Nooit meer opruimen

Martin Heijnsbroek
zaterdag 5 maart 2016

Ik begrijp mijn zoon van 13 wel als hij zijn kamer niet wil opruimen. Opruimen is vervelend, en natuurlijk gooi je dan net die waardevolle aantekening of kabel weg die je over een maand weer nodig hebt. Maar ja, het moet gewoon, je kan niet alles bewaren.

Toen Google begin 2000 de leidende zoekmachine op het internet werd, keek het hier anders tegenaan. Google's business draait om data, en die weggooien was geen optie. In plaats van data op te ruimen en weg te gooien, ontstond de mantra om álle data van bezoekers voor altíjd te bewaren. In ruwe vorm, zonder de data netjes te ordenen in een gestructureerde database.

Het aanschaffen van een traditioneel 'data warehouse' was niet handig. Want die werkt met gestructureerde data en heeft een beperkte capaciteit. En met de groei van Google zou elk data warehouse binnen mum van tijd vollopen. En dan moet er een nog groter en duurder data warehouse komen.

Dus ging Google voor een meer schaalbare oplossing en sloeg het nieuwe, ongestructureerde data gewoon op traditionele harde schijven op. Maar daar zaten wel wat nadelen aan. Doordat er op een gegeven moment miljoenen schijven nodig waren, ging op een gegeven moment ongeveer om het uur een schijf kapot. Daarnaast was het analyseren van de data over zoveel verschillende schijven erg complex en tijdrovend.

Google besefte dat een duurzame oplossing nodig was voor dit big-dataprobleem en kwam tussen 2003 en 2004 twee doorbraken voor het probleem. De eerste doorbraak was het Google File System (GFS), dat data in blokjes van 64MB opbreekt en op verschillende schijven bewaart. De tweede doorbraak was MapReduce, waarmee databewerkingen over verschillende servers parallel konden plaatsvinden en daardoor veel sneller uitgevoerd konden worden.

In dezelfde periode werkte de freelancer Doug Cutting aan een opensourceproject (Nutch) om een zoekmachine te ontwikkelen. Hij worstelde in dit project, net als Google, met de opslag en bewerking van enorme hoeveelheden ongestructureerde data. Op basis van Google's publicaties over GFS en MapReduce, kwam hij tot een opensource-oplossing.

In 2006 zag Yahoo de potentie daarvan en nam Cutting in dienst. Cuttings software werd Hadoop genoemd en bleef ondanks de grote investeringen van Yahoo open source. Waardoor het voor andere giganten als Facebook, Twitter en Linkedin, Ebay en later ook Apple interessant werd om ook met Hadoop aan de slag te gaan.

Dankzij deze spelers ontstond de afgelopen jaren een golf van innovatie rondom Hadoop. Daardoor is het ook voor 'gewone' bedrijven weggelegd om alle webdata te

bewaren en razendsnel te analyseren. En wordt steeds breder het mantra omarmd dat alle data bewaard wordt, in plaats van opgeruimd of weggegooid.