UNIVERSITE DE BORDEAUX

Programmation large échelle
Rapport du TP6
MapReduce Pattern : Top K

Par :
HOCINI Mohamed Fouad
MAURICE Bastien

Enseignant :
David Auber

19 novembre 2017

Ce TP6, nous a permis d'utiliser le pattern TopK, ce pattern consiste à sélectionner les k plus grand/petit ... etc d'un ensemble et dans ce TP nous l'avons utilisé pour trouver les k villes les plus peuplées au monde de notre fichier "worldcitiespop.txt".

1 Exercice 1

Dans cet exercice, nous avons écrit un Mapper et un Reducer avec un k=10. Nous avons avons utilisé un TreeMap qui a comme clé la population et comme valeur les informations de la ville, ce qui nous permet d'avoir un ensemble de ville triées par population.

Le Mapper place chaque ville dans notre treeMap et teste si la taille dépasse k ou non, et si c'est le cas la valeur de la clés (population) la plus petite sera supprimée.

A a fin de cette opération, notre treeMap contiendra le TopK ville recherché parmis toutes les villes de notre fichier "worldcitiespop.txt"

2 Exercice 2

Dans cette partie, nous avons ajouté un Combiner afin d'alléger la charge de travail du Reducer dans le cas ou un grand nombre de Mapper sont utilisés.

Cette partie était la partie la plus dur à réaliser et nous avons eu un peu de problème avec parce que nous n'avons pas eu les résultats attendus directement mais nous avons réussi à la fin à fixer le problème et ça a marcher parfaitement.

Avec l'utilisation des Combiner, chaque ordinateur dans notre cluster a son propre Combiner contrairement au Reducer qui reçoit le nombre d'ordinateur * le nombre de Mapper par ordinateur ...

3 Exercice 3

Dans cette dernière partie, nous avons ajouté k comme paramètre au lieu de le fixer dans notre code, cette valeur "k" passée en argument à l'appel du job est stockée dans la configuration au nom de k ensuite notre treeMap¡¿ récupère cet valeur afin qu'il ne l'a dépasse pas lors de son traitement de notre fichier d'entrée, l'appel au TopK avec les valeurs : 100, 1000, 10000 s'exécute à des vitesses similaires au topK à 10 et les résultats concordent.

4 Résultats

Nous avons fait quelques tests pour voir si nos résultats sont bien ou non, nous citons trois tests ici avec un k=1, 10 , 100 et nous pouvons voir que les résultats dans les figures suivantes :

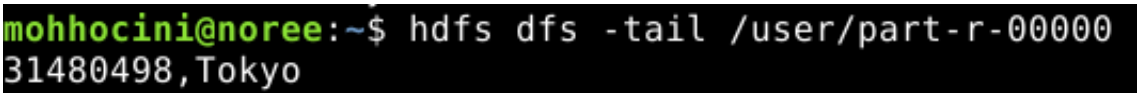

Figure3 : TopK : k=1.

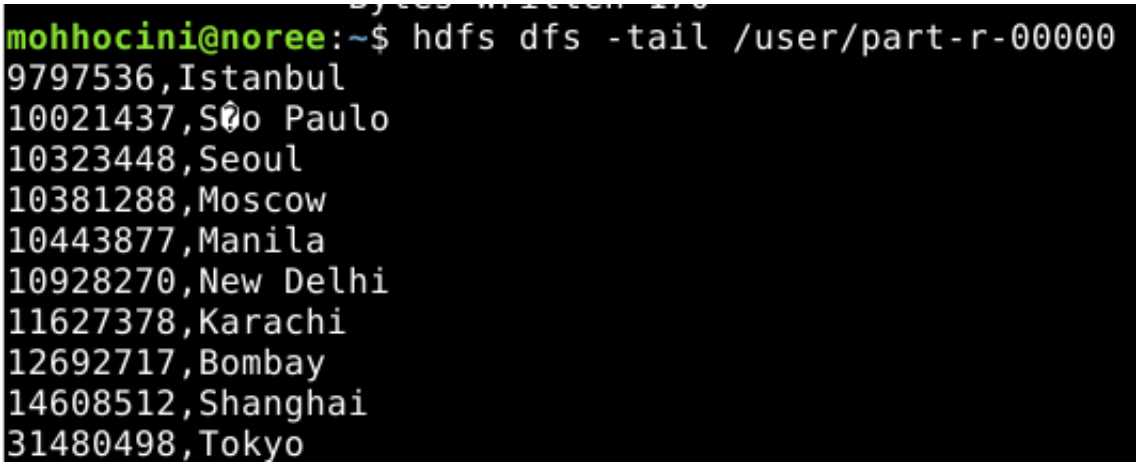

Figure1 : TopK : k=10.

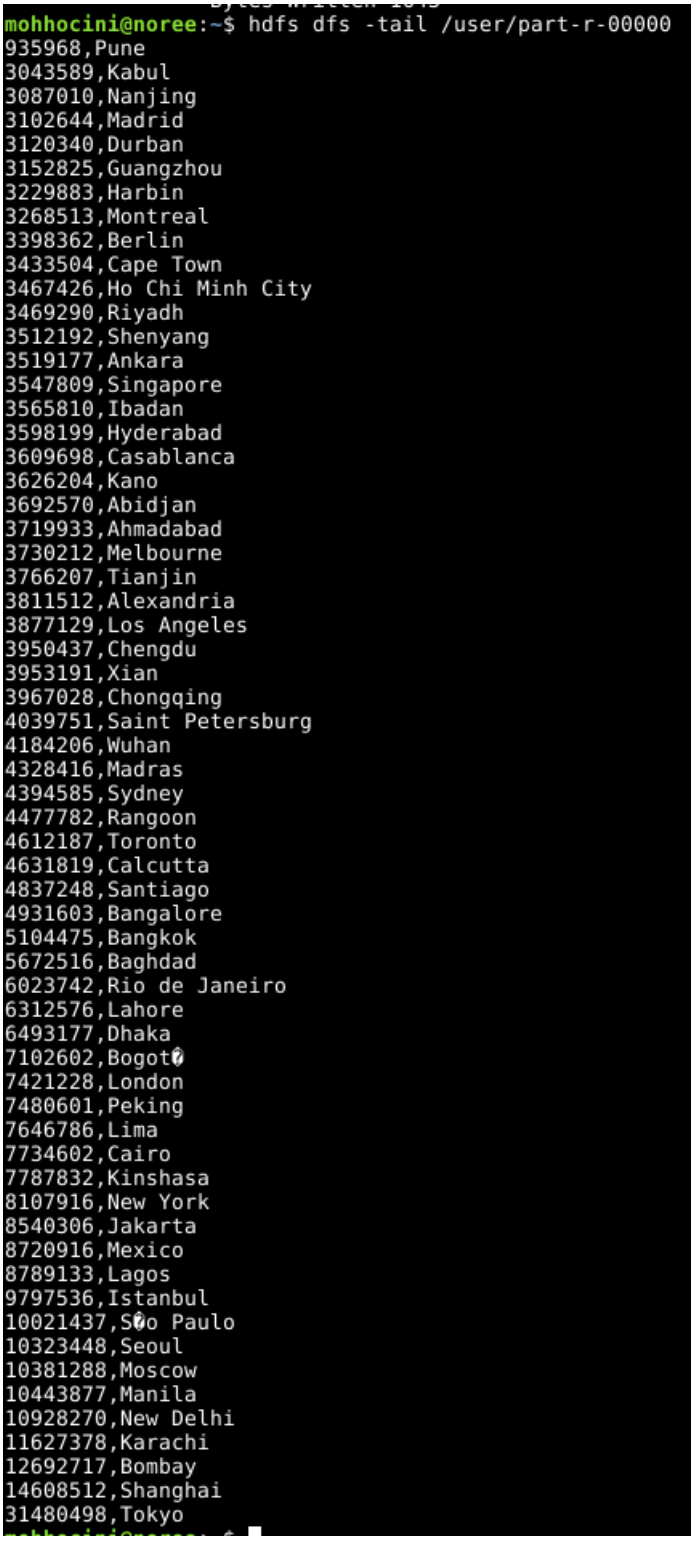

Figure2 : TopK : k=100.