Data-analyse met Elasticsearch
Onderzoeksvoorstel Bachelorproef

Ruben Standaert[1]

Samenvatting
In dit document staat een bachelorproefvoorstel beschreven over de voor- en nadelen van het gebruik van Elasticsearch voor data-analyse. Elasticsearch is een search engine die door veel grote bedrijven actueel gebruikt wordt om data-analyses uit te voeren, om aan logging te doen of om een uitgebreide zoekfunctionaliteit aan te bieden op bijvoorbeeld een webshop. Dit onderzoek is interessant voor Elasticsearch omdat er in toekomstige updates rekening kan gehouden worden met zwakke punten die in dit onderzoek opduiken. Ook zijn er meer en meer bedrijven die zich baseren op data-analyses om belangrijke beslissingen te nemen. Zulke bedrijven krijgen aan de hand van dit onderzoek een beter inzicht of Elasticsearch een goede keuze is voor hen. Om de voor-en nadelen te onderzoeken zullen er aan de hand van een dataset een paar interessante vragen opgesteld worden. Die vragen worden dan beantwoord door het opzetten van een project met Elasticsearch. Uit mijn ervaringen wil ik vooral sterktes en moeilijkheden van Elasticsearch gaan introduceren of bevestigen. Ook zal er gekeken worden of er sprake is van een hoge leercurve en of er voldoende support is. Een bekende kwaliteit van Elasticsearch is dat het opzetten van een nieuw project vlot gaat. Ik verwacht dat ook zelf te ondervinden. Het onderzoek zal worden voorafgegaan door een literatuurstudie. In die literatuurstudie zal er gekeken worden welke voor-en nadelen er reeds bekend zijn. Deze kunnen dan in het onderzoek bevestigd of verworpen worden. Ook zal er gekeken worden welke alternatieven er zijn voor Elasticsearch.

Sleutelwoorden
Data-analyse. Elasticsearch — Search engine

Contact: [3] ruben.standaert.w1083@student.hogent.be

Inhoudsopgave

1. Introductie

Elasticsearch is een zeer populaire search engine. Enkele merkwaardige gebruikers van Elasticsearch zijn Netflix, Facebook, Cisco, Microsoft en Adobe. De search engine staat bekend voor de lage uitvoeringstijd die nodig is om data te zoeken alsook de data-analyse en de logging die ermee mogelijk is. Door het gebruik van nodes en shards is ook de schaalbaarheid een fameuze kwaliteit. Om de voor-en nadelen van Elasticsearch te onderzoeken zal er eerst een literatuurstudie komen. In die literatuurstudie duiken voor- en nadelen op die reeds bekend zijn. Het is de bedoeling om die voor- en nadelen te bevestigen of te verwerpen aan de hand van een project. In dat project zullen er hoogswaarschijnlijk nog extra sterktes of moeilijkheden in Elasticsearch naar boven komen. Als laatste kijken we ook even naar de alternatieven voor Elasticsearch. Op die manier kan Elasticsearch overwegen om enkele moeilijkheden weg te werken. Ook kunnen bedrijven die data-analyses willen doen de voor- en nadelen beter afwegen.

2. Stand van zaken

In de master thesis van Berglund (2013) word er een eerste belangrijke stelling gemaakt over de schaalbaarheid van Elasticsearch. De search engine is zeer schaalbaar omdat de documents verdeeld zijn over verschillende nodes. Bij een specifieke zoekopdracht waarbij men zoekt op het document ID of het type zullen slechts enkele nodes moeten overlopen worden. Dat ideale scenario is echter niet altijd relevant. Afhankelijk van de criteria waarop men zoekt, kan het zijn dat alle nodes moeten overlopen worden. Dat is een limitatie van de schaalbaarheid van Elasticsearch. Dat probleem zou vandaag nog niet opgelost kunnen worden zonder dat er data-verlies voorkomt in de zoekresultaten. Een eerste factor waarmee rekening moet gehouden worden bij het kiezen voor Elasticsearch.
Alex Brasetvik, tech lead van Elastic Cloud, geeft alvast wat voor-en nadelen mee in Brasetvik (2013). Hij legt uit dat Elasticsearch een paar eigenschappen van een databank aan de kant schuift, of vereenvoudigt om aan snelheid te winnen. Want uitvoeringstijd is het belangrijkste concept van Elastic-

search. Een voorbeeld hiervan is concurrency control. Dat wordt gedaan aan de hand van versienummers van de documents. Een optimistische aanpak dat snelheid als voordeel heeft maar in uitzonderlijke gevallen foute resultaten in de data-analyses kan leveren. Van authorizatie is er nog geen sprake. Iedereen heeft dezelfde rechten. Een laatste belangrijk aspect die hij aanhaalt, zijn de out-of-memory errors. Deze worden nog niet zo goed afgehandeld in de huidige versie van Elasticsearch. Zijn conclusie is dat Elasticsearch kan gebruikt worden als voor uw project de limitaties aanvaardbaar zijn. Hij blijkt veel goed te praten aan de hand van de snelheid van Elasticsearch.

3. Methodologie

Een eerste stap in het uitvoeren van dit onderzoek is informatie opzoeken over de werking van Elasticsearch. In die eerste stap zal men al snel kunnen voorspellen of de support goed of slecht zal scoren. Nadien kan er een project opgestart worden. De dataset zal worden geïmporteerd in Elasticsearch en er worden een tiental vragen opgesteld die men aan de hand van een data-analyse in Elasticsearch zal trachten te beantwoorden. Van de eerste stap tot de laatste schrijft de onderzoeker zijn bevindingen op. Wanneer het project afgerond is, wordt er teruggekeken naar de bevindingen. Deze worden dan naast de voor- en nadelen gelegd die uit de literatuurstudie komen waarna er conclusies kunnen worden getrokken.

4. Verwachte resultaten

Een bekende kwaliteit van Elasticsearch is dat het opzetten van een nieuw project vlot gaat. Ik verwacht dat ook zelf te ondervinden. Voor het analyseren van de data wordt er een gemiddeld hoge leercurve verwacht. Het analyseren van de data wordt gedaan aan de hand van verschillende soorten aggregaties. Van die aggregaties verwacht ik dat die veel flexibiliteit bieden maar dat het gebruik ervan enige oefening en opzoekingswerk vraagt. Hoe dat opzoekingswerk verloopt zal veel afhangen van de support dat Elasticsearch te bieden heeft. Aangezien er een uitgebreide uitleg over de architectuur van Elasticsearch te vinden is verwacht ik ook voldoende support te krijgen over het werken met aggregaties.

5. Verwachte conclusies

Ik verwacht te kunnen concluderen dat de voordelen van Elasticsearch als tool voor data-analyses de negatieve punten duidelijk overwegen.

Referenties

Berglund, P. (2013, juni). *Shard Selection in Distributed Collaborative Search Engines* (masterscriptie, University of Gothenburg). Verkregen van https://findwise.com/labs/sites/default/files/reports/shard_selection_elastics.pdf

Brasetvik, A. (2013). Elasticsearch as a NoSQL Database. Verkregen van https://www.elastic.co/blog/found-elasticsearch-as-nosql