Bases de Données Réparties - 8INF803 UQAC

- Automne 2023-

Projet 2 – barème : sur **100 -** pondération : **0.5**

Documents à rendre :

- Tous les fichiers source du projet
- Une vidéo enregistrée en guise de démonstration des différentes étapes de configuration, des différentes phases de créations et des différents tests.

Notes importantes :

- Date de remise du projet : **avant le 11 décembre 2023 à 23H55.**
- Ce travail doit se faire en groupe (minimum 3 et maximum 4 personnes)
- Dans la vidéo démo, tous les étudiants du groupe en question doivent participer à la présentation et allumer leurs webcams pour valider leur présence.
- Un étudiant qui ne participe pas à la démo, sera pénalisé par un zéro.

Partie 1 : Hadoop-MapReduce (50 points)

Télécharger les data sets « title.basics.tsv » et « title.ratings.tsv » à partir du lien suivant :

https://datasets.imdbws.com/

Consulter la documentation des data sets qui se trouve dans le lien suivant :

https://www.imdb.com/interfaces/

Ecrire des programmes MapReduce qui permettent d'obtenir les résultats suivants :

1. Le nombre d'éléments que contient chaque type de titre. Le résultat doit être affiché par ordre décroissant du nombre de films.

2. Le nombre total de films/vidéos sortis par année. Les années doivent être affichées par ordre croissant.

3. Le nombre de films pour chaque intervalle de notes :

[0,1],]1,2],]2,3],]3,4],]4,5],]5,6],]6,7],]7,8],]8,9],]9,10].

Le nombre de films doit être affiché pour chaque intervalle est dans l'ordre donné.

Partie 2 : Apache Spark – Trouver les jobs les plus rémunérateurs en Data Science. (50 points)

Consigne :

1. Créer un fichier python qui contient le code

2. Exécuter le fichier python dans un cluster Spark

3. Après avoir importé les données (des deux fichiers **Salaries-2020-2022.csv** et **Salaries-2023.csv**) dans un data frame, interroger via SQL pour répondre aux

questions suivantes puis visualiser les données quand nécessaire (e.g, courbe, diagramme à barres, etc.) :

a. Remplacer les valeurs de la colonne "employment_type" comme suit :

'FT': 'Full-time', 'PT': 'Part-time', 'C': 'Contract', 'I': 'Internship', 'F': 'Freelance', 'CT': 'Contract', 'FL': 'Freelance'}

b. Quels titres d'emploi (top 10) payent le mieux ?

c. Quels sont les 10 titres d'emploi les plus courants ?

d. Vérifier les emplois les plus et les moins rémunérateurs par année.

e. Quel est le salaire moyen par Job Title ?

f. Dans quels pays (top 3), un ingénieur en ML (ML Enginner) est-il le mieux rémunéré ?

g. Calculer le salaire moyen par niveau d'expérience.

h. Travailler à distance impacterait-il le niveau de salaire pour le même Job Title ?