# Hadoop Assignment 1

Name: Zakaria Balali  
Student number: 636655  
Github source code link: <https://github.com/ZakariaBalali/PDP_Assignments/blob/main/Assignment1/PDPAssignment1.py>

## Prerequisites

-Hadoop VM installed  
-Movielens dataset on VM (Use command “**wget** [**http://witan.nl/hadoop/u.data**](http://witan.nl/hadoop/u.data)**” to download it to your root folder if you don’t have it on your machine)**  
-Putty  
How to run the code

-Start the Hadoop VM  
-Connect to the Hadoop server using putty (127.0.0.1:2222)  
-Login with username and password  
-Create a new file using “Touch (filename).py)”  
-Open the file in the editor using “nano (filename).py”  
-Copy and paste the source code to the new file  
-Save file  
-Run the file using command: “python PDPAssignment1.py -r hadoop --hadoop-streaming-jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar u.data”

## Code and explanation

Step 1: Import libraries (Mrjob and Mrstep)  
Step 2: Define new class MovieRatingCounter  
Step 3: Using the steps method, the following functions are executed:

-Load the columns, split the lines and keep the movieID’s

-Use reducer to sum up the amount of ratings with their movieID’s

-Sort movies by their amount of ratings (high to low)

Text

Description automatically generated

## Result

Text

Description automatically generated