Cluster Galera

Cluster Galera

Le clustering de bases de données est le processus consistant à combiner plusieurs serveurs en les connectant à une seule base de données. Le clustering améliore la disponibilité de votre base de données en répartissant la charge sur différents serveurs. Si un serveur tombe en panne, d'autres sont rapidement disponibles pour continuer à servir.

MariaDB Galera est une solution de clustering multi-maîtres qui vous permet de lire et d'écrire sur n'importe quel nœud du cluster. Avec MariaDB Galera, une modification apportée à un nœud est répliquée sur tous les nœuds. MariaDB Galera prend en charge les moteurs de stockage XtraDB/InnoDB et est disponible uniquement sur Linux.

Configurer le premier serveur

Tout d'abord, connectez-vous au premier serveur et créez un fichier de configuration Galera dans la partie mysqld :

20

1
```
[mysqld]
```

2
```
binlog_format=ROW
```

3
```
default-storage-engine=innodb
```

4
```
innodb_autoinc_lock_mode=2
```

5
```
bind-address=0.0.0.0
```

6
```

```

7
```
# Galera Provider Configuration
```

```
wsrep_on=ON
```

```
wsrep_provider=/usr/lib/galera/libgalera_smm.so
```

```

```

```
# Galera Cluster Configuration
```

```
wsrep_cluster_name="galera_cluster"
```

```
wsrep_cluster_address="gcomm://192.168.0.101,192.168.0.102,192.168.0.103"
```

```

```

```
# Galera Synchronization Configuration
```

```
wsrep_sst_method=rsync
```

```

```

```
# Galera Node Configuration
```

```
wsrep_node_address="192.168.0.101"
```

```
wsrep_node_name="server1"
```

Configurer le deuxième serveur

Le deuxieme serveur est homogène au premier:

20

1
```
[mysqld]
```

2
```
binlog_format=ROW
```

3
```
default-storage-engine=innodb
```

4
```
innodb_autoinc_lock_mode=2
```

5
```
bind-address=0.0.0.0
```

6
```

```

7
```
# Galera Provider Configuration
```

8
```
wsrep_on=ON
```

9
```
wsrep_provider=/usr/lib/galera/libgalera_smm.so
```

10
```

```

11
```
# Galera Cluster Configuration
```

12
```
wsrep_cluster_name="galera_cluster"
```

13
```
wsrep_cluster_address="gcomm://192.168.0.101,192.168.0.102,192.168.0.103"
```

14
```

```

15

```
# Galera Synchronization Configuration
```

16

```
wsrep_sst_method=rsync
```

17

```

```

18

```
# Galera Node Configuration
```

19

```
wsrep_node_address="192.168.0.102"
```

20

```
wsrep_node_name="server2"
```

Démarrage du cluster

Démarrage du premier nœud et vérification

4

1

```
 sudo galera_new_cluster
```

2

```
[root@mariadb01 my.cnf.d]# ps -ef |  grep mysql
```

3

```
mysql 7604 1 4 15:12 ? 00:00:00 /usr/sbin/mysqld --wsrep-new-cluster --wsrep_start_position=0000
```

4

```
root 7650 3887 0 15:12 pts/1 00:00:00 grep --color=auto mysql
```

Vérification :

8

1

```
[root@mariadb01 ~]# mysql -u dba -p
```

2

```
MariaDB [(none)]> show status like '%wsrep_cluster_size%';
```

3

```
+--------------------+-------+
```

4

```
| Variable_name | Value |
```

5

```
+--------------------+-------+
```

6

```
| wsrep_cluster_size | 1 |
```

7

```
+--------------------+-------+
```

8

```
1 row in set (0.00 sec)
```

Pour les autres nœuds, le démarrage se fait par la commande classique :

1

1

```
systemctl start mysqld
```

Safe-To-Bootstrap

Les clusters Galera sont généralement conçus pour fonctionner en continu, il n'est donc pas nécessaire d'arrêter l'ensemble du cluster pendant le fonctionnement normal. Pourtant, s'il est nécessaire d'effectuer une telle procédure, il est important qu'elle se termine en toute sécurité et le plus rapidement possible afin d'éviter les temps d'arrêt prolongés et la perte potentielle de données.

Galera 3.19 inclut deux améliorations importantes au redémarrage de l'ensemble du cluster : la protection « Safe-to-Bootstrap » et la récupération Gcache. Dans cet article, nous allons décrire la première fonctionnalité.

REDÉMARRAGE DE L'ENSEMBLE DU CLUSTER

Tout d'abord, quelques mots sur les redémarrages de cluster en général. Qu'il s'agisse d'un arrêt ordonné ou d'un crash soudain de tous les nœuds, le redémarrage de l'ensemble du cluster est régi par les principes suivants :

Étant donné que l'ancien cluster n'existe plus logiquement, un nouveau cluster logique est en cours de création
Le premier nœud en cours de démarrage doit être amorcé
Il est important de sélectionner le nœud qui a les dernières transactions validées comme premier nœud dans le nouveau cluster

LA PROTECTION SAFE-TO-BOOTSTRAP

Dans un arrêt ordonné, le nœud qui a été arrêté en dernier sera celui qui a la dernière transaction validée et doit être choisi comme premier nœud dans le nouveau cluster. La sélection d'un autre nœud pour ce rôle peut entraîner des erreurs sur la route et ouvrir la possibilité de perdre ces dernières transactions.

Pour faciliter cette décision et éviter les choix dangereux, Galera, à partir de la version 3.19, gardera une trace de l'ordre dans lequel les nœuds sont arrêtés. Le nœud qui a été arrêté en dernier sera marqué comme "Safe-to-Bootstrap". Tous les autres nœuds seront marqués comme dangereux pour l'amorçage.

Lors de l'amorçage du nouveau cluster, Galera refusera d'utiliser comme premier nœud un nœud qui a été marqué comme dangereux pour l'amorçage.

SÉLECTION DU BON NŒUD

La procédure pour sélectionner le bon nœud à partir duquel s'amorcer dépend de la façon dont le cluster s'est terminé : via un arrêt ordonné ou un crash.

En cas d'arrêt ordonné, il suffit de suivre les recommandations de la fonction « Safe-to-Bootstrap ». Recherchez le nœud dont gratate.dat a safe_to_bootstrap : 1 :

```
# GALERA saved state
version: 2.1
uuid:    9acf4d34-acdb-11e6-bcc3-d3e36276629f
seqno:   15
safe_to_bootstrap: 1
```

et utilisez ce nœud.

En cas de crash dur, tous les nœuds auront safe_to_bootstrap: 0 , nous devrons donc consulter le moteur de stockage InnoDB pour déterminer quel nœud a validé la dernière transaction dans le cluster. Ceci est réalisé en démarrant mysqld avec la variable --wsrep-recover , qui produit une sortie comme celle-ci :

```
...
2016-11-18 01:42:15 36311 [Note] InnoDB: Database was not shutdown normally!
2016-11-18 01:42:15 36311 [Note] InnoDB: Starting crash recovery.
...
2016-11-18 01:42:16 36311 [Note] WSREP: Recovered position: 37bb872a-ad73-11e6-819f-
f3b71d9c5ada:345628
...
2016-11-18 01:42:17 36311 [Note] /home/pilou/git/mysql-wsrep-bugs-5.6/sql/mysqld: Shutdown
complete
```

Le nombre après la chaîne UUID sur la ligne « Position récupérée » est celui à surveiller. Choisissez le nœud qui a le nombre le plus élevé et modifiez son gratate.dat pour définir safe_to_bootstrap : 1 :

```
# GALERA saved state
version: 2.1
uuid:    37bb872a-ad73-11e6-819f-f3b71d9c5ada
seqno:   -1
safe_to_bootstrap: 1
```

En faisant cela, vous indiquez à Galera que vous avez volontairement sélectionné ce nœud et cela vous permettra de démarrer à partir de celui-ci.

Reprise sur incident

Récupération sur incident
Contrairement à la réplication MySQL standard, un cluster Galera agit comme une entité logique, qui contrôle le statut et la cohérence de chaque nœud ainsi que le statut de l'ensemble du cluster. Cela permet de maintenir l'intégrité des données plus efficacement qu'avec la réplication asynchrone traditionnelle sans perdre les écritures sécurisées sur plusieurs nœuds en même temps.

Cependant, il existe des scénarios où le service de base de données peut s'arrêter sans qu'aucun nœud ne puisse répondre aux demandes.

Scénario : le nœud A est correctement arrêté

Dans un cluster à trois nœuds (nœud A, nœud B, nœud C), un nœud (nœud A, par exemple) est gracieusement arrêté : à des fins de maintenance, de changement de configuration, etc.

Dans ce cas, les autres nœuds reçoivent un message « au revoir » du nœud arrêté et la taille du cluster est réduite ; certaines propriétés comme le calcul du quorum ou l'incrémentation automatique sont automatiquement modifiées. Dès que le nœud A est redémarré, il rejoint le cluster en fonction de sa wsrep_cluster_address dans my.cnf.

Si le cache d'écriture (gcache.size) sur les nœuds B et/ou C a encore toutes les transactions exécutées pendant que le nœud A était en panne, la jointure est possible via IST . Si IST est impossible en raison de transactions manquantes dans le gcache du donneur, la décision de secours est prise par le donneur et SST est démarré automatiquement.

Scénario : deux nœuds sont correctement arrêtés

Similaire au scénario : le nœud A est arrêté en douceur , la taille du cluster est réduite à 1 — même le seul nœud restant C forme le composant principal et est capable de répondre aux demandes des clients. Pour remettre les nœuds dans le cluster, il vous suffit de les démarrer.

Cependant, lorsqu'un nouveau nœud rejoint le cluster, le nœud C passera à l'état « Donateur/Désynchronisé » car il doit fournir le transfert d'état au moins au premier nœud qui se joint. Il est toujours possible d'y lire/écrire pendant ce processus, mais cela peut être beaucoup plus lent, ce qui dépend de la quantité de données à envoyer pendant le transfert d'état. En outre, certains équilibreurs de charge peuvent considérer le nœud donneur comme non opérationnel et le supprimer du pool. Il est donc préférable d'éviter la situation où un seul nœud est actif.

Si vous redémarrez le nœud A puis le nœud B, vous voudrez peut-être vous assurer que la note B n'utilise pas le nœud A comme donneur de transfert d'état : le nœud A peut ne pas avoir tous les jeux d'écriture nécessaires dans son gcache. Spécifiez node C node comme donneur dans votre fichier de configuration et démarrez le service mysql :

$ systemctl démarrer mysql
Voir également

Documentation Galera : option wsrep_sst_donor
https://galeracluster.com/library/documentation/mysql-wsrep-options.html#wsrep-sst-donor
Scénario : Les trois nœuds sont correctement arrêtés
Le cluster est complètement arrêté et le problème est de l'initialiser à nouveau. Il est important qu'un nœud PXC écrive sa dernière position exécutée dans le grastate.datfichier.

En comparant le numéro de seqno dans ce fichier, vous pouvez voir quel est le nœud le plus avancé (probablement le dernier arrêté). Le cluster doit être amorcé à l'aide de ce nœud, sinon les nœuds qui avaient une position plus avancée devront effectuer le SST complet pour rejoindre le cluster initialisé à partir du moins avancé. En conséquence, certaines transactions seront perdues). Pour amorcer le premier nœud, appelez le script de démarrage comme ceci :

$ systemctl démarrer mysql@bootstrap.service
Noter

Même si vous démarrez à partir du nœud le plus avancé, les autres nœuds ont un numéro de séquence inférieur. Ils devront toujours se joindre via le SST complet car le cache Galera n'est pas conservé au redémarrage.

Pour cette raison, il est recommandé d'arrêter les écritures sur le cluster avant son arrêt complet, afin que tous les nœuds puissent s'arrêter à la même position. Voir aussi pc.recovery.

Scénario : Un nœud disparaît du cluster
C'est le cas lorsqu'un nœud devient indisponible en raison d'une panne de courant, d'une panne matérielle, d'une panique du noyau, d'un crash mysqld, sur mysqld pid, etc.kill -9

Deux nœuds restants remarquent que la connexion au nœud A est interrompue et commencent à essayer de s'y reconnecter. Après plusieurs délais d'expiration, le nœud A est supprimé du cluster. Le quorum est enregistré (2 nœuds sur 3 sont actifs), donc aucune interruption de service ne se produit. Après son redémarrage, le nœud A se joint automatiquement (comme décrit dans Scénario : le nœud A est normalement arrêté).

Scénario : Deux nœuds disparaissent du cluster
Deux nœuds ne sont pas disponibles et le nœud restant (nœud C) n'est pas en mesure de former seul le quorum. Le cluster doit basculer vers un mode non principal, où MySQL refuse de servir les requêtes SQL. Dans cet état, le mysqldprocessus sur le nœud C est toujours en cours d'exécution et peut être connecté, mais toute instruction liée aux données échoue avec une erreur

mysql > sélectionnez * à partir de test . sbtest1 ;
ERREUR 1047 (08 S01): WSREP n'a pas encore préparé le nœud pour l' utilisation de l' application
Les lectures sont possibles jusqu'à ce que le nœud C décide qu'il ne peut pas accéder aux nœuds A et B. Les nouvelles écritures sont interdites.

Dès que les autres nœuds deviennent disponibles, le cluster se reforme automatiquement. Si le nœud B et le nœud C étaient simplement séparés par le réseau du nœud A, mais qu'ils peuvent toujours se joindre, ils continueront à fonctionner car ils forment toujours le quorum.

Si les nœuds A et B tombent en panne, vous devez activer manuellement le composant principal sur le nœud C avant de pouvoir afficher les nœuds A et B. La commande pour ce faire est :

mysql > SET GLOBAL wsrep_provider_options = 'pc.bootstrap=true' ;
Cette approche ne fonctionne que si les autres nœuds sont en panne avant de le faire ! Sinon, vous vous retrouvez avec deux clusters ayant des données différentes.

Références croisées

Ajout de nœuds au cluster

Scénario : tous les nœuds sont tombés en panne sans procédure d'arrêt appropriée
Ce scénario est possible en cas de panne de courant du datacenter ou en cas de bug MySQL ou Galera. En outre, cela peut se produire en raison de la cohérence des données compromise lorsque le cluster détecte que chaque nœud a des données différentes. Le grastate.datfichier n'est pas mis à jour et ne contient pas de numéro de séquence valide (seqno). Cela peut ressembler à ceci :

$ cat /var/lib/mysql/grastate.dat
# État enregistré de GALERA
version : 2.1
uuid : 220dcdcb-1629-11e4-add3-aec059ad3734
numéro de séquence : -1
safe_to_bootstrap : 0
Dans ce cas, vous ne pouvez pas être sûr que tous les nœuds sont cohérents les uns avec les autres. Nous ne pouvons pas utiliser la variable safe_to_bootstrap pour déterminer le nœud qui a la dernière transaction validée car elle est définie sur 0 pour chaque nœud. Une tentative d'amorçage à partir d'un tel nœud échouera à moins que vous ne commenciez mysqldavec le --wsrep-recoverparamètre :

$ mysqld --wsrep-recover
Recherchez dans la sortie la ligne qui signale la position récupérée après l'UUID du nœud (1122 dans ce cas) :

...
... [Remarque] WSREP : Position récupérée : 220dcdcb-1629-11e4-add3-aec059ad3734:1122
...
Le nœud où la position récupérée est marquée par le plus grand nombre est le meilleur candidat bootstrap. Dans son grastate.datfichier, définissez la variable safe_to_bootstrap sur 1 . Ensuite,

amorcez à partir de ce nœud.

Noter

Après un arrêt, vous pouvez boosterrap à partir du nœud qui est marqué comme sûr dans le grastate.datfichier.

...
safe_to_bootstrap : 1
...
Voir également

Documentation Galera
Présentation de la fonctionnalité "Safe-To-Bootstrap" dans Galera Cluster
Dans les versions récentes de Galera, l'option pc.recovery(activée par défaut) enregistre l'état du cluster dans un fichier nommé gvwstate.datsur chaque nœud membre. Comme le nom de cette option l'indique (pc - composant principal), elle n'enregistre qu'un cluster étant dans l'état PRIMAIRE. Un exemple de contenu de : fichier peut ressembler à ceci :

cat /var/lib/mysql/gvwstate.dat
my_uuid : 76de8ad9-2aac-11e4-8089-d27fd06893b9
#vwbeg
view_id : 3 6c821ecc-2aac-11e4-85a5-56fe513c651f 3
bootstrap : 0
membre : 6c821ecc-2aac-11e4-85a5-56fe513c651f 0
membre : 6d80ec1b-2aac-11e4-8d1e-b2b2f6a5018ad 0
membre : 8089-d27fd06893b9 0
#vwend
Nous pouvons voir un cluster à trois nœuds avec tous les membres actifs. Grâce à cette nouvelle fonctionnalité, les nœuds tenteront de restaurer le composant principal une fois que tous les membres commenceront à se voir. Cela permet au cluster PXC de récupérer automatiquement après une mise hors tension sans aucune intervention manuelle ! Dans les logs nous verrons :

Scénario : Le cluster perd son état principal en raison d'un cerveau divisé
Pour les besoins de cet exemple, supposons que nous ayons un cluster composé d'un nombre pair de nœuds : six, par exemple. Trois d'entre eux se trouvent à un endroit tandis que les trois autres se trouvent à un autre endroit et ils perdent la connectivité réseau. Il est recommandé d'éviter une telle topologie : si vous ne pouvez pas avoir un nombre impair de nœuds réels, vous pouvez utiliser un nœud arbitre supplémentaire (garbd) ou définir un poids pc plus élevé pour certains nœuds. Mais lorsque le split brain se produit de quelque manière que ce soit, aucun des groupes séparés ne peut maintenir le quorum : tous les nœuds doivent cesser de répondre aux demandes et les deux parties du cluster essaieront en permanence de se reconnecter.

Si vous souhaitez restaurer le service avant même que le lien réseau ne soit restauré, vous pouvez à nouveau rendre l'un des groupes principal à l'aide de la même commande que celle décrite dans Scénario : deux nœuds disparaissent du cluster

SET GLOBAL wsrep_provider_options = 'pc.bootstrap=true' ;
Après cela, vous êtes en mesure de travailler sur la partie restaurée manuellement du cluster, et l'autre moitié devrait pouvoir se reconnecter automatiquement à l'aide d' IST dès que le lien réseau est restauré.

Avertissement

Si vous définissez l'option d'amorçage sur les deux parties séparées, vous vous retrouverez avec deux instances de cluster vivantes, avec des données susceptibles de diverger les unes des autres. La restauration d'un lien réseau dans ce cas ne les fera pas se rejoindre tant que les nœuds ne seront pas redémarrés et que les membres spécifiés dans le fichier de configuration ne seront pas à nouveau connectés.

Ensuite, comme le modèle de réplication Galera se soucie vraiment de la cohérence des données : une fois l'incohérence détectée, les nœuds qui ne peuvent pas exécuter l'instruction de changement de ligne en raison d'une différence de données - un arrêt d'urgence sera effectué et le seul moyen de ramener les nœuds dans le cluster est via le SST complet