Programação de Sistemas Distribuídos

Mestrado em Engenharia Informática

2016/2017

The Problem

Simple Statistical analysis of tabular data sets.

Short Description

The system consists of a (set of) service(s) to enter and manipulate tabular data sets (numerical values only). The system must provide some form of authentication so that each user can have his own datasets. The user may work with several datasets at the same time. The client will perform several statistical analysis on the data set (either in its entirety or just on a single row/column).

Table 1 - Example Data Set (summary)

	A	B	C
1	68	12	58
2	3	10	96
3	52	14	55
4	25	11	25
5	48	24	78

The system should have the ability to define a set of transformations to apply to the data set, save those transformations and reuse saved transformations with other data sets. This feature is similar to define a macro (user defined transformation)

For instance, given data set M:

- Scale by 2 *then* Add 4 *then* plot dataset as bar chart

Other example, given data set M:

- Add 4 to column 2 *then* add 2 to column 1 *then* scale by 0.5

Macro Transformations are to be considered "heavy" operations and as such must be performed by a pool of worker services[1].

[1] To simulate lengthy processes you may add a random sleep in the implementation of each operation.

Use Cases

1. Register a user
2. Create a new data set from CSV, XML, JSON
3. Modify an existing data set
4. Return a data set as CSV, XML, JSON
5. Show the number of elements in the data set or selected row/column
6. Calculate total row
7. Calculate total column
8. Calculate statistical measures of a row, column, entire data set
    a. geometric mean
    b. median
    c. Mode
    d. Midrange
    e. Variance
    f. standard deviation
9. Perform transformations on the data set (without changing the original data set)
    a. Transpose the dataset
    b. Scale
    c. Add a scalar
    d. Add two data sets
    e. Multiply two data sets
    f. Augment the data set using linear interpolation on the rows or columns
10. Return a chart representation (image binary file) of the dataset
    a. Pie chart of a desired row/column
    b. Line/bar chart of a desired row/column
    c. Line/bar chart of the entire data set
11. Define a macro
12. Load a macro
13. Execute a macro on an existing dataset

Bonus

14. Register new worker in pool registry
15. “Heart beat” of a worker
16. Use OAuth authentication

Assignment

1. Produce the specification of the services of the described system
2. Develop the prototype of the described system

General Considerations

1. The aim of this work in the context of PSIDI is producing the specification of the services and implement the application.
2. The presented solution must follow the principles of service orientation and must make use of the architectural patterns for distributed systems taught in the course. Their use and justification must be clearly indicated in the architecture document.
3. Major importance and weight will be given to the distributed and service oriented nature of the system. That is, it is acceptable that not all business logic is fully implemented if it does not impact the system distributed architecture. For instance, transforming the dataset in UC 8 may return the exact same data of the original dataset or just a static data set without actually performing the transformation on the data.
4. The client application can be any HTTP client (e.g.Advanced REST client). Custom clients (webapps or console) can be developed and will count as bonus.
5. Each team is free to use the technology (programming language and platform) it sees most fit to (technology selection justification must be included in the documentation). The use of node.js is advised.
6. Third party libraries may be used freely but their use must be justified
7. The team may choose which report template to use and which sections to include regarding the specification document

Logistics

1. The assignment is to be made in groups of two or three students
2. The OT classes starting from week 6 and the Lab classes starting from week 10 will be devoted to help the students in carrying out the assignment
3. There will be two deliveries
    a. Phase 1: Specification of services on 11/11/2016
    b. Phase 2: revised specification + working application on 22/12/2016
4. Feedback on Phase 1 will be given by professors during PL lessons.
5. Presentation of final project will be on the **first week of January 2017** according to schedule to be published
6. It is expected that the assignment corresponds to an effort of 38 hours per student for the first delivery and 88 hours per student for the final delivery

Assessment

- Assessment will be done according to the criteria table (in annex) in a scale of 0 to 4 for each criteria.
- Assessment grade may be given with one decimal place
- Final grade is in the scale 0 to 20
- Assessment is individual, as such each student may have a different grade from the other group members

Example data

Table 2 - Example Data Set

	A	B	C	D
1	68	20	12	58
2	22	21	5	3
3	3	6	4	3
4	98	12	12	21
5	3	2	10	96
6	52	2	10	10
7	3	11	11	3
8	48	4	24	78

Table 3 - Statistical measures of the example data set

Entire dataset	
n	32
min	2
max	98
mean	22,97
median	11
mode	3
midrange	50
variance	765,28
std dev	27,66

Table 4 - Statistical measures on rows and columns

	A	B	C	D	min	max	mean	median	mode	midrange	variance	std dev
1	68	20	12	58	12	68	39,50	39	#N/A	40	572,75	23,93
2	22	21	5	3	3	22	12,75	13	#N/A	12,5	77,19	8,79
3	3	6	4	3	3	6	4,00	3,5	3	4,5	1,50	1,22
4	98	12	12	21	12	98	35,75	16,5	12	55	1305,19	36,13
5	3	2	10	96	2	96	27,75	6,5	#N/A	49	1562,19	39,52
6	52	2	10	10	2	52	18,50	10	10	27	384,75	19,62
7	3	11	11	3	3	11	7,00	7	3	7	16,00	4,00
8	48	4	24	78	4	78	38,50	36	#N/A	41	762,75	27,62
min	3	2	4	3								
max	98	21	24	96								
mean	37,13	9,75	11,00	34,00								
median	35	8,5	10,5	15,5								
mode	3	2	12	3								
midrange	50,5	11,5	14	49,5								
variance	1090,11	50,69	32,25	1248,00								
std dev	33,02	7,12	5,68	35,33								

*Table 5 - Augmenting by linear interpolation (on columns **or** on rows)*

	A	B	C	D
1	68	20	12	58
	45	20,5	8,5	30,5
2	22	21	5	3
	12,5	13,5	4,5	3
3	3	6	4	3
	50,5	9	8	12
4	98	12	12	21
	50,5	7	11	58,5
5	3	2	10	96
	27,5	2	10	53
6	52	2	10	10
	27,5	6,5	10,5	6,5
7	3	11	11	3
	25,5	7,5	17,5	40,5
8	48	4	24	78

	A		B		C		D
1	68	44	20	16	12	35	58
2	22	21,5	21	13	5	4	3
3	3	4,5	6	5	4	3,5	3
4	98	55	12	12	12	16,5	21
5	3	2,5	2	6	10	53	96
6	52	27	2	6	10	10	10
7	3	7	11	11	11	7	3
8	48	26	4	14	24	51	78