Big data: architectures and data analytics

MapReduce - Exercises

2

Exercise #17

- Select maximum temperature for each date
  - Input: two structured textual files containing the temperatures gathered by a set of sensors
    - Each line of the first file has the following format

      sensorID,date,hour,temperature\n
    - Each line of the second file has the following format

      date,hour,temperature,sensorID\n
  - Output: the maximum temperature for each date (considering the data of both input files)

3

Exercise #17 - Example

- Input files

  | S1,2016-01-01,14:00,20.5 |
  | S2,2016-01-01,14:00,30.2 |
  | S1,2016-01-02,14:10,11.5 |
  | S2,2016-01-02,14:10,30.2 |

  | 2016-01-01,14:00,20.1,S3 |
  | 2016-01-01,14:00,10.2,S4 |
  | 2016-01-02,14:15,31.5,S3 |
  | 2016-01-02,14:15,20.2,S4 |

- Output

  | 2016-01-01 | 30.2 |
  | 2016-01-02 | 31.5 |

4

Exercise #18

- Filter the readings of a set of sensors based on the value of the measurement
  - Input: a set of textual files containing the temperatures gathered by a set of sensors
    - Each line of the files has the following format

      sensorID,date,hour,temperature\n
  - Output:
    - The lines of the input files associated with a temperature value greater than 30.0

5

Exercise #18 - Example

- Input file

  | S1,2016-01-01,14:00,20.5 |
  | S2,2016-01-01,14:00,30.2 |
  | S1,2016-01-02,14:10,11.5 |
  | S2,2016-01-02,14:10,30.2 |

- Output file

  | S2,2016-01-01,14:00,30.2 |
  | S2,2016-01-02,14:10,30.2 |

6

Exercise #19

- Filter the readings of a set of sensors based on the value of the measurement
  - Input: a set of textual files containing the temperatures gathered by a set of sensors
    - Each line of the files has the following format
      sensorID,date,hour,temperature**\n**
  - Output:
    - The lines of the input files associated with a temperature value less than or equal to 30.0

Exercise #19 - Example

- Input file

  S1,2016-01-01,14:00,20.5
  S2,2016-01-01,14:00,30.2
  S1,2016-01-02,14:10,11.5
  S2,2016-01-02,14:10,30.2

- Output file

  S1,2016-01-01,14:00,20.5
  S1,2016-01-02,14:10,11.5

Exercise #20

- Split the readings of a set of sensors based on the value of the measurement
  - Input: a set of textual files containing the temperatures gathered by a set of sensors
    - Each line of the files has the following format
      sensorID,date,hour,temperature**\n**
  - Output:
    - a set of files with the prefix "high-temp-" containing the lines of the input files with a temperature value greater than 30.0
    - a set of files with the prefix "normal-temp-" containing the lines of the input files with a temperature value less than or equal to 30.0

Exercise #20 - Example

- Input file

  S1,2016-01-01,14:00,20.5
  S2,2016-01-01,14:00,30.2
  S1,2016-01-02,14:10,11.5
  S2,2016-01-02,14:10,30.2

- Output files

high-temp-m-00001

  S2,2016-01-01,14:00,30.2
  S2,2016-01-02,14:10,30.2

normal-temp-m-00001

  S1,2016-01-01,14:00,20.5
  S1,2016-01-02,14:10,11.5