Intégration de données et big data

Durée 2h

© Mourad Ouziri
Mourad.Ouziri@ParisDescartes.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

Recommandations de programmation :
utiliser autant que possible des fonctions scala pour structurer votre code.

QCM (en ligne, *5 pts*)

Exercice 1 (*10 pts*) – Interrogation de données avec Spark

Nous travaillons sur le fichier Personnels.csv (déjà traité en TP !).

Programmez les calculs suivants (avec affichage des résultats) avec Spark *Core* ou Spark SQL :

1. Afficher les nom et prénom de tous les personnels (noms entièrement en majuscules et seulement la première lettre pour les prénoms).
2. Afficher les personnels dont le prénom inclut la sous-chaine « ill ».
3. Afficher les personnels français.
4. Afficher les personnels nés après 2000.
5. Afficher l'âge moyen des personnels par pays.
6. Afficher le nom du personnel le plus âgé.

Exercice 2 (*5 pts*) – Qualité de données

Nous travaillerons sur le fichier *Personnels.csv*.

Les règles de qualité de données sont les suivantes :

- Un nom doit avoir au moins 2 caractères et commencer par une lettre alphabétique.
- Un personnel doit être majeur (18 ans ou plus) l'année de son embauche. Dans le cas contraire, l'année d'embauche est erronée. Ne considérer que l'année de naissance pour le calcul d'âge.
- Les bons codes de pays sont donnés dans le fichier de référence *Villes.csv*.

Travail demandé :

Ecrire du code Spark permettant de nettoyer les données du fichier en supprimant les personnels ne respectant pas au moins une des règles de qualité énoncées ci-dessus. Afficher les données après nettoyage.

Exercice 3 (bonus, *2 pts*) – Mise à jour de données

Mettre à jour les dates de naissances du fichier *Personnels.csv* avec celles données dans le fichier *Maj_Telephone.csv*. Nous supposons que le nom et le prénom des personnels constituent une clé de gestion.

Rendu du travail

Le travail réalisé doit être rendu dans fichier PDF (portant le nom de famille des binômes) indiquant pour chaque question le code écrit et la trace d'exécution sur *spark-shell* montrant le résultat obtenu.

Ce fichier est à déposer dans : https://cloud.parisdescartes.fr/index.php/s/zCJET33GopLjomS

Bonne chance !