République Algérienne Démocratique & Populaire
Ministère de l'Enseignement Supérieur & de la Recherche Scientifique
Ecole Supérieure en Informatique 08 mai 1945 de Sidi Bel Abbes

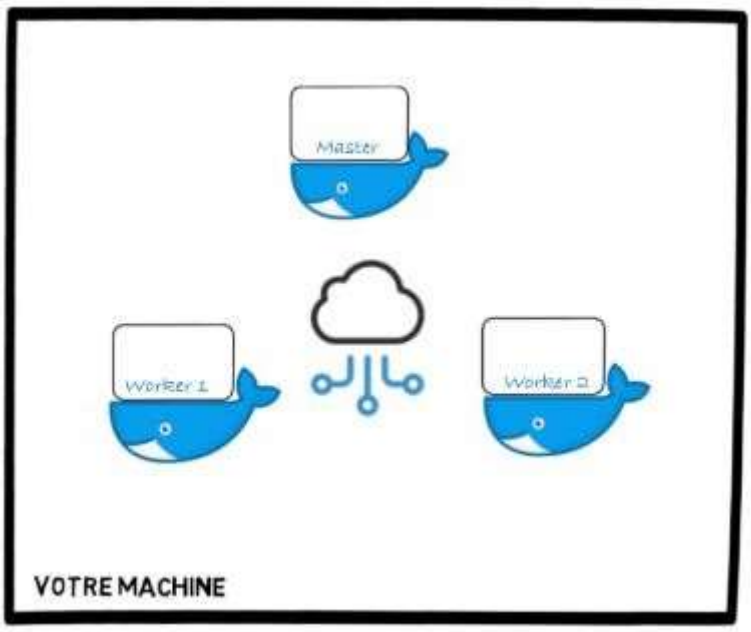
Année universitaire : 2025-2026 **Semestre 1**

3 SC : SIW & ISI **Module : Big Data**

TP 2 : Spark RDD

Installation de Spark sur un cluster :

Nous allons procéder à l'installation de Spark sur un cluster, c'est à dire un ensemble de machines interconnectées, représentées dans notre cas par des conteneurs Docker. L'objectif sera donc de créer un réseau de conteneurs, installer Spark dessus, et lancer les processus sur les différents conteneurs, de façon à obtenir le cluster suivant:



- Créer un réseau qui permettra de connecter les trois nœuds du cluster.

- Créer et lancer trois conteneurs (vous devez utiliser la même image du TP1 et exposer le port 8080 sur le nœud Master).

Configurer Spark :

- Créer le fichier de configuration **slaves** dans le répertoire **/usr/local/spark/conf**.
- Ajouter dans le fichier **slaves** les noms des conteneurs **workers.**
- Démarrer les services Spark sur tous les nœuds.
- Envoyer un programme Python au cluster Spark.

Spark RDD :

Ecrire en Python un programme qui permet de :

(1) créer un RDD à partir du fichier **arbres.csv** et afficher le nombre de lignes à partir du RDD créée.
(2) calculer et afficher la hauteur moyenne des arbres.
(3) afficher le genre du plus grand arbre : le principe est de construire des paires (clé, valeur), avec ici la hauteur des arbres comme clé et leur genre en tant que valeur. Ensuite, on classe les paires par ordre de clé décroissante grâce à **sortByKey** et on garde seulement la première paire.
(4) afficher le nombre d'arbres de chaque genre : le principe est de construire une paire **(genre, 1)** par arbre du fichier, puis de cumuler les valeurs genre par genre avec **reduceByKey**.