Analyzing IMDB Dataset –

Using the dataset from (https://www.kaggle.com/datasets/ashirwadsangwan/imdb-dataset) I am analyzing movies and their actors using PySpark to determine ratings and relationships between movies and actors.

Tools Used
- PySpark: For data manipulation and querying.
- Amazon AWS: Hosted environment for running PySpark on EC2 clusters.
- Data Tables:
  - titles: Contains information about movies and TV shows.
  - principles: Maps actors ("nconst") to titles ("tconst") and their job categories.
  - ratings: Provides average ratings and number of votes for each title.
  - name: Contains details about actors, including birth and death years.

How to Run the Code
1. Create EMR Cluster and create workspace using EMR role
2. Ensure the datasets ("titles", "principles", "ratings", and "name") are loaded with PySpark into workspace.
3. Execute each section of the code in sequence to replicate the results.