Application du paradigme MapReduce aux données ouvertes Cas : Accessibilité des personnes à mobilité réduite aux musées

Billel Arres, Nadia Kabachi, Fadila Bentayeb, Omar Boussaid

Université de Lyon (Laboratoire ERIC)
5 avenue Pierre Mandès-France, 69676 Bron, France
{billel.arres | nadia.kabachi | fadila.bentayeb | omar.boussaid} @univ-lyon2.fr

Résumé. Le modèle *MapReduce* est aujourd'hui l'un des modèles de programmation parallèle les plus utilisés. Définissant une architecture Maître-Esclave, il permet le traitement parallèle de grandes masses de données. Dans ce papier, nous proposons un algorithme basé sur *MapReduce* qui permet, à partir des données publiques du Ministère Français de la Communication et de la Culture, de définir un classement des galeries et musées nationaux selon leurs degré d'accessibilité aux personnes handicapées. Tout en profitant de la puissance et de la flexibilité du paradigme *MapReduce*, les décideurs pourront mettre en place des stratégies efficaces à moindre coût et avoir ainsi une vision plus précise sur les établissements culturels et leurs limites relatives à cette catégorie de personnes. L'algorithme que nous proposons peut être exploité et appliqué à d'autres cas d'études avec des jeux de données plus volumineux.

1 Introduction

Le monde des musées aujourd'hui connaît un engouement sans précédent qui a vu plus de 31 millions de visiteurs se précipiter en 2012 dans les musées nationaux français (NuitDesMusées, 2013). Par ailleurs, et depuis plus d'une dizaine d'années, les ministères et les services publiques des différents pays accordent de plus en plus d'importance à l'ouverture et à la réutilisation de leurs données collectées ou produites au niveau de leurs différents établissements. Or, avec le temps ces données s'accumulent et deviennent très difficiles à stocker, à traiter et à analyser. D'où la nécessité de s'orienter vers de nouvelles solutions et paradigmes de programmation afin de faire face à ces difficultés. Nous proposons dans ce travail un algorithme basé sur le paradigme *MapReduce* (Dean et Ghemawat, 2008) qui permet, à partir des données ouvertes du Ministère français de la communication et de la culture, de déterminer le degré d'accessibilité des personnes handicapées (tout type d'handicape) aux musées nationaux français. L'algorithme définit, à partir de ces données brutes et hétérogènes collectées, une note pour chaque musée selon son degré d'accessibilité, et retourne un classement final des musées par commune et par région de France. Notre algorithme de classement présente plusieurs avantages notamment en terme de rapidité de traitement de grandes quantités de données puisqu'il

ne s'exécute pas de manière classique (mono-poste) mais distribué sur un cluster d'ordinateurs. Nous avons implémenté notre algorithme sur une plateforme *Hadoop* multi-noeuds.

2 Jeu de données

L'algorithme que nous présentons dans ce papier est appliqué aux données publiques [1] récoltées à partir de la grande manifestation annuelle : "La nuit européenne des musées". Elles concernent plus de 3000 musées en Europe dont plus de 1200 établissements français labellisés "Musées de France" par le Ministère de la communication et de la culture. Ces données sont fournies sous le format de fichier .csv, elles regroupent des informations (plus de 48 attributs avec la dernière mise à jour) sur chaque établissement tel que : Le nom du musée, adresse, commune, etc. Mais aussi des données sur différents types d'accessibilités aux personnes handicapés offerts tel que : accès-handicape-moteur, accès-handicape-visuel, accès-handicape-auditif, accès-handicape-intellectuel et acces-handicap-langueDesSignes. Ces dernières sont des données binaires, le but est de traiter ces données là afin d'extraire une note d'accessibilité pour chaque musée, et au final agréger les résultats obtenus géographiquement par ville et par région de France.

3 Présentation de l'algorithme

L'algorithme que nous proposons dans ce travail procède en deux phases : (1) La phase *Map* où les données sont partitionnées et réparties sur les noeuds du cluster et où chaque noeud effectue les traitements nécessaires ; (2) La phase *Reduce* où les résultats intermédiaires de chaque noeud sont regroupés et agrégés.

4 Expérimentation

L'évaluation de l'algorithme a été effectuée sur un cluster d'ordinateurs à deux noeuds identiques. Une machine maitre qui travail en tant que *namenode* et *datanode* en même temps et une machine esclave servant de *datanode* pour les traitements uniquement. Les machines utilisées présentent les mêmes caractéristiques à savoir une mémoire installée de 4Go (RAM), d'un processeur Intel Pentieum - 2.13Hz dual et d'un disque local de capacité égale à 350 Go.

5 Résultats des expérimentations

Les résultats obtenus sont très pertinents et nous renseignent beaucoup d'informations relatives à l'accessibilité des personnes handicapées aux musées nationaux français. L'exploitation de ces résultats nous permet d'établir une analyse selon différents niveaux d'agrégations.

Ces représentations peuvent être très utiles aux décideurs. Par exemple, une comparaison de nos résultats avec les données officielles de la fréquentation des musées Français [2] (Fig. 2), montre une forte corrélation entre ces données.

1. Plateforme française d'ouverture des données publiques (Open Data) : http ://www.data.gouv.fr
2. museostat 2009 - Ministère de la Communication et de la Culture, Direction générale des Patrimoines, France

Algorithm 1 Map (Key,Value)

Input : Jeu de données D
Output : (Key1,Value1)

```
1:  nbr=0 ; i=0 ; deg=0 ; string ville ; string region ; Col{0} ; // Initialisé à la 1ère colonne
2:  for each line in D do
3:     if longueur ligne >= 48 //Nombre d'attributs then
4:        ville=Col{3} ;
5:        region=Col{5} ;
6:        for int j=22 ; j < 27 ; j++ //Parcourir les attributs des types d'accès handicapé do
7:           if Col{j}=1 then
8:              deg++ ;
9:           end if
10:          deg=(deg*100)/6 ; //Calcule du degré (pourcentage %) pour un musée
11:          if deg > 0 then
12:             i=1 ; //Indicateur à 1 si le musée offre au moins un type d'accès
13:          end if
14:       end for
15:       Key1=(ville,region) ; Value1=(1,i,deg) ;
16:       output.collect (Key1,Value1) ;//Encapsulation des résultats en un couple (clé,valeur)
17:    end if
18: end for
```

6 Conclusion

L'objet de l'étude présentée dans cet article est d'étudier la possibilité d'appliquer un nouveau modèle de programmation parallèle aux données publiques. Nous nous sommes en particulier intéressés à l'application du modèle de programmation *MapReduce* pour étudier et analyser l'accessibilité des personnes handicapées aux musées nationaux Français. L'étude appliquée aux données ouvertes a été mené dans une optique exploratoire.

Références

Dean, J. et S. Ghemawat (2008). Mapreduce: simplified data processing on large clusters. *Communications of the ACM - 50th anniversary issue: 1958 - 2008 51*, 107–113.

NuitDesMusées (2013). Ministére de la culture et de la communication, la nuit européenne des musées. *Web page: http://www.nuitdesmusees.culture.fr*.

Summary

Based on official data from the French Ministry of Communication and Culture, we propose in this paper a parallel algorithm as a solution to extract and process these data sets in order to define a ranking of national museums and galleries according to the accessibility degrees for people with disabilities.

Algorithm 2 Reduce (Key1,Value1)

Input : (Key1,Value1)
Output : (Key2,Value2)

```
1:  NbrNoté = 0 ; NbrTot = 0 ; Deg = 0 ;
2:  while values.hasNext() (Répéter le traitement pour toutes les paires (Key1,value1)) do
3:      tot = value1.tokens[0] ; num = value1.tokens[1] ; val = value1.tokens[2] ;
4:      if val > 0 then
5:          Deg=(Deg*NbrNoté + val*num)/(NbrNoté+num) ;
6:          NbrNoté = NbrNoté + num ;
7:      end if
8:      NbrTot=NbrTot+tot ;
9:      Key2=Key1 ; // La clé reste la même
10:     Value2=(NbrTot,NbrNoté,Deg) ;
11:     output.collect (Key2,Value2) ;//Encapsulation des résultats en un couple (clé,valeur)
12: end while
```

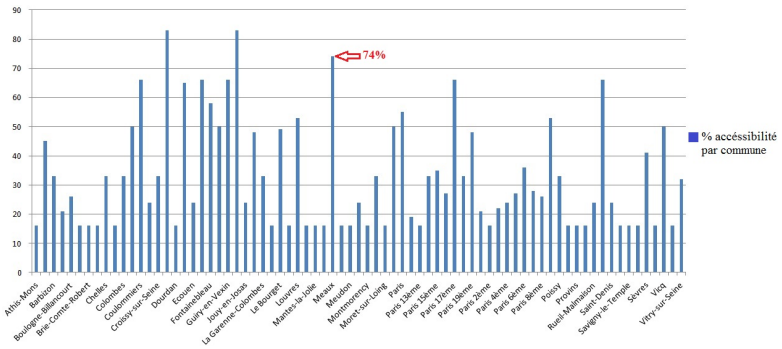

FIG. 1 – *Représentation graphique des résultats obtenus pour la région Ile-de-France.(extrait)*

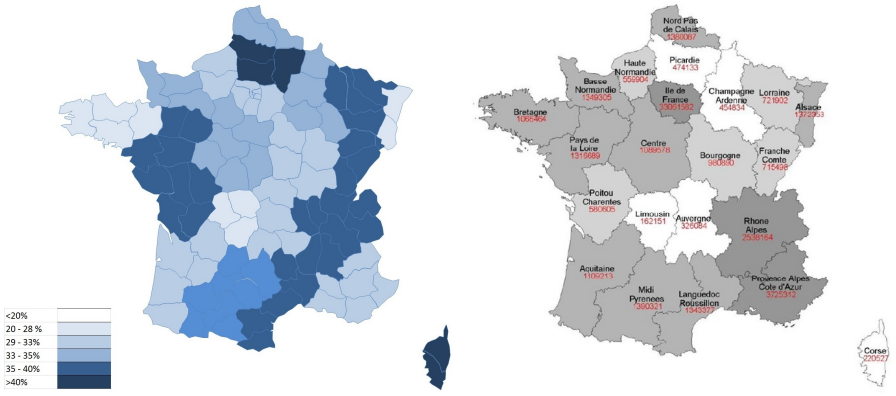

FIG. 2 – *Degré d'accessibilité par région. - Fréquentation régionale des musées.*