

## Base de données du prix d'actions sur Yahoo! Finance

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 *rnr2* du logiciel R.

### 1 Récupération des données

Dans cet exemple, il est proposé d'analyser des données financières sur l'évolution journalière du prix d'une action d'une entreprise à partir d'informations fournies par Yahoo! Finance.

Il est par exemple possible de récupérer les données boursière de l'action Total sur la page <https://fr.finance.yahoo.com/quote/FP.SW/history?p=FP.SW> après avoir fixé l'historique le plus long possible.

Il est bien sûr possible de récupérer ce type de données pour beaucoup d'autres entreprises. Le but de ce projet est d'étudier l'évolution du cours des actions d'un grand nombre d'entreprises à partir des données fournies par Yahoo! Finance. **Une des première étapes de ce projet est donc d'identifier des entreprises qui sont répertoriées sur Yahoo! Finance.** Vous pourrez ensuite stocker l'ensemble des entreprises et l'évolution de leurs prix d'action dans R. En particulier, on veillera à regrouper toutes ces données au sein d'une même variable avant de la convertir au format *HDFS*.

### 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 *rnr2*. 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 chaque entreprise, quelle est son prix d'action le plus haut, le plus bas ? Quelle est la valeur moyenne de son action sur la période observée ?

- Pour un jour donné, quelle est l'entreprise dont le prix d'action est le plus haut, le plus bas ?
- Quelles sont les fréquences du pourcentage de changement du niveau de l'action entre l'ouverture et la fermeture d'une journée sur l'ensemble de données, et pour chaque entreprise ?
- Proposer une typologie des entreprises à l'aide d'une méthode de 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.