Stage M2R CEDRIC

Approche dirigée par les modèles pour la démoralisation de schéma NoSQL

Faten Atigui (ISID), Nicolas Travers (Vertigo, membre associé)

Mots-clés: Diagramme de classes UML; imbrication, éclatement de schémas; systèmes NoSQL

Description :

Les systèmes d'information doivent faire face à une quantité toujours plus grande de données, et cherche à prendre en compte toutes les dimensions de leur éco-système afin de répondre aux exigences du métier. Ce volume toujours plus grand, complexe et dynamique (connu comme les 3V) a mis à mal les techniques traditionnelles de bases de données relationnelles et les entrepôts de données. Ainsi, pour des problèmes de passage à l'échelle, les bases de données NoSQL (HBase[1], Cassandra[2], MongoDB[3], Néo4J[4], etc.) ont vu le jour depuis une dizaine d'années et tentent de répondre à ces besoins. De nouvelles solutions sont proposées chaque année en vue de cibler une optimisation particulière, toutefois, ces fonctionnalités restent ad-hoc.

En conséquence, le choix de la bonne solution NoSQL en fonction des besoins métiers est fondamental pour le système d'information. Il peut avoir d'énormes impacts sur le passage à l'échelle et la pérennité de la solution. Ce choix implique une connaissance précise du besoin, en matière de volumes et dynamicité des données, de diversité d'interrogations et de contraintes sur le système. De plus, être capable de faire la corrélation entre les besoins et les solutions demande une réelle expertise sur le marché de la *Data*, ce qui très souvent s'avère donner une orientation commerciale plutôt que qualitative.

L'objectif de nos travaux de recherche est donc de proposer une approche d'aide au choix d'orientation technologique et de conception d'un système d'information, en reposant sur une méthodologie de modélisation des données, simulation de distribution des données et un modèle de coût adaptatif (pour intégrer de nouvelles fonctionnalités NoSQL). Nous nous intéressons à la fois à un niveau d'abstraction pour la modélisation de SI, mais également à un niveau physique pour favoriser l'optimisation du système. Ces deux approches traditionnellement séparées (architecture ANSI-SPARC) se doivent d'interagir finement pour s'adapter à un contexte d'optimisation extrêmement contraint et complexe.

---

[1] https://hbase.apache.org/

[2] http://cassandra.apache.org/

[3] https://mongodb.com

[4] https://neo4j.com/

Ainsi, dans le cadre d'un stage financé par le CEDRIC en 2018 et effectué par Asma Mokrani, étudiante en M2R Système d'information et Business intelligence au CNAM, nous avons réussi à (i) étudier l'état de l'art, (ii) formaliser notre approche globale, (iii) proposer un protocole d'expérimentation et des tests en utilisant TPC-C. Ces premiers résultats ont été présentés lors d'un workshop franco-russe autour des big data qui a eu lieu le 25 & le 26 octobre 2018 à Paris [4]. D'autre part, avec Asma Mokrani, nous sommes également en train de finaliser un article à soumettre à la revue Ingénierie des systèmes d'information (ISI). Dans la continuité de cette thématique, nous cherchons à élargir le domaine en vue de déposer un projet de recherche permettant à terme, de financer une thèse.

Sujet du stage :

Le stage débutera par l'étude des méthodes de dénormalisation de schéma conceptuels (entités/associations ou diagrammes de classes UML) pour le NoSQL de la littérature [1,2,3,4], ainsi que les différentes solutions NoSQL existantes [5,6,7] pour comprendre les fonctionnalités spécifiques de chacune. Ensuite, l'objectif de ce stage est de :

1. Compléter l'étude de l'état de l'art proposé par [8] ;
2. Proposer une approche dirigée par les modèles permettant de guider le choix du modèle logique et du système NoSQL en se basant principalement sur le modèle conceptuel (diagramme de classes) et les besoins de l'utilisateur formalisés sous forme de requêtes SQL. L'idée est d'étudier les types d'associations spécifiques à UML : agrégation, composition, héritage, etc. ainsi que leur impact sur l'imbrication ou l'éclatement de schémas ;
3. Comparer les résultats issus des recommandations basées sur les modèles conceptuels aux résultats de recommandation basés sur les tests et les expérimentations [8]. Les expérimentations seront testées principalement sur la base MongoDB, mais une ouverture sur HBase est envisagée ;
4. Automatiser le processus de transformation du schéma conceptuel vers le schéma logique et puis en schéma physique.

La finalité est de proposer des orientations d'implémentation pondérées, tout en donnant les avantages et les inconvénients de chaque solution envisagée.

Candidature

Envoyez votre CV, notes de Master, lettre de motivation, lettres de recommandation à faten.atigui@cnam.fr et nicolas.travers@devinci.fr

Profil

Etudiant(e) de Master 2 ou de dernière année d'école d'ingénieur sur un cursus d'informatique
Bon niveau en informatique et plus précisément en systèmes d'information et bases de données, systèmes NoSQL.
Bon niveau de communication scientifique à l'écrit et oral, notamment en anglais

Compétences attendues

L'étudiant retenu devra présenter de bonnes compétences dans le domaine des systèmes de gestion de données massives, l'analyse des données. Elle ou il devra avoir un très bon cursus universitaire et une forte motivation pour la recherche afin de permettre une éventuelle poursuite en thèse à l'issue du stage.

Début et durée du stage : 5 mois à partir de début mars.

Bibliographie

[1] Abdelhedi F, Ait Brahim A, Atigui, Gilles Zurfluh. MDA-Based Approach for NoSQL Databases Modelling. DaWaK 2017: 88-102.

[2] A. Chebotko, A. Kashlev, S. Lu. A Big Data Modeling Methodology for Apache Cassandra. IEEE International Congress on Big Data (BigData Congress). 2015

[3] Robert T. Mason. NoSQL Databases and Data Modeling Techniques for a Document-oriented NoSQL Database. Proceedings of Informing Science & IT Education Conference (InSITE) 2015, 259-268.

[4] Nicolas Travers. Big Data Modelization: From Conception To Optimization. Invited talk - Russian-French Workshop on Big data Applications. 26/10/2018, Paris, France.

[5] RAUT, A. B. NOSQL Database and Its Comparison with RDBMS. International Journal of Computational Intelligence Research, 2017, vol. 13, no 7, p. 1645-1651.

[6] R. Fournier S'niehotta, P. Rigaux, N. Travers. Bases de données documentaires et distribuées. CNAM, 2017 http://b3d.bdpedia.fr/

[7] N. Travers. Maîtrisez les bases de données NoSQL. OpenClassrooms, 2017

[8] A. Mokrani, Approche de dénormalisation de schémas pour implantation NoSQL, CEDRIC, CNAM septembre 2018.