Big data: architectures and data analytics

Spark - Exercises
Exercise #43 – 1

- Critical bike sharing station analysis
- Input:
  - A textual csv file containing the occupancy of the stations of a bike sharing system
    - The sampling rate is 5 minutes
    - Each line of the files contains one sensor reading/sample has the following format
      - stationId, date, hour, minute, num_of_bikes, num_of_free_slots
    - Some readings are missing due to temporarily malfunctions of the stations
      - Hence, the number of samplings is not exactly the same for all stations
    - The number of distinct stations is 100

Exercise #43 – 2

- Input:
  - A second textual csv file containing the list of neighbors of each station
    - Each line of the files has the following format
      - stationId_x, list of neighbors of stationId_x
    - E.g.,
      - s1, s2 s3
      - means that s2 and s3 are neighbors of s1
Exercise #43 – 3

- Outputs:
  - Compute the percentage of critical situations for each station
    - A station is in a critical situation if the number of free slots is below a user provided threshold (e.g., 3 slots)
    - The percentage of critical situations for a station $S_i$ is defined as $(number\ of\ critical\ readings\ associated\ with\ S_i)/(total\ number\ of\ readings\ associated\ with\ S_i)$

Exercise #43 – 4

- Store in an HDFS file the stations with a percentage of critical situations higher than 80% (i.e., stations that are almost always in a critical situation and need to be extended)
  - Each line of the output file is associated with one of the selected stations and contains the percentage of critical situations and the stationId
  - Sort the stored stations by percentage of critical situations
Exercise #43 – 5

- Compute the percentage of critical situations for each pair (timeslot, station)
  - Timeslot can assume the following 6 values
    - [0-3]
    - [4-7]
    - [8-11]
    - [12-15]
    - [16-19]
    - [20-23]

Exercise #43 – 6

- Store in an HDFS file the pairs (timeslot, station) with a percentage of critical situations higher than 80% (i.e., stations that need rebalancing operations in specific timeslots)
  - Each line of the output file is associated with one of the selected pairs (timeslot, station) and contains the percentage of critical situations and the pair (timeslot, stationId)
  - Sort the result by percentage of critical situations
Exercise #43 – 7

- Select a reading (i.e., a line) of the first input file if and only if the following constraints are true
  - The line is associated with a full station situation
    - i.e., the station $S_i$ associated with the current line has a number of free slots equal to 0
  - All the neighbor stations of the station $S_i$ are full in the time stamp associated with the current line
    - i.e., bikers cannot leave the bike at Station $S_i$ and also all the neighbor stations are full in the same time stamp
  - Store the selected readings/lines in an HDFS file and print on the standard output the total number of such lines