Uitdagings vir die hantering van groot datastelle

Authors:
Ian D. van der Linde[1]
Eduan Kotzé[1]

Affiliations:
[1]Department of Computer Science and Informatics, University of the Free State, South Africa

Correspondence to:
Ian van der Linde

Email:
vanderlindeid@ufs.ac.za

Postal address:
PO Box 339, Bloemfontein 9300, South Africa

How to cite this abstract:
Van der Linde, I.D. & Kotzé, E., 2015, 'Uitdagings vir die hantering van groot datastelle', *Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie* 34(1), Art. #1335, 1 page. http://dx.doi.org/10.4102/satnt.v34i1.1335

Note:
A selection of conference proceedings: Student Symposium in Science, 06 and 07 November 2014, Science Campus, University of South Africa. Organising committee: Mr Rudi W. Pretorius and Ms Andrea Lombard (Department of Geography, University of South Africa) and Dr Hertzog Bisset (South African Nuclear Energy Corporation [NECSA]).

Challenges relating to the handling of big data. Present-day methods of data processing seemingly cannot keep up with recent trends in data generation. This research specifically investigates the viability of Neo4j and Apache Cassandra as possible tools to process and store real-time Twitter data in an entirely NoSQL environment.

Daar word beraam dat die hoeveelheid data wat wêreldwyd gestoor word, min of meer gelykstaande is aan 300 eksagrepe (300 miljoen teragrepe). Hierdie getal groei teen ongeveer 28% per jaar, en bied groot uitdagings vir ondernemings wat by hul mededingers wil bybly. Data word geklassifiseer as groot data sodra dit oor die volgende drie kenmerke beskik: *volume* (data is groter as dit wat normaalweg deur relasionele databasisse gestoor en hanteer kan word), *spoed* (data word teen 'n hoë en permanente tempo ingevoer) en *verskeidenheid* (data word gelyktydig vanaf verskillende bronne ingevoer).

Om hierdie data te hanteer, is daar nuwe databasisse ontwikkel wat nie verhoudings tussen tabelle gebruik nie, en wat nie uitsluitlik aan die volledige ANSI SQL-standaard voldoen nie. Hulle staan bekend as NoSQL- (*not only SQL*) databasisse. Hulle bestaan nie uit die algemene tabulêre, verwantskapgedrewe strukture soos gevind in relasionele databasisse nie. Inteendeel, hulle gebruik strukture soos sleutelwaardestoorplek, dokumentdatabasisse, grafiekdatabasisse, en wyetabeldatabasisse (sonder verwantskappe en ongenormaliseerd).

Hierdie navorsing ondersoek die vermoë van twee NoSQL-databasisse om intydse data vanaf Twitter te stoor en te hanteer op 'n verspreide, hoëwerkverrigtingomgewing met betroubare verdraagsaamheid teenoor foute.

Die eerste van hierdie databasisse is Neo4j, 'n verspreide en gedupliseerde grafiekdatabasis wat geen SQL gebruik nie. Entiteite word gestoor en gemanipuleer met behulp van 'n unieke taal wat as Cypher bekend staan. Hierdie taal vertoon die verwantskap tussen nodusse in die grafiek op 'n grafiese wye, wat dit maklik leesbaar en verstaanbaar maak in die konteks waarin dit voorkom. Neo4j stoor 'n volledige kopie van die grafiek as geheel op elke sisteem in die groepering van rekenaars waarop dit geïnstalleer is, wat beteken dat die maksimum grootte van die grafiek nie die stoorplek van een enkele rekenaar in die groep kan oorskry nie.

Die tweede NoSQL-databasis is Apache Cassandra wat oorspronklik deur Facebook ontwikkel is en later aan die Apache Foundation geskenk is vir vrye ontwikkeling en beskikbaarstelling. Dit is ontwerp vir sisteme wat buitengewone hoeveelhede invoere moet hanteer in 'n verspreide en gedupliseerde omgewing. In teenstelling met Neo4j, verdeel Cassandra die datastel op in afskortings wat op elke rekenaar in die versameling gestoor is. Dit beteken dat die kapasiteit van die datastel as geheel uitgebrei word deur slegs addisionele rekenaars by te voeg ('n funksionaliteit waartoe Neo4j nie in staat is nie). Cassandra gebruik CQL, 'n taal wat soortgelyk is aan SQL wat slegs 'n klein gedeelte van SQL-funksionaliteit bied om sodoende vinniger te kan funksioneer.

Hierdie twee databasisse se werkverrigting, instandhouding en vermoë om op 'n liniêre wyse te skaal, is dan teen mekaar opgeweeg om die doeltreffendste sisteem op te stel om intydse Twitterdata te stoor en te verwerk.