

R is de meest populaire omgeving voor data-analyse

De gereedschappen van de data scientist

Data scientists hebben de keuze uit een groot aantal specifieke softwarehulpmiddelen.

Binnen dat enorme aanbod zijn er enkele pakketten die steeds boven blijven drijven, sommige zelfs al meer dan 30 jaar lang.

door: RICHARD KEIJZER / R.KEIJZER@AUTOMATISERINGIGDS.NL

Zoveel mensen, zoveel wensen. Dat geldt ook voor het gebruik van software voor het analyseren van data. Sommige data scientists hebben genoeg aan een spreadsheet als Excel, waarin ze hun data kunnen opbergen en er beperkte analyses op kunnen loslaten. Gebruikers met grotere datasets lopen al snel aan tegen de beperkingen van Excel, ondanks alle geavanceerde functies die Microsoft er in de loop van de tijd aan heeft toegevoegd. Zij wijken uit naar professionele pakketten of zelf ontwikkelde software.

R

De ontwikkelingen op het gebied van tools voor data-analyse gaan daarvoor snel. Toch zijn er wel een aantal pakketten die blijvertjes zijn, R bijvoorbeeld. "Voor het halen van trends uit grote hoeveelheden gegevens zijn er speciale hulpmiddelen; ik denk dat het Project R het meeste gewicht in de schaal legt", zegt Albert Bogaard van Ortec. "Na de analyse heb je een goede tool nodig om de gevonden verbanden in een mooi plaatje te zetten. Voor die visualisatie speelt een hulpmiddel als Spotfire een belangrijke rol."

"Data scientists ontwikkelen vaak hun eigen software, die ze onderbrengen in bibliotheken die ook voor vakbroeders te gebruiken zijn. Die libraries worden gemaakt op basis van bekende pakketten zoals R en MatLab. Door het instellen van parameters kunnen die pakketten worden afgestemd op een bepaald gebruiksdoel", zegt Mark van de Sanden, softwaredeskundige van SURFSara.

R is ook een van de twee softwaretools waarin studenten aan de Universiteit van Tilburg worden geschoold. In Tilburg loopt al een masteropleiding Data Scientist. Komend studiejaar wordt er, in samenwerking met Technische Universiteit Eindhoven, een bachelorleergang aan toegevoegd. "De studenten krijgen een breed palet voorgeschoteld. De belangrijkste softwaretools waarmee wordt gewerkt zijn R en Python", aldus een woordvoester van de Universiteit Tilburg.

Ook tools die in staat zijn om bestaande software om te zetten in een parallelle vorm zijn in trek. Daarmee kan oude code geschikt worden gemaakt voor moderne, parallel werkende hardware. Vooral bij gebruik van een supercomputer is dat van het grootste belang, anders kan een situatie ontstaan dat een van de processoren zeer zwaar wordt belast met code die voor één rekenkern is geschreven, terwijl de andere processoren niets staan te doen.

Samenwerking tussen apps

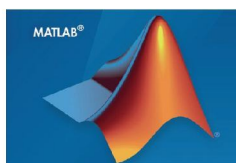
R is geen one-size-fits-all-oplossing. "Het is goed om onderscheid te maken tussen de verschillende tools. Gaat het om data-analyse dan heb je heel andere hulpmiddelen tot je beschikking dan bij het visualiseren van grote hoeveelheden gegevens", zegt Abhinav Sharma, die als data-analist heeft gewerkt bij Facebook.

De data scientists kiezen daarom vaak hun eigen mix van hulpmiddelen, software waarmee ze zelf het best werken. "Daar is op zich niets mis mee, alleen krijg je dan wel het probleem dat de ene module minder goed kan samenwerken met een ander", zegt Gregory Piatetsky van

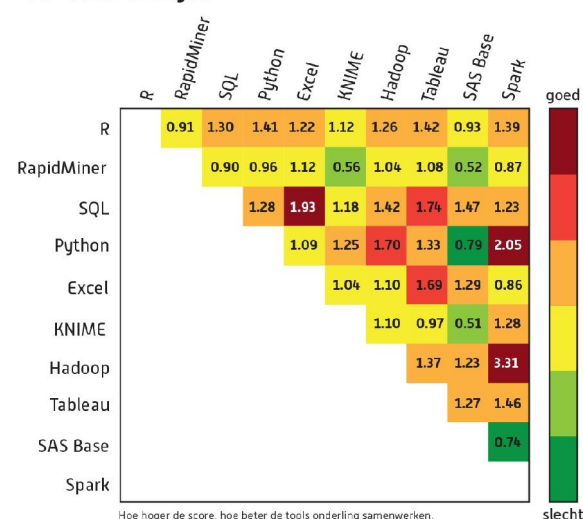
MATLAB

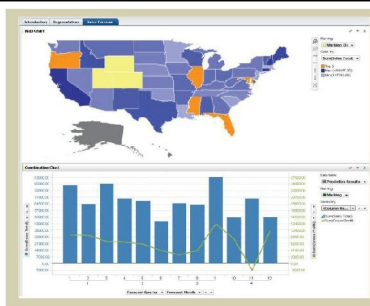
De omgeving Matlab kan het best worden gekenschetst als een numerieke verwerkingsumgeving, waarin met matrices en vectoren kan worden gewerkt. De uitkomsten van de berekeningen zijn relatief eenvoudig om te zetten naar mooie grafieken. De gebruiker hoeft geen ingewikkelde toeren uit te halen om bijvoorbeeld een 3D-grafiek te maken. Een enkel commando is voldoende om van een draadgrafiek (wireframe) naar een volumeweergave (solid model) te gaan.

Er zijn interfaces aanwezig voor andere programmeeromgevingen zoals C, C++, Fortran en Python. Matlab wordt veel gebruikt door mensen met een wetenschappelijke achtergrond. Het is opvallend dat dit pakket nog steeds zo in trek is, in aanmerking genomen dat het al in 1984 op de markt kwam. De software werkt onder de meeste populaire systemen, zoals Windows, Linux en OS X.



Samenwerking tussen softwarepakketten voor data-analyse





SPOTFIRE

De wortels van dit product liggen bij het Human-Computer Interaction Lab van de Universiteit van Maryland. De code werd enkele jaren uitsluitend gebruikt in academische kringen, tot

deze in 2007 werd gekocht door het bedrijf Tibco. Sinds die tijd wordt het product als TIBCO Spotfire op de markt gebracht. Er zijn verschillende uitvoeringen leverbaar, zowel voor de desktop als voor de cloud.

Gebruikers zijn tevreden met Spotfire, zo blijkt uit onderzoek in de VS. De deelnemers geven de software een 4,1 op een schaal van 5 punten. Vooral de geo-functies van het pakket, waarbij data gevisualiseerd worden op een landkaart, vallen erg in de smaak.

KDnuggets. Hij constateert dat gebruikers door onwetendheid over de beschikbare tools, vaak juist die pakketten kiezen die niet goed op elkaar aansluiten. Zijn bedrijf onderzoekt daarom jaarlijks de meest populaire software voor Big Data, data mining en data science. "Wij verzamelen meningen van gebruikers in onze poll en destilleren daaruit een matchwaarde. Hoe hoger de waarde, des te beter sluiten de diverse tools op elkaar aan", stelt Piatetsky [zie de figuur op pagina 10]. Volgens de gebruikers passen softwaremodules uit dezelfde bloedgroep goed op elkaar, en gaat het vaak mis bij een combinatie tussen commerciële softwarepakketten en open source. Piatetsky: "Er is bijvoorbeeld een prima match tussen Excel en SQL, terwijl het open source hulpmiddel KNIME maar erg slecht overweg kan met SAS base. Ook de combinatie tussen SAS base en RapidMiner is bar slecht."



R is verreweg de meest populaire omgeving voor de data scientist. Deze mengeling van een pakket met bijbehorende programmeertaal is ontworpen door Ross Ihaka en Robert Gentleman aan de universiteit van Auckland (Nieuw-

Zeeland). De software wordt nu onderhouden door het R core team. De ontwerpers kozen voor de letter R omdat hun beide voornamen daarmee beginnen. R draait op vrijwel alle platforms, te weten Linux, OS X, Unix en Windows.

R is een gereedschapskist met allerlei statistische tools. Daarmee kunnen reeksen data worden vergeleken, bewerkt en gecorreleerd. Als een gebruiker een bepaalde tool mist, kan hij die zelf maken met behulp van de bijbehorende programmeertaal. Omdat het programma open source is, is de kans groot dat iemand anders het gewenste tool al heeft gemaakt. Dat kan dan meestal gratis worden gedownload.

"R is een werkpaard dat nog relatief onbekend is. Ik zie mensen in mijn omgeving heel vaak gebruik maken van Excel, onder het motto dat ze daarmee prima uit de voeten kunnen. Of, als hun baas veel geld heeft, gebruiken ze een super-luxe statistisch pakket. Wat dat betreft is R een betere optie. Het kost bijna niets en de functionaliteit is stukken groter dan die van Excel", zegt Lisa Federer, onderzoeker bij de NIH Library, een onderdeel van de National Institutes of Health in de VS.

Mensen die net beginnen met R hebben baat bij contact met een lokale gebruikersvereniging. Nederland kent er drie, in Amsterdam, Twente en Nijmegen. De Twente R User Group (TRUG) bestaat uit docenten en studenten die allemaal bezig zijn met R. Een van de doelen is om de leden attent te maken op nieuwe ontwikkelingen en soms obscure functies binnen R. Door daar samen over te praten ontstaat meer begrip van de mogelijkheden.

WATSON

IBM verkoopt verschillende diensten op basis van zijn 'deep learning systeem' Watson. Watson Analytics is van deze diensten degene die het dichtst komt bij de wensen van de data scientist. Op basis van een grote berg gegevens kan de IBM supercomputer uitspraken doen over bijvoorbeeld klantgedrag.

De computer levert de resultaten 'uit de muur', zodat de gebruiker zelf niet al te veel hoeft te doen. Het invullen van een ingewikkeld spreadsheetmodel – met alle mogelijke fouten van dien – is niet meer nodig. De ruwe data kunnen worden gevoed aan Watson, die in staat is om de gegevens glad te strijken.

Watson Analytics wordt gepresenteerd in de vorm van een zogeheten freemium dienst. Dat wil zeggen dat de basisfunctionaliteit – het vergaren en ordenen van gegevens, beantwoorden van eenvoudige vragen en het tekenen van simpele grafieken – gratis is. Wie meer wil, zal daarvoor moeten betalen.



De programmeertaal Python is begin jaren 90 ontworpen en ontwikkeld door Guido van Rossum, destijds verbonden aan het Centrum voor Wiskunde en Informatica. Bij de bouw van Python stond de leesbaarheid voorop. Zo wordt structuur aangebracht door indentatie, ofwel regelinspringing, in plaats van bijvoorbeeld accolades uit talen zoals C. Statements worden simpelweg beëindigd door het eind van de regel. Variabelen krijgen geen typedeclaratie. Daardoor zijn de programma's makkelijk te schrijven en is de listing overzichtelijk.

Gebruikers van Python wordt het zo makkelijk mogelijk gemaakt. Het is zelfs mogelijk om vanuit andere pakketten een Python-script op te nemen, simpel door de gewenste handelingen uit te voeren. De leercurve van Python is daarmee laag, zeker in vergelijking met C. De taal leent zich dan ook goed voor gebruik door statistici, die niet van huis uit gewend zijn om te programmeren.

Python wordt geleverd met een uitgebreide bibliotheek om van alles en nog wat standaard te kunnen bewerken. Het is erg eenvoudig om in Python herbruikbare code te schrijven. Doordat veel applicaties die mensen schrijven gratis aan anderen ter beschikking wordt gesteld, groeien de mogelijkheden van de bibliotheek voortdurend.