Enterprise

Een heldere strategie en veel creativiteit liggen ten grondslag aan de software van goed draaiende ondernemingen. Vooral voor degene, die het moeten hebben van een sterke aanwezigheid op het internet. Bij bol.com is dat niet anders. De IT-afdeling van zo'n honderd ontwikkelaars heeft een dagtaak aan het up-to-date houden en innoveren van de techniek.

TechRally bij bol.com levert veel nuttigs op

Creativiteit zonder productiespanning

Om de creativiteit een extra impuls te geven heeft bol.com onlangs samen met technologiepartner Xebia een 'TechRally' gehouden, waarbij de ontwikkelaars volledig de vrije hand kregen om nieuwe ideeën te ontwikkelen. Even uit de praktijk van alledag en zonder de spanning van de productie.

"De TechRally komt erop neer dat je een stel ontwikkelaars in een hok zet en ze volledig de vrije hand geeft. Het is voor de 'fun' van de ontwikkelaars, om ze even helemaal los te laten gaan. Aan het einde van de dag presenteren de teams – bij bol.com waren dat zes teams van zes à zeven personen – wat zij hebben ontwikkeld aan het hele bedrijf. Het is ook een toetsing van waar je afdeling toe in staat is. Tijdens de Rally blijkt dat je meer kunt dan je denkt. Bol.com is niet de enige onderneming, die de ontwikkelaars deze uitdaging biedt; ook 'grote jongens' als Amazon en Google bieden hun ontwikkelaars tijd en ruimte om in een ontspannen sfeer nieuwe ideeën te ontwikkelen en kansen te genereren", vertelt Martin Weidner, teamleider development Java bij bol.com.

"Samen met Xebia hebben we een aantal thema's opgesteld, waarbinnen de ontwikkelaars hun eigen gang konden gaan. Aanvankelijk waren we een beetje bang dat sommige veel te ingewikkeld zouden zijn om daar binnen een dag iets mee te kunnen doen, maar achteraf viel dat mee en hebben we hele leuke resultaten gezien. We hadden vijf thema's: een open API of interface, waar andere ontwikkelaars buiten bol.com tegenaan kunnen ontwikkelen; visualisatie van de statistische data uit de logfiles; aanbevelingen op basis van de logfiles; het genereren van regressietests op basis van de logfiles en; geautomatiseerde klantgegevens."

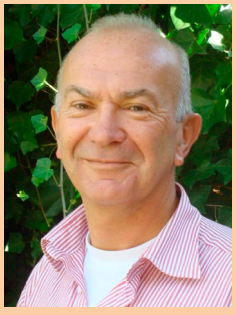
IT-architect Niels Basjes van bol.com vult aan: "We hebben deze doelen bij de mensen neergelegd en vervolgens aan de mensen overgelaten wat ze daarbinnen leuk zouden vinden om te bouwen".

Open API

"Dat levert hele verrassende resultaten op", aldus Martin enthousiast. Neem de open API, waarmee derden tegen onze shop aan kunnen programmeren en onze catalogus kunnen benaderen. We hebben ook gekeken hoe het werkt als we daar zelf tegenaan programmeren. Wat zijn de valkuilen? Een aantal mannen heeft bijvoorbeeld een shop voor schoolboeken gebouwd. Daar kunnen we wellicht later iets mee. Zo zouden scholen kunnen kijken of ze bepaalde boeken met een korting kunnen kopen. Een ander thema dat erg goed is gevallen, is de visualisatie van de statistische data uit de logfiles. Dit leverde tijdens de demo een kaart van Nederland op, waar je realtime kon zien wanneer en waar in Nederland welke producten worden besteld."

Niels: "Een andere opdracht was: bouw eens aanbevelingen voor de klanten op basis van de logdata. We hebben een cluster neergezet met Hadoop, een kleine drie weken aan productie logfiles en de teams succes gewenst. En het werd ook een succes. Die logfiles werden in een klein halfuurtje door het cluster verwerkt. Het was indrukwekkend dat ze op een klantnummer konden inprikken en daar reële aanbevelingen uit konden destilleren. In feite hebben ze in een dag de basisimplementatie neergelegd. Ik vind dat heel bijzonder."

Martin: "Er zijn bestaande grote, dure software-



Robert de Ruiter
is hoofdredacteur van Java Magazine.

De resultaten worden in de 'bioscoop' aan de medewerkers getoond.

Je proeft tijdens de TechRally dat er een productieve sfeer heerst bij alle teams.

pakketten. Zo'n TechRally is een mooie gelegenheid om te kijken of dat niet anders kan. Een ander thema hebben we gewijd aan testen. Met een aantal ervaren testers van Xebia is bekeken of we op basis van de logfiles regressietests konden genereren. In plaats van steeds nieuwe tests te moeten schrijven. Aan de logs kun je immers zien waar mensen naartoe klikken, wat de goede paden zijn, waar ze fouten krijgen. Met Xebium, een Xebia testtool die is gebaseerd op Selenium, hebben we die logfiles nagespeeld en daar heel nuttige dingen mee kunnen doen."

Niels: "Met de data in het cluster hebben we tests en experimenten gedaan om te zien welke geautomatiseerde klantsegmentatie we daaruit konden afleiden. Dat heeft mooie plaatjes opgeleverd, maar die materie bleek iets te complex om in een dag tijd echt visuele resultaten op te leveren."

Bioscoop

Bol.com heeft in haar pand in Utrecht een grote zaal, die 'de bioscoop' wordt genoemd. Daarin vond de TechRally plaats. "Dat ziet er dan heel productief uit. De mensen zitten lachend te werken aan hun Macs en laptops. Je proeft dat er een hele productieve sfeer ontstaat. Niet alleen dankzij de ontwikkelaars, maar het hele bedrijf kwam een kijkje nemen bij de ontwikkelaars. Van alle andere afdelingen kwamen regelmatig mensen binnenlopen. De ontwikkelaars zelf kwamen van bol.com en Xebia. In de dagelijkse routine heeft Xebia bij bol.com voornamelijk een ondersteunende functie, die is ontstaan toen bij bol.com Scrum werd geïntroduceerd. De dagelijkse werkzaamheden bij bol.com worden door eigen IT-ers verricht, maar voor de TechRally waren de teams aangevuld met 'Xebianen'.

Een overzicht van het cluster rapport in Hadoop.

Rob Dielemans, manager software development bij Xebia, legt uit: "Van Xebia doet in principe iedereen mee aan de TechRally en van de klant mag meedoen wie dat leuk vindt. Al eerder hebben we dit zo georganiseerd bij MTV en daar is dat toen ook in goede aarde gevallen. De extra dimensie voor ons is dat we dan ook met de klantdata kunnen werken. Bij bol.com is dat drie gigabyte per dag en dat is dan ook nog compressed. Dat is natuurlijk een hele andere situatie dan wanneer je met de ontwikkelaars vanuit het niets nieuwe software ontwikkelt. Voor onze ontwikkelaars geeft de TechRally ook een enorme stimulans en ze leren de klant beter kennen. Dat kun je ook doen tijdens een dagje zeevissen, maar de TechRally vinden we zinvoller en zeker niet minder leuk."

De TechRally vindt heel duidelijk plaats zonder enige productiedruk, maar ze levert zeker resultaten op. Zo heeft het gebruik van Hadoop bij bol.com hele nieuwe inzichten opgeleverd. Het cluster bestond uit twaalf door de IT-afdeling afgedankte desktops. Pentium 4 en dat soort spul. Met dit cluster kon je in een half uurtje bijna drie weken aan data verwerken. "Dat was wel een eye-opener. Dit is een tak van technologie die ook in de productiesfeer heel veel mogelijkheden gaat bieden", zegt Niels. En ook de experimenten met de Open API hebben verschillende user stories opgeleverd, die in een backlog zijn opgenomen en waar in de nabije toekomst nog iets mee zal worden gedaan.

De extra dimensie is dat er met de klantdata kan worden gewerkt.

Opbrengst

De TechRally levert niet direct iets op, maar heeft toch grote voordelen. Martin: "Het leuke van een e-commercebedrijf, waar de IT het kloppende hart is, is dat je moet blijven innoveren. Zo'n TechRally is een uitgelezen kans om op het technische vlak te innoveren. Er komen dingen uit, waar je niet veel mee kunt. Dat moet je dan ook accepteren. Maar het levert altijd nieuwe inzichten op."

Niels: "We hebben een website met ontzettend veel traffic. Hoe beter je naar de ruwe data kijkt en hoe dieper je daarop inzoomt, hoe meer je ervan kunt leren en hoe beter je de klant ermee kunt ondersteunen. We hebben een zelflerend systeem, dat kijkt naar het gedrag van de klanten en helpt de klant tijdens het zoeken door lijstjes met zoektermsuggesties te produceren. Daarvoor ploeg je maanden aan logdata door om daar de patronen in te herkennen. Dat is een van de redenen, waarom we nu Hadoop in productie hebben: een toolkit in Java om een rekenklus in dusdanig veel stukjes te hakken dat je de klus kunt verspreiden over een verzameling computers. Je kunt op deze manier twee- tot drieduizend computers parallel aan het werk zetten. Qua kosten om op te schalen is dat heel voordelig, want je kunt er standaard kasten voor gebruiken. Bij bol.com lopen we in Nederland met de Hadoop-ontwikkelingen voorop." (Friso van Vollenhoven gaat hier vanaf pagina 25 verder op in, red).

Een TechRally staat uiteraard niet helemaal los van het bedrijfsbeleid. Het heeft weinig zin om een nieuw computerspelletje te ontwikkelen, als je daar binnen de onderneming niets mee kunt. Bol.com volgt een strategie, die is opgebouwd rond een aantal pilaren (bijvoorbeeld 'mobility'). Binnen deze pilaren is echter alle ruimte om technologische vernieuwingen aan te brengen. In de praktijk wordt dit geregeld in Scrumteams, die regelmatig met de business overleggen. Dat is wel een groot verschil met de TechRally, want daar is de business afwezig. "Zie het maar als een brainstormsessie binnen de onderneming. Daar heb je natuurlijk ook altijd een zekere omkadering nodig. Die rol wordt tijdens de TechRally verzorgd door de thema's."

Java

Bol.com heeft ook een vrij grote Oracle-club (de WebLogic Application Server is het platform voor de back-end-infrastructuur), maar deze TechRally was puur een Java-actie. "Nou", zegt Niels met een grijns, "volgens mij heb ik ook nog wel wat stukken Ruby-code zien passeren. Maar bij bol.com werken vooral Java-jongens. Java is voor bol.com ook het grootste platform." Xebia vindt dat open source in zijn algemeenheid grote voordelen heeft. Mede daarom is voor Hadoop is gekozen. "En de kwaliteit is er niet minder om. Tijdens de voorbereidingen voor de TechRally bleek opeens de performance iets terug te lopen. Toen we naar de oorzaak gingen zoeken bleek dat van één van machines in het cluster de voeding was doorgebrand. Het proces had dus iets langer geduurd, maar werd ondanks die dode machine wel met succes afgerond. Dat is toch geweldig: dat je met dat oude ijzer nog zo veel nuttige dingen kunt doen". «

Wie kent het niet?

Bol.com heeft in de twaalf jaar van haar bestaan in Nederland een enorme vlucht doorgemaakt. De bedoeling was om de activiteit in heel Europa uit te rollen, maar het grote succes is typerend voor ons land. Het is een van de weinige dotcom-bedrijfjes die echt van de grond zijn gekomen nadat investeerders waren gevonden en zeer succesvolle reclamecampagnes zijn gevoerd. Er zijn ook nauwelijks nog Nederlanders, die niet van het bedrijf hebben gehoord. Bij een recente steekproef bleek dat de naamsbekendheid van bol.com in Nederland 100% is. "Dat is voor ons ook een enorme stimulans. Als iemand hoort dat je bij Bol.com werkt, krijg je bijna altijd enthousiaste reacties. Dat is natuurlijk hartstikke leuk", zegt Martin.

De internetwinkel heeft ruim 3 miljoen actieve klanten in Nederland en België die in 2010 goed waren voor bijna 16 miljoen verkochte artikelen. Bol.com is daarmee marktleider op het gebied van online verkoop van boeken, entertainment, elektrische apparaten en speelgoed tevens de grootste internetwinkel van Nederland en België. Bezoekers van het warenhuis hebben met een muisklik toegang tot meer dan 5 miljoen artikelen, waaronder nieuwe en tweedehands Nederlandstalige en buitenlandse boeken, muziek, dvd's en games, maar ook notebooks, software, pc-accessoires, elektronica, mobiele telefonie, lcd- en plasma televisies, een breed assortiment elektronische huishoudelijke artikelen, speelgoedartikelen, e-readers met tienduizenden bijbehorende e-books en een fotoalbumservice. In februari 2011 werd bol.com plaza gelanceerd. Hiermee is bol.com de eerste winkel die haar klanten de mogelijkheid biedt om ook artikelen van andere aanbieders via de winkel te bestellen.

NoSQL bij bol.com

In mijn vorige artikel in het Java Magazine hebben we het gehad over verschillende mogelijk redenen om de relationele database wereld te verlaten en een NoSQL alternatief in te zetten. Dat was een ietwat theoretische verhandeling. Deze keer gaan we kijken naar een concreet voorbeeld van een NoSQL oplossing. Gelukkig heb ik van mijn opdrachtgever bol.com toestemming gekregen om het een en ander te vertellen over de technologie die daar wordt gebruikt.

Achtergrond

Vrijwel alle websites houden webserver logs bij. Dit zijn de logfiles waar alle webrequests van alle gebruikers in staan. Bij websites met een significant bezoekersvolume zijn de webserverlogs een rijke bron aan informatie. Hieruit kunnen gebruikersprofielen worden opgemaakt, statistische waarheden over bezoekpatronen worden gehaald, peformance analyses worden gedaan en noem maar op.

Ook bol.com is voor een aantal features afhankelijk van het analyseren van recente en historische webserver logfiles.

Logfile analyse

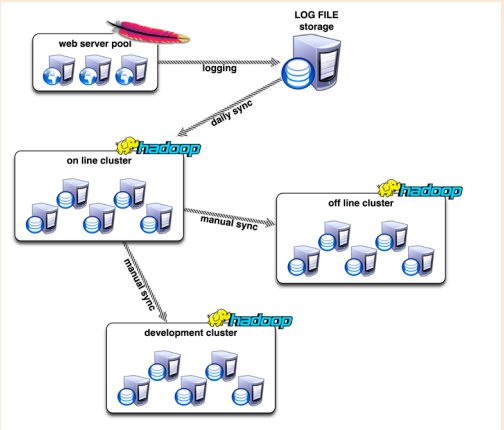
Er wordt jaarlijks ongeveer 1,5 TB aan webserver logs gegenereerd, iets meer dan 4 GB per dag. Op deze data worden verschillende analyses gedaan. Niet alle analyses gebruiken de volledige historische data die beschikbaar is, sommige slechts de meest recente drie maanden, oftewel 360 GB. Sommige analyses worden dagelijks gedaan en dienen als input voor andere systemen en sommige zijn ad hoc analyses die eenmalig worden gedraaid. Voor de verschillende analyses wordt altijd een gedeelte van de logs in een batch proces verwerkt.

Er schuilt niet per se een zware technische uitdaging in het inlezen en verwerken van textfiles. Probleem is echter dat het vanaf een bepaald datavolume lastiger wordt gemaakt, omdat het niet op een economische manier mogelijk is om de volledige data (bijvoorbeeld 360 GB log files) in het geheugen te laden. Door deze beperking is het niet mogelijk om op een makkelijke manier een datastructuur bij te houden voor het groeperen of sorteren van logevents, iets dat veel gebeurt in de analyses.

Een voor de hand liggende oplossing voor bovenstaand probleem is een database. Het is echter zo dat de meeste bewerkingen uitgaan van een full data analysis, oftewel full table scans in op de database. Dat is niet iets waar een relationele database in uitblinkt. Ook het opslaan en bijhouden van meerdere jaren aan logevents kan een probleem opleveren. Een database van meer dan 1 TB mag fors worden genoemd en we hebben het hier over 1,5 TB per jaar.

Oplossing

Als alternatief voor batchverwerkingen op basis van een relationele database heeft bol.com gekozen voor het inzetten van Apache Hadoop, een oplossing voor batchverwerking van grote hoeveelheden data op een cluster van machines. Hadoop combineert opslag en batchprocessing op een cluster van machines en levert op die manier een platform dat zich uitstekend leent voor het verwerken van records.



Bovenstaand is een high level diagram van de oplossing. De webserver logs worden dagelijks naar een Hadoop cluster gesynchroniseerd. Op deze omgeving worden alle dagelijkse, geautomatiseerde batchjobs gedraaid. Dit zijn de jobs waarvan het resultaat nodig is voor user facing functionaliteit. Het online cluster kan ook wel worden gezien als productieomgeving. Het tweede cluster, hier genoemd offline cluster, bevat ook een kopie van de logdata. Deze wordt af en toe, wanneer nodig in sync gebracht met de meest recente versie op het productiecluster. Op deze tweede omgeving worden nieuwe versies van de productiejobs getest, maar ook kunnen hier ad hoc analyses worden gedraaid en rapportages gemaakt voor interne doeleinden. Uiteindelijk is er een developmentomgeving die niet kritiek is en kan worden gebruikt voor allerhande test- en ontwikkeltaken.

Hadoop bestaat in de basis uit een storage component, het Hadoop Distributed Filesystem (HDFS),

Hadoop is een platform dat zich goed leent voor het verwerken van records.

Friso van Vollenhoven
is NoSQL specialist bij Xebia.

Een voordeel van Hadoop is dat er altijd wordt uitgegaan van full data analysis.

en een processing component, Hadoop MapReduce. Het framework distribueert zowel storage als processing over meerdere machines op een manier die lineair schaalbaar is. Dat laatste is een belangrijke eigenschap, want het betekent dat de capaciteit van een cluster afhankelijk is van het aantal machines (en niet van hoe zwaar de machines zijn uitgerust). Hierdoor is het mogelijk om de capaciteit uit te breiden al naar gelang de vraag groeit door simpelweg hardware toe te voegen aan de omgeving.

Het is dus niet zo dat er bij aanvang meteen geïnvesteerd moet worden in een hoeveelheid hardware die klaar is voor de toekomstige behoefte aan capaciteit. In het geval van databaseservers is dit vaak wel het geval. Een bijkomend voordeel van de Hadoop oplossing is dat er altijd wordt uitgegaan van een full data analysis, dus alle data wordt bij iedere verwerking bekeken. Om deze reden is het niet nodig om van tevoren te bedenken welke indexen er nodig zijn of op welke manier data moet genormaliseerd of gedenormaliseerd. Er is geen databaseschema.

Deze NoSQL-oplossing kan worden gebruikt voor rapportages, ad hoc analyses en het bouwen van (zoek)indexes voor user facing elementen.

Conclusie

Door de inzet van Apache Hadoop slaagt bol.com erin om op een economische manier om te gaan met het datavolume dat dagelijks wordt gegenereerd. De oplossing biedt ook veel kansen die er eerder niet waren. Een full data analysis op een paar honderd GB aan data kan in een half uur en is heel flexibel, omdat er geen schema is of vastgelegde representatie van de data.

Ook prettig is het feit dat de oplossing volledig rust op open source software. Het uitschalen van storage en processing hebben daardoor een transparant kostenplaatje: enkel de hardware en stroomkosten. «

Tijdens de TechRally werd ook regelmatig groepenoverleg gevoerd.