TP 01 : Initiation au framework hadoop et au patron MapReduce

1. Objectifs:

- Le premier objectif de ce TP est de configurer et d'installer Hadoop en mode single-node afin de pouvoir rapidement effectuer des opérations simples en utilisant Hadoop MapReduce et le système de fichiers distribués Hadoop (HDFS).
- Deuxième objectif est de vous familiariser avec l'utilisation de HDFS en mode *single-node* pour le traitement des données. Vous allez travailler avec un ensemble un fichier TSV (515 319 essais cliniques) téléchargé depuis le site clinicaltrials.gov, qui contient des informations sur divers essais cliniques. Vous allez analyser ces fichiers en utilisant MapReduce pour effectuer diverses tâches telles que l'identification des données manquantes, le calcul de distributions et la séparation des fichiers selon des critères spécifiques.
- L'installation et la configuration d'un cluster Hadoop.
- Cluster Hadoop vs single-node : apprendre à mesurer le temps d'exécution de chaque job MapReduce dans Hadoop afin d'optimiser les performances et mieux comprendre le fonctionnement du traitement distribué.

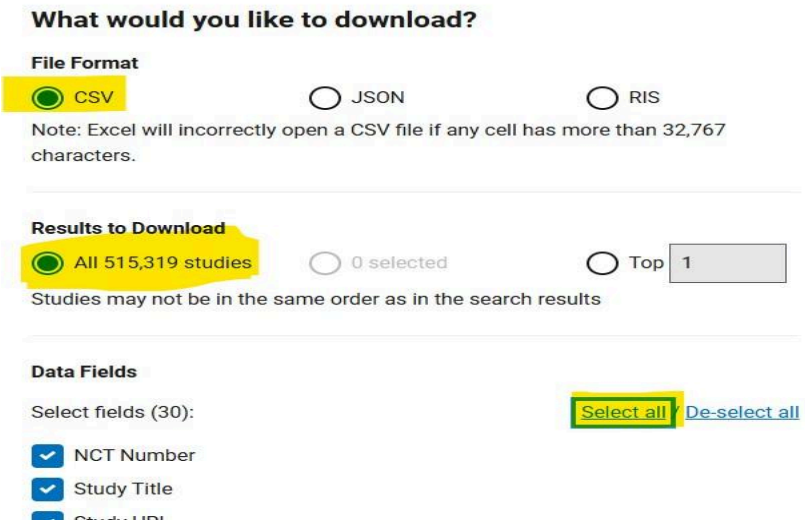
➔ Ce Tp est noté, n'oubliez pas de soumettre votre compte rendu.

2. Installer un cluster Hadoop en single-node:

Pour installer un **cluster Hadoop en single-node** sur votre machine, suivez les instructions de la documentation officielle d'Apache Hadoop.

https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html.

données: https://clinicaltrials.gov/search



3. Hadoop HDFS en mode single-node

Vous devez télécharger les données essais cliniques depuis clinicaltrials.gov. Un fichier contient les détails des essais cliniques, tels que la localisation, la phase de l'essai, les dates de début et de fin, les traitements testés, etc. Prenez le temps de comprendre vos données en consultant le site.

https://clinicaltrials.gov/study-basics/glossary

Une fois le fichier téléchargé, vous le organisez dans un répertoire appelé input/clinical_trials/.

Vous devez calculer et noter (dans votre compte rendu) le temps d'exécution de chaque job en utilisant l'une des méthodes suivantes :

- Utilisez les informations des logs Hadoop dans le ResourceManager UI.
- Utilisez la commande time dans le terminal avant d'exécuter le job.
- Ajoutez du code dans votre programme MapReduce.

Remarque **:**

Vous êtes libres de choisir d'autres structures de clé-valeur si vous trouvez que cela est plus adapté à votre logique.

Q1 : Écrivez un programme MapReduce qui permet de détecter les essais dont la colonne 'Start Date' est vide ou contient des valeurs nulles.

Consignes :

- Pour chaque ligne de chaque, vérifiez si la colonne 'Start Date' est vide ou nulle.
- Si la colonne 'Start Date' est vide ou nulle, le programme doit marquer cet essai comme contenant des données manquantes.
- Vous devrez émettre une clé (l'id de l'essai) et une valeur (indiquant que des données manquantes ont été trouvées dans ce fichier).
- La fonction de réduction doit récupérer les essais émis par le mappage et identifier ceux qui contiennent des données manquantes dans la colonne 'Start Date'.
- À la fin de l'exécution du programme MapReduce, les essais contenant des valeurs manquantes dans la colonne 'Start Date' doivent être déplacés dans un nouveau fichier dans input//clinical_trials/essais_avec_valeurs_mq.

Q2 : Écrire un programme MapReduce pour déterminer combien d'essais cliniques appartiennent à chaque phase (par exemple, Phase 1, Phase 2, Phase 3).

Consignes :

1. Le mapper doit émettre une paire clé-valeur pour chaque essai clinique. La clé sera la phase (par exemple, "Phase 1"), et la valeur sera le nombre "1", représentant un essai clinique dans cette phase.
2. Le **reducer** doit agréger les résultats pour chaque phase en additionnant les "1" associés à chaque clé (phase). Cela donnera le nombre total d'essais cliniques dans chaque phase.
3. Le résultat doit être un fichier JSON avec les phases comme clés et le nombre d'essais dans chaque phase comme valeurs.

4. Cluster Hadoop de N machines

>----------prochaine séance----------<