Indexation et recherche de contenu utilisant MapReduce de Google

Rapport final

Elodie Corbel, Kévin M'Ghari,
Mickaël Olivier, Clarisse Renou

Encadrant : Alexandru Costan

Résumé

Le résumé est limité à 10 lignes au maximum.

1 Remerciements

Nous tenous tout d'abord à remercier Alexandru Costan, professeur à l'INSA de Rennes, nous encadrant lors de ce projet. Il a sû nous donner de précieux conseils et nous guider tout au long de l'année aussi bien que nous donner un sujet d'étude riche et intéressant. Nous remercions aussi le personnel de l'INRIA, nous aillant accueilli lorsque nous allions rendre visite à Monsieur Costan.

2 Introduction

Dans le cadre des études pratiques, en troisième année au département informatique à l'INSA de Rennes, nous avons été amené à réaliser un projet tout au long de l'année. Celui que nous avons choisi porte sur l'indexation et la recherche de contenu utilisant MapReduce de Google[1]. Nous avons choisi ce sujet car la recherche de contenu et Google sont pour nous quelque chose d'inévitable à partir du moment où un utilisateur utilise un ordinateur. En effet, Google, connu surtout pour son très populaire moteur de recherche, est un phénomène à lui tout seul, il représente 6,4% du traffic Internet mondial en 2010[2]. Connaître et utiliser un modèle de programmation tel que MapReduce conçu par Google était donc pour nous très motivant.

Le principe de la recherche de contenu est assez simple. Il existe des documents structurés ou non dans lesquels des personnes souhaitent effectuer une requête auprès d'un serveur. Le serveur renvoie donc les documents dans lesquels se trouvent les mots sur lesquelles porte la requête. L'indexation permet d'améliorer la rapidité et les performances d'une recherche de contenu. En effet, celle-ci identifie les éléments significatifs du document afin de permettre un accès plus rapide à ceux-ci en créant un index. Lors d'une recherche, le moteur de recherche va d'abord chercher les informations dans l'index puis rend les documents à l'utilisateur.

Nous avons donc pour mission de faire un moteur de recherche avec son système d'inxation utilisant MapReduce de Google.

Dans un premier temps, nous allons vous expliquer les solutions que nous avons choisi et la façon dont nous avons décomposé le travail. Puis, nous vous expliquerons

comment nous avons conçu notre moteur de recherche. Et enfin, nous expliciterons les résultats obtenus.

3 Etude du projet

Afin de déterminer ce que nous devions faire, dans une première partie de l'année, nous avons étudier la littérature existante sur le fonctionnement de MapReduce de Google afin de mieux comprendre ce que nous devions faire. Puis, nous avons découpé notre travail. Nous allons donc dans une première partie expliquer le fonctionnement de MapReduce et la solution choisie pour implémenter ce modèle de programmation. Puis, nous parlerons de la répartition du travail.

3.1 Présentation de MapReduce

MapReduce est un modèle de programmation popularisé par Google. Il est utilisé pour l'indexation de contenu. Il se repose sur deux fonctions une fonction Map et une fonction Reduce. La fonction Map prend en entrée une clé et des valeurs associées. Par exemple, pour un fichier, la clé peut être un numéro de ligne et la valeur le texte de la ligne. Cette fonction Map ensuite rend une clé intermédiaire et une valeur intermédiaire. Dans l'exemple du Wordcount dans laquelle on veut compter le nombre de chaque mot, la clé intermédiaire peut être un mot et la valeur 1. La fonction Reduce ensuite prend les clés et valeurs intermédiaires données par la fonction Map et aggrégège le résultat afin de donner une clé finale et un résultat final. Dans l'exemple du Wordcount, la clé serait un moteur et le résultat serait le nombre d'occurences de ce mot. Il est très efficace pour le traitement de données importantes mais le

FIGURE 1 – Schema illustratif MapReduce

résultat n'est pas immédiat. Il est donc utilsé en tâche de fond.

3.1.1 Titre de sous-sous-section

Titre de paragraphe Exemple de référence à une figure au format PostScript encapsulé (figure 2). Cette figure a été créée à l'aide de `xfig`[1] après exportation du fichier fig vers le format Encapsulated PostScript.

1. disponible sous Unix/Linux.

FIGURE 2 – Exemple d'inclusion d'une figure EPS

3.2 Encore un titre de sous-section

Exemple de liste à puces :

- ligne de remplissage pour visualiser la mise en page. Ligne de remplissage pour visualiser la mise en page ;
- ligne de remplissage pour visualiser la mise en page. Ligne de remplissage pour visualiser la mise en page.

Ligne de remplissage pour visualiser la mise en page. Ligne de remplissage pour visualiser la mise en page.

4 Conclusion

LaTeX c'est facile pour produire des documents standard et nickel ! Et BibTeX pour les références, c'est le pied.

Références

[1] Alexandru COSTAN : *Sujets des etudes pratiques 2012-2013*, chapitre Sujet 8. Institut National des Sciences Appliquees de Rennes, 2012.

[2] Wikipedia l'encyclopedie LIBRE : Google. `http://fr.wikipedia.org/wiki/Google`.