Movie Data Analysis

Problem Statement

1. Setup hadoop cluster with yarn, hive and spark in local using docker or cloud aws/gcp/azure.
2. You have given three files **movies.csv**, **ratings.csv** and **tags.csv**
3. Load all three files in **HDFS** location
4. Write spark job to solve below mentioned problem statements
    a. Show the aggregated number of ratings per year
    b. Show the average monthly number of ratings
    c. Show the rating levels distribution
    d. Show the 18 movies that are tagged but not rated
    e. Show the movies that have rating but no tag
    f. Focusing on the rated untagged movies with more than 30 user ratings,
    show the top 10 movies in terms of average rating and number of
    ratings
    g. What is the average number of tags per movie in tagsDF? And the
    average number of tags per user? How does it compare with the
    average number of tags a user assigns to a movie?
    h. Identify the users that tagged movies without rating them
    i. What is the average number of ratings per user in ratings DF? And the
    average number of ratings per movie?

j. What is the predominant (frequency based) genre per rating level?

k. What is the predominant tag per genre and the most tagged genres?

l. What are the most predominant (popularity based) movies?

m. Top 10 movies in terms of average rating (provided more than 30 users reviewed them)

5. Make sure to store the output of each problem statement in single csv with header in output **HDFS** path