



Évaluation 3

Titre du cours

Big Data 1

Numéro du cours

420-J35-RO

Programme

Big Data

Groupe

478

Prénom et nom de l'enseignant

Abderrazak Sahraoui

Date

2023-09-29

APERÇU

- Cette évaluation vise à mesurer votre habileté à programmer des processus ETL en utilisant des frameworks Hadoop, MapReduce, MRjob et Python et/ou tout autre outils que vous jugez pertinent pour du big data.
- Cette évaluation est à faire en équipe de deux à trois étudiants.
- Ne partagez pas votre copie avec les autres équipes.

PARTIE 1

1. Créer un programme (Ev3_eqX_Partie1.py) MapReduce en Python pour lire les données du fichier LivresAuteursLangue.txt et pour effectuer les opérations suivantes :
2. Nettoyer les données :
 - a. Enlever les informations étiquetées inconnues. (si l'auteur est marqué comme inconnu, il faudrait supprimer ce mot du fichier. Même chose pour la langue originale ou le titre original.
 - b. Enlever la partie langue originale et la partie titre original si la langue originale est français.
3. Attribuer un numéro à chaque livre en utilisant une séquence de 1 à 100.
4. Diriger les sorties de votre programme vers un fichier LivresPartie1_eqX.txt

Note 1 : remplacer X dans les noms des programmes par le numéro de votre équipe.

Note 2 : Utiliser le protocole mrjob ByteValueProtocol pour vos sorties afin de permettre un affichage correct des caractères accentués. Voir exemple en Annexe.

Note 3 : Si votre fichier de sortie LivresPartie1_eqX.txt affiche un encodage UTF-16 LE. Enregistrer votre fichier avec un encodage UTF-8 dans VSCode. Voir procédure en annexe.

PARTIE 2

1. Créer un programme (Ev3_eqX_Partie2.py) MapReduce en Python pour lire les données du fichier LivresPartie1_eqX.txt et pour effectuer les opérations suivantes :
2. Transformer les données pour obtenir le format de la table **livres3** vue en cours sur Hive :

```

Livres3.txt
1  74,Français:La Trilogie de l'Empire#Anglais:The Empire Trilogy,Raymond E. Feist#Janny Wurts
2  76,Français:Le Cycle de l'Âge de la Mort#Anglais:The Death Gate Cycle,Margaret Weis#Tracy Hickman
3  78,Français:La Trilogie de l'Éveil,Pauline Alphen
```

Dans ce format, on regroupe le titre en français et le titre original en une collection de type map. Les auteurs sont regroupés en une collection de type array.

3. Diriger les sorties de votre programme vers un fichier LivresPartie2_eqX.txt

Note 1 : remplacer X dans les noms des programmes par le numéro de votre équipe.

Note 2 : Utiliser le protocole mrjob ByteValueProtocol pour vos sorties afin de permettre un affichage correct des caractères accentués. Voir exemple en Annexe.

Note 3 : Si votre fichier de sortie LivresPartie2_eqX.txt affiche un encodage UTF-8 mais n'affiche pas correctement le contenu, passer en affichage avec encodage UTF-16 LE. Voir procédure en annexe.

PARTIE 3

1. Créer un programme (Ev3_eqX_Partie3.py) MapReduce en Python pour lire les données du fichier LivresPartie1_eqX.txt et pour effectuer les opérations suivantes :
2. Transformer les données pour obtenir le format de la table **livres4** du cours sur Hive :

```
livres4.txt
1  74,La Trilogie de l'Empire,Raymond E. Feist#Janny Wurts,Anglais#The Empire Trilogy
2  76,Le Cycle de l'Âge de la Mort,Margaret Weis#Tracy Hickman,Anglais#The Death Gate Cycle
3  78,La Trilogie de l'Éveil,Pauline Alphen
```

Dans ce format, on garde le titre en français séparé du titre original qui est stocké en tant que type struct composé de deux champs langue original et titre original. Les auteurs sont regroupés en une collection de type array.

3. Diriger les sorties de votre programme vers un fichier LivresPartie3.txt

Note 1 : remplacer X dans les noms des programmes par le numéro de votre équipe.

Note 2 : Utiliser le protocole mrjob ByteValueProtocol pour vos sorties afin de permettre un affichage correct des caractères accentués. Voir exemple en Annexe.

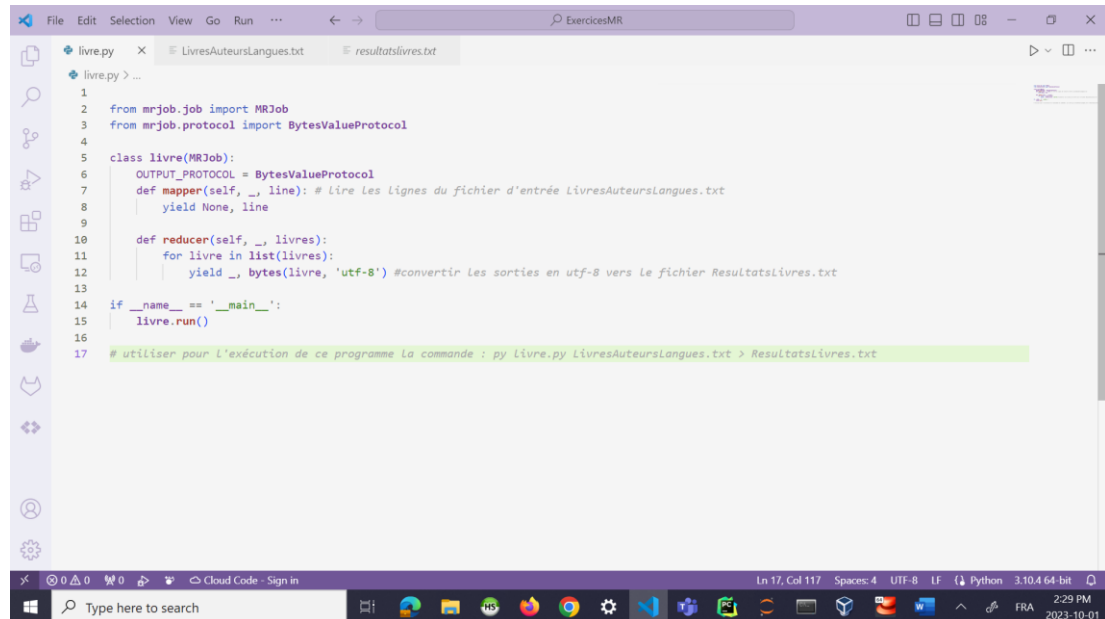
Note 3 : Si votre fichier de sortie LivresPartie3_eqX.txt affiche un encodage UTF-8 mais n'affiche pas correctement le contenu, passer en affichage avec encodage UTF-16 LE. Voir procédure en annexe.

PARTIE 4

1. Créer un script HiveQL (ev3_Partie4_eqX.hql) pour créer une base de données LibrairieEv3EqX contenant une table livresev3eqX. La table doit avoir un schéma similaire au schéma de livres3 ou livres4 du cours sur Hive.
2. Charger dans votre table les données du fichier LivresPartie2_eqX.txt si votre table est de type livres3 ou celles de LivresPartie3_eqX.txt si votre table est de type livres4.
3. Créer un script de requêtes d'interrogation ev3_Requetes_eqX.hql avec les requêtes suivantes :
 1. Afficher les livres ayant deux auteurs.
 2. Afficher les livres ayant une version originale en espagnol.
 3. Afficher les livres n'ayant pas de version originale autre que le français.

ANNEXE 1

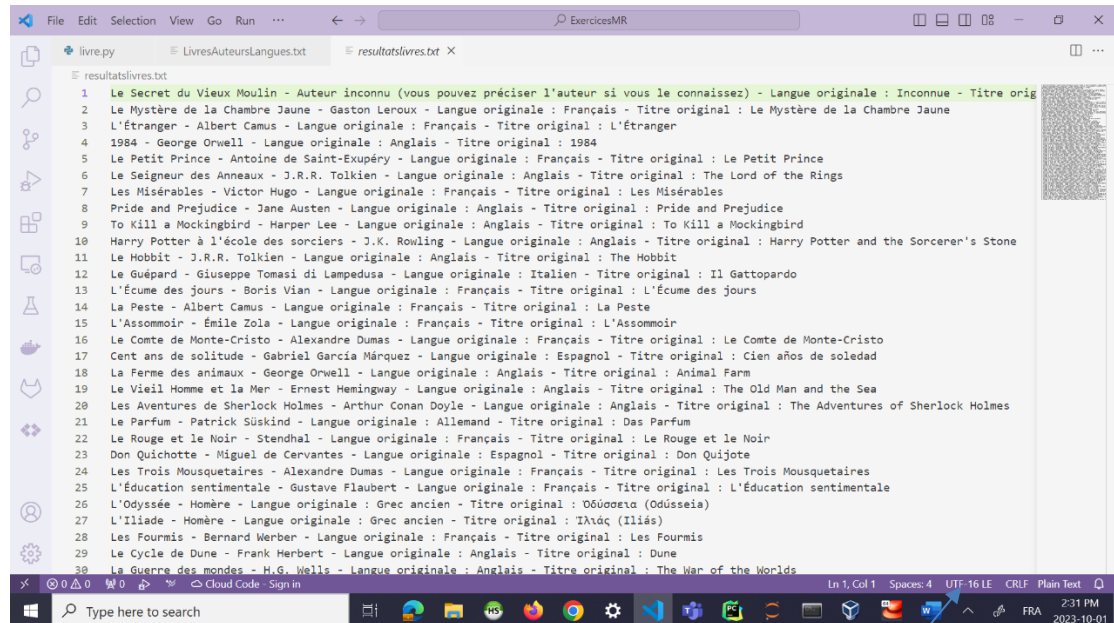
Utiliser le protocole mrjob ByteValueProtocol pour vos sorties afin de permettre un affichage correct des caractères accentués. Voir exemple ci-dessous :



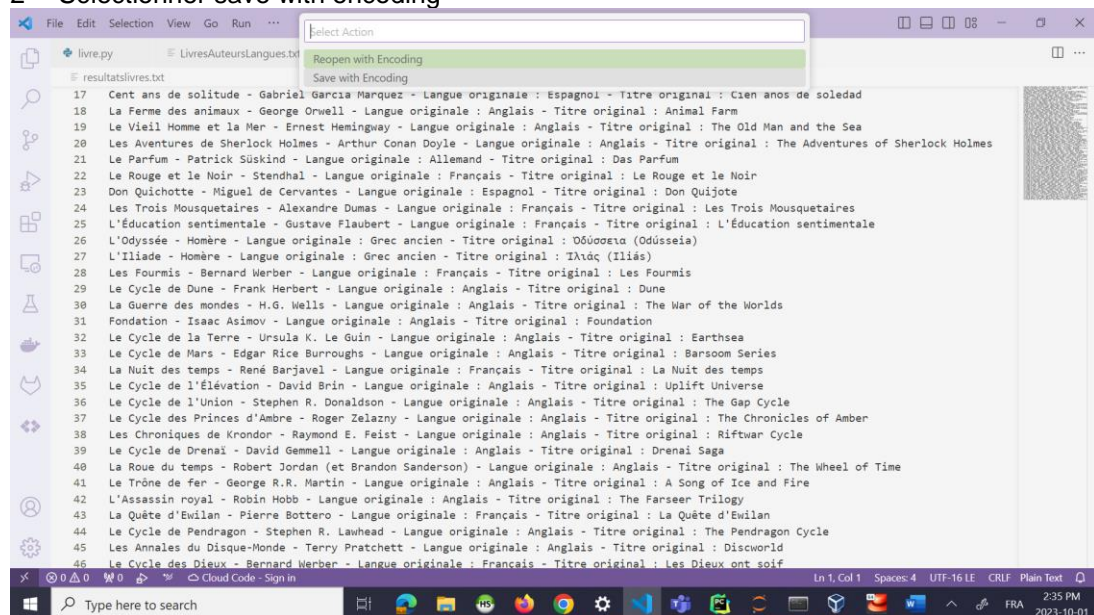
```
1 from mrjob.job import MRJob
2 from mrjob.protocol import ByteValueProtocol
3
4
5 class livre(MRJob):
6     OUTPUT_PROTOCOL = ByteValueProtocol
7     def mapper(self, _, line): # Lire Les lignes du fichier d'entrée LivresAuteursLangues.txt
8         yield None, line
9
10    def reducer(self, _, livres):
11        for livre in list(livres):
12            yield _, bytes(livre, 'utf-8') #convertir Les sorties en utf-8 vers Le fichier ResultatsLivres.txt
13
14    if __name__ == '__main__':
15        livre.run()
16
17 # utiliser pour l'exécution de ce programme la commande : py livre.py LivresAuteursLangues.txt > ResultatsLivres.txt
```

ANNEXE 2

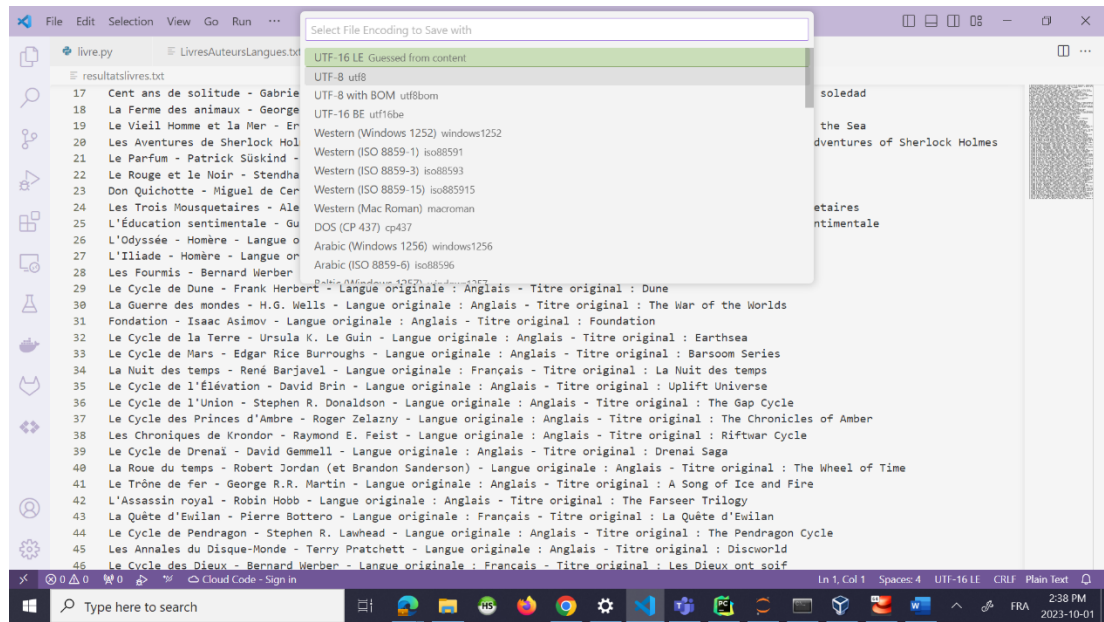
Si votre fichier de sortie LivresPartie1_eqX.txt affiche un encodage UTF-16 LE. Enregistrer votre fichier avec un encodage UTF-8 dans VSCode.



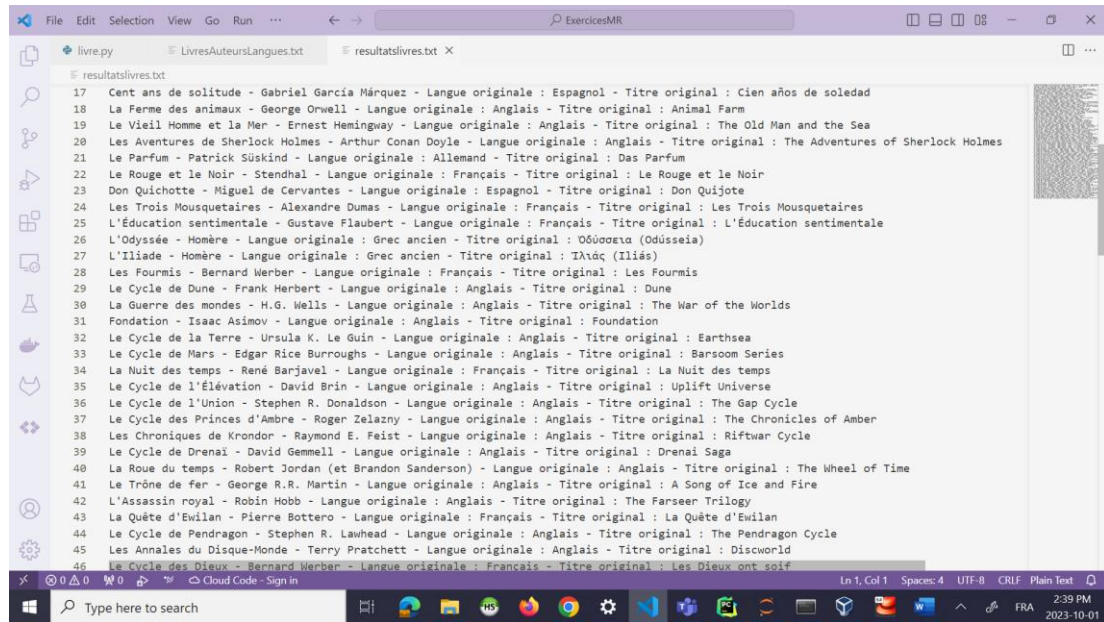
- 1- Cliquer sur le texte UTF-16 LE qui se trouve sur la barre violette en bas de l'écran
- 2- Sélectionner save with encoding



- 3- Sélectionner utf-8

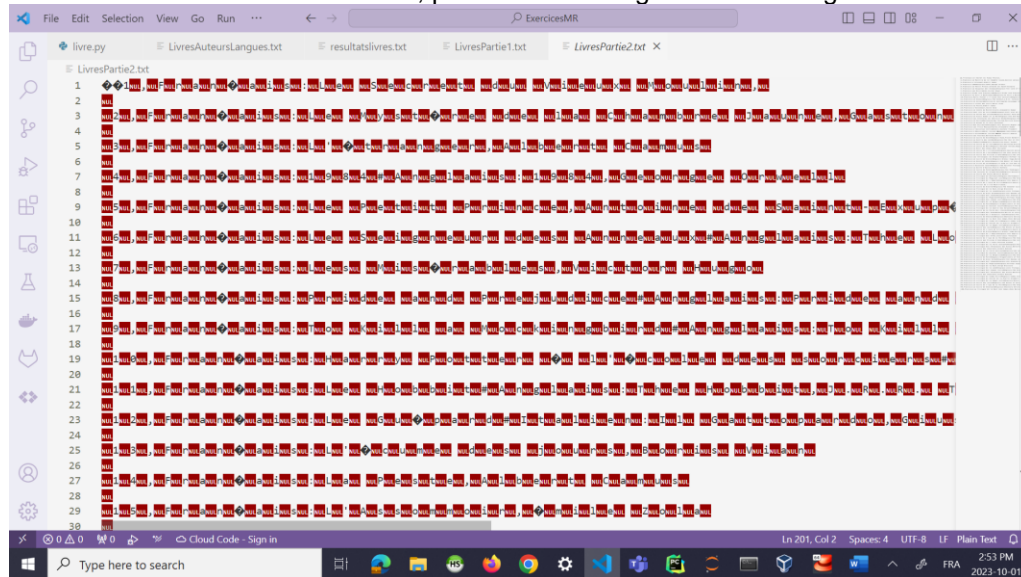


Et votre fichier prendra l'encoding utf-8

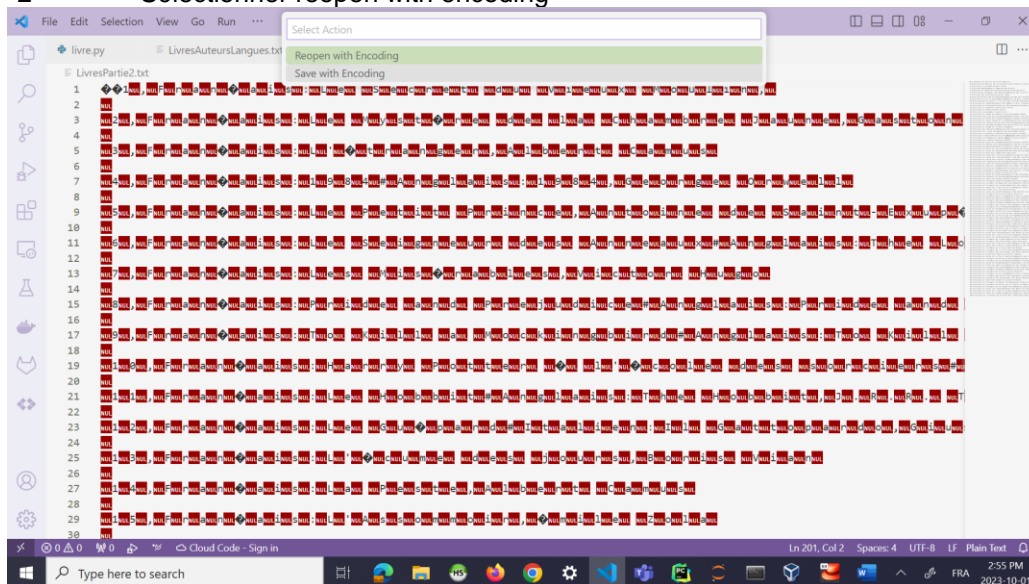


ANNEXE 3

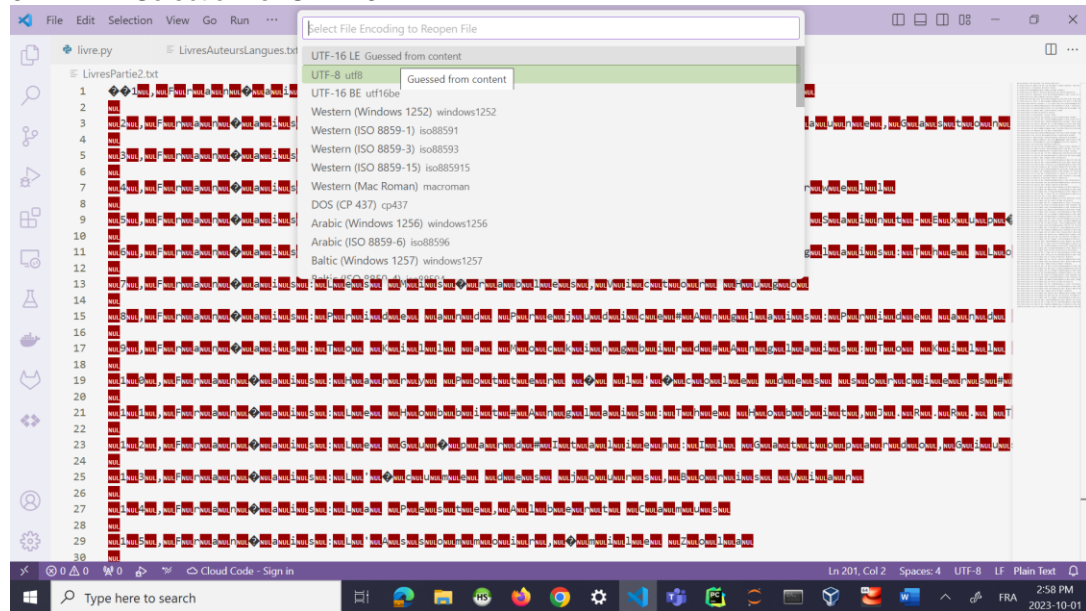
Si votre fichier de sortie LivresPartie3_eqX.txt affiche un encodage UTF-8 mais n'affiche pas correctement le contenu, passer en affichage avec encodage UTF-16 LE.



- 1- Cliquer sur le texte UTF-8 qui se trouve sur la barre violette en bas de l'écran
- 2- Sélectionner reopen with encoding



3- Sélectionner UTF-16 LE



Votre fichier s'affiche correctement

