Projet: Conception et Déploiement d'un Système de Bases de Données Réparties en Production

Auteur : Badr TAJINI - Course Distributed databases (Project) - EFREI - 2023/2024

Contexte :

Les entreprises modernes opérant dans des domaines tels que la finance, la santé, le commerce en ligne, et les services publics, recueillent et gèrent d'énormes volumes de données. Un accès rapide, sécurisé et fiable à ces données géographiquement réparties est une condition essentielle à la prise de décision, à l'efficacité opérationnelle et à l'amélioration de l'expérience des clients.

Objectif du Projet :

Concevoir et déployer une solution de bases de données réparties utilisant Oracle ou PostgreSQL avec Citus pour un domaine d'activité choisi. La solution doit optimiser la performance des requêtes, assurer la cohérence des données, et être conforme aux normes de sécurité et de confidentialité des données, telles que le RGPD.

Tâches :

1. **Choix du Secteur et Analyse des Besoins :**
    - Sélectionnez un secteur d'activité pertinent et analysez les besoins spécifiques en matière de données.
    - Identifiez les exigences en termes de volume de données, de fréquence d'accès, de répartition géographique des utilisateurs et de contraintes réglementaires.
2. **Conception de l'Architecture de la Base de Données :**
    - Déterminez les modalités de fragmentation verticale, horizontale ou mixte adaptées aux cas d'utilisation du secteur.
    - Planifiez la répartition des nœuds Citus sur différents data centers ou régions cloud pour une distribution géographique.
3. **Implémentation et Configuration :**

- Mettez en place l'infrastructure avec Oracle ou PostgreSQL et Citus, en configurant les nœuds et les règles de distribution des données.
- **Question de recherche (PostgreSQL)** : Étude des options de réplication disponibles dans Azure Database for PostgreSQL Hyperscale (Citus) :
  - Quelles sont les stratégies de réplication intégrées fournies par Azure pour PostgreSQL Hyperscale?
  - Comment ces stratégies peuvent-elles être personnalisées pour répondre à des scénarios de haute disponibilité spécifiques?

4. **Développement de Fonctionnalités Spécifiques :**

- Créez des vues, procédures stockées et triggers qui correspondent aux besoins métier du secteur.
- **Question de recherche (PostgreSQL)** : Étude des mécanismes de sécurité d'Azure pour PostgreSQL et Citus :
  - Quelles sont les pratiques de sécurité par défaut fournies par Azure pour les bases de données PostgreSQL avec Citus ?
  - Comment ces pratiques peuvent-elles être évaluées ou améliorées pour garantir la conformité avec le RGPD ?

5. **Optimisation et Tests :**

- Optimisez les performances des requêtes et réduisez la latence en utilisant des outils de profilage et de benchmarking.
- Exécutez des tests de charge pour évaluer la capacité du système à gérer de gros volumes de transactions simultanées.
- *Exemples* :
  - **Optimisation des Performances des Requêtes** :
    - Utilisez l'outil `EXPLAIN` pour analyser les plans d'exécution des requêtes et identifier les goulots d'étranglement.
    - Expérimentez avec différentes stratégies de clé de distribution pour équilibrer la charge entre les nœuds dans Citus.
    - Utilisez des index, des jointures parallèles, et optimisez les requêtes pour tirer parti de la distribution des données.
  - **Utilisation d'Outils de Profilage et de Benchmarking** :

        - Utilisez `pg_stat_statements` pour suivre et identifier les requêtes les plus coûteuses en - termes de ressources.
        - Employez `pgBench` pour simuler une charge de travail et obtenir des mesures de performance.

6. **Documentation :**
    - Rédigez une documentation technique (**Scripts SQL**) complète pour l'administration du système et la gestion des données.

Calendrier (Deadline)

- Date limite : **24/05/2024 (23h59 heure Paris).**

Livrables et Livraison:

- **Équipes** :
    - Former des équipes de deux personnes maximum, afin de favoriser la collaboration et la diversité des compétences. Attribuer des étapes et des rôles spécifiques à chaque étudiant de l'équipe. La réussite du projet repose sur une communication efficace et une bonne répartition des tâches au sein de l'équipe.
    - Pour remédier au nombre impair de la cohorte, il est possible d'accepter une seule équipe de trois personnes.
- **Méthode de livraison** :
    - Un rapport détaillé illustrant l'analyse des besoins, la conception de l'architecture, le déploiement, les fonctionnalités et les bénéfices du système pour le secteur choisi.
    - Scripts de création et de configuration de la base de données.
        - Fichier PDF : **projet_bdr2023_NAME1_NAME2.zip**

Critères d'Évaluation :

- Pertinence et profondeur de l'analyse des besoins du secteur.
- Solidité de l'architecture de la base de données et de la stratégie de distribution des données.
- **(Recherche)** - Efficacité des mécanismes de sécurité et de conformité avec le RGPD.

Conditions à Respecter :

1. Utilisation de PostgreSQL avec Citus ou Oracle comme système de gestion de base de données réparties.
2. **(Recherche ouverte)** - Le projet doit inclure une stratégie détaillée pour la gestion des données conformément au RGPD (Annexe pour plus d'informations).
3. **Pondération plus élevée** si la solution est hébergée sur le cloud. Cette solution permet de bénéficier des fonctionnalités d'Azure Database for PostgreSQL Hyperscale (Citus).

Annexe

Pour la deuxième question "**Conditions à Respecter**", il est recommandé de se référer à cette approche, illustrée ci-dessous, qui reflète une méthodologie appliquée dans des environnements de production. Cette approche vise à garantir une gestion sécurisée, conforme et résiliente des données personnelles au sein d'un système de base de données distribuées. L'objectif sous-jacent est de stimuler une réflexion approfondie et d'encourager l'application d'hypothèses pertinentes dans le cadre d'une future gestion exhaustive de vos projets en tenant compte des exigences du Règlement Général sur la Protection des Données (RGPD).

Stratégie pour la gestion des données conformément au RGPD :

1. **Évaluation des Risques et Identification des Données :**
    - **Analyse des Types de Données :** Examiner et classer les différents types de données personnelles gérées.
    - **Évaluation des Risques :** Analyser les risques associés à chaque type de donnée pour s'assurer de la conformité au RGPD.
2. **Techniques de Sécurisation des Données :**
    - **Chiffrement et Anonymisation :** Rechercher et implémenter les meilleures méthodes pour sécuriser les données, incluant le chiffrement et l’anonymisation.
    - **Méthodes de Contrôle d'Accès :** Évaluer et appliquer des modèles de contrôle d'accès pour protéger les accès aux données sensibles.
3. **Gestion du Consentement et des Droits des Utilisateurs :**
    - **Systèmes de Gestion du Consentement :** Intégrer des systèmes pour gérer efficacement le consentement des utilisateurs.
    - **Implémentation des Droits des Sujets de Données :** Faciliter la gestion des droits des utilisateurs, comme l'accès, la rectification et la suppression de leurs données.
4. **Architectures de Haute Disponibilité et de Récupération :**

- **Stratégies de Backup :** Développer des stratégies pour la sauvegarde et la restauration des données en conformité avec le RGPD.
- **Plans de Continuité des Activités :** Élaborer des plans pour assurer la continuité des opérations et la récupération rapide après des sinistres.

5. **Audit, Surveillance et Conformité :**

   - **Outils de Surveillance :** Identifier et mettre en œuvre des outils pour surveiller et alerter en cas de violations de données.
   - **Protocoles d'Audit :** Créer des procédures d'audit pour vérifier régulièrement la conformité au RGPD.

D'autres ressources supplémentaires :

- https://learn.microsoft.com/fr-fr/azure/cosmos-db/postgresql/tutorial-shard
- https://learn.microsoft.com/fr-fr/azure/cosmos-db/postgresql/tutorial-design-database-multi-tenant