Projet en groupe de INF 356 & IGE 356 : Ingénierie des données

I- Première partie

On souhaite établir des statistiques sur la longueur des mots dans un document, ou un ensemble de documents, à partir d'une application *Map-Reduce*. Proposez un algorithme dans le paradigme *Map - Reduce* pour les 3 analyses suivantes :

1. **Question 1 :** Compter le nombre de mots de chaque longueur présente dans le texte (en vue d'établir un histogramme des longueurs de mots).

2. **Question 2 :** Compter le nombre de mots de 1 à 5 caractères (inclus), de 6 à 10 caractères (inclus), de 11 à 15 caractères (inclus) et de plus de 15 caractères présents dans le texte.
   Remarque : si on génère plusieurs fichiers de sorties, il est nécessaire que l'ordre des noms de fichiers soit celui des sous-ensembles de longueurs de mots.

3. **Question 3 :** Obtenir les listes de mots de 1 à 5 caractères (inclus), de 6 à 10 caractères (inclus), de 11 à 15 caractères (inclus) et de plus de 15 caractères présents dans le texte. Il n'est pas demandé de trier les mots à l'intérieur d'une liste, ni d'éliminer les doublons.

NB : Dans tous les cas on décrira les paires clé-valeur et le pseudo-code des traitements utilisés à chaque étape de la solution *Map-Reduce*.

II- Deuxième partie

Faire une étude comparative de deux systèmes de gestion de données noSQL Cassandra et MongoDB. Vous devez ressortir le principe de fonctionnement de chacun d'entre eux en spécifiant les différences dans leurs paradigmes. Expliquer les différents opérations (Insertion, Suppression et Sélection) pour chacun de ces systèmes de gestions de données.

III- Troisième partie

Faite une étude comparative des écosystèmes Hadoop et Spark. Quelles sont les similitudes et les différences entre ces deux écosystèmes ?

NB : Date limite pour remettre le projet : **Mardi le 30 Avril 2024 avant minuit**. Devoir à faire en groupe de 3.