Big data: architectures and data analytics

Spark - Exercises

Exercise #37

- Maximum values
  - Input: a textual csv file containing the daily value of PM10 for a set of sensors
    - Each line of the files has the following format

      sensorId,date,PM10 value (µg/m^3)\n
  - Output: the maximum value of PM10 for each sensor
    - Store the result in an HDFS file

Exercise #37 - Example

- Input file

  ```
  s1,2016-01-01,20.5
  s2,2016-01-01,30.1
  s1,2016-01-02,60.2
  s2,2016-01-02,20.4
  s1,2016-01-03,55.5
  s2,2016-01-03,52.5
  ```

- Output

  ```
  (s1,60.2)
  (s2,52.5)
  ```

Exercise #38

- Pollution analysis
  - Input: a textual csv file containing the daily value of PM10 for a set of sensors
    - Each line of the files has the following format

      sensorId,date,PM10 value (µg/m^3)\n
  - Output: the sensors with at least 2 readings with a PM10 value greater than the critical threshold 50
    - Store in an HDFS file the sensorIds of the selected sensors and also the number of times each of those sensors is associated with a PM10 value greater than 50

Exercise #38 - Example

- Input file

  ```
  s1,2016-01-01,20.5
  s2,2016-01-01,30.1
  s1,2016-01-02,60.2
  s2,2016-01-02,20.4
  s1,2016-01-03,55.5
  s2,2016-01-03,52.5
  ```

- Output

  ```
  (s1,2)
  ```

Exercise #39

- Critical dates analysis
  - Input: a textual csv file containing the daily value of PM10 for a set of sensors
    - Each line of the files has the following format
      sensorId,date,PM10 value ($\mu g/m^3$)\n
  - Output: an HDFS file containing one line for each sensor
    - Each line contains a sensorId and the list of dates with a PM10 values greater than 50 for that sensor

7

Exercise #39 - Example

- Input file

      s1,2016-01-01,20.5
      s2,2016-01-01,30.1
      s1,2016-01-02,60.2
      s2,2016-01-02,20.4
      s1,2016-01-03,55.5
      s2,2016-01-03,52.5

- Output

      (s1,[2016-01-02, 2016-01-03])
      (s2,[2016-01-03])

8

Exercise #39 bis

- Critical dates analysis
  - Input: a textual csv file containing the daily value of PM10 for a set of sensors
    - Each line of the files has the following format
      sensorId,date,PM10 value ($\mu g/m^3$)\n
  - Output: an HDFS file containing one line for each sensor
    - Each line contains a sensorId and the list of dates with a PM10 values greater than 50 for that sensor
    - Also the sensors which have never been associated with a PM10 values greater than 50 must be included in the result (with an empty set)

9

Exercise #39 bis - Example

- Input file

      s1,2016-01-01,20.5
      s2,2016-01-01,30.1
      s1,2016-01-02,60.2
      s2,2016-01-02,20.4
      s1,2016-01-03,55.5
      s2,2016-01-03,52.5
      s3,2016-01-03,12.5

- Output

      (s1,[2016-01-02, 2016-01-03])
      (s2,[2016-01-03])
      (s3,[])

10