MP1. Big Data & Ecommerce

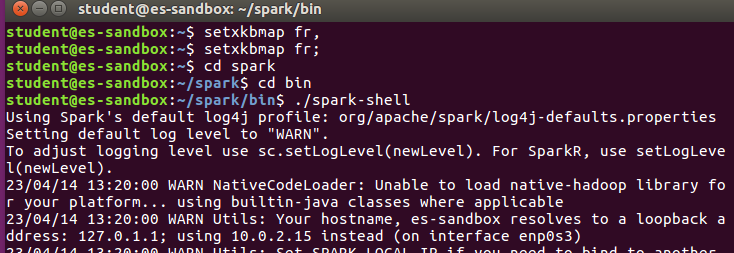
2022-2023

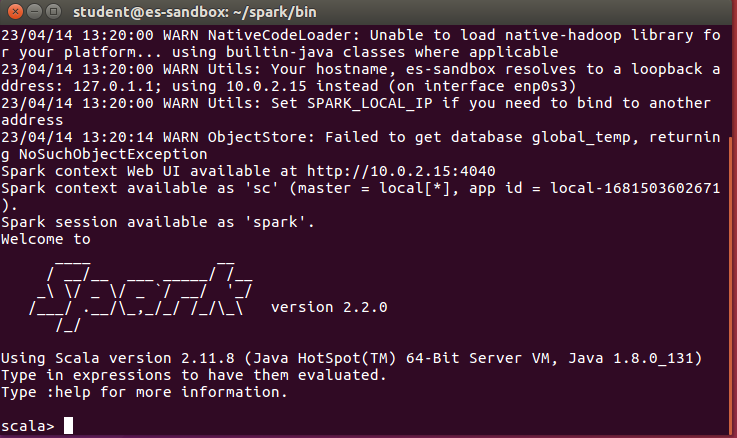
IHEC CARTHAGE

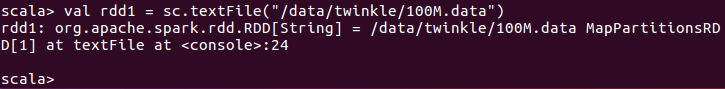
Spark TP 1

(SCALA & Python)

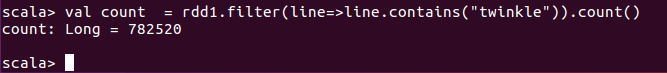
1. Charger le fichier texte simple /data/twinkle/100M.data fourni sous le répertoire racine en utilisant un RDD nommé rdd1.







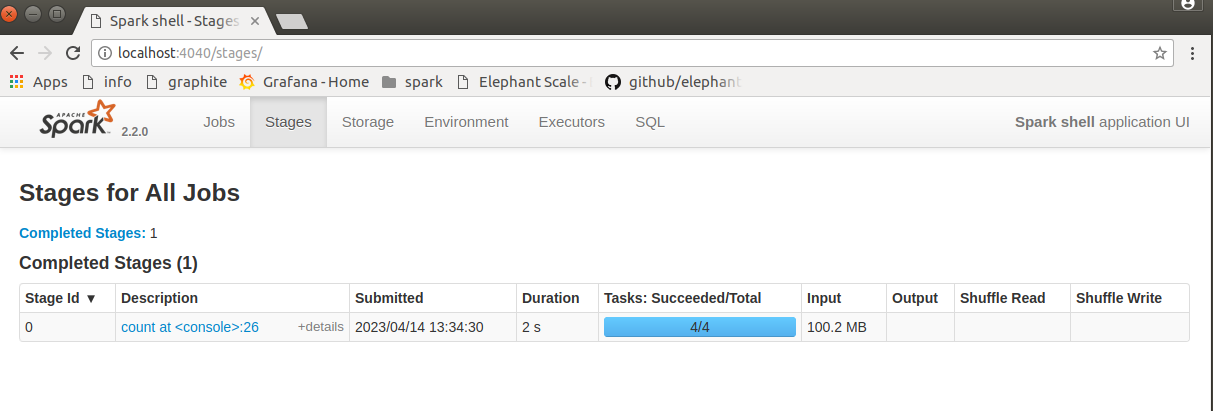
2. Déterminer le nombre de lignes qui contiennent le mot 'twinkle'. 3. Après avoir taper ces commandes



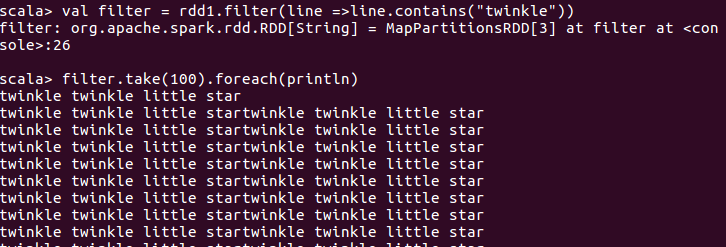
a) Aller à l'interface utilisateur du shell Spark

b) Inspecter la section 'Stages' dans l'interface utilisateur.

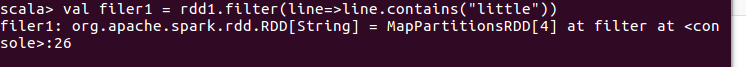
c) Comment le filtre est-il exécuté ? Pouvez-vous expliquer le comportement ?

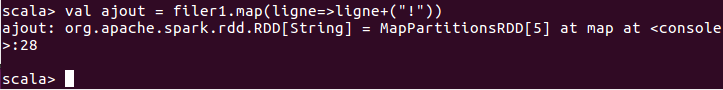


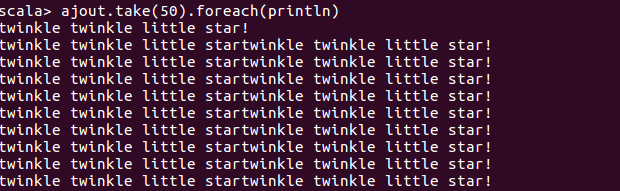
4. Afficher 100 lignes contenant le mot 'twinkle'.



5. Afficher 50 lignes contenant le mot « little » en ajoutant à la fin de chaque ligne « ! ».



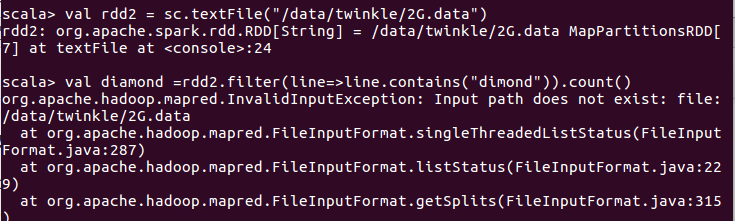




6. Démarrer le shell avec plus de mémoire

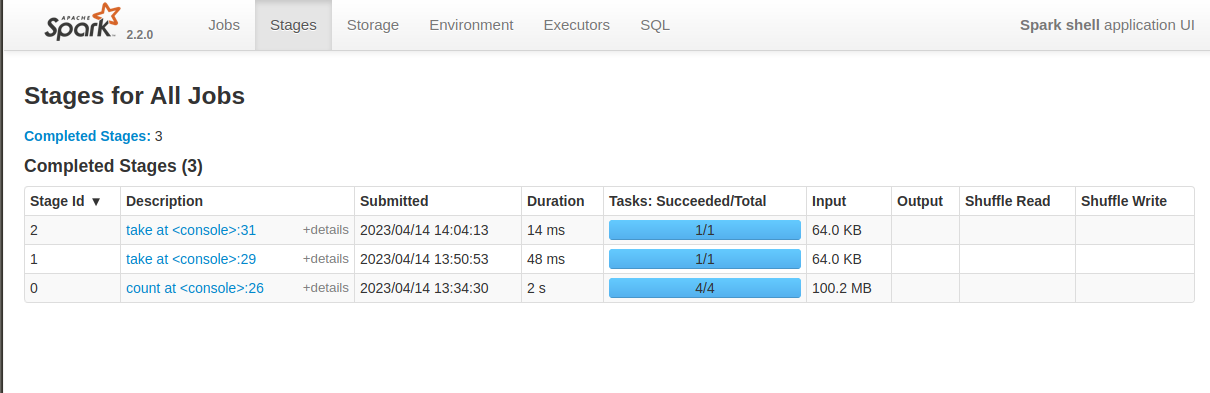
7. Charger le fichier (plus grand) /data/twinkle/2G.data en utilisant un RDD nommé rdd2.

8. Compter le nombre de lignes qui contiennent le mot "diamond"



9. Combien de « tâches » sont utilisées dans le calcul ci-dessus

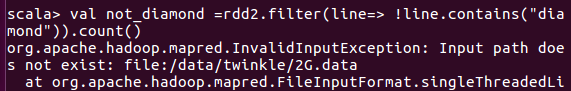
Pour ce calcul spark a utilisé 4 tâches d’après la section stages comme il est mentionné ci-dessus



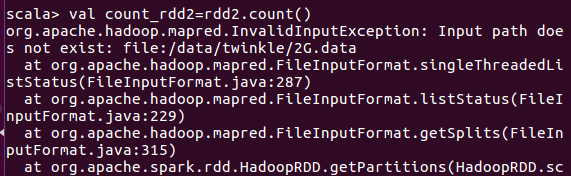
10.Pouvez-vous expliquer le nombre de tâches ?

La création de partitions dépend du nombre de partitions créées. Dans ce cas, RDD2 a créé 4 partitions, dont 3 pour l'opération filter() et 1 pour l'opération count().

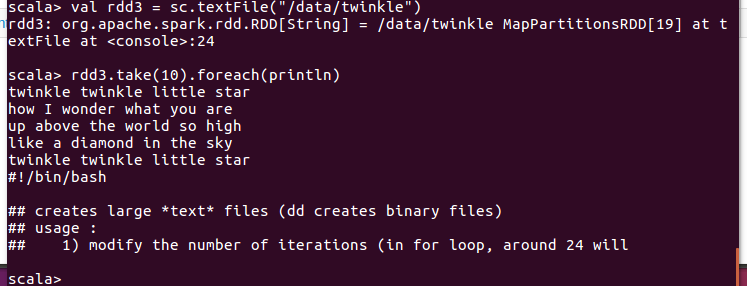
11.Compter le nombre de lignes qui ne contiennent pas le mot "diamond"



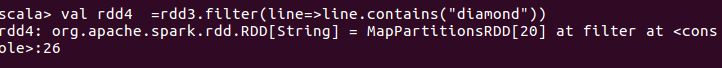
12.Vérifier que la somme des lignes contenant le mot diamond et celle qui ne contiennent pas est égale au nombre de ligne de rdd2.



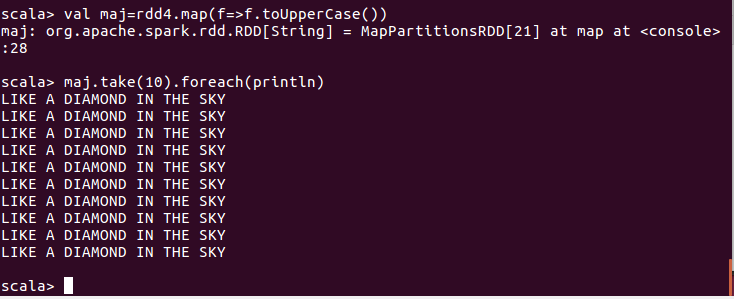
13.Charger tous les fichiers \* .data dans le répertoire /data/twinkle en utilisant un RDD nommé rdd3.



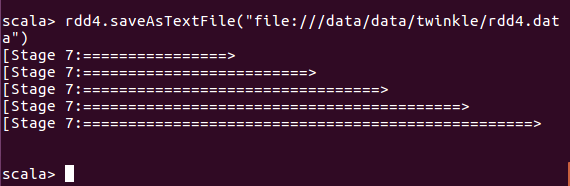
14.Créer un nouveau RDD, rdd4, en filtrant le rdd3 pour le mot «diamond».

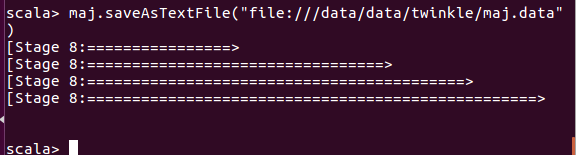


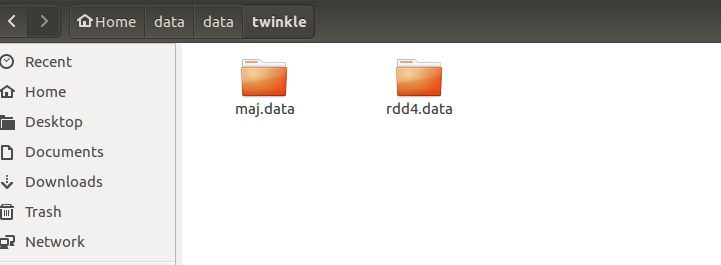
15.Afficher 20 lignes des données du rdd4 tel que tout le contenu est en majuscule.



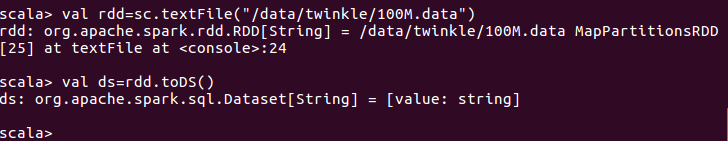
16. Enregistrer le résultat du filtre, rdd4, sur le disque et vérifier le contenu du répertoire de sortie.



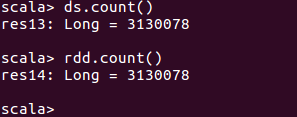




17.Charger le fichier texte simple /data/twinkle/100M.data fourni sous le répertoire racine en utilisant un rdd puis un dataset.



18.Calculer le nombre de lignes avec chaque modèle de données puis comparer entre eux.

****