Spark Solutions Exercise 39 bis

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 (μg/m³)\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)

Exercise #39 bis - Example

Input file

```
$1,2016-01-01,20.5

$2,2016-01-01,30.1

$1,2016-01-02,60.2

$2,2016-01-02,20.4

$1,2016-01-03,55.5

$2,2016-01-03,52.5

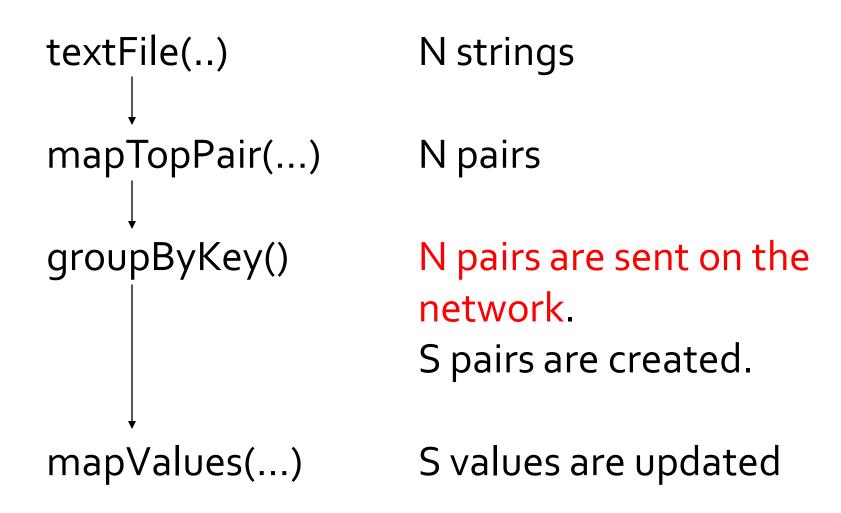
$3,2016-01-03,12.5
```

Output

```
($1, [2016-01-02, 2016-01-03])
($2, [2016-01-03])
($3, [])
```

- From each input reading
 - if PM10>50 -> Return (sensorId, date)
 - if PM10<=50 -> Return (sensorId, null)
- Group together values by key
 - Return (sensorId, [list of values])
 - One pair per distinct sensorld is returned
- Remove null from the lists of values

- N: number of input lines
- S: number of distinct sensors
- M: number of input lines with PM10>50
- Sc: number of distinct sensors associated with PM10>50 at least on time
- M<N</p>
- Sc<S</p>
- S<<N</p>



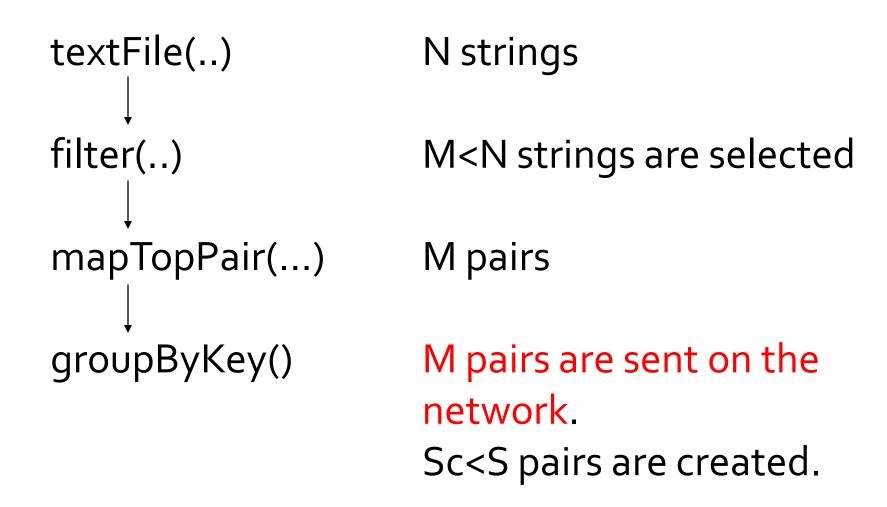
- Select only readings with PM10>50
- From each of the selected input readings
 - Return (sensorId, date)
- Group together values by key
 - Return (sensorId, [list of dates])
 - One pair per distinct sensorId is returned

- From each input reading
 - Return sensorld
- Remove duplicates
- Apply subtract to remove the sensorIds that are associated with PM10>50 at least one time
- For each of the selected sensorIds
 - Return (sensorIds, []), where [] is the empy list

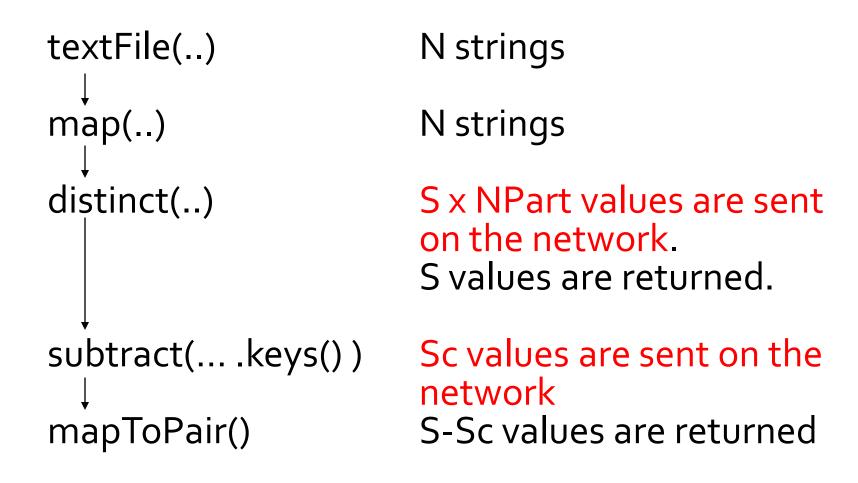


- N: number of input lines
- S: number of distinct sensors
- M: number of input lines with PM10>50
- Sc: number of distinct sensors associated with PM10>50 at least on time
- NPart: number of partitions
- M<N</p>
- Sc<S</p>
- S<<N</p>

Exercise #39 bis – Solution V2 – Part 1



Exercise #39 bis – Solution V2 – Part 2



Exercise #39 bis – Solution V2 – Part 3

union(..)

S pairs are returned

Comparison

- N: number of input lines
- S: number of distinct sensors
- M: number of input lines with PM10>50
- Sc: number of distinct sensors associated at least on time with PM10>50
- NPart: number of partitions
- M<N</p>
- Sc<S</p>
- S<<N</p>

Comparison

- Solution V1 sends on the network
 - N pairs (sensorId, value) on the network
- Solution V2 sends on the network
 - M pairs (sensorId, value) +
 - S x NPart sensorlds +
 - Sc sensorlds

Comparison

- S<<N</p>
- If we suppose that PM10>50 is a rare event
 - M<<N</p>

- Hence, Solution V2 sends less data on the network than Solution V1 if M<<N
 - if $M << N => N > M + S \times NPart + Sc$