

Tout ce qu'il faut savoir sur les bases de données SQL distribuée

Les bases de données SQL distribuées font beaucoup parler d'elles en ce moment et ce pour une bonne raison. Les bases de données relationnelles, auxquelles les entreprises du monde entier confient l'exécution d'applications critiques n'ont toutefois pas pu suivre les pionniers de l'engagement client et du commerce en ligne. Les bases de données non SQL, dont la capacité à supporter les applications Web modernes induisant des charges de travail massives a suscité un vif intérêt, sacrifient néanmoins l'homogénéité et l'intégrité des données. Les bases de données SQL distribuées réunissent le meilleur des deux mondes.

En effet, une base de données SQL distribuée est constituée de plusieurs instances de base de données interopérables (des nœuds dans un cluster), où chacune stocke et interroge non pas la totalité mais un sous-ensemble de données. Il suffit donc d'ajouter des instances pour augmenter la capacité de stockage et/ou les performances de requête. Et pour garantir une forte homogénéité, l'écriture de toutes les copies de données (répliques) est bien entendu synchronisée simultanément.

Voici quelques domaines pour lesquels les bases de données SQL distribuées sont une solution adaptée :

- Sites d'e-commerce : Une grosse opération commerciale ou une demande saisonnière peut générer un pic de fréquentation record sur un site d'e-commerce. Souvenez-vous des légions d'internautes qui se sont retrouvés face à une erreur 404 pendant la journée Amazon Prime Day de 2018 Un vrai cauchemar... En prévision de ces événements générateurs de pics d'affluence, vous devez vous assurer de disposer d'une capacité suffisante pour traiter la surfréquentation.
- Uber Eats ou GrubHub enregistrent l'essentiel de leurs commandes à l'heure des repas plutôt qu'au petit-déjeuner. Puisqu'il est notoire que la plupart des transactions se concentrent sur les créneaux 11h30-14h et 17h30-19h30, il n'y a pas lieu de fonctionner à capacité maximale à toute heure du jour et de la nuit.
- Les médias en streaming : Les entreprises de communication telles que Verizon et AT&T qui fournissent des services payants et gratuits de vidéo à la demande doivent mettre à jour d'énormes volumes de données pour que des millions de clients puissent recevoir en temps voulu les bons programmes, aux formats et aux tarifs convenus.
- Assurance maladie : Les inscriptions n'étant ouvertes que quelques semaines par an, c'est sur une période de 6 semaines que près de 8,3 millions de personnes s'inscrivent à l'assurance maladie sur healthcare.gov. Une forte capacité est indispensable pour traiter cette énorme affluence sur un nombre de semaines aussi réduit.

Comment les entreprises gèrent la montée en charge

Jusqu'à récemment, les options de montée en charge des bases de données relationnelles étaient peu nombreuses. La première, à savoir le redimensionnement des équipements, consistait à remplacer un serveur existant par un serveur de capacité supérieure. Et lorsque le plus puissant des serveurs matériels est devenu insuffisant, certaines entreprises ont opté pour des appliances matérielles du type [Oracle Exadata](#).

La seconde option, c'est-à-dire la montée en charge en partitionnant automatiquement les données (le « sharding »), a été adoptée par les géants de la technologie doués d'une expertise approfondie en ingénierie et par tous ceux qui ne voulaient pas, ou qui ne pouvaient pas, investir des millions de dollars dans des appliances matérielles. Certes plus économique que le redimensionnement, le partitionnement automatique des données est généralement dédié au support de requêtes et d'applications spécifiques, mais sa portée est limitée et il tend à être fragile et difficile à mettre à jour.

Les bases de données SQL distribuées offrent une évolutivité sans compromis pour toutes les applications, sans avoir à dépenser une fortune en appliances matérielles ni à subir les limites et le fardeau technique du partitionnement.

Les charges de travail des bases de données sont assez souvent variables. Les bases de données SQL distribuées sont conçues pour adapter la capacité aux besoins saisonniers et métiers. Elles vous permettent de fonctionner avec le minimum de ressources nécessaires en période calme et d'en ajouter rapidement pour gérer les pics d'activité. Cette élasticité est cruciale pour quantité d'applications.

Cette capacité d'adaptation immédiate à des conditions changeantes est une différence majeure par rapport à d'autres itérations SQL. Les bases SQL distribuées permettent d'ajouter des nœuds à la demande et de redimensionner de manière horizontale. Elles évitent aussi de devoir modifier votre clé de partition ou le mode de classement des données. Croyez-moi : personne n'a envie de passer sa vie à partitionner.

Certaines bases de données SQL distribuées peuvent aussi répartir les données entre plusieurs instances de base de données, en vérifiant constamment que la distribution des données est efficace et optimale. Face à des changements de comportements clients ou à de nouvelles applications, elles évaluent l'intérêt de transférer quelques données ici et là pour optimiser les performances.

Cette flexibilité atténue par ailleurs les problèmes d'indisponibilité et de coût. Par exemple, 100 instances peuvent être nécessaires pour atteindre vos conditions de trafic maximal. Mais si vous avez besoin qu'elles ne fonctionnent de concert que 1 % du temps seulement, vous n'êtes pas prêt à payer 24 heures sur 24, 7 jours sur 7. En fonction de la solution qu'ils utilisent, les clients peuvent redimensionner à la hausse ou à la baisse, heure par heure, pour comprimer au maximum leurs frais. C'est aussi simple à faire dans le Cloud que d'appuyer sur un bouton pour augmenter la capacité ou la réduire par mesure d'économie lorsque le trafic s'équilibre.

De plus, une distribution homogène des données est assurée chaque fois que des nœuds sont ajoutés et supprimés, ou lorsque des données viennent s'ajouter de manière inégale. Cette approche de la distribution automatique des données évite de devoir gérer des partitions et permet de maintenir la disponibilité des données en cas de perte de nœuds.

Les bases de données SQL distribuées sont conçues selon le principe que plus les instances sont nombreuses dans une base de données distribuée, plus le risque que certaines tombent en panne est élevé. C'est un fait. Ainsi, parce qu'elles sont distribuées et répliquées sur une poignée d'autres nœuds, les données restent disponibles en cas de panne d'infrastructure. La base de données suppose que des instances vont et viennent sans préavis et gère automatiquement les changements de topologie, ce qui est un gros avantage pour les entreprises qui n'ont donc plus à craindre ni les coûts, ni une perte de chiffre d'affaires, ni une situation embarrassante ni une atteinte à la réputation en raison d'une indisponibilité.

La fin des compromis

Les équipes chargées des bases de données ont besoin de données durables, sécurisées, homogènes et que les bases de données elles-même soient constamment disponibles et évolutives. C'est beaucoup demander, certes, mais c'est d'une importance capitale pour la réussite de toute entreprise.

Jusqu'à récemment, peu de bases de données garantissaient un tel niveau de fonctionnalité. L'évolutivité s'obtenait au prix de beaucoup de contraintes, de restrictions et de compromis, en renonçant notamment à l'intégrité des données et des transactions. Ces contraintes étaient notamment liées au nombre de défis techniques en jeu pour exécuter des transactions sur de nombreuses instances de base de données en toute sécurité (voir par exemple la conformité ACID – Atomicité, Cohérence, Isolation et Durabilité). Les bases de données SQL distribuées marquent la fin de ces compromis.

L'avenir s'annonce radieux

Les bases de données SQL distribuées continuent d'évoluer, répondant aux demandes d'un marché qui exige des données plus fiables, transparentes, précises, fournies à une vitesse toujours plus élevée et à une échelle véritablement mondiale. Cette toute nouvelle technologie conçue pour le Cloud est par définition native Cloud. Et les communautés de la technologie et des affaires conviendront que la migration vers le Cloud est fondamentalement un train en marche, comme le démontre [une récente enquête auprès de décideurs informatiques](#).

En ce début 2021, je vois qu'à l'avenir les bases de données SQL distribuées s'étendront sur plusieurs centres de données et deviendront transparentes pour les applications, partout, et que ce soit en écriture ou en lecture. Pas uniquement dans un centre de données, mais à une échelle véritablement mondiale. Telle sera la prochaine phase de maturité. L'avenir des bases de données SQL distribuées s'annonce donc très prometteur et nous n'avons pas fini d'en entendre parler.