PROGRAMMING ASSIGNMENT 4

ANALYZING THE MOVIELENS DATASET USING SPARK
VERSION 1.0

DUE DATE: Wednesday May 1st, 2024 @ 8:00 pm

OBJECTIVE

The objective of this assignment is to gain experience in developing Spark programs. As part of this assignment, you will be working with the MovieLens dataset that describe ratings and free-text tagging activities from MovieLens, a movie recommendation service. This dataset was created by GroupLens and primarily hosted at Kaggle. You will be using Apache Spark (version 3.5.0) to implement this assignment.

This assignment must be done individually and will account for **15 points** towards your cumulative course grade. There are a few components to this assignment, and the points-breakdown is listed in the remainder of the text. This assignment may be modified to clarify any questions (and the version number incremented), but the crux of the assignment and the distribution of points will not change.

1 Cluster setup [3 points]

As part of this assignment you are responsible for setting up your own HDFS and Spark cluster with Spark running on every node.

Refer to these documents and videos:

Shared Script -
If you have not yet downloaded and run the script that will generate Hadoop and Spark configurations, please do so by obtaining the .tar from the General channel on Microsoft Teams.

The guide to get started is also on Teams.

Infospaces videos -
Running jobs using spark shell for debugging : https://infospaces.cs.colostate.edu/watch.php?id=186

Compiling and creating jar using SBT and submitting job on Spark standalone cluster: https://infospaces.cs.colostate.edu/watch.php?id=185

2 Analysis of the MovieLens Dataset

You should develop Spark programs that leverage the DataFrame construct to process the main *and* supplementary datasets to answer the following questions. Questions Q1-Q5 account for **1 point each while Q6-Q7 account for 2 points each. Cumulatively, Q1-Q7 account for 9 points.**

Q1. How many movies were released for every year within the dataset?
The title column of movies.csv includes the year each movie was published. Some movies might not have the year, in such cases you can ignore those movies.

Q2. What is the average number of genres for movies within this dataset?

Q3. Rank the genres in the order of their ratings? Again, a movie may span multiple genres; such a movie should be counted in all the genres.

Q4. What are the top-3 combinations of genres that have the highest ratings?

Q5. How many movies have been tagged as "comedy"? Ignore the "case" information (i.e. both "Comedy" and "comedy" should be considered).

Q6. What are the different genres within this dataset? How many movies were released within different genres? A movie may span multiple genres; in such cases, that movie should be counted in all the genres?

Q7. Be creative and come up with your own data analytics question and answer for MovieLens dataset

[3 points]
You should include a PDF report that substantiates the results from your analysis. Please only include a PDF document (no Word or OpenOffice or Google Docs please).

3 Dataset Description

For this assignment you will be using the MovieLens 20M Dataset.

The datasets describe ratings and free-text tagging activities from MovieLens, a movie recommendation service. It contains 20,000,263 ratings and 465,564 tag applications across 27,278 movies. These data were created by 138,493 users between January 09, 1995 and March 31, 2015. The dataset was generated on October 17, 2016.

Users were selected at random for inclusion. All selected users had rated at least 20 movies. No demographic information is included. Each user is represented by an id, and no other information is provided.

The data are contained in six files.

- tags.csv that contains tags applied to movies by users:
  - userId
  - movieId
  - tag
  - timestamp

- ratings.csv that contains ratings of movies by users:
  - userId
  - movieId
  - rating
  - timestamp

- movies.csv that contains movie information:
  - movieId
  - title
  - genres

- links.csv that contains identifiers that can be used to link to other sources:
  - movieId
  - imdbId
  - tmbdId

- genome-scores.csv that contains movie-tag relevance data:
  - movieId
  - tagId
  - relevance

- genome-tags.csv that contains tag descriptions:
  - tagId
  - tag

References:

1. https://www.kaggle.com/grouplens/movielens-20m-dataset

2. F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19 (December 2015), 19 pages.

4 Provided Resources

You can download the dataset here (http://www.cs.colostate.edu/~csx55/ml-20m.zip)

5 Deductions

There will be a **15-point deduction** if any of the restrictions below are violated.

1. You should not implement this assignment as a stand-alone program.
2. You should not implement this assignment using anything other than Hadoop MapReduce. Implementing your own framework or using a 3rd party framework (that is not Hadoop) to implement this assignment is not allowed.

Version Change History

This section will reflect the change history for the assignment. It will list the version number, the date it was released, and the changes that were made to the preceding version. Changes to the first public release are made to clarify the assignment; the spirit or the crux of the assignment will not change.

Version	Date	Comments
1.0	4/3/2024	First public release of the assignment.