

# Database HA state of the 2020 tooling: een vergelijkende studie en proof-of-concept

## Onderzoeksvorstel Bachelorproef 2020-2021

Elias Ameye<sup>1</sup>

### Samenvatting

In deze bachelorproef zal er onderzoek gedaan worden naar de huidige staat, anno 2020, van open-source database high availability (HA) tooling. In het eerste deel van dit onderzoek zal er geduid worden wat open-source database HA tooling juist is en waarom dit wordt gebruikt. Hierna wordt een vergelijkende studie uitgevoerd over de verschillende, huidige "state of the art" open-source database HA oplossingen/tools. Uit deze studie en in samenwerking met Inuits zal gekeken worden welke open-source database HA oplossing/tool er kunnen uitgerold worden binnen de infrastructuur van Inuits, om zo ook de komende 10 jaar aan de vraag van database HA te kunnen voldoen. Na de vergelijkende studie zal wordt in deze bachelorproef als proof-of-concept een PostgreSQL (pgSQL) cluster opgezet. Deze cluster zal volledig geautomatiseerd en reproduceerbaar zijn met behulp van de configuration management tool: Puppet. De vraag naar PostgreSQL (pgSQL) wordt steeds groter, waardoor een onderzoek naar open-source database high availability (HA) tooling zeker niet onmisbaar is.

### Sleutelwoorden

Systeembeheer — Open-source — PostgreSQL (pgSQL) — database HA

### Co-promotor

Jan Collijs<sup>2</sup> (Inuits)

Contact: <sup>1</sup> elias.ameye@student.hogent.be; <sup>2</sup> /;

## Inhoudsopgave

1	Introductie
2	State-of-the-art
3	Methodologie
4	Verwachte resultaten
5	Verwachte conclusies
	Referenties

## 1. Introductie

Database high availability (HA) staat voor de garantie van het behouden van gegevens in geval er zich een defect of storing voordoet aan de databank (SQL) server. Een storing aan een databank (SQL) server kan te wijten zijn aan verschillende factoren. Het verlies van netwerkconnectie kan leiden tot het falen van een server, maar ook een defect in de software of hardware kan ernstige gevolgen hebben voor dataverlies. Ook omgevingsfactoren zoals temperatuur moeten ook in rekening genomen worden. En een menselijke vergissing kan altijd gebeuren. Investeren in high availability geeft meer zekerheid voor data en biedt verschillende mogelijkheden voor failover en systeembescherming (IBM, 2019). Met behulp van clusters kan er een actieve en één of meerdere standby instanties van de databank (SQL) server zijn. Deze standby instanties zullen dezelfde gegevens bevatten als de actieve server (Bermingham, 2019). Wanneer

dan op één locatie een server uitvalt, kan een standby instantie van de databank (SQL) server inspringen waardoor dataverlies en server downtime gereduceerd worden.

Als proof-of-concept wordt er in dit onderzoek een PostgreSQL (pgSQL) cluster opgezet. PostgreSQL is een open-source, object-relacioneel databank systeem (PostgreSQL, 2020). Bij Inuits, een Belgisch open-source bedrijf met verschillende vestigingen in Europa, merken ze een stijging in de vraag naar het PostgreSQL verhaal. Deze cluster wordt volledig geautomatiseerd en zal reproduceerbaar zijn, dit aan de hand van Puppet, een configuration management tool. Aan de hand van deze cluster zal dan getoond worden hoe database high availability (HA) in werking treedt.

## 2. State-of-the-art

Over database high availability (HA) is er veel informatie te vinden. Een Google search naar "database high availability" levert in minder dan één seconde 733.000.000 resultaten op. Wanneer ik hier open sourceaan toevoeg, zijn er nog steeds 433.000.000 resultaten beschikbaar. Er zijn verschillende studies te vinden over high availability bij MySQL. De top drie open-source databanken van 2019 zijn, in volgorde van top 3, MySQL met 31.7%, PostgreSQL met 13.4% en MongoDB met 12.2% (Anderson, 2020) van het totaal aantal open-source databank gebruikers. Er zijn ook heel veel fora die antwoorden bieden op vragen van personen omtrent database high availability (HA), open-source en SQL servers.

### 3. Methodologie

In de eerste fase van het onderzoek zal er een vergelijkende studie gebeuren over de huidige, anno 2020, database high availability (HA) tooling. Deze verschillende tools/oplossingen zullen dan met elkaar vergeleken worden. In de tweede fase van het onderzoek wordt de focus gelegd op het opzetten van de PostgreSQL (pgSQL) cluster. Deze zal vooraf gegaan worden door een voorbereidende studie over PostgreSQL (pgSQL). Aan de hand hiervan zal er gewerkt worden aan het opbouwen van de PostgreSQL (pgSQL) cluster. Vooraleer dit geautomatiseerd wordt, zal de opbouw manueel verlopen. Wanneer deze manueel een succes is, zal er gewerkt worden aan de automatisatie van de PostgreSQL (pgSQL) cluster. De opbouw zal gebeuren via virtuele machines (VirtualBox) waarop Linux-distributies staan. In het onderzoek wordt Ubuntu gebruikt. Deze keuze kan wijzigen naargelang het verloop van het onderzoek. De opbouw zal telkens grondig gedocumenteerd worden. Alle commando's zullen overlopen worden. Hierna zal er een inleidende studie zijn over Puppet. Hierna zal er via Puppet gewerkt worden om deze PostgreSQL (pgSQL) cluster te reproduceren. Ook hier zal alles grondig gedocumenteerd worden. Na het opzetten van de PostgreSQL (pgSQL) cluster zullen er verschillende experimenten zijn die de high availability (HA) zullen testen.

### 4. Verwachte resultaten

Uit het onderzoek zal duidelijk blijken dat database high availability (HA) tooling mogelijk is binnen een PostgreSQL (pgSQL) cluster. Wanneer de PostgreSQL server zal uitvallen, zal er een standby instantie van de PostgreSQL (pgSQL) het werk van de actieve server overnemen. Hierdoor zal er geen downtime of dataverlies zijn. De data zal beschikbaar blijven en blijft onverstoord.

De PostgreSQL (pgSQL) cluster zal ook geautomatiseerd en reproduceerbaar zijn. Hierdoor zal het opzetten van deze cluster in een nieuwe omgeving niet veel tijd en moeite kosten.

### 5. Verwachte conclusies

Uit het onderzoek zal blijken dat database high availability (HA) een blijvend topic is waar voldoende aandacht aan besteedt moet worden. In kleine bedrijven hoeft high availability (HA) geen al te grote prioriteit te hebben, maar naarmate een bedrijf groeit, zal high availability (HA) steeds belangrijker worden. Een storing of downtime van de databank (SQL) server kan grote gevolgen hebben voor bedrijven. Gevolgen zoals verlies van vertrouwen bij klanten, verlies van inkomen, verlies van informatie.

Uit dit onderzoek zal ook blijken dat PostgreSQL (pgSQL) een volwaardig alternatief is van MySQL bij het opzetten van clusters.

### Referenties

- Anderson, K. (2020, januari 17). *2019 Open Source Database Report*. DZone. Verkregen 31 december 2020, van <https://dzone.com/articles/2019-open-source-database-report-top-databases-pub>
- Bermingham, D. (2019, mei 9). *Clustering for SQL Server High Availability*. Big Data Quarterly (BDQ). Verkregen 31 december 2020, van <https://www.dbta.com/BigDataQuarterly/Articles/Clustering-for-SQL-Server-High-Availability-131639.aspx>
- IBM. (2019). *High availability for databases*. International Business Machines Corporation (IBM). Verkregen 31 december 2020, van [https://www.ibm.com/support/knowledgecenter/SSANHD.7.6.1.2/com.ibm.mbs.doc/gp\\_highavail/c\\_ctr\\_ha\\_for\\_databases.html](https://www.ibm.com/support/knowledgecenter/SSANHD.7.6.1.2/com.ibm.mbs.doc/gp_highavail/c_ctr_ha_for_databases.html)
- PostgreSQL. (2020). PostgreSQL. Verkregen 31 december 2020, van <https://www.postgresql.org/>