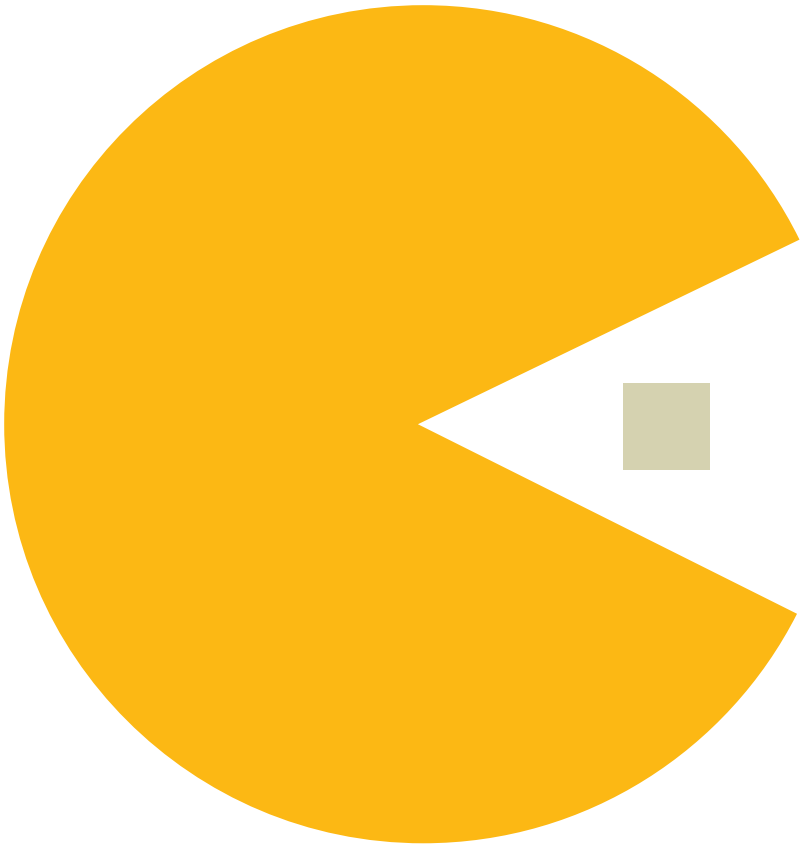

VAN BUIKGEVOEL NAAR FEITEN

BIG DATA KUN JE DAT ETEN?

Big data is hip. Uit onderzoek blijkt dat vrijwel alle enterprises big-data-initiatieven op de agenda hebben staan. Big data is happening; iedereen weet dát je er iets mee moet, maar nog niet iedereen heeft voor ogen wát dat dan is. Waardoor je soms de uitspraak 'we doen big data' hoort. Alsof het doen van data (kan dat überhaupt?) een doel an sich is. Het is alsof een restaurant zegt 'we doen voedsel'.

De term big data verwijst naar verzamelingen data, die zo groot zijn dat ze met traditionele databasesystemen niet goed te verwerken zijn. Big data richt zich op het opslaan en analyseren van veelsoortige databronnen, zowel realtime als near realtime, en in grote hoeveelheden. Het zijn de drie bekende v's: *variety, velocity* en *volume*, om daarmee de belangrijkste v te bereiken: *value*. Value verkrijg je als je de inzichten uit je data-analyse gebruikt om toegevoegde waarde aan de klant te leveren. Big data gaat derhalve om veel meer dan tooling. Het gaat vooral om de inzet van de juiste skills om de data te vertalen naar de beoogde waarde. Het blijkt dat gebalanceerde samenwerking van een data-scientist, een data-engineer, een IT-toolingexpert en een domeinexpert essentieel is. Richt je dit niet bewust in, dan eindig je wellicht met een mooie tool en fancy plaatjes, maar zonder nieuwe klantwaarde.

Teamwerk

Vanuit de kant van IT-tooling zorg je voor het geschikte big-dataplatform. Een agile platform dat veel data kan opslaan, schaalbaar is, snel aanpasbaar, kostenefficiënt en waarmee je data kunt doorzoeken en analyseren met de noodzakelijke snelheid. Vanuit de kant van data-engineering zorg je dat veel verschillende databronnen makkelijk gekoppeld kunnen worden (via api's, scripts of databasekoppelingen) en dat de juiste queries zorgen voor de juiste bewerkte datasets bij de juiste personen. De juiste semantieken moeten ontwikkeld worden om data-analyse binnen de organisatie gemeengoed

te maken. Vanuit de kant van de data-scientist gebruik je wiskunde, statistiek, algoritmes en visualisatietechnieken om data te correleren, voorspellingen te doen maar vooral ook om hypotheses te genereren die de bruikbare intelligentie ontsluiten. De domeinexpert is uiteindelijk de 'value creator'. Hij filtert de hypotheses en bepaalt of deze al dan niet vertaald worden naar nieuwe business-rules. Hij kent de business als geen ander.
Pragmatici denken al deze rollen te kunnen combineren. Het is echter gebleken dat dit nagenoeg onmogelijk is. Het versterken van deze rolverdeling is vooral vanuit het tijdperspectief succesvol gebleken. Uiteindelijk is de relatie data-scientist en domeinexpert het meest van belang; hier zit het echte concurrentievoordeel.

Restaurant

Voorheen was data-analyse als de vendingmachine naast de koffieautomaat waar je in diverse vakjes lang houdbare snacks vindt om de middagtrek te stillen. Gestructureerde input en output: vakje A1 geeft tegen 80 cent een zakje paprikachips, vakje D3 geeft voor datzelfde bedrag een Mars. En op een dag komt er een strategisch initiatief om big data te gaan doen. Met andere woorden: we starten een restaurant.
Hoe beginnen we? Daar waar we bij de vendingmachine voldoende hadden aan een service-engineer (die maandelijks de lege vakjes bijvulde en eens in het jaar vakje D3 herprogrammeerde van 80 cent voor de Mars naar één euro voor een Snelle Jelle) hebben we nu een team nodig met verschillende skills. We hebben een keukenbrigade nodig, moeten slim kunnen inkopen, gastvrij kunnen bedienen, en op tijd kunnen leveren. Essentieel is dat je iemand hebt die vanuit de ingrediënten echt lekker eten kan maken: de data-scientist.

Grasduinen

Terug naar big data. 'We doen het' omdat we denken dat er value in zit die toegevoegde waarde levert, te verzilveren is. Je wilt grasduinen in die grote hoeveelheden data om via statistische methoden, correlaties, visualisaties en machine-learning nieuwe inzichten te verkrijgen en nieuwe business-rules te creëren. Daarvoor is teamwork nodig. Zowel binnen je restaurant (kok, sommelier, bediening als een soepel team met verschillende rollen), maar meer nog in het samenspel met je gasten. Succesvolle big-datatrajecten betrekken vanaf het begin de gast bij de zoektocht naar het lekkerste recept. Want big data betekent zoeken en vinden, zoeken en weggooien, zoeken en fine-tunen. Het is een continu proces tussen de kok, sommelier en de fijnproever/gast. Ofwel, tussen de data-scientist die slim en snel kan zoeken in de grote hoeveelheden data, en de domeinexpert die de huidige producten of diensten kent en die hypotheses kan opstellen op basis van de input van de scientist.

> "DE DATA-SCIENTIST MOET VANUIT DE INGREDIËNTEN ECHT LEKKER ETEN KUNNEN MAKEN"

Hypothesen

Je zoekt naar inzichten in grote hoeveelheden data die je niet eerder met elkaar gecorreleerd hebt. Je weet dus nog niet precies naar welk inzicht je op zoek bent. Je weet niet welke vragen je moet stellen om meer value uit je data te halen. De business-rules komen niet direct vanuit de datavisualisatie en queries.
Het formuleren van hypotheses (een stelling die met 'true' of 'false' beantwoord kan worden) helpt hierbij. Hypotheses helpen out-of-the-box te denken, want je hoeft niet uit te tekenen wat voor grafiek of dashboard je wil zien, je hoeft alleen maar stellingen te poneren op basis van bestaande domeinkennis. Genereer daarom zoveel mogelijk hypotheses, of ze nu te valideren zijn of niet. Next step: iedereen laten stemmen. Is deze hypothese true of false? Hiermee creëer je een eigenaarschap omtrent business-rule-definities – iedereen wil meedenken over de hypotheses, iedereen wil de antwoorden op de hypothese weten.
Vervolgens duikt de data-scientist in de data om het antwoord op de hypothese te krijgen. Dit kan hij niet alleen, hij moet het hoofd van de domeinexpert 'leeglepelen', want alle data vereist interpretatie, domeinkennis en het begrijpen van onderliggende randvoorwaarden. En daarmee ga je van buikgevoel ('dit gerecht is lekker') naar feiten ('een combinatie van 15 procent zuur met 25 procent zout leidt in 90 procent van de smaaktesten tot een leeggegeten bord').
Deze feiten zijn belangrijk. Zij helpen bij de inschatting van de levensvatbaarheid van de data. Oftewel de 5e v: *viability*. De meeste data-scientists gaan ervan uit dat slechts 5 procent van de relevante variabelen verantwoordelijk is voor 95 procent van de waarde die je uit data kunt halen. De hypothesevalidaties zorgen dat deze 5 procent gevonden wordt. Ze helpen ook bij het vinden van (onbekende) inputs om een model te maken, en bij het bepalen welke variabelen relevant zijn voor je model. Ze laten zien welke inputs variabel of constant zijn, en welke daarvan voorspellende waarde hebben.

Rol voor de CIO

Hoe kan de CIO in deze big-data-ontdekkingstocht een rol spelen en liever nog versnellen? De vraag is natuurlijk primair of hij dit wil en kan. Indien het antwoord hierop 'ja' is, kan hij vooral sturen middels het formuleren van een big-datadienst alsof het een restaurant is. Meer dan de tooling alleen dus. Daarna is het zoeken naar geschikte onderwerpen en stakeholders binnen de organisatie om het big-datagebeuren van de grond te krijgen. Zijn van oudsher onafhankelijke rol komt daarbij goed van pas. Voorwaarde is wel dat er binnen de IT-afdeling intrinsieke interesse is voor het businessproces, anders is beginnen zinloos: dan staat er een fancy keuken in plaats van een restaurant. ✖

MARIANNE FARO is Principal Consultant binnen Itility en geeft leiding aan het competence center Analytics.