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

Répartition de la population par âge et par pays

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

Dans cet exemple, il est proposé d'analyser les histogrammes de répartition de la population par âge dans plusieurs pays du monde. Ces données peuvent être importées depuis l'International Data Base (IDB), qui est produite par l'International Programs Center, US Census Bureau (IPC, 2000). Elles peuvent être téléchargées a l'URL :
`https://www2.census.gov/programs-surveys/international-programs/about/idb/idbzip.zip`

En particulier on pourra s'intéresser au fichier `IDBext194.txt`

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*.**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 pays donné, calculer des statistiques de base sur l'âge de la population du type moyenne, médiane, variance...
- Pour une classe d'âge donnée, calculer des statistiques de base sur l'âge de la population du type moyenne, médiane, variance...
- Proposer une typologie des pays à 2 classes à l'aide d'une méthode de classification non-supervisée.
- Utiliser le modèle de régression linéaire pour prédire l'histogramme de répartition de la population dans un pays en fonction des données des années précédentes.

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.