Datakwaliteit en big data tooling

In de tweede aflevering van zijn big data-serie gaat Klaas Jan Mollema in op data governance. Dit is de beheersingsstrategie waarbij datakwaliteit, datamanagement, procesmanagement en risicomanagement op elkaar afgestemd worden. Het heeft als doel ervoor te zorgen dat de gegevensinvoer aan precieze normen voldoet. Aan bod komen de rol van de informatieprofessional rond datakwaliteit en de beschikbare tools die moeten zorgen voor betrouwbaarheid in het ecosysteem van big data-omgevingen.

Klaas Jan Mollema MSc.

In de eerste aflevering van deze serie verkenden we de wereld van big data-management en werd inzichtelijk gemaakt hoe de visie op data veranderd is. De consequenties voor de informatieprofessional en de IT'er zijn duidelijk: ook al wordt veel informatiespecialistisch werk geautomatiseerd, er is nog steeds sturing nodig om te komen tot betrouwbare en duurzame opslag en analyse van data. Kritisch en conceptueel duiden van data en informatie, uitvoeren van statistische analyses, toepassen van snelle opslagmethodieken en bedenken van slimme bevragingstechnieken zijn met die grote hoeveelheid, ontstaanssnelheid en variëteit aan data belangrijker vaardigheden dan ooit.

Datakwaliteit

Aan die vaardigheden ligt een nog belangrijker informatiespecialistische activiteit ten grondslag: het bewaken van de inhoud. Datakwaliteit wordt gezien als dé trend voor 2017, aangezien steeds meer organisaties beslissingen nemen op basis van data uit business intelligence-toepassingen. Hoewel informatieprofessionals altijd al de inhoudelijke validiteit en kwaliteit van informatie hoog in het vaandel hadden, lijkt het nu ook in de datamanagement- en IT-wereld herkend te worden.

Voor het bewaken van de datakwaliteit is de IT-afdeling vaak verantwoordelijk gemaakt.[1] Hoewel dat vanzelfsprekend lijkt, bestaat er geen logisch verband tussen IT-beheer en datakwaliteit. Datakwaliteit is onderdeel van een groter begrip: data governance.[2]

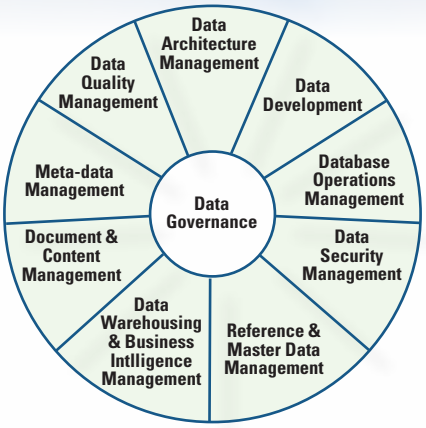
Data governance is een beheersingsstrategie die ervoor zorgt dat de invoer van gegevens door een teamlid of door geautomatiseerde processen aan precieze normen voldoet.[3] De Global Data Management Community DAMA vat de activiteiten rond data governance samen in afbeelding 1. Denk bijvoorbeeld aan het aansluiten op de business, de definitie van data-eenheden en het bewaken van de integriteit van data en het datamodel. Het bewaken van de kwaliteit is dus niet zozeer een technische activiteit, maar vooral ook een informatie-inhoudelijke. Het gaat om het vastleggen van datadefinities, het bijhouden van woordsystemen en het leggen van inhoudelijke relaties zodat de inhoud van de informatie voldoende is gestandaardiseerd.

De inhoudelijke verantwoordelijkheid over de data zou moeten worden belegd bij een functie op managementniveau, zodat medewerkers lager in de organisatie eenduidige definities hanteren van de gegevens in de datasets.

Hoe vast te stellen?

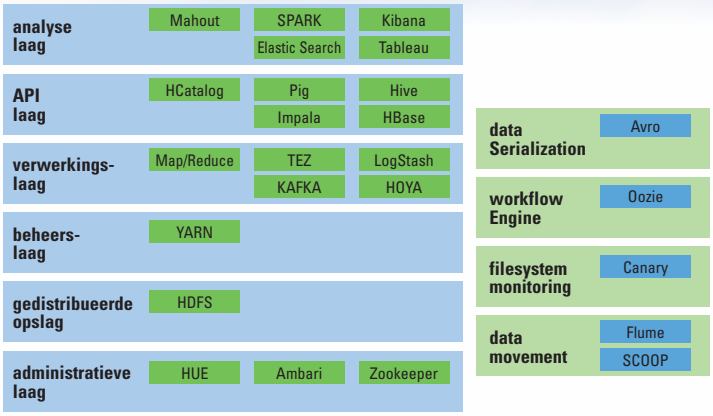
Om de kwaliteit van data vast te stellen wordt gekeken naar de volgende aspecten:

- *Processen*: Op welke manier wordt data ingevoerd en gecontroleerd? Hoe verhoudt de data zich tot de werkelijke situatie op de werkvloer waar de data wordt gecreëerd? Is de data compleet, op standaarden gebaseerd, consistent, accuraat en voorzien van een tijdstempel?
- *Organisatie*: Wie is verantwoordelijk voor dataprocessen en de inhoudelijke kwaliteit.
- *Mensen*: Zijn medewerkers op de hoogte van consequenties van slecht ingevoerde data? Zijn afdelingen op de hoogte van elkaars datagebruik? Draagt het totaal aan databestanden bij aan het behalen van het bedrijfsdoel?
- *Systemen*: Bestaat er één universele versie van de waarheid en is de bron van de waarheid opgeslagen? Controle van de volledigheid, geldigheid, consistentie, tijdigheid en nauwkeurigheid die gegevens geschikt maakt voor een specifiek gebruik.



Bron: www.dama.org/content/body-knowledge

Figuur 1: Data Governance volgens DAMA international



Figuur 2: Hadoop Ecosysteem met lagen en bijbehorende tools

Overigens lijkt het er steeds vaker op dat de 'universele waarheid', zoals we die kennen uit de klassieke informatiesystemen, niet meer bestaat. Denk bijvoorbeeld aan hotelboeking- en vliegticketsites waarbij elke gebruiker andere prijzen te zien krijgt op basis van onder andere zijn zoekgeschiedenis. Door de koppeling met interesseprofielen zal de gebruiker steeds vaker gepersonaliseerde informatie voorgeschoteld krijgen. Hierbij is het lastig om één versie van de waarheid te vast te leggen.

Data-opslag en -analyse

Een ander aspect van data governance is het voorzien in betrouwbare opslag en analyse. In aflevering 1 van de big data-serie werden al globaal de onderdelen van een gedistribueerd bestandssysteem geschetst. Apache Hadoop biedt zo'n omgeving aan, namelijk het HaDoop FileSystem (HDFS).

Elke nieuwe gebruiker die zich inleest over Hadoop, wordt overstelpt door een groot scala aan namen en termen. Het open source-karakter van de software heeft ertoe geleid dat er rond het bestandssysteem verschillende tools zijn ontstaan die goed op elkaar aansluiten.

Om helderheid te verschaffen deel ik de softwarematige tools binnen het Hadoop Ecosysteem op in verschillende lagen, zoals te zien is in figuur 2. Deze lagen (grijs met groen in figuur 2) hebben elk een eigen taak in het opslaan en verwerken van de data. In de *administratieve laag*, een soort control panel, vindt het beheer van de servers plaats. De *gedistribueerde opslag* van gegevens zorgt voor de verspreiding van de bestanden over de verschillende servers en de snelle toegang tot de opgeslagen data. De *beheerslaag* voor de servercapaciteit houdt bij hoeveel geheugen, diskruimte, processorkracht is gebruikt en verdeelt taken over servers die nog niet veel te doen hebben. De *verwerkingslaag* krijgt de gegevens binnen en zet ze klaar voor analyse (zie aflevering 1: 'Map/Reduce'). De *Advanced Programming Interface (API) laag* biedt toegang tot de data en zorgt ervoor dat de data bevraagd kan worden. En tot slot is er de *analyselaag*, waar gegevens in relatie tot elkaar opgeslagen kunnen worden om vervolgens bijvoorbeeld in een grafiek te worden gezet. In sommige programma's is het zelfs mogelijk om historische data te analyseren en met behulp van een algoritme de toekomst te voorspellen.

Andere tools

Elk van deze lagen heeft één of verschillende tools (groen/blauw in de figuur 2) die bijdragen aan een deel van het dataproces. Vaak zijn de tools gespecialiseerd in één deel van het proces, zoals 'real time processing' of 'machine learning'.

Naast deze lagen zijn er nog tools die het bestandssysteem monitoren, tools die de werkdruk bijhouden, tools die zorgen voor de inhoudelijke transformatie van data en tools die bijdragen aan het import-/exportproces vanuit een tool naar je big data-omgeving (geel met blauw in figuur 2).[4]

Naast het Hadoop Ecosysteem zijn er ook data-clouddiensten die dergelijke functionaliteiten aanbieden. Aflevering 3 van deze serie zal ingaan op de verschillende aanbieders en de voor- en nadelen van een Hadoop-installatie versus clouddiensten. Zowel de lokale Hadoop-installatie als de clouddiensten bieden betrouwbare dataopslag aan met veel back-upmogelijkheden. Tevens worden in beide omgevingen analysemogelijkheden aangeboden om met de data te werken.

Vacatures in big data

Hoewel voorgaande alinea doet vermoeden dat het big data-speelveld alleen is weggelegd voor de techneuten onder ons, geeft een zoektocht naar vacatures een ander beeld. Natuurlijk zijn er technische vacatures rond het modelleren van data en het beheer van big data-opslag. Maar de datascientist bijvoorbeeld is in opkomst en diens functie lijkt veel op die van een informatieprofessional. Als datascientist vind je de verborgen patronen in de data. Door middel van statistische analyse en sterke (SQL) queries kom je tot antwoorden op business-vragen die je vervolgens visueel inzichtelijk maakt. En datavisualisatie maken is soms al een functie op zich. De verwachting is dat rond het beheer van datakwaliteit zowel op strategisch, tactisch als operationeel niveau banen zullen ontstaan. Elk bedrijf wenst immers veilige, toegankelijke en betrouwbare data om haar beslissingen op te nemen. <

Noten

1] Zie: tinyurl.com/zs5q8gz.

2] Zie ook https://www.dama.org/content/body-knowledge.

3] Bron definitie: Wikipedia.

4] De uitgebreide variant van afbeelding 2 vind je op www.zijlmo.nl/mo/shared/hadoopecosysteem.pdf.

Klaas Jan Mollema MSc. (www.zijlmo.nl/mo) is specialisatiecoördinator Business Data Management aan de opleiding Informatica van Hogeschool Leiden.