Exercice : Manipulation de données avec Scala et Spark

L'objectif de cet exercice est d'apprendre à manipuler des fichiers, des chaînes de caractères, des tableaux, et des RDDs (Resilient Distributed Datasets) en Scala, tout en se familiarisant avec des opérations de jointures sur les RDDs dans Apache Spark.

Étape 1 : Lecture des fichiers texte avec Spark

Commande :

```
val files = sc.wholeTextFiles("/tmp/*.txt").collect.toList
```

Questions :

- Que fait cette commande ?
- Quel est le type de la variable `files` ?
- Imaginez qu'il y ait trois fichiers dans le répertoire `/tmp/` : `file1.txt`, `file2.txt` et `file3.txt`. Donnez un exemple du contenu possible de la variable `files`.

Étape 2 : Manipulation de chaînes de caractères

Commande :

```
var ch = "hello a b  c   bye"
ch.split("\\s+")
ch.split(" ")
```

Questions :

- Quelle est la différence entre ces deux opérations de découpage (`split("s+")` et `split(" ")`) ?
- Si `ch` contient `"hello scala world"`, quelles seront les sorties respectives des deux commandes ?

Étape 3 : Accès aux éléments dans un tableau de tuples

Commande :

```
var fruit = Array(("apple", "2"), ("orange", "3"), ("pear", "1"))
fruit(2)._2
```

Questions :

- Que renvoie cette commande et pourquoi ?
- Ajoutez le tuple `("banana", "4")` au tableau `fruit` et accédez à la valeur associée à la banane.

Étape 4 : Opérations sur les RDDs avec Spark

Commandes :

```
val a = Array((1, 2), (3, 4), (3, 6))
val b = Array((3, 9))
val A = sc.parallelize(a)
val B = sc.parallelize(b)
val J = A.join(B).collect
val L = A.leftOuterJoin(B).collect
```

Questions :

- Expliquez ce que font les opérations `join` et `leftOuterJoin` dans ce contexte.
- Quelle est la différence entre les résultats de `join` et `leftOuterJoin` ? Donnez les résultats obtenus pour les collections `A` et `B`.

Étape 5 : Défi supplémentaire

1. Modifiez les RDDs `A` et `B` pour inclure d'autres paires (comme `(5, 7)` pour `A` et `(1, 10)` pour `B`), et réexécutez les opérations `join` et `leftOuterJoin`. Quelles sont les nouvelles sorties ?
2. Écrivez une commande Scala qui calcule la somme des deuxièmes éléments de tous les tuples dans le RDD `A`.

Livrable attendu

Pour chaque étape, vous devez :

- Exécuter les commandes Scala et afficher les résultats obtenus.
- Fournir une explication claire de chaque commande et de ses résultats.

Bonne chance dans la réalisation de cet exercice !