TP 1 : Manipulation du système de fichiers distribué HDFS

Problématique

Une entreprise dispose d'un cluster Hadoop configuré avec le système de fichiers distribué **HDFS** afin de stocker et gérer de grands volumes de données de manière fiable et distribuée. Votre mission consiste à **prendre en main HDFS** en réalisant les opérations de base de gestion des fichiers à l'aide des commandes HDFS.

L'objectif est de comprendre comment les données sont organisées, stockées et manipulées dans HDFS par rapport à un système de fichiers classique.

Objectifs du TP :

À l'issue de ce TP, l'étudiant sera capable de :

- créer et organiser des répertoires dans HDFS,
- transférer des fichiers entre le système local et HDFS,
- consulter et analyser les fichiers stockés dans HDFS,
- effectuer des opérations de gestion des fichiers dans HDFS,
- comprendre les notions de blocs et de réplication.

Environnement de travail

- Cluster Hadoop exécuté via Docker
- Accès au conteneur NameNode
- Utilisation des commandes HDFS (hdfs dfs)

Partie 1 — Création de l'espace de travail HDFS

1. Créer dans HDFS un répertoire principal nommé **/tp-hdfs**.
2. Créer dans HDFS la structure de répertoires suivante :

   /tp-hdfs

   ├── input

   ├── output

   └── archive

3. Vérifier que la structure de répertoires a été correctement créée.

Partie 2 — Transfert de données vers HDFS

4. Créer localement un fichier texte contenant au minimum dix lignes de texte.

5. Copier ce fichier depuis le système de fichiers local vers le répertoire **/tp-hdfs/input** dans HDFS.
6. Lister le contenu du répertoire **/tp-hdfs/input** afin de vérifier la présence du fichier.

Partie 3 — Consultation et analyse des fichiers HDFS

7. Afficher le contenu du fichier stocké dans HDFS sans le rapatrier en local.
8. Afficher les informations détaillées du fichier HDFS (taille, propriétaire, date, facteur de réplication).
9. Expliquer brièvement :
    - pourquoi HDFS découpe les fichiers en blocs,
    - le rôle du facteur de réplication dans la tolérance aux pannes.

Partie 4 — Gestion des fichiers dans HDFS

10. Copier le fichier présent dans **/tp-hdfs/input** vers le répertoire **/tp-hdfs/archive**.
11. Renommer le fichier copié dans le répertoire **/tp-hdfs/archive**.
12. Déplacer le fichier original du répertoire **/tp-hdfs/input** vers le répertoire **/tp-hdfs/output**.
13. Vérifier la présence des fichiers dans les répertoires **/tp-hdfs/output** et **/tp-hdfs/archive**.

Partie 5 — Suppression et nettoyage

14. Supprimer un fichier présent dans HDFS.
15. Supprimer le répertoire **/tp-hdfs/archive** ainsi que son contenu.
16. Vérifier que les éléments supprimés ne sont plus accessibles dans HDFS.

Travail demandé

- Réaliser l'ensemble des questions
- Noter les commandes HDFS utilisées
- Ajouter des captures d'écran justifiant chaque étape importante

Livrables

- Un document PDF ou Word contenant :
    - les commandes utilisées,
    - les captures d'écran,
    - de courtes explications associées aux questions.