PLAN DE COURS
Automne 2015

UQAC

8INF803 • Bases de données réparties (Gr 1)

Enseignant:	Edmond La Chance
Bureau:	P4-6570
Courriel:	edmond.lachance@gmail.com
Page web du cours:	https://github.com/edmondlachance/8INF803

Contenu général

L'utilisation simultanée d'un grand nombre de serveurs de bases de données, reliés par des connexions réseau à haut débit, rend maintenant possible le traitement de requêtes autrement hors de portée d'un serveur unique. Par exemple, le moteur de recherche Google tire ses performances de sa capacité à indexer un très grand nombre de pages web à travers un réseau de nœuds, chacun possédant une portion de la base de données complète. Sa fiabilité tient au fait qu'un même élément d'information se retrouve dans plusieurs de ces nœuds, diminuant les effets d'une panne de l'un d'entre eux. Cependant, cette décentralisation apporte aussi son lot de défis. Par exemple, la mise à jour d'un élément doit s'effectuer avec succès sur chacune des copies. De la même manière, la fragmentation de la base de données doit minimiser les communications inter-nœuds.

Ce cours se veut une introduction aux concepts fondamentaux des bases de données réparties. On y discutera de l'architecture générale d'un système distribué, des techniques de fragmentation des bases de données et de décomposition de requêtes. On verra ensuite comment un tel système peut gérer des transactions atomiques semblables à celles d'un système centralisé, et comment les éventuelles pannes peuvent être récupérées. Ensuite, le cours placera une attention particulière à des technologies modernes utilisées dans de nombreux systèmes actuels, tels les tables de hachage distribuées, l'algorithme MapReduce et le traitement de requêtes XML. Finalement, nous verrons aussi l'architecture Lambda. Une alternative aux SGBD incrémentaux. L'architecture Lambda a un fonctionnement très différent, un peu plus semblable à git qu'à un SGBD conventionnel.

Objectifs du cours

Au terme de ce cours, l'étudiant aura acquis les compétences suivantes:

- Définir les concepts importants relatifs aux bases de données réparties
- Appliquer les techniques de fragmentation horizontale et verticale pour transformer un modèle relationnel classique en un modèle distribué et justifier ses choix
- Décomposer une requête classique en un plan de requêtes distribué équivalent, et appliquer des techniques d'optimisation de requêtes
- Expliquer les concepts fondamentaux de l'optimisation de requêtes
- Décrire le fonctionnement et utiliser les fonctions de base des technologies modernes de bases de données réparties

- Modéliser et interroger des sources de données utilisant des modèles non relationnels, telles les bases de données en graphes, en arbres, et le traitement parallèle MapReduce.
- Comprendre le fonctionnement de l'architecture Lambda pour construire des applications Bigdata. Le scaling horizontal versus le scaling vertical.

Sujets abordés

Le cours se divise en deux grandes parties.

1. La première moitié du cours porte sur les bases de données relationnelles réparties: après quelques rappels sur les bases de données relationnelles centralisées, on y présentera les propblèmes propres au cas distribué (fragmentation des relations, allocation des fragments, exécution et optimisation des requêtes).
2. La seconde moitié est composée de séances indépendantes présentant différentes alternatives au modèle relationnel, par exemple: les bases de données en graphes, les bases de données XML, les tables de hachage distribuées et l'algorithme de traitement MapReduce. Pour finir, nous parlerons de l'architecture Lambda. Une architecture qui permet de construire des applications BigData.

Évaluation

L'évaluation consistera en les éléments suivants:

- Deux travaux à faire en équipes de deux à quatre personnes, à remettre les **20 octobre** et **15 décembre**. Les travaux seront composés de questions à développement, ainsi que d'exercices de programmation faisant appel aux technologies et aux concepts présentés en classe. L'usage de documentation externe (manuel du cours, ressources sur le web) est encouragé, en autant que les étudiants répondent par eux-mêmes aux questions. Chacun des travaux compte pour 25 points. Tout travail remis en retard sans motif valable sera pénalisé de 10% par jour de retard.
- Un examen intra à la séance du **20 octobre**. Cet examen compte pour 25 points.
- Un examen final à la séance du **8 décembre**. Cet examen compte pour 25 points.

La note finale, sur 100, est la somme de toutes ces évaluations. La note de passage est fixée à **60%** ou D.

Date limite d'abandon sans mention d'échec

20% de l'évaluation aura été transmise à l'étudiant avant la date limite d'abandon sans mention d'échec, soit le Lundi 9 novembre.

Qualité du français écrit

Tout travail remis doit être conforme aux exigences de la politique institutionnelle en matière de maitrise du français écrit. Comme il en est fait mention dans le Manuel de gestion (3.1.1-012).[1] Tout travail dont la qualité du français serait jugée non conforme par l'enseignant pourra être pénalisé jusqu'à concurrence de 10% du résultat maximal prévu.

Évaluation de la qualité de l'enseignement

Ce cours sera évalué en fonction de la Procédure relative à l'évaluation des activités aux programmes d'études de cycles supérieurs (Manuel de gestion, 3.1.2-008).

Formule pédagogique

Le cours sera dispensé en classe par un professeur. Il n'y aura pas de séance de travaux pratiques associées à ce cours.

Situation du cours dans le programme

Le cours est optionnel pour les étudiants inscrits aux programmes de cycles supérieurs. Aucun cours ne lui est préalable.

Références

Aucun livre ou recueil de notes n'est obligatoire pour ce cours. Les diapositives utilisées durant les séances et les lectures préalables seront rendues disponibles en format électronique sur le site du cours. Le cours suivra d'assez près le contenu du manuel suivant:

- M.T. Özsu, P. Valduriez. (2011). Principles of Distributed Database Systems, 3rd edition. Springer, ISBN 0-13-659707-6.

Et ce livre pour l'architecture Lambda.

- Nathan Marz (2015). Big Data: Principles and best practices of scalable realtime data systems

Les références suivantes sont suggérées pour des compléments à la matière vue en classe. Des articles de journaux et de conférences scientifiques seront également au programme; les références seront publiées sur le site du cours.

- I. Robinson, J. Webber, E. Eifrem. (2013). Graph Databases. O'Reilly, ISBN 978-1-449-35626-2.

[1] http://www.uqac.ca/direction_services/secretariat_general/manuel/index.pdf

- S. Abiteboul, O. Benjelloun, T. Milo. (2008). The Active XML project: an overview. *The VLDB Journal*, 17, 1019-1040.
- J. Lin, C. Dyer. (2010). Data-Intensive Text Processing with MapReduce. Morgan and Claypool, ISBN 1-60-845342-1.
- T. White. (2012). Hadoop: The Definitive Guide. O'Reilly, ISBN 978-1-449-31152-0.
- S. Ceri, G. Pelagatti. (1984). Distributed Databases: Principles and Systems. McGraw-Hill, ISBN 0-07-010829-3.