TP: PySpark pour analyse de données de films.

Objectif

Utiliser PySpark pour manipuler un jeu de données sur les films, effectuer des transformations simples et réaliser des analyses.

Étapes

1. Prise en main: Préparation des données

- Charger le fichier CSV contenant les données des films dans un DataFrame PySpark.
- Nettoyer les données en supprimant les lignes vides ou invalides.
- Convertir les dates dans le format `yyyy-MM-dd`.

2. Intermediaire: Manipulation des données & analyse

- **Filtrage :**
  - Les films ayant une note très basse (tomatometer_rating < 20).
  - Les films sortis (en cinéma) après l'année 2000.
- **Moyennes :**
  - La note moyenne des films par studio.
  - La note moyenne des films par directeur.

3. Avancé: Utilisation de fonctions avancées

- Diviser les genres multiples d'une colonne en genres individuels.
- Calculer la durée moyenne des films pour chaque genre.