Grote gegevens en denderende data: hoe slaan websites hun data op?

Florian Dejonckheere

Hogeschool Gent
florian@dejonckhee.re

30 september 2018

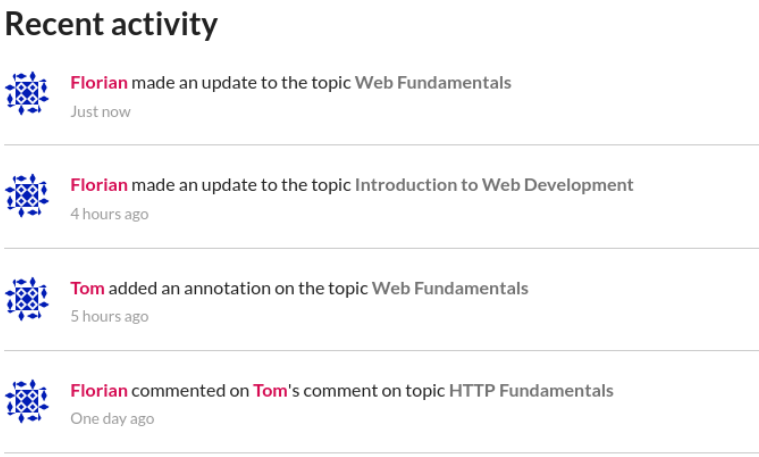
De laatste jaren zijn grootschalige sociale mediaplatformen zoals Twitten en Facebook sterk opgekomen. Door het toenemende gebruik ervan is de digitale voetafdruk van de gemiddelde internetgebruiker ook massief gestegen. Deze websites hebben dan ook te maken met een zondvloed van gegevens die opgeslagen moeten worden, klaar om ze snel naar de gebruiker te sturen. Maar hoe worden die gegevens precies opgeslagen? We bekijken het voorbeeld van Open Webslides.

Open Webslides

Open Webslides is een online collaboratieplatform voor leerkrachten en leerlingen. Hierbij wordt er door beide partijen aan dezelfde cursus gesleuteld, zodat studenten deze beter kunnen leren en leerkrachten gemakkelijk feedback kunnen verwerken.

Het deel van Open Webslides waar we geïnteresseerd in zijn, is de feed van de *recente activiteit*. Deze feed ziet er een beetje uit zoals de feed op Facebook of Twitter. De meest recentste berichten staan bovenaan, de minder recentere berichten gaan eronder. Natuurlijk gaat het hier niet over foto's en evenementen, maar wel over wie iets veranderd heeft op de cursus, en wat precies. Zo kan je in één oogopslag zien wanneer de leerkracht een hoofdstuk heeft toegevoegd, of dat een van je medestudenten een vraag heeft gesteld bij een onduidelijke zin.

Maar hoe slaan we deze gegevens precies op? Bij elk item hoort een cursus, een auteur en de actie die uitgevoerd werd. En natuurlijk ook het tijdstip van de actie, waarop we ook moeten sorteren vooraleer de lijst naar de gebruiker te sturen.



Figuur 1: *Recent activiteits-feed*

Databanken

Er kan gekozen worden uit de vele, al bestaande databanktoepassingen. Maar eerst moeten we kiezen welk type databank het best past: een *relationele*, een *document*- of een *graafdatabank*. Bij een relationele databank worden de gegevens opgeslaan in tabellen: elke tabel heeft kolommen, en elk item heeft een waarde voor deze kolommen. Dit is de traditionele manier van gegevens opslaan. Het laat ons toe om gemakkelijk te zoeken in alle gegevens, en

deze te sorteren. Dat komt goed van pas als we alle items willen sorteren op tijdstip. Het combineren van gegevens in verschillende tabellen – bijvoorbeeld een tabel voor auteur, en een andere tabel voor cursussen – gaat zeer snel in een relationele databank.

Een document-databank daarentegen, slaat alle gegevens op als een document: een ondoorzichtig stuk data waar er niet in gezocht kan worden. Om dit te omzeilen, steken we alle gegevens die nodig zijn in eenzelfde document. In ons geval wil dit dus zeggen dat de auteur, cursus en acties in hetzelfde document zitten. Document-databanken zijn typisch zeer goed in het afhandelen van massieve datastromen. Dat kan in ons voordeel spelen als we een systeem overwegen met duizenden gebruikers.

Ten laatste kunnen we ook kiezen voor een graafdatabank. Een graaf is een verzameling van punten of objecten, en die verbonden zijn door lijnen. Het doorkruisen van lijnen van object tot object is een zeer snelle actie. Dat komt goed uit voor de gelinkte datastructuur van Open Webslides.

Aangezien relationele databanken niet de focus zijn van het onderzoek, moet er uiteindelijk gekozen worden tussen een document- en een graafdatabank. Hiervoor gaan we een schema opstellen, en dat in elke databank invoegen. We doornemen ook een aantal realistische scenarios om deze databank te gebruiken. Aan de hand van deze twee elementen kunnen we testen welke databank het beste ophoudt aan de (gesimuleerde) stress van duizenden gebruikers. Het verdict: de graafdatabank is 10 tot in bepaalde gevallen zelfs 40 keer trager dan de document-databank. Een duidelijke overwinning dus!

Conlusie

Het is dus duidelijk voor het Open Webslides project: document-databanken bieden dus de meest efficiënte oplossing.

We hebben nu een kijkje genomen achter de schermen in de wereld van data-opslag op grote schaal. We zagen hoe bedrijven zoals Facebook en Twitter de massieve vloed aan data opslaan. Het verschil tussen de databanktoepassingen is nu ook duidelijk: relationele, document- en graafdatabanken.

Mensen staan zelden stil bij de technologie die achter social media websites zitten, en welke hordes de ontwikkelaar heeft moeten nemen om deze technologieën naar een groot publiek te brengen. Denk daar maar eens over na, bij de volgende gedachteloze scroll van je Facebook feed!