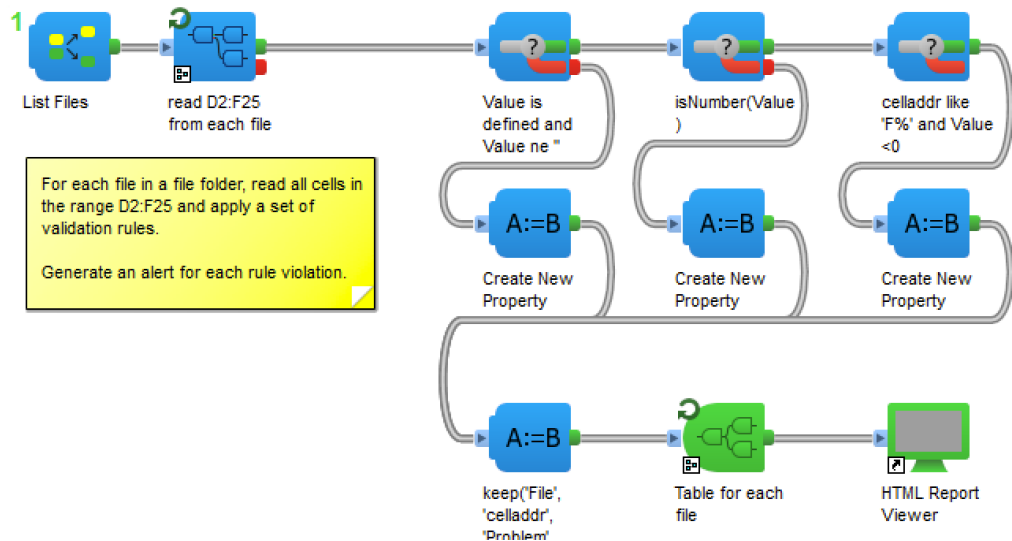


10 In a Large Folder, Find Files with Typos

Purpose

Automatically process a set of files in a folder, look for mis-formatted data in a cell range. For each file that contains bad data, generate an automated alert.



Workflow

In this example we read all files from the folder `data/process_reports`. For each file in the folder, we read all cells in the range D2:F25 (i.e. columns D,E,F) on the first worksheet and apply a set of validation rules. We generate an alert for each rule violation.

Validation rules: 1) cells must not be blank, 2) values must be numeric, 3) values in column F must be positive.

A typical file may be in either .XLS or .XLSX format and looks like this:

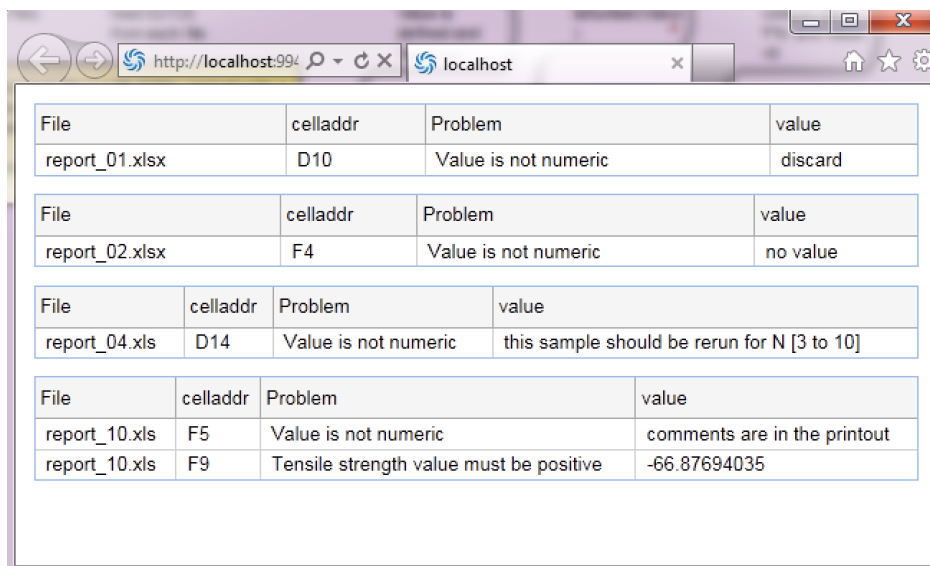
	A	B	C	D	E	F
	Source File	Date	Time Sample Number	N	Tensile.Strength	
1	StatProc_2013-11-08_168_drp_10.txt	Friday, November 08, 2013	11:21	10	1	68.82250515
2	StatProc_2013-11-08_925_drp_10.txt	Friday, November 08, 2013	11:21	10	2	65.77934742
3	StatProc_2013-11-08_781_drp_10.txt	Friday, November 08, 2013	11:21	10	3	66.55046057
4	StatProc_2013-11-08_622_drp_10.txt	Friday, November 08, 2013	11:21	10	4	67.61689702
5	StatProc_2013-11-08_525_drp_10.txt	Friday, November 08, 2013	11:21	10	5	66.61504265
6	StatProc_2013-11-08_851_drp_10.txt	Friday, November 08, 2013	11:21	10	6	67.91185585
7	StatProc_2013-11-08_331_drp_10.txt	Friday, November 08, 2013	11:21	10	7	69.16738621
8	StatProc_2013-11-08_745_drp_10.txt	Friday, November 08, 2013	11:21	10	8	66.87694035
9	StatProc_2013-11-08_998_drp_10.txt	Friday, November 08, 2013	11:21	10	9	65.69047442
10	StatProc_2013-11-08_236_drp_10.txt	Friday, November 08, 2013	11:21	10	10	68.70761393
11	StatProc_2013-11-08_238_drp_10.txt	Friday, November 08, 2013	11:21	10	11	68.75399543
12	StatProc_2013-11-08_690_drp_10.txt	Friday, November 08, 2013	11:21	10	12	67.87046558
13	StatProc_2013-11-08_172_drp_10.txt	Friday, November 08, 2013	11:21	10	13	69.90987216
14	StatProc_2013-11-08_32_drp_10.txt	Friday, November 08, 2013	11:21	10	14	70.39292514
15	StatProc_2013-11-08_772_drp_10.txt	Friday, November 08, 2013	11:21	10	15	67.22185408
16	StatProc_2013-11-08_409_drp_10.txt	Friday, November 08, 2013	11:21	10	16	68.32628776
17	StatProc_2013-11-08_401_drp_10.txt	Friday, November 08, 2013	11:21	10	17	65.19077462
18	StatProc_2013-11-08_586_drp_10.txt	Friday, November 08, 2013	11:21	10	18	68.25196604
19	StatProc_2013-11-08_14_drp_10.txt	Friday, November 08, 2013	11:21	10	19	69.88293775
20	StatProc_2013-11-08_456_drp_10.txt	Friday, November 08, 2013	11:21	10	20	67.91607401
21	StatProc_2013-11-08_414_drp_10.txt	Friday, November 08, 2013	11:21	10	21	67.81585381
22	StatProc_2013-11-08_491_drp_10.txt	Friday, November 08, 2013	11:21	10	22	67.49583408
23	StatProc_2013-11-08_787_drp_10.txt	Friday, November 08, 2013	11:21	10	23	66.57299342
24	StatProc_2013-11-08_362_drp_10.txt	Friday, November 08, 2013	11:21	10	24	68.45683603
25						

Using the **Cell Range Reader** component, we read the same D2:F25 range from each file and put cell values into Pipeline Pilot data records with the property "value".

Using PilotScript filters we identify rule violations and create annotation text for each violation. The annotations are then aggregated into a table that is displayed in a web browser window, organized by file name and cell location on the worksheet.

Results

Here is the result of the protocol.



The screenshot shows a web browser window with the address bar displaying 'http://localhost:994' and a tab labeled 'localhost'. The main content area contains a table with four columns: 'File', 'celladdr', 'Problem', and 'value'. The table is divided into four sections by horizontal lines. The first section has two rows, the second has two rows, the third has two rows, and the fourth has two rows.

File	celladdr	Problem	value
report_01.xlsx	D10	Value is not numeric	discard
File	celladdr	Problem	value
report_02.xlsx	F4	Value is not numeric	no value
File	celladdr	Problem	value
report_04.xls	D14	Value is not numeric	this sample should be rerun for N [3 to 10]
File	celladdr	Problem	value
report_10.xls	F5	Value is not numeric	comments are in the printout
report_10.xls	F9	Tensile strength value must be positive	-66.87694035