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

Three input datasets

1. citiesRegion
   - Type: static, csv file
   - Fields: city, region
2. citiesPopulation
   - Type: static, csv file
   - Fields: id (of the city), city, population
3. bookings
   - Type: dynamic, stream
   - Fields: timestamp, value
   - Each entry with value x indicates that someone booked a hotel in the city with id x

Requirements

- For all queries: limit unnecessary recomputations as much as possible!
- Q1: compute the total population for each region
- Q2: compute the number of cities and the population of the most populated city for each region

Requirements

- Q3: Print the evolution of the population in Italy year by year until the total population in Italy overcomes 100M people
  - Assume that the population evolves as follows:
    - In cities with more than 1000 inhabitants, it increases by 1% every year
    - In cities with less than 1000 inhabitants, it decreased by 1% every year
  - The output on the terminal should be a sequence of lines
    - Year: 1, total population: xxx
    - Year: 2, total population: yyy
    - …
  - You may round the population of each city to the nearest integer during your computation

- Q4: compute the total number of bookings for each region, in a window of 30 seconds, sliding every 5 seconds