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

Base de données du réseau routier californien

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 le graphe du réseau routier californien. Les données à récupérer se situent dans un fichier nommé

`roadNet-CA.txt.gz`

Le fichier est accessible sur la page

`http://snap.stanford.edu/data/roadNet-CA.html`

Une fois décompressé, il peut être chargé dans R (dans la variable `reseau`) 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 de la commande

```
reseau_bigdata <- to.dfs(reseau)
```

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 y a-t-il de nœuds, d'arêtes ? Est-ce que les arêtes sont orientées ou non orientées ?
- Quel est le nombre moyen de voisins par nœuds ? Donner la loi empirique du nombre de voisins par nœuds. Même questions avec les voisins situés à une distance maximale k pour k un entier (2 par exemple).

- Proposer/étudier une méthode d'échantillonnage des arêtes puis des nœuds (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.