Université de Bordeaux
M2 Master MAS-MSS - UE : Projet Données Massives
Année 2018-2019

Base de données d'un site de vente en ligne

Le but de ce projet informatique est de traiter des données à travers le paradigme de programmation *MapReduce*, la structure de fichiers *HDFS* et à l'aide du package `rmr2` du logiciel R.

1 Récupération des données

Les données à récupérer se situent dans une archive nommée
`ECT.tar.gz`
Les fichiers sont accessibles sur la page
`http://prof.ict.ac.cn/BigDataBench3/wp-content/uploads/2014/04/four_data/ECT.tar.gz`
L'archive est composée de deux fichiers représentant des paniers d'acheteurs sur un site de vente en ligne. Une fois décompressée, les deux fichiers peuvent être chargés dans R (dans les variables `OS_ORDER` et `OS_ORDER_ITEMS`) à l'aide de la commande `read.table`.

2 Traitement des données

Ce projet informatique est très libre, en particulier toutes les initiatives personnelles et les approches innovantes seront fortement appréciées. Néanmoins, l'un des buts étant de vous familiariser avec la problématique de la programmation *MapReduce*, il vous sera demandé d'implémenter chaque question en utilisant le modèle de programmation *MapReduce* à travers le package `rmr2`. Afin de vérifier la cohérence de vos résultats, vous pouvez éventuellement les comparer avec un traitement standard des données en R.

Attention : pour le traitement des données en *MapReduce*, on utilisera directement les données converties au format *HDFS* à l'aide des commandes

```
OS_ORDER_bigdata <- to.dfs(OS_ORDER)
OS_ORDER_ITEMS_bigdata <- to.dfs(OS_ORDER_ITEMS)
```

Il n'est pas autorisé d'utiliser des pré-traitements des données en R avant de les convertir en *HDFS* !

Voici quelques questions qui pourront être traitées avec *RHadoop* :

- Combien il y a-t-il d'acheteurs ? de ventes ?
- Quel est le montant moyen dépensé par acheteur ? Représenter graphiquement la répartition des montants d'achats par vente puis par acheteur.

- Quel est le nombre moyen de produits par paniers ?
- Proposer/étudier une méthode d'échantillonnage des acheteurs (tirage aléatoire d'un échantillon d'acheteurs).

Nous insistons sur le fait que les questions précédentes ne sont que des suggestions et n'ont pas vocation à être exhaustives. Toute prise d'initiative sera appréciée.

3 Travail à effectuer

Il est demandé de nous envoyer un compte-rendu sous la forme d'un fichier Rmarkdown (et le .pdf associé) avec vos codes *R* correctement commentés.