NoSQL, Insertion de masse sur une architecture distribuée

Travail de Bachelor réalisé en vue de l’obtention du Bachelor HES

par :

Simon DI RELLA

Conseiller au travail de Bachelor :

Alexandros KALOUSIS, Directeur de travail

Genève, le < date de dépôt >

Haute École de Gestion de Genève (HEG-GE)

Filière Informatique de Gestion

Déclaration

Ce travail de Bachelor est réalisé dans le cadre de l’examen final de la Haute école de gestion de Genève, en vue de l’obtention du titre d’Informaticien de Gestion.

L’étudiant atteste que son travail a été vérifié par un logiciel de détection de plagiat.

L’étudiant accepte, le cas échéant, la clause de confidentialité. L'utilisation des conclusions et recommandations formulées dans le travail de Bachelor, sans préjuger de leur valeur, n'engage ni la responsabilité de l'auteur, ni celle du conseiller au travail de Bachelor, du juré et de la HEG.

« J’atteste avoir réalisé seul le présent travail, sans avoir utilisé des sources autres que celles citées dans la bibliographie. »

Fait à Genève, le < date >

Simon Di Rella

< Signez la déclaration ici >

Remerciements

Je remercie mon directeur de travail « Alexandros Kalousis » pour m’avoir suivi durant la rédaction et le développement de ce travail de Bachelor.

J’en profite également pour remercier la Haute Ecole de Gestion pour la qualité des modules proposés, ainsi qu’à la facilité à l’accès à l’enseignement, tant financier qu’aux horaires proposés pour les étudiants qui sont en emploi.

Résumé

Ce travail est ciblé sur les bases de données dîtes « NoSQL », il sera présenté en deux parties.

La première sera un travail écrit sur divers sujets des bases de données « NoSQL ». Il traitera sur l’apparition de cette nouvelle technologie, ces différences avec une base de données relationnelle traditionnelle, les diverses conceptions et points techniques, son architecture et la spécification de deux bases de données « NoSQL » qui sont « MongoDB » et « CouchDB ».

Cette spécification me permettra d’aborder la deuxième partie, qui est la mise en place de ces bases de données sur un environnement clustérisé afin d’obtenir une architecture distribuée. Elle sera complétée par un développement dans le langage de programmation « Java », qui permettra d’insérer en masse des données mockées, afin d’observer le comportement de ces deux bases.

Table des matières

[Déclaration i](#_Toc381706995)

[Remerciements ii](#_Toc381706996)

[Résumé iii](#_Toc381706997)

[Liste des tableaux iii](#_Toc381706998)

[Liste des figures iii](#_Toc381706999)

[1. Introduction 3](#_Toc381707000)

[2. Titre du premier chapitre 3](#_Toc381707001)

[3. Titre du second chapitre 3](#_Toc381707002)

[3.1 Titre de second niveau 3](#_Toc381707003)

[3.2 Titre de second niveau 3](#_Toc381707004)

[3.2.1 Titre de troisième niveau 3](#_Toc381707005)

[3.2.2 Titre de troisième niveau 3](#_Toc381707006)

[3.2.2.1 Titre de quatrième niveau 3](#_Toc381707007)

[3.2.2.1.1 Titre de cinquième niveau 3](#_Toc381707008)

[4. Conclusion 3](#_Toc381707009)

[Bibliographie 3](#_Toc381707010)

[Annexe 1 : Titre de l’annexe 3](#_Toc381707011)

Liste des tableaux

[Tableau 1 : Titre du tableau 3](#_Toc381707012)

Liste des figures

[Figure 1 : Titre de la figure 3](file:///Y:\BODER\1-Cours\160-travail_de_bachelor\Modeles_TDB_HEG_2014.dotx#_Toc381707014)

# Introduction

Lorsque nous pensons « **Base de donnée** », nous pensons infailliblement à notre vieille base de données basée sur un modèle relationnel, incontournable depuis la fin des années 80.

Après 30 ans de domination, une nouvelle génération de base de données commence à pointer le bout de son nez. Suite à une convergence de divers points comme la baisse des prix, la hausse des performances des machines et de la bande passante, certaines entreprises ayant des besoins spécifiques se sont lancées dans la création de base de données répondant à leurs besoins. Souvent pour faciliter la montée en charge ou la recherche rapide d’informations.

Lors d’un événement consacré à l’architecture distribuée en 2009, le terme « **NoSQL** » fut inventé. Il est une abréviation de « **Not Only SQL** ». On pourrait polémiquer sur ce terme, mais il a l’avantage de faire parler de lui. En réalité, ces nouvelles bases de données ne se séparent pas du langage SQL, puisque certaines d’entre elles utilisent une variante pour accéder aux données. On pourrait dire qu’elles sont non relationnels, et donc, que certaines données sont redondantes.

En résumé, le « **NoSQL** » n’est ni une régression, ni une évolution du modèle traditionnel, mais tout simplement une nouvelle conception adaptée à certain besoins spécifiques. Comme l’on pourrait choisir une voiture, rapide mais inconfortable, à une spacieuse mais peu réactive, notre choix se fera pour combler au maximum les besoins du moment.

# Base de données relationnelle

## Histoire

Le concept de base de données relationnelle fut inventé par « **Edgard Franck Codd** », En 1948, il commença à travailler pour IBM en tant que programmeur. Il obtint un doctorat en Computer Science qui lui permit de travailler par la suite comme chercheur aux laboratoires de recherche de San José. En 1969, il publia un article qui posa les fondements du modèle relationnel, ainsi que les bases mathématiques et algébriques de relations.

En 1974, le laboratoire de San Jose se lança dans le développement d’un prototype expérimental, appelé « **System R** », basé sur les concepts de « **Edgard Franck Codd** ». Afin de pouvoir manipuler les données, ils développèrent un langage nommé « **Structured English Query Language** **–** **Sequel** », qui fut changé en « **SQL** » dû au terme déjà déposée par une compagnie britannique pour désigner une marque.

Les fondements des bases de données relationnelles étaient posés. Suite aux bonnes performances du « System R », Oracle, DB2 et les descendants d’Ingres implémentèrent ce concept et rendirent le rendirent plus populaire.

## Les principes

Un des principes est d’établir un modèle qui permet de séparer clairement la représentation logique des données, de son organisation physique. La représentation logique des données permet de modéliser un ensemble de données par l’utilisation de tables et de leurs relations, ce qui force l’utilisateur à se plier à ces concepts. Toutefois, le concept évite la redondance des données et permet l’application des opérations algébriques, telles que l’intersection, la jointure ou le produit cartésien.

# Le Big Data et les bases « NoSQL »

Au fur et à mesure de l’augmentation du volume des données dans certains secteurs, et surtout celui d’Internet, le modèle relationnel perdait en performance dans la gestion des données à traiter. Dopé par l’augmentation de la bande passante, la baisse des coûts des machines et la généralisation du réseau, de nouveaux concepts on fait leur apparition, comme la virtualisation et l’informatique distribuée.

Un nouveau défi vu le jour : comment traiter un volume de plusieurs téraoctets, tout en ayant des performances optimales.

C’est en octobre 1997, que l’expression « Big Data » apparu à travers des articles scientifiques, qui bien sûr désigne le fait d’avoir des quantités tellement énormes de données qu’il devient difficile de garantir un traitement efficace et performant. Le défi était lancé. Plusieurs entreprises commencèrent à rechercher et développer de nouvelles technologies utilisant les nouveaux concepts liés l’évolution d’Internet.

Les premières bases de données dîtes « NoSQL » virent le jour. Couplées aux concepts d’architecture distribuée et virtualisation, que nous verrons plus en détail dans d’autres chapitres, elles solutionnèrent le problème.

Il va de soi que cette nouvelle manière de traiter les données est spécifique à ce genre de problématique, et n’est pas une évolution des bases de données traditionnels. Elle amène une nouvelle dimension sur le choix de la basse de données à utiliser lors d’un nouveau projet. Elle n’intègre en aucun cas tous les avantages que peut fournir une base de données traditionnelle, en plus de ses avantages.

En résumé, il faut garder à l’esprit que les bases de données « NoSQL » sont un autre moyen de traiter les données, et le choix de la base de données n’est que plus varié. Il faudra cependant faire une analyse des besoins afin de choisir la plus appropriée.

Afin de mieux comprendre les différences

# dddd

La mise en page du rapport de travail de Bachelor comprend une marge de 2,5 cm en haut et en bas, ainsi qu’à droite. À gauche, une marge de 3,5 cm est requise pour permettre la reliure.

Les paragraphes seront rédigés en utilisant le style « **Corps de texte** » : Arial 11 points, interligne de 1,5, justifié, espace au-dessus de 9 pt, en-dessous de 6 pt.

Il n’est pas nécessaire d'insérer des retours à la ligne supplémentaires entre les différents paragraphes.

Pour les titres, utilisez les styles de paragraphe « **Titre 1**» à « **Titre 5** ». Evitez d'avoir plus de niveaux dans votre travail de Bachelor. Les styles « Titre 6 » et suivants ne sont pas paramétrés.

Le style « **titre non numéroté centré** » est utilisé pour les titres suivants : déclaration, remerciements, résumé, liste des tableaux, liste des figures, bibliographie, annexes.

Tous les éléments entourés de crochets < > doivent être modifiés ou effacés et tous les crochets doivent être supprimés.

N’oubliez pas de modifier le pied de page pour y inscrire le titre de votre travail de Bachelor, ainsi que votre nom et votre prénom. Vous devrez le faire à deux endroits : sur la page contenant la déclaration et sur cette page-ci.

Pour les listes à puces, utilisez le style de paragraphe « **puce TB**» comme dans l'exemple ci-dessous :

* Énumération
* Énumération
* Énumération

Ajoutez si nécessaire des sauts de page dans votre document (onglet Insertion > bouton Saut de page). Insérez toujours un saut de page lorsque vous changez de chapitre de niveau 1.

Faites attention à la cohérence dans la numérotation des pages. Avant l'introduction, les pages sont numérotées en chiffres romains. A partir de l'introduction, la numérotation reprend à 1 en chiffres arabes et doit être continue sur tout le document, y compris dans les annexes.

Pour les figures, ajoutez une légende automatique sur le modèle ci-dessous (onglet Références > bouton Insérer une légende). Choisissez l’étiquette « Figure » et la position « Au-dessus de la sélection » pour la légende. Groupez les éléments. Écrivez ensuite la source en-dessous de l'image, selon le modèle ci-dessous et appliquez le style « **source** » à ce paragraphe.

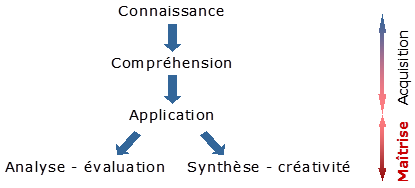


Figure 1 : Titre de la figure

(Bertholet 2004, p. 22)

Procédez de la même manière pour les tableaux avec la légende et la source, en choisissant l’étiquette « Tableau », comme dans l'exemple ci-dessous.

Tableau 1 : Titre du tableau

|  |  |  |
| --- | --- | --- |
| **Année** | **Ventes** | **Commentaire** |
| 1999 |  |  |
| 2000 |  |  |
| 2001 |  |  |
| 2002 |  |  |

(Office de la statistique 2003, p. 26-27)

Les légendes permettent de générer les tables des figures et des tableaux. Après l'ajout d'une figure ou d'un tableau, pensez à actualiser ces tables.

Poursuivez votre texte immédiatement après la figure ou le tableau avec le style « **Corps de texte** ».

Les notes de bas de page[[1]](#footnote-1) sont indiquées par un appel de note. Le texte de la note de bas de page aura le style « **Note de bas de page** » : texte en retrait de 0,5 cm à gauche, police Arial, 10 points, interligne simple, alignement à gauche.

Les citations de moins de 2 lignes peuvent être insérées à l’intérieur du texte, en étant encadrées par des « guillemets » et suivies du nom de l’auteur, de l’année de publication et du numéro de la page d’où la citation est tirée, le tout entre parenthèses (Auteur 2012, p. 54). Lorsqu’une citation est plus longue, il est préférable de la mettre en retrait, comme suit, en appliquant le style de paragraphe « **Citation** » au paragraphe, sans oublier les « guillemets » :

« Le style des citations est Arial, 11 points, italique, texte justifié, retraits à droite et à gauche de 0,5 cm par rapport aux paragraphes réguliers, espace avant et après de 6 points, interligne simple. » (Sartre 2012, p. 66)

Utilisez une tabulation pour aligner la citation à droite.

Reprenez ensuite votre texte normalement avec le style « **Corps de texte** ».

# Titre du premier chapitre

Votre texte...

Votre texte...

# Titre du second chapitre

Votre texte…

Votre texte…

## Titre de second niveau

Votre texte...

Votre texte...

## Titre de second niveau

Votre texte...

Votre texte...

### Titre de troisième niveau

Votre texte...

### Titre de troisième niveau

Votre texte...

Votre texte...

#### Titre de quatrième niveau

Votre texte...

##### Titre de cinquième niveau

Evitez d'aller plus loin que ce cinquième niveau.

# Conclusion

Votre texte…

Votre texte…

Bibliographie

Le style de paragraphe appliqué dans les références doit être « **références bibliographiques** » : Arial 11 points, interligne simple, justifié, espace en dessus 6 pt, espace en dessous 6 pt. Voici un exemple :

BRUCHEZ, Rudi, 2015. Les bases de données NoSQL et le Big Data (2ème édition), Paris : Eyrolles, 2015. Droit en action. ISBN 978-2-212-14155-9.

Annexe 1 : Titre de l’annexe

Votre texte ou le document mis en annexe.

Évitez de mettre deux annexes sur la même page. Faites plutôt un saut de page puis indiquez le numéro de la nouvelle annexe et son titre.

1. Note de bas de page. [↑](#footnote-ref-1)