentatielaag is overigens geen kant en klare GUI, maar levert alleen maar de bouwstenen om deze te kunnen vullen in de vorm van diverse API's.

> Het cache-mechanisme is het kroonjuweel van Mondrian

*Dimensielaag.* Hier vindt het parsen, valideren en uitvoeren van de MDX plaats. De MDX-implementatie is op een paar kleine, goed gedocumenteerde details na volledig gelijk aan die van Microsoft. Dit heeft weer als bijkomend voordeel dat er een groot aantal boeken en online documentatie voor deze taal beschikbaar is.

*Sterlaag.* Hierin wordt het cachen van de geaggregeerde gegevens geregeld. Het cache-mechanisme is het kroonjuweel van Mondrian, omdat hiermee een goede performance wordt bereikt, zelfs als de onderliggende database hier niet voor is




**Afbeelding 1:** Mondrian Architectuur.

geoptimaliseerd. Een en ander hangt uiteraard wel af van de hoeveelheid geheugen die aan het Mondrian proces kan worden toegewezen.

*Datalaag.* Deze wordt gevormd door een SQL database naar keuze. Op de Mondrian site wordt een flinke lijst met databases genoemd die in elk geval compatibel zijn met het product, maar ook als de database niet in deze lijst voorkomt zal het waarschijnlijk wel werken: de enige vereiste is dat de database beschikt over een standaard JDBC driver.

De keuze voor een ROLAP-oplossing is gemaakt omdat hiermee de noodzaak vervalt om een eigen opslagstructuur te ontwikkelen, terwijl het op deze manier mogelijk is om OLAP-analyses uit te voeren op constant wijzigende data, zonder de overhead van en vertraging door een cube processor. Ook levert dit een compact en gemakkelijk te installeren product op.

## Gebruik Mondrian

Om met een eigen database aan de slag te kunnen, dient allereerst de Workbench gebruikt te worden voor het maken van een JDBC-connectie naar de database en het definiëren van het schema, bestaande uit cubes, dimensies en measures. Ook zaken als named sets, user defined functions, calculated members, virtual cubes en dimensies, hierarchieën en property's kunnen opgenomen worden in het schema. Nadat het schema compleet is kan het worden getest door middel van het afvuren van MDX query's. Het schema is een XML file dat vervolgens door jPivot gebruikt kan worden als metadata om de ingevoerde MDX statements te vertalen. Hoewel jPivot een prima hulpmiddel is voor de doorgewinterde MDX kenner en een bruikbare set OLAP-bewerkinge n aanbiedt (zie afbeelding 2), is het (juist door het ontbreken van een visuele selectiemogelijkheid) niet bepaald geschikt voor business users. Ondanks deze tekortkoming is het product met een gestage opmars bezig, zelfs zo gestaag dat Pentaho/Mondrian in de 2007 versie van de OLAP Survey van Nigel Pendse is opgenomen.

## Datamining met WEKA

Datamining is een verzameling technieken om op geautomatiseerde wijze patronen en verbanden te ontdekken in (grote) gegevensverzamelingen. Het doel van deze exercitie is deze

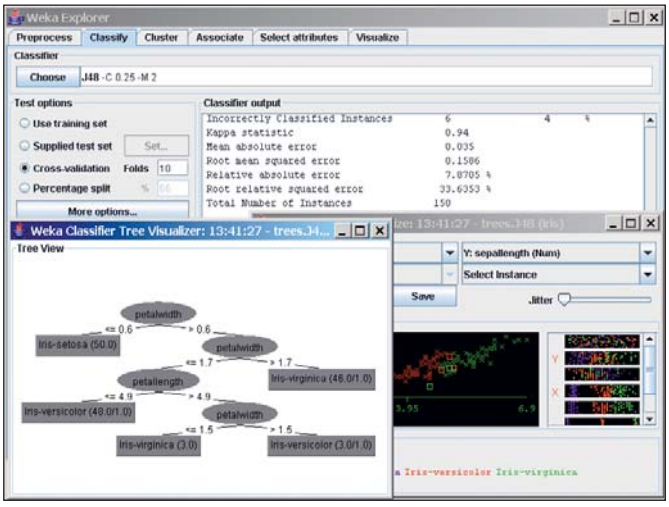
verbanden te gebruiken om voorspellingen te doen over nieuwe data die niet in de oorspronkelijke verzameling aanwezig waren. Er hangt vaak nog een waas van geheimzinnigheid over het onderwerp, wat deels veroorzaakt wordt door de vaak complexe en ondoorgrondelijke algoritmen die worden gebruikt. Gelukkig wordt het zelf experimenteren met datamining met behulp van producten als MS Analysis Services en Weka heel laagdrempelig, waardoor het hopelijk een bredere toepassing krijgt dan momenteel het geval is. Datamining wordt bijvoorbeeld al gebruikt bij mobiele operators om de 'churn' te minimaliseren, bij financiële instellingen om de kredietwaardigheid van klanten te beoordelen of om fraude te detecteren, en in de medische wereld om de waarschijnlijkheid van een hartaanval te voorspellen in het geval van pijn in de borst.

Ondanks het feit dat het WEKA-project al enige tijd geleden is geadopteerd door Pentaho, zullen nog weinig mensen dit product kennen. Weka wordt, zoals veel OS software, vooral gebruikt in de wetenschappelijke wereld en het onderwijs en heeft zich daar een prominente positie verworven. Weka is oorspronkelijk ontwikkeld aan de universiteit van Waikato in Nieuw Zeeland en is het acroniem voor Waikato Environment for Knowledge Analysis.

De Weka is echter ook een vogel die alleen in Nieuw Zeeland voorkomt en het symbool voor het project. Weka is van origine een set van datamining ('machine learning') algoritmen ontwikkeld in TCL/TK, C en Makefiles, waarna in 1997 is besloten om het hele product opnieuw te ontwikkelen in Java. Het grote voordeel hiervan is dat Weka op elk denkbaar platform draait, zo lang er maar een Java Virtual Machine is geïnstalleerd. Er zijn twee 'major' versies: de 'book' versie (3.4, behorend bij het leerboek 'Data Mining' van Ian Witten en Eibe Frank) en de 'developer' versie die elke dag wordt bijgewerkt (3.5.6 op het moment van schrijven).

## Architectuur WEKA

Weka bestaat uit een aantal verschillende onderdelen. Allereerst natuurlijk de algoritmen die het hart vormen van het systeem.




**Afbeelding 2:** Weka Explorer.

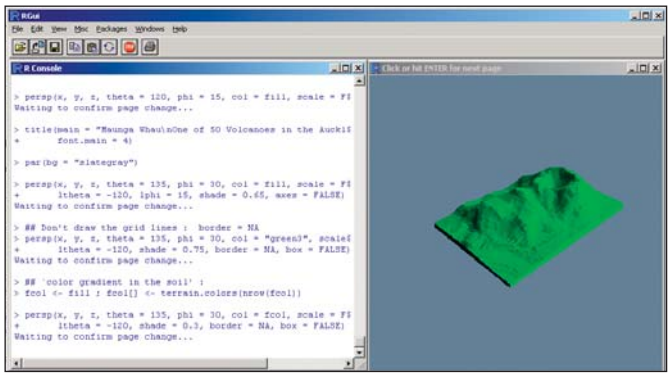
Hieromheen is een workbench ontwikkeld waarmee het eigenlijke werk gedaan wordt. De workbench bevat vier applicaties, waarvan de Explorer en de Experimenter de belangrijkste zijn. Het menu bevat daarnaast nog tools om zogenaamde ARFF files (Weka's eigen Attribute Relational File Format) en SQL data te bekijken en een aantal visualisatiekeuzes, waarmee data geplot of beslissingsbomen in beeld gebracht kunnen worden.
In de meeste gevallen zal gestart worden met de Explorer (zie afbeelding 2) die allereerst gebruikt wordt voor het preprocessen van data. Deze data kunnen afkomstig zijn uit een flat file of een database, waarbij het ARFF format het meest voor de hand ligt vanwege de geïntegreerde viewer/editor. Om de aldus verkregen dataset te verkennen, is een groot aantal cluster-, classificatie- en associatiealgoritmen beschikbaar. Met de Experimenter kunnen meerdere datasets met meerdere algoritmen in batch worden geanalyseerd om het statistisch beste algoritme te achterhalen. De resultaten van een run kunnen worden weggeschreven in ARFF- of CSV-formaat, of via een JDBC interface naar een database.
Belangrijk bij het opzetten van een datamining-proces is het trainen en testen van het model. Trainen betekent het afleiden van patronen uit een dataset; testen is vervolgens het valideren van deze patronen met behulp van een andere dataset. Weka biedt voorzieningen om zowel de train- als de testdata uit hetzelfde bestand te lezen, waarbij bijvoorbeeld 80 procent van de data random wordt geselecteerd om te trainen, en 20 procent om te testen.

## Gebruik WEKA

Het installeren van Weka is werkelijk een fluitje van een cent. Simpelweg downloaden en uitvoeren van de installer is voldoende, dus het draait binnen een paar minuten. Weka is voorzien van uitgebreide help- en tutorial files om snel met het product aan de slag te kunnen. Let echter wel op: een eenvoudig te gebruiken datamining-product betekent niet dat de gebruiker plotseling een specialist is op dit gebied. Integendeel: zonder de nodige theoretische (statistische) kennis is Weka niet bruikbaar. De meegeleverde documentatie dient dan ook alleen om het product toegankelijker te maken, niet om algemene datamining-kennis op te doen.
Voor de geïnteresseerde lezer is het boek 'Data Mining, Practical Machine Learning Tools and Techniques' van Witten en Frank een uitstekende introductie; voor mensen met minder tijd die snel een overzicht willen krijgen is http://www.thearling.com/dmintro/dmintro_2.htm een prima online alternatief. Voor wie echter weet waar hij of zij mee bezig is, biedt Weka met zijn makkelijk te gebruiken interface en de duidelijke structuur waarmee het pakket is opgezet een prima datamining tool.

## Statistiek: R

Een beetje vreemd om een pakket 'R' te noemen, ook al is het een Open Source-variant van 'S'. Dat zal wel aan de gewoonte in de Unix-wereld liggen om alle commando's zo kort mogelijk te houden. 'R' is een geïntegreerde verzameling tools voor het



**Afbeelding 3:** Rgui.

bewerken en visualiseren van data. R is ook een taal die door gebruikers uitgebreid kan worden met nieuwe functies. De makers van R noemen het niet zozeer een statistisch pakket, als wel een omgeving waarbinnen statistische methodes zijn geïmplementeerd. Het sterke punt van R bestaat uit de hoge kwaliteit grafische output, zowel op het scherm als in geprinte vorm. Hier is ook een belangrijk onderscheid met een pakket als Weka te vinden, dat sterk is in de analyse maar in de presentatie wat minder te bieden heeft dan R.

## Architectuur R

R is in wezen een taal inclusief een run-time omgeving met visualisatie-opties, een debugger en toegang tot systeemfuncties. R kan ook programma's uitvoeren die zijn opgeslagen in script files. De functionaliteit van R wordt beschikbaar gesteld via zogenaamde packages, waarvan de 'base' package de meest voorkomende statistische methoden en grafiekmogelijkheden herbergt. Tijdens een sessie kunnen aanvullende packages geladen worden om bijvoorbeeld de beschikking te krijgen over SPSS import/export-functies. Buiten de standaard binnen het R-project ontwikkelde packages is er een fikse bibliotheek aan externe packages voorhanden. Ook kunnen eigen packages ontwikkeld worden, mocht hieraan behoefte zijn. R onderscheidt zich (naast uiteraard het prijskaartje) van pakketten als SAS en SPSS doordat alle tussenresultaten van bewerkingen en analyses automatisch worden toegekend aan objecten.
Deze objecten kunnen vervolgens weer worden gebruikt in vervolgstappen of voor het genereren van output. Alles gebeurt op een redelijk hoog abstractieniveau. Een vector x definiëren is niets anders dan x <- c(10.4, 5.6, 3.1, 6.4, 21.7) invoeren op de command line. Daarna is 'x' als object beschikbaar en kan met plot(x) worden afgedrukt. Doordat R eveneens over standaard programmeerconstructies als looping en conditionele uitvoering beschikt, is het vrij eenvoudig om complexe analyses op te zetten die uit een groot aantal stappen bestaan.

## Gebruik R

Ook hier geldt dat de installatie zeer eenvoudig is. Wie R start opent daarmee de Rgui (zie afbeelding 3) die gebruikt kan worden om de console te starten. Wie hier een mooie drag and drop interface, een overdaad aan menu-opties of handige wizards verwacht, komt bedrogen uit. Een '>' teken waarachter de R commando's kunnen worden ingevoerd is het enige hulpmiddel. Geen pakket voor dummy's dus, maar gebruikers die een tool als R zinvol toe kunnen passen vallen niet in deze categorie. Toch is R niet moeilijk om te gebruiken. Wie ooit een introductie in programmeren en statistiek heeft gevolgd zal verrast zijn door het gemak waarmee data gemanipuleerd, geanalyseerd en gevisualiseerd kunnen worden. Wie R installeert krijgt er eveneens een flinke set documentatie bij die bijzonder duidelijk is opgezet. Ook zijn er vele (leer)boeken waarin R wordt gebruikt en voor elke functie is via help('functienaam') snel aanvullende informatie op te vragen, inclusief duidelijke voorbeelden. R wordt dan ook niet voor niets veel als leermiddel toegepast binnen de academische wereld.

## Conclusies

Het doel was om te beoordelen in hoeverre Open Source software beschikbaar én bruikbaar is voor geavanceerde data-analyse.
In dit artikel zijn daarom voor OLAP, datamining en statistische analyse drie van de bekendere oplossingen belicht waarmee in elk geval het antwoord van de beschikbaarheid is gegeven. Ook zijn de besproken producten goed bruikbaar voor het doel waarvoor ze zijn ontwikkeld, waarbij wel enkele kanttekeningen zijn te plaatsen. Hoewel Mondrian een krachtige ROLAP-tool is, bevat de meest voor de hand liggende Open Source front-end jPivot te weinig functionaliteit om het ook concurrerend te maken met de grote commerciële OLAP-pakketten. Met name de MDX drempel zal voor velen te hoog zijn. Weka en R zijn pakketten die duidelijk een andere doelgroep hebben dan de gemiddelde business user, maar wel zeer krachtige, uitbreidbare hulpmiddelen zijn. Tussen R en Weka zit wel een stuk overlap: voor zaken als clustering en classificatie kunnen bijvoorbeeld beide pakketten worden ingezet. In het geval van neurale netwerken zal eerder Weka gebruikt worden, terwijl regressie meer iets voor R is. Ze kunnen echter ook naast elkaar gebruikt worden, bijvoorbeeld om in R gebruik te maken van de krachtige grafische mogelijkheden om de output van Weka te visualiseren. Beide pakketten worden ook wereldwijd in datamining- en statistiekcursussen gebruikt, waardoor er veel laagdrempelig introductiemateriaal voorhanden is. Dit, in combinatie met het Open Source karakter en de platformonafhankelijkheid van de software, maakt het beide zeer aantrekkelijke aanvullingen op veel bestaande standaard BI-implementaties. Toch die statistiekboeken maar weer eens onder het stof vandaan halen ...

## Referenties

Mondrian: http://mondrian.pentaho.org
Weka: http://www.cs.waikato.ac.nz/ml/weka
R: http://www.r-project.org

**Jos van Dongen**
Jos van Dongen (jvdongen@tholis.com) is Senior Consultant bij Tholis Consulting.