# Instruction

Download small MovieLens dataset from this page https://grouplens.org/datasets/movielens/, direct link http://files.grouplens.org/datasets/movielens/ml-latest-small.zip

Read the “README.txt” file included in the zip to understand the data.

Develop an Apache Spark (http://spark.apache.org/) application (or many copies of the application, each per programming language) to answer to the following questions. Don’t limit your imagination and understanding.

- Programming languages can be Scala, Python or both (each application per language).
- You can use RDD, DataFrame, Dataset. We expect to see **at least one answer with RDD, one answer with Dataset**.
- Develop the application as you would do in a **real life project** (e.g. include tests if needed; use engineering best practices for structuring and implementation; think about the performance, scalability, and maintenance; use local git repo and include all the commit history when you send back the solution).

Include documents with your application, feel free to explain how you understand the source data, how you understand the problem, your approach to the question, caveats & limitations of the solution. Include a “README.md” file to explain how to compile & run the application.

# Questions

## General questions

1. How many “Drama” movies (movies with the "Drama" genre) are there?
2. How many unique movies are rated, how many are not rated?
3. Who give the most ratings, how many rates did he make?
4. Compute min, average, max rating per movie.
5. Output dataset containing users that have rated a movie but not tagged it.
6. Output dataset containing users that have rated AND tagged a movie.
7. Describe how you would find the release year for a movie (refer to the readme for information).
8. Enrich movies dataset with extract the release year. Output the enriched dataset.
9. Output dataset showing the number of movies per Genre per Year (movies will be counted many times if it's associated with multiple genres).

## QA specific question

1. Describe how you would write tests to ensure that all movies have a release year? Write these tests.
2. Write tests to ensure that the rating & tag happened on or after the year that the movie was released (here we can only check the release year against the year of rating & tag).

3. Write tests to ensure that at least 50% of movies have more than one genres.