POLITECNICO
MILANO 1863

Apache Spark

Alessandro Margara
alessandro.margara@polimi.it
https://margara.faculty.polimi.it

Rules

- Rename the SparkGroupXX.java file replacing XX with the number of your group

- Write in the comment on top of the class your group number and the name of all group members

- Submit only a single java file with your solution
  - Submitted from the contact email provided in the group registration document

Assumptions

Four input datasets

1. profs
   - Type: static, csv file
   - Fields: prof_name, course_name
2. courses
   - Type: static, csv file
   - Fields: course_name, course_hours, course_students
3. videos
   - Type: static, csv file
   - Fields: video_id, video_duration, course_name
4. visualizations
   - Type: dynamic, stream
   - Fields: timestamp, value
   - Each entry with value v indicates that someone watched video with video_id equal to v

Requirements

- For all queries: limit unnecessary recomputations as much as possible!
- For streaming queries: write the results on the console, showing only the results that changed since the last evaluation

Requirements

- Q1: compute the total number of hours of lecture for each prof
- Q2: for each course, compute the total number of visualizations of videos of that course, computed over 1 minute, updated every 10 seconds
- Q3: for each video, compute the total number of visualizations of that video with respect to the number of students in the course in which the video is used