POLITECNICO
MILANO 1863

Apache Spark

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

Rules

- Rename the WebAccessAnalysisGroupXX.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

Two input datasets

1. accessLog
   - Type: static, csv file
   - Fields: UserID, PageURL, AccessTime, Duration
2. streamingData
   - Type: dynamic, stream
   - Fields: timestamp, value
   - Each entry with value v indicates that someone visited the page with PageURL 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 top 5 most popular pages from the accessLog, that is, the 5 pages that received the highest number of visualizations

- Q2: for each user u and for each popular page p (as computed in query Q1), compute the number of times user u visited page p

- Q3: for each popular page p (as computed in query Q1), compute the number of accesses to p within streamingData in a window of size 10 seconds, slide 4 seconds

- Q4: for each popular page p (as computed in query Q1), and for each window w (as computed in query Q3), compute the difference between the number of accesses to p within the accessLog and within w