Big Data in the cloud: je bedrijf in de wolken?

Data is niet meer weg te denken uit de hedendaagse maatschappij. Denk maar aan de vele mails die bedrijven hebben verzonden in verband met de invoering van de GDPR; hoeveel privacy-overeenkomsten heb je niet opnieuw moeten aanvaarden? Dit geeft een indicatie dat bedrijven data-analyses uitvoeren. Doet een bedrijf dit niet, dan mist het kostbare informatie over zijn klanten. Data is een belangrijke bron van informatie: The world's most valuable resource is no longer oil, but data. Per minuut worden gigantische hoeveelheden data geproduceerd. Waar moeten deze data naartoe? Wie gaat deze data behandelen zodat enkel zinvolle informatie opgeslagen wordt? Maar het belangrijkste: hoe en waar gaan we de data verwerken? En dat vormt de focus van deze scriptie.

Traditionele systemen voldoen niet meer of zijn te duur om de gigantische hoeveelheden data te verwerken of op te slaan. Bedrijven moeten vaak elders op zoek naar een manier om hun data op te slaan en er vervolgens analyses mee uit te voeren. Momenteel bestaan er heel veel verschillende mogelijkheden om big data op te slaan en te verwerken, denk hierbij aan Hadoop, Spark, Flink, Samza, Dremel... Veel van die systemen zijn heel groot en vragen veel tijd voor installatie en configuratie. Enerzijds is het daardoor voor KMO's niet opportuun om hiermee data-analyses uit te voeren, maar anderzijds missen ze wel kostbare informatie. Vanuit dat idee vertrok IntoData met de vraag om onderzoek te doen naar enkele oplossingen voor dit probleem. De oplossing van dit probleem bevindt zich in cloudplatformen.

Cloudplatformen

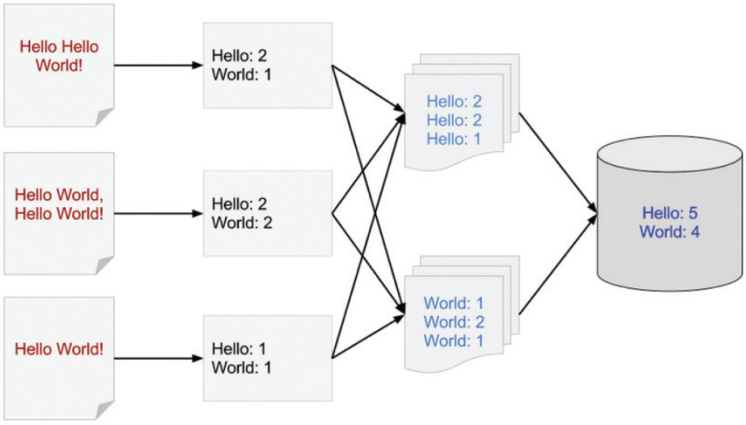
Een cloudplatform is zoals een programma op je eigen pc, alleen draait dit programma in de cloud. Wat is daar nu zo handig aan? Wel, het kost veel minder tijd en geld om servers op te starten en te configureren in de cloud dan zelf voor alles te zorgen. Bijvoorbeeld: een Apache Spark server opzetten in de Google Cloud duurt nog geen vijf minuten, je eigen server opzetten kost al snel een of twee dagen. Momenteel zien we drie grote spelers in de cloud: Google, Amazon en Azure (Microsoft). Elk van deze spelers biedt zijn eigen cloudplatformen aan voor heel veel verschillende toepassingen. IntoData heeft ervoor gekozen om Google BigQuery, Google Cloud Dataproc en Amazon EMR verder uit te diepen, aangezien deze drie systemen veelbelovend leken voor een (relatief) lage prijs. Dit werd gedaan aan de hand van een vergelijkende studie, waarbij elk van de tools werd afgetoetst aan enkele requirements en negen queries werden uitgevoerd op elk van de tools. Een query is een zoekopdracht in een databank die eventueel enkele gegevens teruggeeft.

Apache Spark

Google Cloud Dataproc en Amazon EMR zijn twee cloudplatformen gebaseerd op Apache Spark, een framework dat toelaat om gigantische hoeveelheden data te verwerken op meer dan één server. Elke server doet een stukje van de verwerking, waarna de resultaten van

elke server samengebracht worden tot het gewenste resultaat. Dit is beter gekend als het MapReduce-principe. Het grote voordeel van MapReduce is dat de data niet mooi opgeruimd hoeft te zijn voor de verwerking start. De data mag onzuiverheden bevatten zoals niet ingevulde waarden of verkeerd opgemaakte waarden.

Figuur 1 legt eenvoudig het principe van Apache Spark uit. Links zie je een aantal bestanden die de woorden 'Hello' en 'World' bevatten. In de eerste stap worden deze woorden gesplitst en wordt er geteld hoeveel keer een woord voorkomt. In de tweede stap worden deze aantallen samengevoegd tot één resultaat per woord. Dat resultaat wordt uiteindelijk in stap drie opgeslagen.



Figuur 1: MapReduce-principe (Sato, 2012)

Dremel

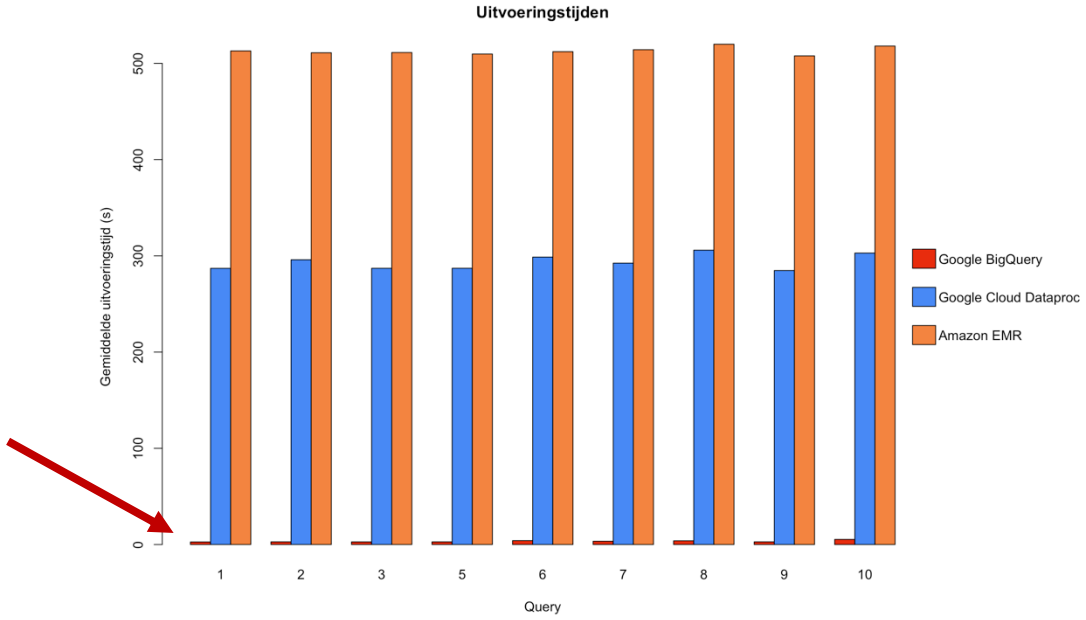
Google BigQuery is gebaseerd op Dremel. Dremel is qua principe vergelijkbaar met Apache Spark, aangezien het ook toelaat om gigantische hoeveelheden data te verwerken over verschillende servers. Maar Dremel slaat de data op een andere manier op dan Apache Spark, waardoor er een grote snelheidswinst optreedt. Dit is het best te vergelijken met een simpele tabel: Dremel slaat data op per kolom, terwijl Apache Spark data opslaat per rij. Daardoor kan Dremel de data beter comprimeren (of verkleinen) omdat Dremel zeker weet welke soort data in een kolom zit, terwijl Apache Spark dit niet weet bij een volledige rij. Bijvoorbeeld: in een volledige rij kunnen cijfers en tekst door elkaar voorkomen, terwijl in een kolom enkel cijfers zitten. Een nadeel van Dremel is dat de data reeds mooi opgeruimd moet zijn alvorens ze opgeslagen kan worden.

Cijfers

Requirements

Elk van de tools werd aan een aantal requirements afgetoetst, hierbij werd een score gegeven van één tot vijf, gebaseerd op hoe goed elke tool scoort op het requirement. Google BigQuery scoort gemiddeld 4/5, Google Cloud Dataproc en Amazon EMR scoren gemiddeld 3/5. Uit statistische analyses van de scores bleek dat geen van de drie tools beter scoort: er waren geen significante verschillen aantoonbaar.

Uitvoeringstijden queries



Figuur 2: Grafiek uitvoeringstijden

De negen queries in dit onderzoek werden een aantal keer uitgevoerd op een dataset van boetes uit de Verenigde Staten, goed voor zo'n 9GB aan data. Figuur 2 toont een overzicht van de gemiddelde uitvoeringstijd per query. Daaruit blijkt dat Google BigQuery de absolute winnaar is op het gebied van uitvoeringstijd. Een query duurt er gemiddeld drie seconden (zie pijl) in vergelijking met Google Cloud Dataproc, met gemiddeld vijf minuten en Amazon EMR, met gemiddeld acht minuten.

Conclusie

Zijn cloudplatformen nu een rendabele oplossing voor een KMO? Dit onderzoek wees uit dat elk van de drie tools zijn eigen use case heeft, maar ook dat ze zeker geschikt zijn voor KMO's. Een analyse op een cloudplatform kost een paar honderd euro, wat in vergelijking met een eigen server relatief weinig is. Je kan de server op de cloudplatformen namelijk verwijderen wanneer je analyse afgerond is, terwijl je je eigen server niet zomaar kan verwijderen, waardoor die bijgevolg vaker staat te draaien voor niets.
Kortom: gigantische hoeveelheden data verwerken met weinig configuratiewerk en weinig kosten, waarom zou je niet in de cloud werken?

Bibliografie

Sato, K. (2012). *An Inside Look at Google BigQuery.* Google.