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

Onderzoeksvoorstel Bachelorproef 2020-2021

Elias Ameye[1]

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[2] (Inuits)

Contact: [1] elias.ameye@student.hogent.be; [2] /;

Inhoudsopgave

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 PostgresSQL (pgSQL) cluster opgezet. PostgreSQL is een open-source, object-relationeel 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 öpen sourceäan toevoeg, zijn er nog steeds 433.000.000 resultaten beschikbaar. Over database high availability zijn er dus veel artikels beschikbaar die nuttig kunnen zijn voor dit onderzoek. Er zijn ook heel veel fora die antwoorden bieden op vragen van personen omtrent database high availability (HA), open-source en SQL servers.

Over high availability (HA) zijn er voldoende artikels, blogs... Eén artikel spreekt over high availability clustering en hoe dit kan aangepakt worden. Het bespreekt de cluster

architectuur en wat de best practice is voor high availability binnen een cluster. De conclusie die hier getrokken wordt is dat het primaire doel van een high availability (HA) systeem is om te voorkomen en elimineren van alle single points of failure. Deze moeten beschikken over meerdere geteste actieplannen. Dit zodat ze in geval van storing, verstoring en defect in dienstverlening direct, gepast en onafhankelijk kunnen reageren. Zorgvuldige planning + betrouwbare implementatiemethoden + stabiele softwareplatforms + degelijke hardware-infrastructuur + vlotte technische operaties + voorzichtige managementdoelstellingen + consistente databeveiliging + voorspelbare redundantiesystemen + robuuste back-upoplossingen + meerdere herstelopties = 100% uptime (Singer, 2020). In veel artikels zien we meer of mindere punten, maar het zijn wel vaak dezelfde die altijd terugkomen. In een ander artikel wordt een highly available (HA) infrastructuur gekenmerkt door: 1. Hardware redundantie; 2. Software en applicatie redundantie; 3. Gegevens redundantie; 4. Elimineren van storingspunten (Jevtic, 2018).

Er is een artikel die vier van de meest gebruikte database high availability (HA) tools/oplossingen oplijst en deze ook uitlegt. Deze vier zijn ”PgPool-II”, ”PostgreSQL Automatic Failover (PAF)”, ”RepMgr [Replication Manager] ”, en ”Patroni”(Akhtar, 2020). Dit artikel is zeker de moeite waard om door te nemen omdat in dit artikel eigenschappen van deze vier oplossingen vergeleken worden met elkaar. De website zelf biedt ook veel links aan naar gerelateerde artikels over PostgreSQL en high availability (HA). MySQL zelf toont op hun website verschillende manieren van hoe zij high availability implementeren in een cluster (MySQL, 2021). Deze info zal nuttig zijn voor in dit onderzoek. Via MySQL zal ik vergelijkingen kunnen maken en zal ik eventueel lijnen kunnen doortrekken naar PostgreSQL.

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.

Over het opzetten van een PostgreSQL (pgSQL) cluster zijn er ook veel artikels te vinden. In een bepaalde blog wordt er een high available (HA) PostgreSQL cluster („How to Set Up a Highly Available PostgreSQL Cluster Using Patroni on Ubuntu 16.04”, 2019) opgezet aan de hand van Patroni op een Ubuntu Linux-distributie. Deze blog zal zeer interessant zijn voor dit onderzoek omdat er ook in getest wordt. Dit zal handig zijn om te kunnen vergelijken bij het opzetten en testen van de proof-of-concept .

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. Dit zal gebeuren via virtuele machines. En Wanneer de virtuele machine met de PostgreSQL server op, uitval, 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.

Uit dit onderzoek zullen ook best practises blijken die geïmplementeerd kunnen worden om op die manier het risico op verlies van gegevens te verminderen. De kans om offline te zijn zullen lager liggen met een high available systeem.

Via Puppet op een snelle, moeiteloze manier een high available (HA) PostgreSQL (pgSQL) cluster kunnen opzetten en configureren. Dit zal kunnen doordat de cluster geautomatiseerd en reproduceerbaar zal zijn.

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. Zonder de implementatie van een high available (HA) architectuur kan een storing of downtime van de databank (SQL) server grote gevolgen hebben op een bedrijf/organisatie. Gevolgen zoals verlies van vertrouwen bij klanten, verlies van inkomen, verlies van informatie. Door middel van hardware-, software- en gegevensredundantie en het elimineren van mogelijke storingspunten gaan we high availability kunnen blijven garanderen. De kost en tijd die gïnvesteerd moet worden bij het onderhouden van een high available systeem zal lager liggen dan de kost en tijd die geïnvesteerd moet worden in geval van een downtime.

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

Referenties

Akhtar, H. (2020, augustus 10). *PostgreSQL High Availability: The Considerations and Candidates*. Verkregen 2 januari 2021, van https://www.highgo.ca/2020/08/10/postgresql-high-availability-the-considerations-and-candidates/

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

How to Set Up a Highly Available PostgreSQL Cluster Using Patroni on Ubuntu 16.04. (2019, februari 19). Alibaba Cloud. Verkregen 2 januari 2021, van https://www.alibabacloud.com/blog/how-to-set-up-a-highly-available-postgresql-cluster-using-patroni-on-ubuntu-16-04_594477

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

Jevtic, G. (2018, juni 22). *What is High Availability Architecture? Why is it Important?* PhoenixNAP. Verkregen 2 januari 2021, van https://phoenixnap.com/blog/what-is-high-availability

MySQL. (2021). *MySQL Enterprise High Availability*. MySQL. Verkregen 2 januari 2021, van https://www.mysql.com/products/enterprise/high_availability.html

PostgreSQL. (2020). PostgreSQL. Verkregen 31 december 2020, van https://www.postgresql.org/

Singer, D. (2020, augustus 6). *What is High Availability? A Tutorial*. Liquid Web. Verkregen 2 januari 2021, van https://www.liquidweb.com/kb/what-is-high-availability-a-tutorial/