SPARK

# Spark RDD and SPARK RDD API [Python]

Use movelens data set, using movies.csv and ratings.csv files as input data and solve below problems only using Spark RDD, no DataFrame to be used. Use PySpark.

#### Load movies data into spark rdd tuple with (movieid, title, genres) format

#### Count number of movies based on Genres fields, use Adventure, Romance, Comedy as specific genres, get the count of movies on each genres

#### Load ratings.csv into tuple (userid, movieid, rating, timestamp) format

#### Remove the rating which is lower than 1.0 from the ratings.csv

#### Use reduceByKey, count the ratings for the movies based on movieid, for example, if a movie received 10 ratings, the output should be a tuple (movieId, 10)

#### Use Fold or FoldByKey, find the sum/or avg rating received by a movie in rating.csv

#### Convert timestamp in rating.csv into Python date time object

#### Join ratings and movies based on movie id, so that we could see movie title and genres for the given movie

#### Use sort on question number 5, sort the movie rating count by descending order so that highest rated movie should be on the top

#### Use filter on question number 5, if any movie received less than 100 ratings, then filter out the movie

# Apache Log file using RDD [Scala]

## Use file from <http://www.almhuette-raith.at/apache-log/access.log> as input

## Find count of Web GET, POST method in the log file

## Find the count of 404, 200 responses

## Find the count of files who length greater than 100 KB [content length is available after status code like 200 **11894**]

#### 

# 