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 onderzoek gedaan worden naar de huidige staat van open-source database high availability (HA) tooling. De vraag naar PostgreSQL (pgSQL) wordt steeds groter, waardoor een onderzoek naar open-source database high availability (HA) tooling zeker niet onmisbaar is. In het eerste deel van dit onderzoek zal er geduid worden wat open-source database high availability (HA) tooling precies is en wat de toepassingen ervan zijn. Hierna wordt een vergelijkende studie uitgevoerd over de verschillende open-source database high availability (HA) oplossingen. Na de vergelijkende studie wordt als proof-of-concept een PostgreSQL (pgSQL) cluster opgezet met behulp van de automation tool Puppet. Dit zorgt voor automatisering en reproduceerbaarheid. In dit onderzoek zal ik worden bijgestaan door Ruben Demey, Global IT Operations Manager bij ST Engineering iDirect.

Sleutelwoorden
Systeembeheer — Open-source — PostgreSQL — Databases — High availability

Co-promotor
Jan Collijs[2] (Inuits)

Contact: [1] elias.ameye@student.hogent.be; [2] rdem@idirect.net;

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 server. Een storing aan een databank server kan te wijten zijn aan verschillende factoren. Voorbeelden hiervan zijn het verlies van netwerkconnectie en een defect in de software of hardware van de databankserver. Ook menselijke factoren en omgevingsfactoren moeten in rekening genomen worden. Voorbeelden hiervan zijn een menselijke vergissing en een wijziging in temperatuur. Investeren in high availability geeft meer zekerheid over de beschikbaarheid van data en biedt verschillende mogelijkheden voor failover en systeembescherming (IBM, 2019). Met behulp van clusters kan er één actieve en een of meerdere standby instanties van de databank server zijn. Een cluster is een groep van servers en computers die samenwerken met elkaar alsof het één systeem is. Deze standby instanties zullen, in het ideale geval, dezelfde gegevens bevatten als de actieve server (Bermingham, 2019). Wanneer dan een actieve server faalt, kan een standby instantie inspringen waardoor dataverlies en server downtime gereduceerd worden.

2. State of the art

Singer spreekt over high availability (HA) clustering als een groep van servers die applicaties en services ondersteunen die op een betrouwbare manier gebruikt kunnen worden met een minimaal aantal aantal downtimes. Hij bespreekt de cluster architectuur en wat de best practice is voor high availability (HA) binnen een cluster. De conclusie die hier getrokken wordt, is dat het primaire doel van een high availability (HA) systeem het voorkomen en elimineren van alle single points of failure zijn. Dit systeem moet 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). Ook Jevtic heeft het over veel van de punten die hierboven reeds zijn aangehaald. Jevtic spreekt over een highly available architectuur waarin meerdere componenten samenwerken om een ononderbroken service gedurende een bepaalde periode te garanderen. Dit omvat ook de reactietijd op verzoeken van gebruikers. Jevtic kenmerkt een highly available (HA) infrastructuur aan de hand van: 1. Hardware redundantie; 2.

Software en applicatie redundantie; 3. Gegevens redundantie; 4. Elimineren van storingspunten (Jevtic, 2018). Akhtar heeft vier van de meest gebruikte database high availability (HA) oplossingen opgelijst. Deze vier zijn ”PgPool-II”, ”PostgreSQL Automatic Failover (PAF)”, ”RepMgr [Replication Manager] ”, en ”Patroni”(Akhtar, 2020). Akhtar vergelijkt deze verschillende oplossingen kort met elkaar. Akhtar definieert high availability (HA) als niet alleen de continuïteit van een bepaalde service, maar volgens hem gaat high availability (HA) ook over het vermogen van een systeem om een (hogere) werkdruk te kunnen schalen en te beheren. Dit systeem moet volgens Akhtar de gemiddelde werkdruk, maar ook de piekmomenten aankunnen. Aldus Andersen zijn de top drie open-source databanken van 2019, 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.

3. Methodologie

In de eerste fase van het onderzoek zal er een vergelijkende studie gebeuren over de huidige, database high availability (HA) oplossingen. Deze verschillende tools/oplossingen zullen dan met elkaar vergeleken worden. Hierbij wordt gekeken naar welke elementen er allemaal (meermaals) voorkomen. Hiervan komt er een lijst die gebruikt zal worden om te schiften tussen de verschillende oplossingen. Op deze manier zal er dan een oplossing gekozen worden. Via deze methode wordt er gekeken om maximum 3 verschillende oplossingen uit te werken, waarbij één oplossing gebruikt zal worden bij de proof of concept. In de tweede fase van het onderzoek wordt de focus gelegd op het opzetten van de PostgreSQL (pgSQL) cluster als proof of concept. 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 zal vooraf gegaan worden door een literatuurstudie 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. De opbouw zal gebeuren via virtuele machines (VirtualBox) waarop Linux-distributies staan. In het onderzoek zal dus gebruik gemaakt worden van Linux-servers. De keuze van Linux-distributie zal onderbouwd worden in dit onderzoek. De opbouw van de cluster zal telkens grondig gedocumenteerd worden. Alle commando’s zullen hierbij overlopen worden. Hierna zal er een inleidende literatuurstudie zijn over Puppet en zal er aansluitend 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 zal er getest worden of de gebruikte database high availability (HA) oplossing functioneel is. Deze testen zullen uitgebreid beschreven worden.

4. Verwachte resultaten

Uit het onderzoek zal blijken dat verschillende database high availability (HA) oplossingen mogelijk zijn binnen een PostgreSQL (pgSQL) cluster. De opbouw van deze cluster zal gebeuren aan de hand van virtuele machines. Wanneer de virtuele machine, waarop de PostgreSQL server staat, uitvalt, zal er een standby instantie van deze server het werk van de actieve, uitgevallen server overnemen. Hierdoor zal er geen downtime of dataverlies zijn. De data zal beschikbaar en onverstoord blijven.

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

Door gebruik te maken van Puppet zal de reproduceerbaarheid van de high available (HA) PostgreSQL (pgSQL) cluster zeer eenvoudig en snel moeten verlopen.

5. Verwachte conclusies

Uit het onderzoek zal blijken dat database high availability (HA) een blijvend topic is waar voldoende aandacht aan besteed moet worden in kleine en grote bedrijven. 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 zal er high availability gegarandeerd worden. De kost en tijd die geïnvesteerd moet worden in het onderhouden van een high available systeem zal lager liggen dan de kost en tijd die geïnvesteerd moet worden in geval van een downtime. Met de proof of concept wordt dan aangetoond dat database high availability (HA) eenvoudig te implementeren valt in een PostgreSQL (pgSQL) cluster.

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/