Big data for internet applications

RDDs of numbers

RDDs of numbers

- Spark provides specific actions for RDD containing numerical values (integers or floats)
- RDDs of numbers can be created by using the standard methods
  - parallelize
  - transformations that return an RDD of numbers
- The following specific actions are also available on this type of RDDs
  - sum(), mean(), stdev(), variance(), max(), min()

RDDs of numbers: actions

- All the examples reported in the following are applied on inputRDD that is an RDD containing the following double values
  - [1.5, 3.5, 2.0]

RDDs of numbers: Summary

Action	Purpose	Example	Result
sum()	Return the sum over the values of the inputRDD	inputRDD.sum()	7.0
mean()	Return the mean value	inputRDD.mean()	2.3333
stdev()	Return the standard deviation computed over the values of the inputRDD	inputRDD.stdev()	0.8498
variance()	Return the variance computed over the values of the inputRDD	inputRDD. variance()	0.7223
max()	Return the maximum value	inputRDD.max()	3.5
min()	Return the minimum value	inputRDD.min()	1.5

RDDs of numbers: example

- Create an RDD containing the following float values
  - [1.5, 3.5, 2.0]
- Print on the standard output the following statistics
  - sum, mean, standard deviation, variance, maximum value, and minimum value

DoubleRDD actions: example

```
# Create an RDD containing a list of float values
inputRDD = sc.parallelize([1.5, 3.5, 2.0])

# Compute the statistics of interest and print them on
# the standard output
print("sum:", inputRDD.sum())
print("mean:", inputRDD.mean())
print("stdev:", inputRDD.stdev())
print("variance:", inputRDD.variance())
print("max:", inputRDD.max())
print("min:", inputRDD.min())
```