Contrôle de connaissances

Intégration de données et Big Data

© Mourad Ouziri
mourad.ouziri@u-paris.fr

Quelques liens utiles :

Spark Scala API doc : https://spark.apache.org/docs/2.3.0/api/scala/index.html#package
Scala API doc : https://www.scala-lang.org/api/2.11.10/#package
Java API doc : https://docs.oracle.com/javase/7/docs/api/

Recommandation de bonnes pratiques de programmation :

- Programmer les traitements et calculs le plus possible dans des fonctions *scala*.
- Structurer les données sous forme d'objets.

QCM 1 (*5 pts, sans documents*) – Intégration de données

QCM 2 (*8 pts, sans documents*) – Big Data

Exercice 1 (*7 pts, avec documents*) – Programmation Spark *(durée 1h)*

Soit la liste de (employé, entreprise-employeur, salaire) suivante :

```
("Bob – EDF - 3000 €", "Marie – edf - 1500 €", "Imene – EDF - 9000 $", "Paul - Allianz – 4500 $", "Pierre - LaPoste - 1500 €", "Samir - Laposte - 3500 $", "Thi - LaPoste - 1500 €", "Eva - allianz - 6000 $", "Alice - edf - 5000 €", "Alice - Allianz - 1700 €")
```

Nous admettons le taux de change suivant : 1€ = 2$

Cette liste est à récupérer au format *.scala* de Moodle.

Travail demandé :

Programmer les calculs suivants (avec affichage des résultats demandés) avec Spark *Core* (*RDD*) :

1. Afficher le nom et le salaire des employés percevant un salaire inférieur ou égal à 2000 €. *(2 pts)*
2. Calculer le nombre d'employés par catégorie de salaires. Les catégories définies sont : la catégorie des salaires modestes (salaires strictement inférieurs à 2000 €), la catégorie des salaire moyens (salaires entre 2000 € et 4000 €, inclus) et la catégorie des salaires confortables (salaires strictement supérieurs à 4000 €). *(2 pts)*
   Le résultat attendu est le tableau des couples *(Catégorie-Salaire, Effectif)* suivant :
   [(Salaires-Modestes, 5), (Salaires-Moyens, 3), (Salaires-Confortables, 2)].
3. Rechercher les entreprises (leur nom seulement) qui ne paient que des salaires modestes. *(3 pts)*

Rendu du travail de l'exercice :

Le code réalisé doit être rendu dans le dépôt ci-dessous en un seul fichier nommé *votre-nom.scala* incluant la trace d'exécution (résultat ou message d'erreur) pour chaque question : https://cloud.parisdescartes.fr/index.php/s/ac9oJDn3kCtGiNN

Bonne chance !