

TME 3 : Jointures parallèles sous Spark (suite) - Traduction de SQL vers MR

Prérequis :

Consulter la rubrique Informations pratiques/Notes sur les TD-TME de la page de l'UE pour les instructions de démarrage.

Pour ce TME, **se connecter au serveur Spark distant (instructions de la sous-rubrique Utilisation à distance)**.

Les identifiants et mots de passe sont communiqués pendant le TME.

Prendre la peine de vérifier l'**interface web** après exécution des instructions qui entraînent une action (*count* par exemple) et **noter** les observations.

Exercice 1: Jointures

Reprendre l'exercice sur les jointures de la planche précédente et tester sur le cluster. Vous pourrez varier les paramètres de partitionnement en entrée de `sc.textFile` et `reduceByKey` et constater la différence.

Exercice 2 : Jointures

Télécharger le fichier `network.gz` depuis `/home/etudiant/data`. Il s'agit d'une table de tripletts (u,v,poids) indiquant qu'il existe un arc de u vers v d'importance poids.

Formuler une jointure en MR qui retrouve les chemins de longueur 2 puis 3.

Pour la dernière tester la différence entre une exécution mon-passe et une avec plusieurs étapes.

Le prochain exercice est à faire en local (vous déconnecter du cluster)

Exercice 3 : Allez sur <https://github.com/YSmart/YSmart> et suivre les instructions pour installer l'outil YSmart. Tester sur les exemples fournis.