Planification du déploiement d'une plateforme Cassandra-Spark et estimation de sa tolérance limite

7 juillet 2018

1 Objectifs

Nous présentons dans ce document, le choix des caractéristiques physiques des machines qui doivent constituer un cluster Cassandra-Spark pour qu'il puisse stocker et garantir une consultation et un traitement rapide des données. Ensuite, nous montrons sur un exemple la démarche à suivre pour estimer le nombre de machines nécessaire au début de la mise en place d'un cluster Cassandra pour qu'il soit parfaitement opérationnel jusqu'à une date fixée à partir de laquelle son extension deviendra indispensable pour le maintenir en bon état de fonctionnement. Puis, nous montrons comment estimer l'espace de stockage utile à ne pas dépasser dans un cluster Cassandra. Nous terminerons par une estimation de la tolérance limite à la perte des machines dans un cluster Cassandra en fonction de la taille des données qu'il contient déjà.

2 Choix des machines physiques

Les performances d'un cluster Cassandra-Spark sont étroitement liées au nombre de cores CPU, à la taille de la mémoire RAM, au type de disque dur utilisé et à la bande passante du réseau.

1. Le nombre de cores CPU par machine doit être au moins égal à 8 (de préférence 16 cores ou plus).
2. Il est recommendé d'utiliser les disques durs de type SSD.
3. La taille de la mémoire est liée à celle du disque dur. Par exemple, 1 TB de disque doit correspondre à au moins 32 GB de mémoire RAM (de préférence 64 GB de RAM ou plus). Ainsi, une machine de 2 TB de disque dur devrait avoir au moins 64 GB de RAM.
4. Le débit du trafic réseau entre les machines du cluster doit être au moins égal à 1 GB/s.

3 Compaction des données

Un facteur à prendre impérativement en compte lors du dimensionnement d'un cluster Cassandra est l'opération interne de compaction des données. Pour mieux comprendre la

compaction, il faut savoir que chaque table est définitivement enregistrée dans le disque dur sous un format immutable appelé "SSTable". Un SSTable n'est donc pas modifiable. Ainsi, lors de l'insertion de nouvelles lignes ou des mises à jour ou des suppressions de certaines lignes anciennes dans une table, un nouveau SSTable contenant ces modifications est créée et enregistré dans le disque dur. De cette manière, plusieurs SSTables peuvent être associés à une même table. Lors de la lecture d'une table, toutes les SSTables associés à cette table sont consultés et seuls les dernières versions des lignes et des colonnes demandées sont retournées au client. Le temps de lecture devient par conséquent long lorsque le nombre de SSTables à consulter augmente.

Pour améliorer ses performances, Cassandra réalise périodiquement des compactions des SSTables afin de réduire leur nombre. Durant la compaction, le cluster est toujours disponible en lecture et en écriture. En effet, Cassandra réalise la consolidation de toutes les SSTables associées à chaque table à partir de leurs copies. La compaction nécessite donc une duplication des données. Il faut donc de l'espace libre dans le disque dur pour que cela se déroule dans de bonnes conditions.

Nous terminons cette section en disant que la compaction est le mécanisme interne de Cassandra qui consiste à réaliser les actions suivantes.

i) Création d'une nouvelle SSTable pour chaque table dont le contenu est une fusion des contenus de tous les SSTables associés à cette table tout en tenant compte des modifications apportées.

ii) Suppression de tous les anciens SSTables associés à chaque table pour libérer de l'espace disque dur et permettre désormais un accès rapide au contenu d'une table uniquemement à partir du nouveau SSTable créée pour cette table.

4 Dimensionnement initial du cluster

Nous souhaitons mettre en place un cluster Cassandra capable de stocker progressivement durant 4 années les données de 12 expérimentations arrivant avec un débit journalier de 270 MO/jour. Après ces 4 années, il faudrait ajouter des nouvelles machines à ce cluster pour l'agrandir afin de pouvoir y introduire de nouvelles données. Pour la détermination de la taille initiale de ce cluster, nous considérons un certains nombre d'hypothèses.

4.1 Hypothèses

1. Chaque machine a un disque dur de 2 TB.
2. Environ 90% du disque dur de chaque machine est réellement utilisable pour le stockage des données.
3. Dans chaque machine, 50% de l'espace disque dur utilisable est libre et résevé pour la compaction. Dans ce cas, on dit que le facteur de compaction est de 2.
4. Chaque donnée est répliquée 3 fois dans le cluster.

5. Trois vues de la table de base contenant les données de chaque expérimentation sont créées pour permettre et optimiser les temps d'éxécution de certaines requêtes répondant aux besoins des utilisateurs.

4.2 Estimation du nombre de machines

La capacité totale nécessaire pour le stockage des données mentionnées dans la section précédente est de :

$$270\,\text{MO} \times 365 \times 4 \times 12 \times 3 \times 4 \times 2 = 113529600\ \text{MO}$$

où, nous avons tenu compte du nombre total de jours de l'expérimentation dans 4 ans qui est de 365×4, du facteur de réplication qui est de 3, du nombre des expérimentations qui est 12, du nombre total de tables (la table de base et ses 3 vues) qui est de 4 et du facteur de compaction qui est de 2.

La capacité disque dur utilisable de chaque machine est de :

$$\frac{2000000\,\text{MO} \times 90}{100} = 1800000\,\text{MO}$$

Le nombre de machines nécessaire est de :

$$\frac{113529600}{1800000} = 63.072$$

Il faut donc au moins 64 machines ayant chacune 2 TB d'espace de stockage disque dur pour enregistrer progressivement les mesures des expérimentations à effectuer dans les 4 prochaines années. Pour réduire les dépenses en énergie, on peut commencer avec un cluster Cassandra constitué de 16 machines et l'étendre chaque année en lui ajoutant 16 machines supplémentaires.

5 Calcul de l'espace disque dur utilisable

Dans un cluster Cassandra en cours de fonctionnement depuis un certain temps, il faut par moment vérifier que l'espace total utilisable pour le stockage des données n'est pas atteint. Cette vérification permettra de planifier certaines opérations de maintenance comme la suppression des données qui ne serviront plus ou l'ajout de nouvelles machines dans le cluster pour augmenter sa capacité de stokage et pour réduire son temps d'éxécution des requêtes.

Considérons par exemple un cluster Cassandra constitué de 64 machines ayant chacune un disque dur de 2 TB. Le calcul de l'espace total limite utilisable dans ce cluster pour le stockage des données peut se faire en suivant les étapes décrites ci-dessous.

1. Calcul de l'espace total disque dur du cluster :

$$2\,\text{TB} \times 64 = 128\,\text{TB}.$$

2. Calcul de l'espace total disque dur utilisable du cluster. Il est recommendé de déduire 10% de l'espace total disque dur du cluster.

$$\frac{128\,\text{TB} \times 90}{100} = 115.2\,\text{TB}.$$

3. Calcul de l'espace disque dur utilisable pour le stockage des données dans le cluster. Il est recommendé de laisser libre environ 50% de l'espace total disque dur utilisable du cluster pour la compaction des SSTables.

$$\frac{115.2\,\text{TB} \times 50}{100} = 57.6\,\text{TB}.$$

Pour un tel cluster, il faut donc vérifier par moment que la taille des données qu'il contient ne dépasse pas 57.6 TB.

6 Estimation de la tolérance limite aux pannes

Dans un cluster Cassandra, il peut arriver qu'une machine tombe en panne. Dans ce cas, toutes les données sont toujours présentes dans le cluster à condition qu'on ait pris le soin de répliquer chacune d'elles au moins trois fois tel qu'il est recommendé. Cependant, si une machine est morte ou est déconnectée du cluster durant plus d'un certain temps qui est fixé à 10 jours par défaut, les données détenues par la machine morte seront redistribuées dans le cluster à partir de leur réplicas présent dans les machines restantes de sorte que le facteur de réplication de chaque donnée soit respecté. Ainsi, toute la charge du cluster repose désormais sur les machines encore en vie. Il semble crucial de se poser la question suivante.

"*Quel est le nombre limite de machines que l'on peut se permettre de perdre consécutivement par intervalles de 10 jours sans les remplacer et conserver toute la charge et un bon fonctionnement du cluster constitué des machines en vie restantes ?*"

Dans cette section, nous proposons une formule qui permet d'estimer ce nombre limite de machines en fonction de la taille du cluster et de la charge en données qu'il contient. Nous supposons que les machines ont des caractéristiques physiques identiques et que la charge est uniformément répartie entre ces machines. Nous utiliserons dans la suite les notations ci-dessous.

1. N : Nombre total de machines constituant le cluster.
2. V : Volume limite autorisé pour le stockage des données dans le cluster. Précisons que la capacité V est calculée selon la procédure décrite dans la section 5 à partir de l'espace disque dur de chaque machine.
3. X : Volume de données en cours détenu par chaque machine.
4. M : Nombre limite de machines qui peuvent consécutivement mourir par intervalles de 10 jours sans risquer de perdre les données.

Si on perd M machines de cette manière, le cluster constitué par le reste des machines encore en vie a une capacité limite de stockage qui vaut

$$V \times \left(1 - \frac{M}{N}\right).$$

La charge totale en données détenues par le cluster avant la perte des M machines vaut

$$N \times X.$$

Pour que les $N - M$ machines en vie restantes puissent supporter cette charge tout en garantissant un bon fonctionnement du cluster, il faut que l'inégalité suivante soit satisfaite

$$N \times X \leq V \times \left(1 - \frac{M}{N}\right).$$

Une résolution simple de cette inégalité par rapport à l'inconnue M nous donne la condition suivante :

$$M \leq N \times \left(1 - \frac{N \times X}{V}\right).$$

Cette condition peut encore s'exprimer sous la forme suivante

$$M \leq (1 - \alpha) \times N, \tag{1}$$

où $\alpha \in [0, 1]$ est le taux de remplissage de chaque machine, ce qui signifie que chaque machine contient une quantité de données

$$X = \alpha \times \left(\frac{V}{N}\right).$$

En se servant de la formule (1), nous estimons le nombre M dans quelques situations particulières.

1. Pour un cluster de $N = 16$ machines qui est rempli au taux $\alpha = 1/2$, le nombre limite de machines est $M = 8$.
2. Pour un cluster de $N = 16$ machines qui est rempli au taux $\alpha = 1/3$, le nombre limite de machines est $M = 11$.
3. Pour un cluster de $N = 64$ machines qui est rempli au taux $\alpha = 2/3$, le nombre limite de machines est $M = 22$.
4. Pour un cluster de $N = 64$ machines qui est rempli au taux $\alpha = 3/4$, le nombre limite de machines est $M = 16$.

7 Conclusion

Dans ce document, nous avons proposé quelques caratéristiques physiques que doivent avoir les machines dans un cluster Cassandra-Spark. Puis, à travers quelques exemples nous avons présenté les différentes étapes nécessaires pour le bon dimensionnement initial d'un cluster Cassandra tout en indiquant comment surveiller régulièrement la charge en données qu'il supporte afin de planifier certaines opérations de maintenance permettant d'éviter une perte quelconque des données.