Offre de stage Master 2 - CIMI

Laboratoire / Université : IRIT/ Université Paul Sabatier Toulouse III

Equipe de recherche : PYRAMIDE

Encadrant : Riad Mokadem (riad.mokadem@irit.fr)

Titre du stage : Conception et implémentation d'une stratégie de réplication de données basée sur l'apprentissage par renforcement dans les systèmes Cloud

1. Contexte et problématique

La popularité croissante des services et applications interconnectés (par exemple Internet des objets et les réseaux sociaux) ont conduit à la génération de gros volumes de données. Un des défis pour les applications est de pouvoir stocker et analyser ces données hétérogènes et réparties avec des coûts raisonnables d'infrastructure. Dans ce contexte, l'approche «Cloud Computing» permet de réduire considérablement ces coûts, soit en se basant sur des serveurs composés de machines à bas prix (Clouds privés), soit en louant des services auprès de fournisseurs Cloud suivant le modèle « pay-as-you-go » (Clouds publics). Pour les applications analysant ces données, les problèmes d'accès et de disponibilité de données sont très importants. Une technique bien connue pour traiter ces problèmes est la réplication de données qui consiste à stocker plusieurs copies de données, appelées répliques, sur plusieurs sites. Elle vise à : (i) augmenter la disponibilité des données, (ii) réduire la consommation de la bande passante et (iii) gérer efficacement la tolérance aux pannes [1]. De nombreuses stratégies de réplication de données ont été proposées dans les environnements cloud. Elles visent à obtenir les meilleures performances du système tout en satisfaisant un contrat de niveau de service (SLA), établi entre un fournisseur de cloud et ses locataires, i.e., les consommateurs. Principalement, un SLA comprend des objectifs de niveau de service (SLO) du locataire, par exemple, la disponibilité et la performance, qui doivent être satisfaits par le fournisseur. D'un autre coté, le fournisseur Cloud vise à maximiser son profit économique [2]. Il est alors important d'ajuster le nombre de répliques de manière dynamique afin de prendre en compte la rentabilité du fournisseur.

Afin d'assurer le dimensionnement automatique des ressources, de nombreux fournisseurs de Cloud se basent sur la réplication de données basée sur des seuils à cause de sa nature intuitive. A titre d'exemple, un seuil de temps de réponse, intégré dans le SLA, est préalablement négocié entre le fournisseur et ses locataires. Dans ce contexte, certains travaux se basent sur l'observation des valeurs de métriques afin de les comparer par la suite à des seuils fixés d'avance [1]. D'autres travaux [3] combinent l'approche des seuils avec la théorie de contrôle permettant l'obtention de seuils dynamiques en se basant sur une modélisation mathématique de la charge de travail. Enfin, certains travaux se basent sur la prédiction des valeurs de métriques tels que le score de réplication par intervalle [4] ou encore la charge de travail [5] afin de les comparer à des seuils prédéfinis. Cette prédiction s'appuie sur l'utilisation de techniques telles que les séries chronologiques ou encore sur l'exploitation du journal de requêtes afin de prédire les périodes à forte charge de travail et les données qui seront les plus populaires dans le futur [6]. En conséquence, des ressources peuvent être allouées à l'avance, par exemple la création de nouvelles répliques. Cependant, le choix des métriques à considérer et la fixation de seuils de manière efficace nécessite une intervention humaine afin de fixer le seuil pour chaque métrique et une connaissance approfondie des tendances actuelles de la charge de travail, ce qui n'est pas facile à réaliser.

2. Objectifs et résultats attendus

Afin d'éviter l'intervention humaine lors de la définition des seuils, nous pourrons considérer une réplication de données basée sur l'apprentissage par renforcement [7]. Dans les algorithmes d'apprentissage par renforcement tel que le Q-learning, un agent autonome dispose d'un certain nombre

d'actions possibles permettant le changement de l'état d'un environnement. Il reçoit alors une récompense (ou une pénalité) pour chacune de ses actions. Ensuite, cet agent doit mémoriser la séquence des actions qui maximise sa récompense totale. Néanmoins, cette approche nécessite une période d'apprentissage.

Seuls quelques travaux de dimensionnement automatique basés sur l'apprentissage par renforcement dans le Cloud sont dédiés à l'interrogation de bases de données relationnelles. La plupart se sont intéressé aux systèmes NoSQL [8]. Les méthodes existantes doivent alors être adaptées au contexte des bases de données relationnelles avec notamment, la prise en compte de nombreuses tâches dépendantes et des relations intermédiaires qui peuvent être stockées sur le disque.

L'objectif de ce stage est la conception d'une stratégie de réplication de données efficace basée sur l'apprentissage par renforcement. La stratégie proposée pourra s'appuyer sur un agent informatique qui pourra mémoriser certaines actions lui permettant de privilégier la création rentable (pour le fournisseur) d'une réplique d'une relation, tout en satisfaisant les objectifs des locataires. Il est donc important de proposer, puis d'implémenter via simulation [9], une stratégie de réplication permettant de répondre aux problématiques classiques telles que : (i) quelles données répliquer ? (ii) quand répliquer ces données ? (iii) où répliquer ces données mais aussi à des problématiques spécifiques aux environnements Cloud tels que (iv) déterminer le nombre de répliques nécessaires afin de satisfaire simultanément les objectifs du locataire, i.e., objectifs SLO, avec un profit économique pour le fournisseur de Cloud.

3. Mots clés

Gestion de données, Systèmes Cloud, Réplication de données, Apprentissage par renforcement, Modèle de coûts, Modèle économique, Performances.

4. Bibliographie

[1]. R. Mokadem, A. Hameurlain. A Data Replication Strategy with Tenant Performance and Provider Economic Profit Guarantees in Cloud Data Centers. Journal of Systems and Software (JSS), Elsevier, V. 159, (2020).

[2] Armbrust, M., Stoica, I., Zaharia, M., Fox, A., Grifh, R., Joseph, A.D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin. A.: A view of cloud computing. Communications of the ACM. V. 53(4). pp. 50-58, (2010)

[3] H. Ghanbari, B. Simmons, M. Litoiu, G. Iszlai. Exploring alternative approaches to implement an elasticity policy. IEEE Int. Conf. on Cloud Computing (CLOUD), pp. 716–723. (2011)

[4] Li, C., Wang, Y., Chen, Y., and Luo, Y. Energy efficient fault-tolerant replica management policy with deadline and budget constraints in edge-cloud environment. Journal of Network and Computer Applications, V. 143 : pp. 152–166, (2019)

[5] Hsu, T.-Y. and Kshemkalyani, A. D. A Proactive, Cost-aware, Optimized Data Replication Strategy in Geodistributed Cloud Datastores. In Proceedings of the 12th IEEE/ACM International Conference on Utility and Cloud Computing, UCC'19, pp. 143–153, New York, NY, USA (2019)

[6] Liu, J., Shen, H., Narman, H. S., Lin, Z., and Li, Z. Popularity-aware Multi-failure Resilient and Cost-effective Replication for High Data Durability in Cloud Storage. IEEE Transactions on Parallel and Distributed Systems, V. 30 (10), pp. 2355-2369, (2018).

[7] L. Ferreira, F. Coelho, J. Pereira. Self-tunable DBMS Replication with Reinforcement Learning. Remke A., Schiavoni V. (eds) Distributed Applications and Interoperable Systems. DAIS 2020. Lecture Notes in Computer Science, V. 12135. Springer, Cham, pp. 131-145, (2020)

[8] A. Naskos, A. Gounaris, I. Konstantinou. Elton: a cloud resource scaling-out manager for nosql databases. 34th IEEE Int. Conf. on Data Engineering (ICDE), IEEE, pp.1641–1644. (2018)

[9] R.N. Calheiros, R. Ranjan, A. Beloglazov, C.A.F. De Rose, R. Buyya. CloudSim: A Toolkit for Modeling and Simulation of Cloud Computing Environments and Evaluation of Resource Provisioning Algorithms. Software: Practice and Experience. V. 41, N. 1, pp. 23-50. (2010)