Intégration de données

Durée 2h45

© Mourad Ouziri
Mourad.Ouziri@ParisDescartes.fr

Recommandations de programmation : Utiliser autant que possible

- *des objets pour structurer vos DataSet/DataFrame/RDD (éviter les chaînes de caractères)*
- *des fonctions scala pour structurer votre code.*

La qualité de code en dépend et sera prise en compte dans l'évaluation de votre travail.

QCM en ligne *(6 pts, environ 20mn)*

Répondre aux deux QCM en ligne suivants :
https://forms.gle/84bKgh18TcZEW3kU8
https://forms.gle/fPAZLGqPAKEEKYy68

Exercice – Programmation Spark *(14 pts, environ 2h20)*

Nous travaillons sur le fichier de données *Salaries.csv.* Le taux de change à prendre en considération est : 1 EUR = 1.1 USD

Il vous est demandé de programmer les calculs suivants (avec affichage des résultats) avec l'API Spark *Core* :

1. Charger les données du fichier dans un RDD d'objets.
2. Afficher (seulement le *nom et le pays*) les salariés français. Le nom complet du pays doit être affiché (remplacer FR par France, UK par Royaume Uni et USA par Etats-Unis).
3. Afficher (seulement le *nom et le salaire*) les salariés ayant un salaire inférieur ou égal à 3000 EUR.
4. Afficher (seulement le *nom*, la *date de naissance* et la *date d'embauche*) les salariés embauchés en étant mineurs.
5. Calculer (et afficher) le salaire moyen de tous les salariés.
6. Calculer (et afficher) le salaire moyen par pays. Afficher le *nom du pays* et le *salaire moyen*.

Rendu du travail

Le travail réalisé pour l'exercice 1 doit être rendu dans un seul fichier PDF (portant le nom du groupe) et indiquant pour chaque question le code *Spark-scala* réalisé et la trace d'exécution dans *spark-shell* montrant le résultat obtenu (ou la trace de l'erreur le cas échéant).

Ce fichier est à déposer dès la fin de l'épreuve dans : https://cloud.parisdescartes.fr/index.php/s/ifYKC45g7ZjbEso

Bonne chance !