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

Prénoms à la naissance en France

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 fichiers disponibles aux adresses suivantes :
- `http://francois.guillem.free.fr/data/sourceprenoms.rda`
- `http://francois.guillem.free.fr/data/prenoms.rda`

contiennent des données sur les prénoms à la naissance en France (par département) entre 1946 et 2006. Ces fichiers portent sur un peu plus de 10000 prénoms. Chaque observation est une ligne qui correspond à un prénom, un nombre d'occurrences, un département, une année.

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 de la commande

```
prenoms_bigdata <- to.dfs(prenoms)
sourceprenoms_bigdata <- to.dfs(sourceprenoms)
```

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

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

- Pour un prénom donné, calculer des statistiques de base sur l'année de naissance du type moyenne, médiane, variance pour la France entière, puis par département...
- Pour une année donnée, déterminer le prénom donné le plus de fois, le moins de fois par département...

- Pour un prénom donné, déterminer s'il existe des corrélations plus ou moins fortes entre les départements à partir de la forme de la répartition de ce prénom au cours des années.
- Est-il possible de regrouper de façon cohérente les histogrammes des prénoms entre les départements en utilisant une méthode classification non-supervisée ?

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.