Big data: architectures and data analytics

Spark - Exercises
Exercise #40

- Order sensors by number of critical days
  - 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)
  - Output: an HDFS file containing the sensors ordered by the number of critical days
    - Each line of the output file contains the number of days with a PM10 values greater than 50 for a sensor s and the sensorId of sensor s

Exercise #40 - 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
  
  2, s1
  1, s2
Exercise #41

- Top-k most critical sensors
  - 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
      \[\text{sensorId}, \text{date}, \text{PM10 value (μg/m}^3)\]\n    - The value of k
      - It is an argument of the application

Exercise #41

- Top-k most critical sensors
  - Output:
    - An HDFS file containing the top-k critical sensors
      - The "criticality" of a sensor is given by the number of days with a PM10 values greater than 50
      - Each line contains the number of critical days and the sensorId
Exercise #41 - 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

- k = 1
- **Output**
  - 2, s1

Exercise #42

- **Mapping Question-Answer(s)**
  - **Input:**
    - A large textual file containing a set of questions
      - Each line contains one question
      - Each line has the format
        - QuestionId, Timestamp, TextOfTheQuestion
    - A large textual file containing a set of answers
      - Each line contains one answer
      - Each line has the format
        - AnswerId, QuestionId, Timestamp, TextOfTheAnswer
Exercise #42

- **Output:**
  - A file containing one line for each question
  - Each line contains a question and the list of answers to that question
    - QuestionId, TextOfTheQuestion, list of Answers

Exercise #42 - Example

- **Questions**
  - Q1,2015-01-01,What is ..?
  - Q2,2015-01-03,Who invented ..

- **Answers**
  - A1,Q1,2015-01-02,It is ..
  - A2,Q2,2015-01-03,John Smith
  - A3,Q1,2015-01-05,I think it is ..
Exercise #42 - Example

- Output

\((Q_1([\text{What is ..?}], [\text{It is .., I think it is ..}]))\)
\((Q_2([\text{Who invented ..?}], [\text{John Smith}]))\)