Travail Dirigé 3 - Travailler avec les rdd

Nicolas Martin - Spark core batch

Objectif : Maîtriser les traitements avec les RDD. Comprendre leur utilité et leurs spécificités.

Datasets utilisés :

- Arbres de paris - arbresalignementparis2010.csv - http://bit.ly/arbresparis
- Familles d'arbres - tree_familly.csv - MyGes

Questions :

1. Etudiez le Fichier. Que pouvez-vous en dire ?
2. Enlevez le header du fichier, puis Comptez le nombre d'arbres de la ville de Paris
3. Enlevez le header d'une autre manière, affichez les 20 premiers types d'arbres (le type d'arbre est représenté par la troisième colonne). Certains arbres n'ont pas de types
4. Affichez tous les types d'arbres sans doublons
5. Trouvez la taille totale de tous les arbres en utilisant .sum(), la taille est représentée par la huitième colonne
6. Trouvez la taille totale de tous les arbres en utilisant .reduce(), la taille est représentée par la huitième colonne
7. Calculez la taille moyenne des arbres. Vous pouvez aussi le faire avec reduce
8. Comptez le nombre d'arbres par type en utilisant countByValue
9. Comptez le nombre d'arbres par type en utilisant reduceByKey, ordonner le résultat par type d'arbre dans l'ordre alphabétique
10. Jointure avec les familles d'arbres, comptez les arbres par famille
11. Jointure avec les familles d'arbres en broadcastant les familles d'arbres, comptez les arbres par famille