SCIENCES SORBONNE UNIVERSITÉ

Travaux dirigés
Exercices de Map-Reduce

Jonathan Lejeune

Objectifs

Ce document a pour objectif de vous familiariser avec le concept du Map-Reduce Pour chaque question qui vous demande un programme Map-Reduce précisez à chaque fois :

- ce que fait un map
- ce que fait un reduce
- ce que représente les différentes clés/valeurs transmises avant et après chaque phase de map et reduce
- le nombre de reduce
- le cas échéant la politique de partitionnement et de tri pour la phase de shuffle

Vous donnerez vos solutions en pseudo-code. Les fonctions à spécifier ont les prototypes suivant :

- la fonction map

```
Map(<TypeCléMap> key, <TypeValeurMap> value){
   //ici votre pseudo code map
}
```

- la fonction reduce

```
Reduce(<TypeCléReduce> key, liste de <TypeValeurReduce> values){
   //ici votre pseudo code reduce
}
```

- la fonction de partitionnement

```
IdReduce getPartition(<TypeCléReduce> key, <TypeValeurReduce> value,
                                                entier nbReduce){
   //retourne l'identifiant reduce de 0 à nbReduce-1 pour <key,value>
}
```

- la fonction de tri

```
entier compare(<TypeCléReduce> key1, <TypeCléReduce> key2){
    //retourne négatif si key1 < key2
    //retourne nul si key1 = key2
    //retourne positif si key1 > key2
}
```

Il est tout à fait possible de considérer qu'un type de clé ou de valeur n'est pas un type basique (un entier, un réel ou une chaîne de caractère) mais un type structuré agrégeant plusieurs champs.

Exercice 1 – StereoPrix

StereoPrix est une entreprise de grande distribution et souhaite faire des statistiques sur les ventes. Elle possède un ensemble de données stocké sur un système HDFS. Ces données sont stockées dans des fichiers textes. Chaque ligne d'un fichier correspond à la vente d'un produit et on peut y trouver des informations comme :

- la date et l'heure de vente
- le nom du magasin où le produit a été vendu
- le prix de vente
- la dénomination du produit
- la catégorie du produit (ex : fruits et légumes, électroménager, jouet, ...)

Question 1

Écrivez un programme map-reduce permettant de calculer *Le chiffre d'affaire de l'entreprise c'est à dire la somme total des ventes de l'année 2018.*

Question 2

Écrivez le programme map-reduce permettant de calculer *le chiffre d'affaire généré pour chaque catégorie.*

Question 3

Écrivez le programme map-reduce permettant de calculer *le produit le plus vendu par catégorie.*

Exercice 2 – Noodle

L'entreprise « Noodle » , gestionnaire d'un moteur de recherche de pages web, souhaite effectuer des statistiques sur l'ensemble des requêtes des usagers. L'entreprise possède plusieurs serveurs dans le monde. À chaque nouvelle requête reçue, une ligne de log est sauvegardée dans un fichier du serveur sur lequel la connexion s'est faite. Cette ligne de log est formatée de la manière suivante :

- le premier mot indique la date de la connexion au format JJ_MM_AAAA_HH_MM_SS
- le deuxième mot indique l'adresse IP du client ayant fait la requête
- le reste de la ligne contient les mots-clés de la requête. Chaque mot-clé est séparé par le caractère +.

Nous donnons comme exemple la ligne ci-dessous où des espaces ont été ajoutés pour une meilleure lisibilité.

```
02_11_2012_12_32_10 132.227.045.028 musique+orientale
```

Ceci indique que le client 132.227.45.28 a fait une requête comportant les mots clés « musique » et « orientale » le 02/11/2012 à 12h 32min 10sec.

Question 1

Écrivez le programme map-reduce permettant de calculer par tranche horaire d'une demi-heure d'une journée type le mot clé le plus recherché dans la tranche ainsi que le nombre total de requêtes reçues dans la tranche.

Par exemple, voici le format d'un fichier de sortie

```
entre 00h00 et 00h29  <Le mot le plus recherché dans cette tranche horaire>  <le nombre  total  de requêtes reçues dans cette tranche horaire>
entre 00h30 et 00h59  <Le mot le plus recherché dans cette tranche horaire>  <le nombre  total  de requêtes reçues dans cette tranche horaire>
entre 01h00 et 01h29  <Le mot le plus recherché dans cette tranche horaire>  <le nombre  total  de requêtes reçues dans cette tranche horaire>
 ...
entre 23h00 et 23h29  <Le mot le plus recherché dans cette tranche horaire>  <le nombre  total  de requêtes reçues dans cette tranche horaire>
entre 23h30 et 23h59  <Le mot le plus recherché dans cette tranche horaire>  <le nombre  total  de requêtes reçues dans cette tranche horaire>
```

Question 2

Modifiez le programme précédent pour avoir ces mêmes informations mais classifiées par mois de l'année. Nous souhaitons un fichier de sortie par mois. Par exemple un fichier représentant le mois de février ne dois contenir que des informations sur les jours du mois de février.

Exercice 3 – Un cas réel d'utilisation : Last.fm

Last.fm est un site web de radio en ligne et de musique communautaire offrant différents services à ses utilisateurs comme par exemple l'écoute ou le téléchargement gratuit de musiques. Il existe plus de 25 millions d'utilisateurs qui utilisent Last.fm tous les mois générant ainsi beaucoup de données à traiter. L'analyse de données la plus courante se fait sur les informations que les utilisateurs transmettent au site lorsqu'ils écoutent une musique. Grâce à ces informations, il possible de produire entre autres des hit-parades.

Un titre peut être écouté de deux manières différentes par un utilisateur :

- soit en local sur son propre ordinateur et les informations d'écoute sont envoyées directement au serveur de Last.fm
- soit via une web radio sur le site même. Dans ce cas, l'utilisateur a la possibilité de passer le titre sans l'écouter.

Le système logue pour chaque utilisateur et pour chaque titre le nombre de fois où l'utilisateur a écouté le titre en local, le nombre de fois où l'utilisateur a écouté le titre en ligne et le nombre de fois où l'utilisateur l'a passé sans l'écouter. Le tableau ci-dessous en donne un exemple.

UserId	**TrackId**	**LocalListening**	**RadioListening**	**Skip**
111115	222	0	1	0
111113	225	3	0	0
111117	223	0	1	1
111115	225	2	0	0
111120	221	0	0	1

L'objectif de cet exercice est de calculer pour chaque titre :

1. le nombre de personnes qui l'ont écouté au moins une fois (en local ou en radio)
2. le nombre de fois où il a été écouté et passé sans écoute.

Pour ceci nous allons procéder en trois jobs MapReduce :

- le job 1 et 2 qui calculeront respectivement 1) et 2) et qui pourront s'exécuter en parrallèle
- le job 3 qui fusionnera les résultats des deux jobs précédents

Question 1

Écrivez le programme map-reduce du job 1. Dans le cas de l'exemple donné sa sortie devra être :

TrackId	**#listener**
222	1
223	1
225	2

Question 2

Écrivez le programme map-reduce du job 2. Dans le cas de l'exemple donné sa sortie devra être :

TrackId	#listening	#skips
221	0	1
222	1	0
223	1	1
225	5	0

Question 3

Écrivez le programme map-reduce du job 3. Dans le cas de l'exemple donné sa sortie devra être :

TrackId	#listener	#listening	#skips
221	0	0	1
222	1	1	0
223	1	1	1
225	2	5	0

Exercice 4 – Calcul de Π en *MapReduce*

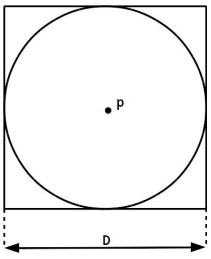
Nous souhaitons calculer la valeur de π grâce à une méthode de Monte Carlo. Cette approche probabiliste du calcul de π considère :

- un point p de coordonnées (x, y) dans un repère orthonormé $(O, \overrightarrow{i}, \overrightarrow{j})$;
- un cercle de diamètre D centré sur le point p ;
- un carré de coté D centré sur le point p.

L'aire A_{cercle} du cercle peut s'exprimer :
$A_{cercle} = \pi * rayon^2 \Leftrightarrow A_{cercle} = \pi * \frac{d^2}{4} \Leftrightarrow \pi = 4 * \frac{A_{cercle}}{d^2} \Leftrightarrow \pi = 4 * \frac{A_{cercle}}{A_{carre}}$

En discrétisant par un nombre suffisamment grand de points l'aire du carré, on peut obtenir une bonne approximation du rapport des aires et donc de la valeur de π. Ainsi, la méthode consiste à tirer aléatoirement des points se situant dans le carré, puis à calculer la proportion des points appartenant au cercle. La valeur de π s'obtient alors en multipliant par 4 ce ratio.

Pour savoir si un point P de coordonnées (x, y) appartient au cercle, il suffit de calculer sa distance par rapport au centre du cercle (le point p) en utilisant le théorème de Pythagore. Pour rappel, la distance entre deux points A et B de coordonnées respectives (x_A, y_A) et (x_B, y_B) est égale à $\sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$. Si cette distance est supérieure au rayon du cercle alors le point n'est pas dans le cercle.



Question 1

Dans le cadre de ce problème, à quoi servirai les fichiers de données d'entrée ici ?

Question 2

Écrivez le programme map-reduce permettant de calculer une estimation de Pi à l'aide de la méthode de Monte-Carlo exposée ci-dessus.