Data Ingestor Auto Pilot – DIAP

Easy Mass-Import of EXTERNAL Data (Files) into SAS

Stephan Weigandt

Sr Enterprise Data Engineer – Customer Advisory Support

FIRD – Field Innovation Research & Development

Stephan.Weigandt@sas.com

Introduction

Stephan Weigandt

Sr Enterprise Data Engineer
Field Innovation Research & Development, US Customer Advisory

- Former SAS Partner (with various companies since 1998), joined SAS in 2016,
  - 24-years of architecting and creating automated solutions around CRM and closed loop predictive environments with focus on data processing and integrating.
  - Co-created, designed and implemented a fully automated, closed loop predictive platform (building 350 models quarterly, which only needs to be supervised by 1.5 FTEs, including ETL, building model datasets, running QC on data and models, and creating automated PPTs for management consumption)
  - Engaged across many industries and participated in customer references.
- Everything is possible (…with SAS ☺).
  - And…if I have to do something twice, I will probably automate it.
  - I enjoy streamlining and standardizing, “everything data” with the client needs always on top of the mind.
- Love wearing hats and easily adjusted to the Californian lifestyle including doing yoga and more and more surfing.
- Certified in:
  - AWS Certified Solutions Architect – Associate
  - SAS V6 and V8 Certified Professional Advanced
  - Masters in Physics
- SAS Computing Award winner 2000 (Europe, Dublin)

What is the situation?
→ COVID-19

SAS helping and supporting local governments and researchers

- many urgent requests to make external data available for dashboards and/or modeling
- Data provided via websites or folder structures with missing and/or dirty metadata

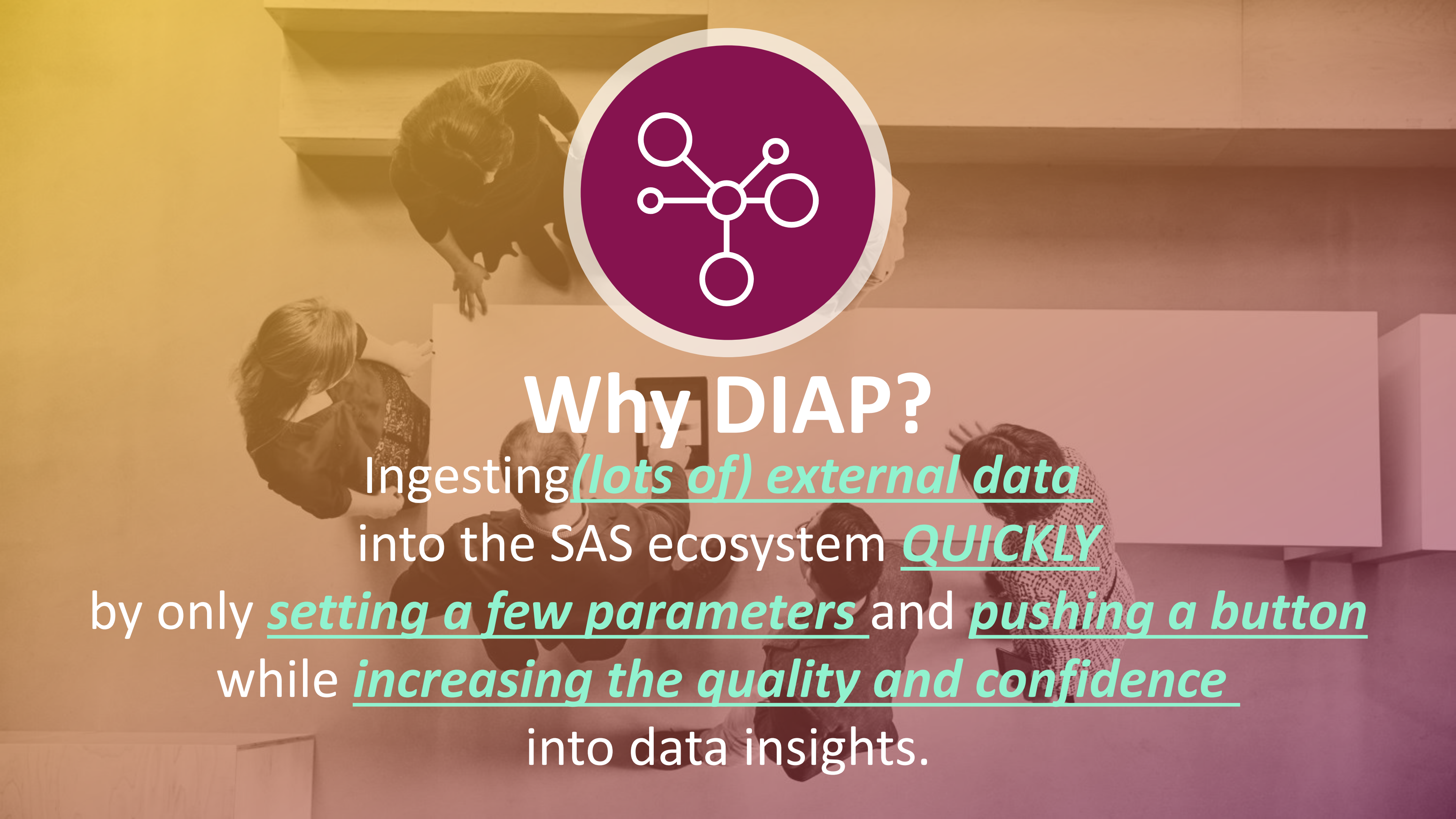
Time and Quality is of essence

Literally Life and Death could be on the line.

Standard OOB tools quickly run out of options

- “difficult” files
- time consuming (one file at a time)

→ from taking hours to days to being a stumble block



Why DIAP?

Ingesting***(lots of) external data***
into the SAS ecosystem ***QUICKLY***
by only ***setting a few parameters*** and ***pushing a button***
while ***increasing the quality and confidence***
into data insights.

What are the Main Pain Points with External Data…

Overwhelm with variety of data

- Files can be distributed over many and complex distributed directories
- XLSX, CSV, TXT, JMP, SHP, JSON,
- Fixed Width, XML
- variety of possible separators

Challenging Metadata

- identify and deal with Metadata issues
    - file names and variable names can be too long and contain “weird” characters
    - dealing with missing OR duplicate variable names

Consistency and Standardization

- create reference tables
- no load date or load time tracking
- profiling option

How did we solve it?

Automation is key

DIAP allows to set a handful of parameters and let the machine do the job

- has power to traverse a folder structure
- Import controlled into SAS Viya
- consuming all the files in folder
- Separator determination
- review and create metadata
- Keep track of Load Date/Time

How did we solve it?

Automation is key

DIAP allows to set a handful of parameters and let the machine do the job

has power to traverse a folder structure

- just provide top level directory name
- no limit in depth and complexity

consuming all the files in folder

- XLSX, CSV, TXT, JMP, SHP, JSON,
- Fixed Width
- XML

review and create metadata

- create reference tables
- identify and deal with Metadata issues
  - weird characters in names
- create profiles (integrates with "Autoprofiling on VA" solution)

Import controlled into SAS Viya

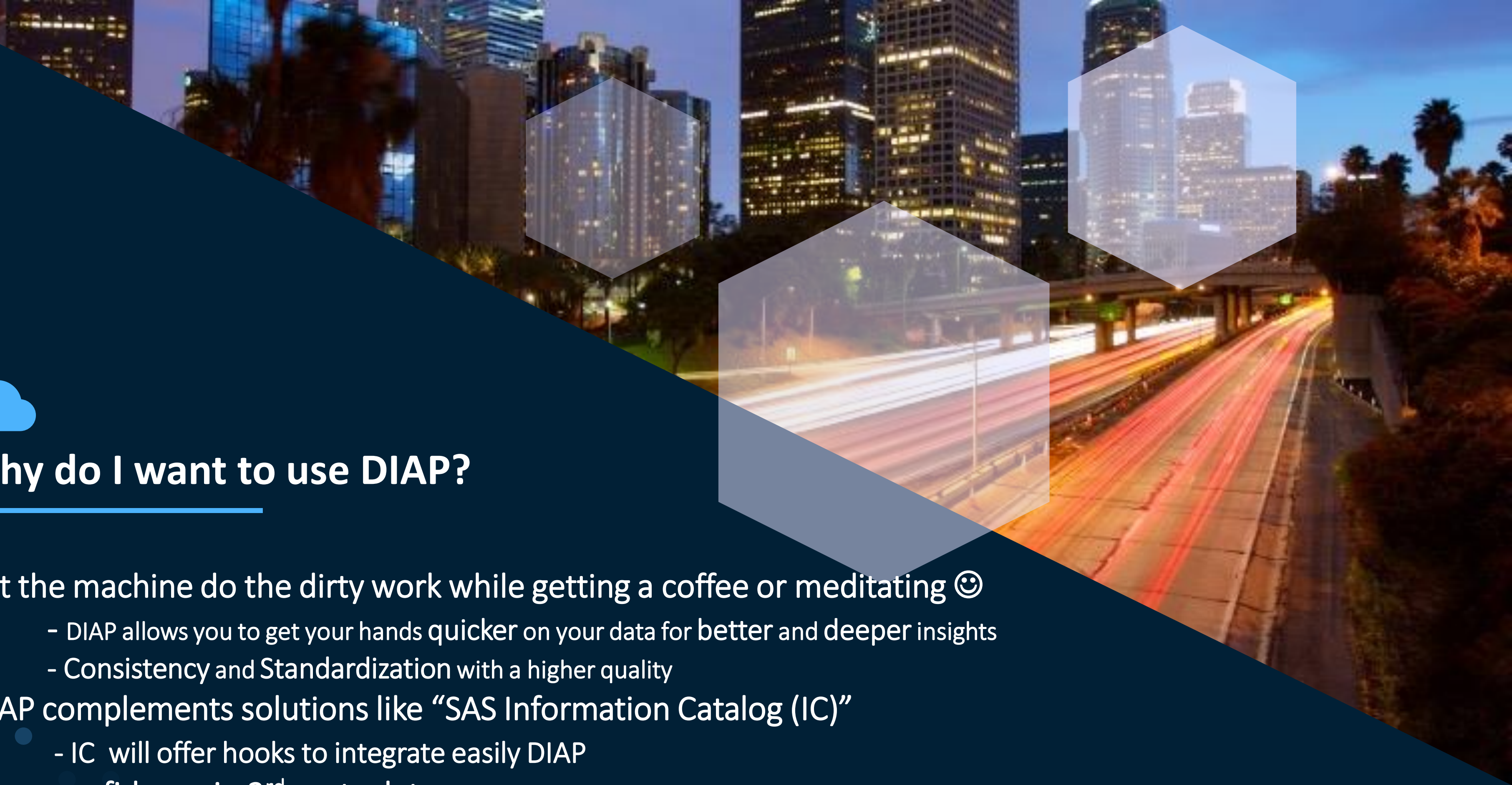
- Use data right away for dashboards/models
- Provide consistency for downstream referencing

Separator determination

- Tab, CSV, Semicolon, Pipe

Keep track of Load Date/Time

- Only upload modified files since last upload



Why do I want to use DIAP?

Let the machine do the dirty work while getting a coffee or meditating ☺

- DIAP allows you to get your hands quicker on your data for better and deeper insights
- Consistency and Standardization with a higher quality

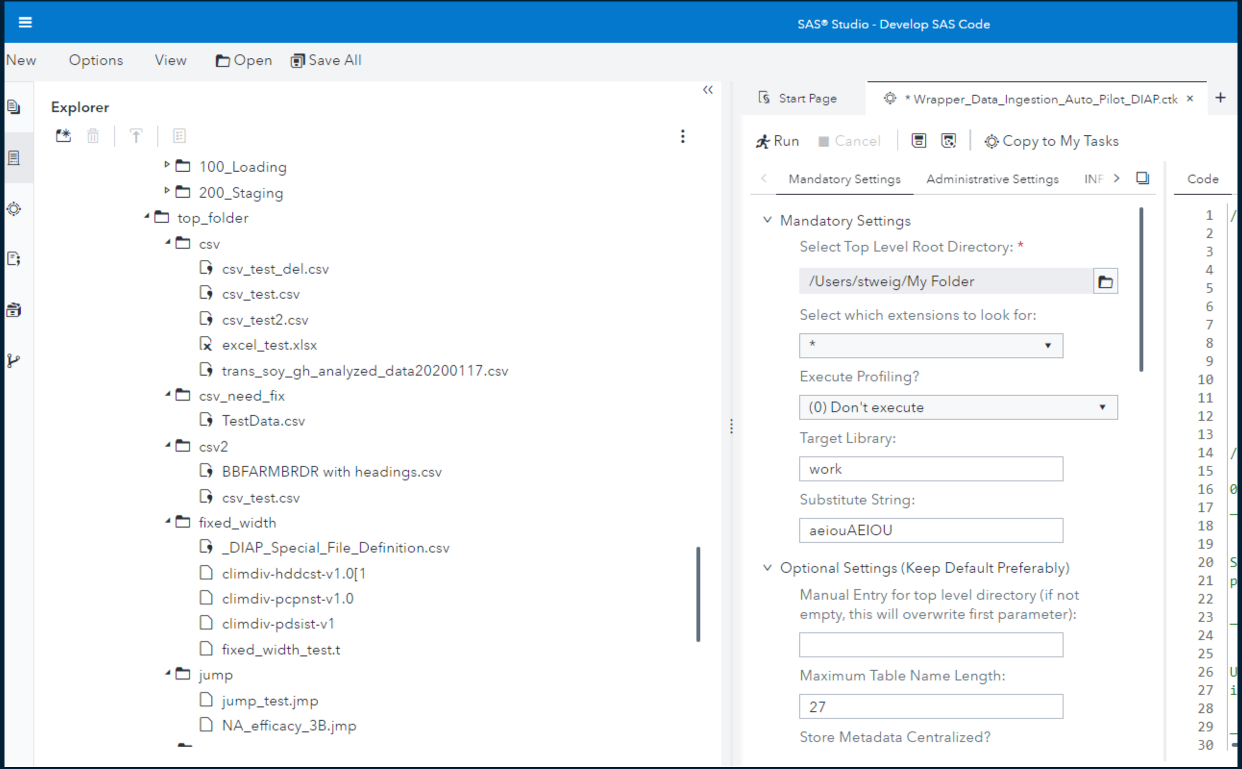
DIAP complements solutions like "SAS Information Catalog (IC)"

- IC will offer hooks to integrate easily DIAP

More confidence in 3rd party data

Avoid Data Swamp

Let’s start the journey…
Interacting with DIAP via Task UI



- 100_Loading
- 200_Staging
- top_folder
  - csv
    - csv_test_del.csv
    - csv_test.csv
    - csv_test2.csv
    - excel_test.xlsx
    - trans_soy_gh_analyzed_data20200117.csv
  - csv_need_fix
    - TestData.csv
  - csv2
    - BBFARMBRDR with headings.csv
    - csv_test.csv
  - fixed_width
    - _DIAP_Special_File_Definition.csv
    - climdiv-hddcst-v1.0[1
    - climdiv-pcpnst-v1.0
    - climdiv-pdsist-v1
    - fixed_width_test.t
  - jump
    - jump_test.jmp
    - NA_efficacy_3B.jmp

Mandatory Settings

Select Top Level Root Directory: *
/Users/stweig/My Folder

Select which extensions to look for:
*

Execute Profiling?
(0) Don't execute

Target Library:
work

Substitute String:
aeiouAEIOU

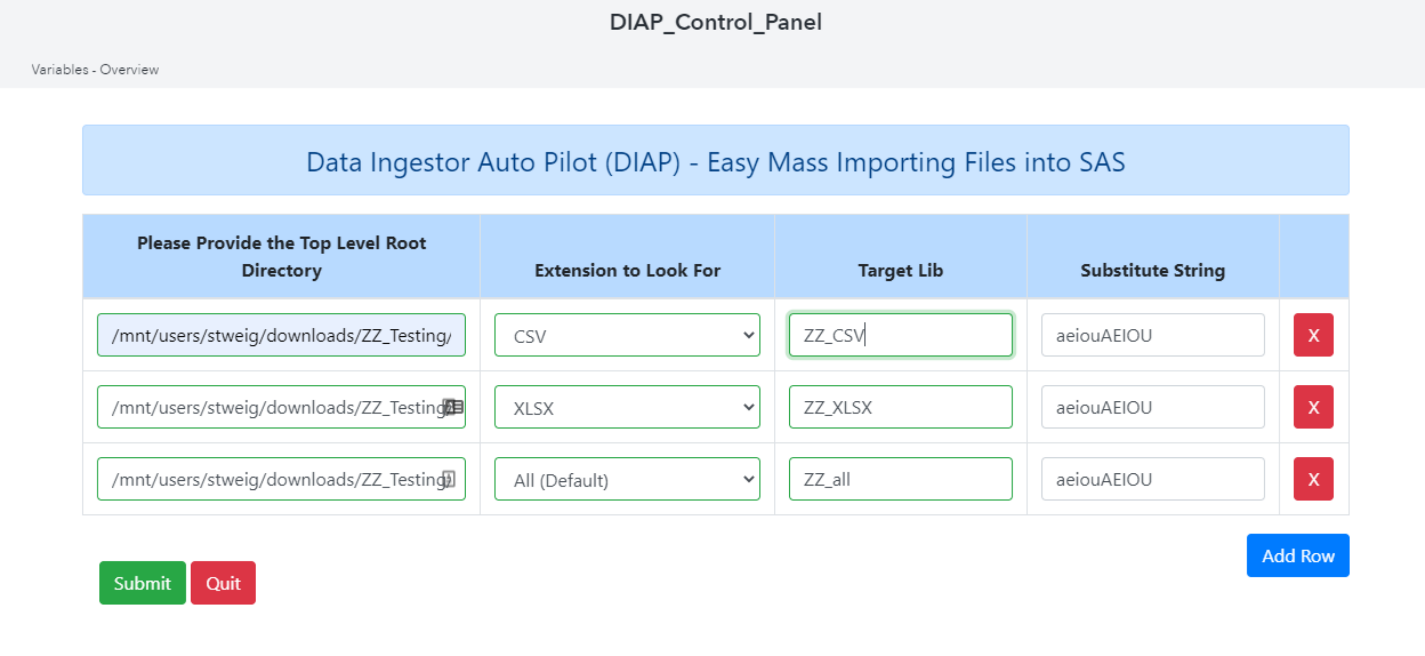
Optional Settings (Keep Default Preferably)

Manual Entry for top level directory (if not empty, this will overwrite first parameter):

Maximum Table Name Length:
27

Store Metadata Centralized?

“Heavier” version within VA – Interacting with DIAP via customized HTML within VA and Using Jobexecution



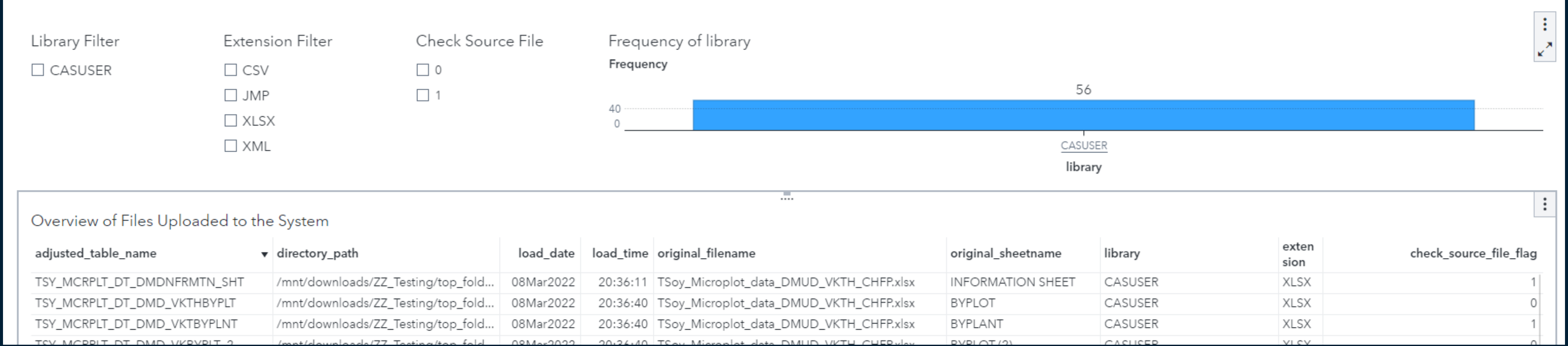
DIAP_Control_Panel

Variables - Overview

Data Ingestor Auto Pilot (DIAP) - Easy Mass Importing Files into SAS

Please Provide the Top Level Root Directory	Extension to Look For	Target Lib	Substitute String	
/mnt/users/stweig/downloads/ZZ_Testing/	CSV	ZZ_CSV	aeiouAEIOU	X
/mnt/users/stweig/downloads/ZZ_Testing	XLSX	ZZ_XLSX	aeiouAEIOU	X
/mnt/users/stweig/downloads/ZZ_Testing	All (Default)	ZZ_all	aeiouAEIOU	X

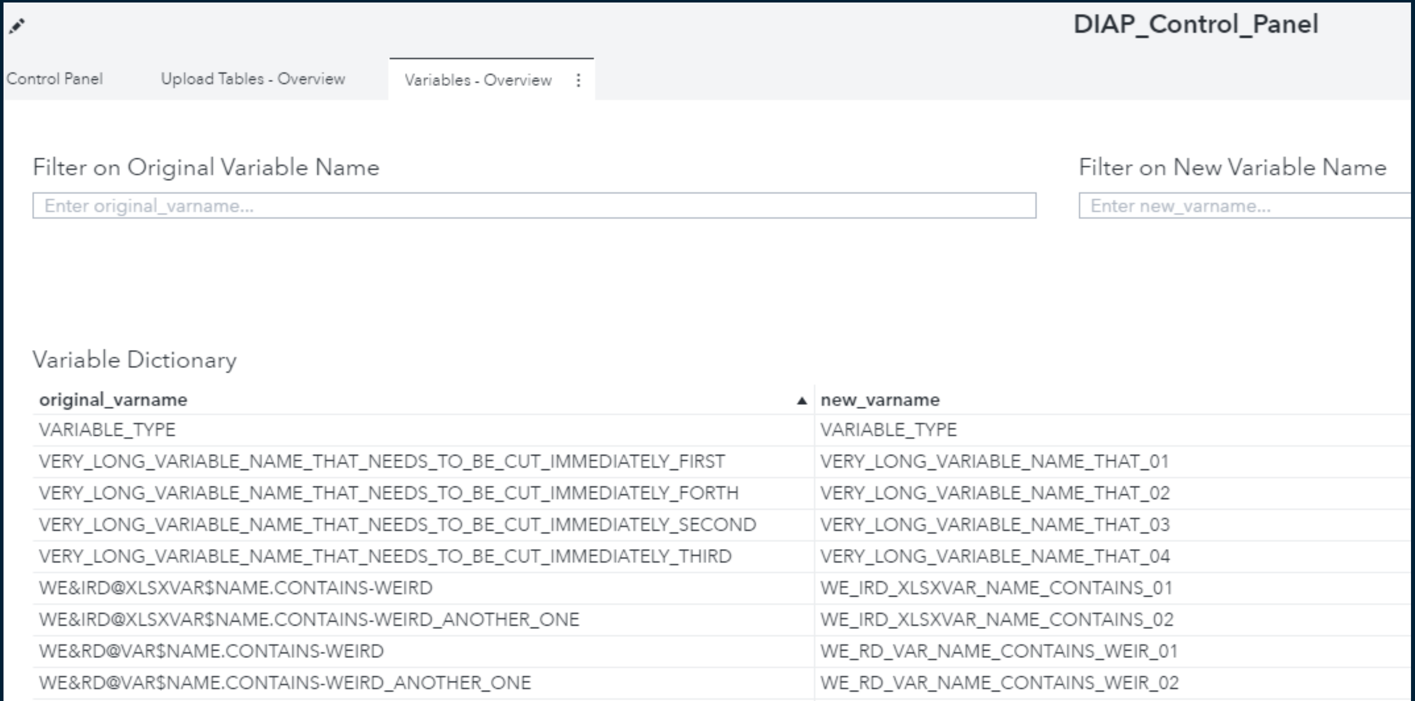
Reference Table - Upload Log Table



Overview of Files Uploaded to the System

adjusted_table_name ▾	directory_path	load_date	load_time	original_filename	original_sheetname	library	extension	check_source_file_flag
TSY_MCRPLT_DT_DMDNFRMTN_SHT	/mnt/downloads/ZZ_Testing/top_fold...	08Mar2022	20:36:11	TSoy_Microplot_data_DMUD_VKTH_CHFP.xlsx	INFORMATION SHEET	CASUSER	XLSX	1
TSY_MCRPLT_DT_DMD_VKTHBYPLT	/mnt/downloads/ZZ_Testing/top_fold...	08Mar2022	20:36:40	TSoy_Microplot_data_DMUD_VKTH_CHFP.xlsx	BYPLOT	CASUSER	XLSX	0
TSY_MCRPLT_DT_DMD_VKTBYPLNT	/mnt/downloads/ZZ_Testing/top_fold...	08Mar2022	20:36:40	TSoy_Microplot_data_DMUD_VKTH_CHFP.xlsx	BYPLANT	CASUSER	XLSX	1
TSY_MCRPLT_DT_DMD_VKBYPLT_2	/mnt/downloads/ZZ_Testing/top_fold...	08Mar2022	20:36:40	TSoy_Microplot_data_DMUD_VKTH_CHFP.xlsx	BYPLOT (2)	CASUSER	XLSX	0

Reference Table - Variable Overview Table



Filter on Original Variable Name

Enter original_varname...

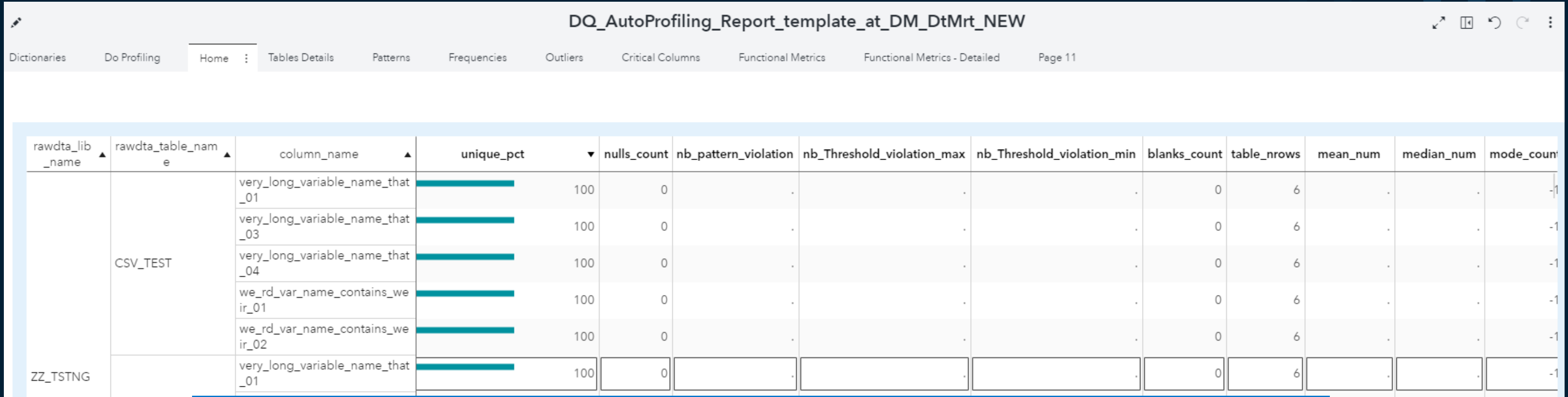
Filter on New Variable Name

Enter new_varname...

Variable Dictionary

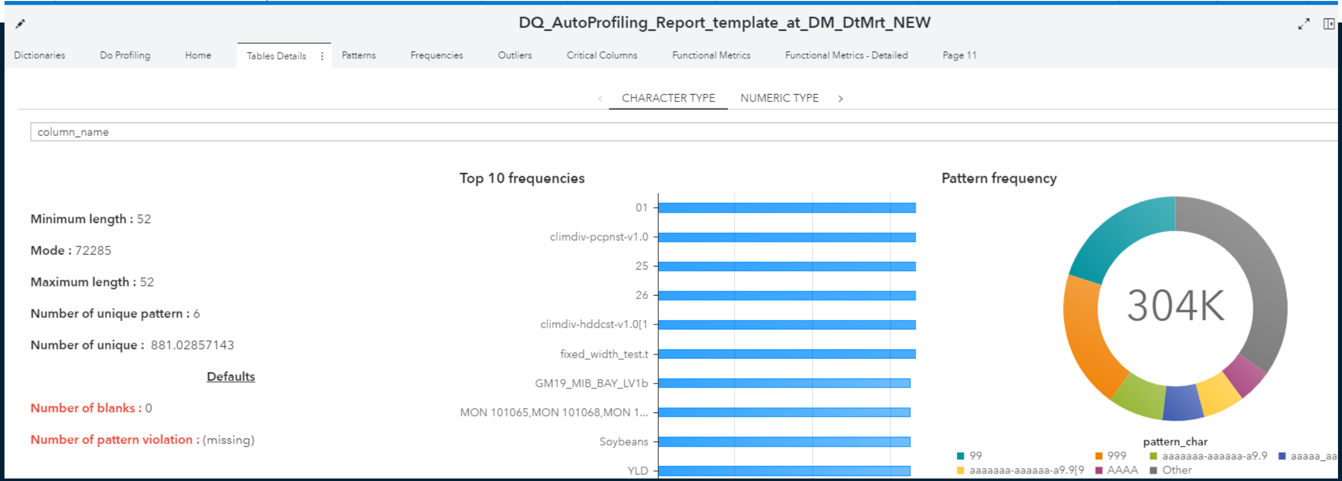
original_varname ▲	new_varname
VARIABLE_TYPE	VARIABLE_TYPE
VERY_LONG_VARIABLE_NAME_THAT_NEEDS_TO_BE_CUT_IMMEDIATELY_FIRST	VERY_LONG_VARIABLE_NAME_THAT_01
VERY_LONG_VARIABLE_NAME_THAT_NEEDS_TO_BE_CUT_IMMEDIATELY_FORTH	VERY_LONG_VARIABLE_NAME_THAT_02
VERY_LONG_VARIABLE_NAME_THAT_NEEDS_TO_BE_CUT_IMMEDIATELY_SECOND	VERY_LONG_VARIABLE_NAME_THAT_03
VERY_LONG_VARIABLE_NAME_THAT_NEEDS_TO_BE_CUT_IMMEDIATELY_THIRD	VERY_LONG_VARIABLE_NAME_THAT_04
WE&IRD@XLSXVAR$NAME.CONTAINS-WEIRD	WE_IRD_XLSXVAR_NAME_CONTAINS_01
WE&IRD@XLSXVAR$NAME.CONTAINS-WEIRD_ANOTHER_ONE	WE_IRD_XLSXVAR_NAME_CONTAINS_02
WE&RD@VAR$NAME.CONTAINS-WEIRD	WE_RD_VAR_NAME_CONTAINS_WEIR_01
WE&RD@VAR$NAME.CONTAINS-WEIRD_ANOTHER_ONE	WE_RD_VAR_NAME_CONTAINS_WEIR_02

Integration with Profiling Solution (AutoProfiling from SAS Europe)



DQ_AutoProfiling_Report_template_at_DM_DtMrt_NEW

rawdta_lib_name ▲	rawdta_table_name ▲	column_name ▲	unique_pct ▼	nulls_count	nb_pattern_violation	nb_Threshold_violation_max	nb_Threshold_violation_min	blanks_count	table_nrows	mean_num	median_num	mode_coun
ZZ_TSTNG	CSV_TEST	very_long_variable_name_that_01	100	0	.	.	.	0	6	.	.	-1
		very_long_variable_name_that_03	100	0	.	.	.	0	6	.	.	-1
		very_long_variable_name_that_04	100	0	.	.	.	0	6	.	.	-1
		we_rd_var_name_contains_weir_01	100	0	.	.	.	0	6	.	.	-1
		we_rd_var_name_contains_weir_02	100	0	.	.	.	0	6	.	.	-1
		very_long_variable_name_that_01	100	0	.	.	.	0	6	.	.	-1



DQ_AutoProfiling_Report_template_at_DM_DtMrt_NEW

Minimum length : 52

Mode : 72285

Maximum length : 52

Number of unique pattern : 6

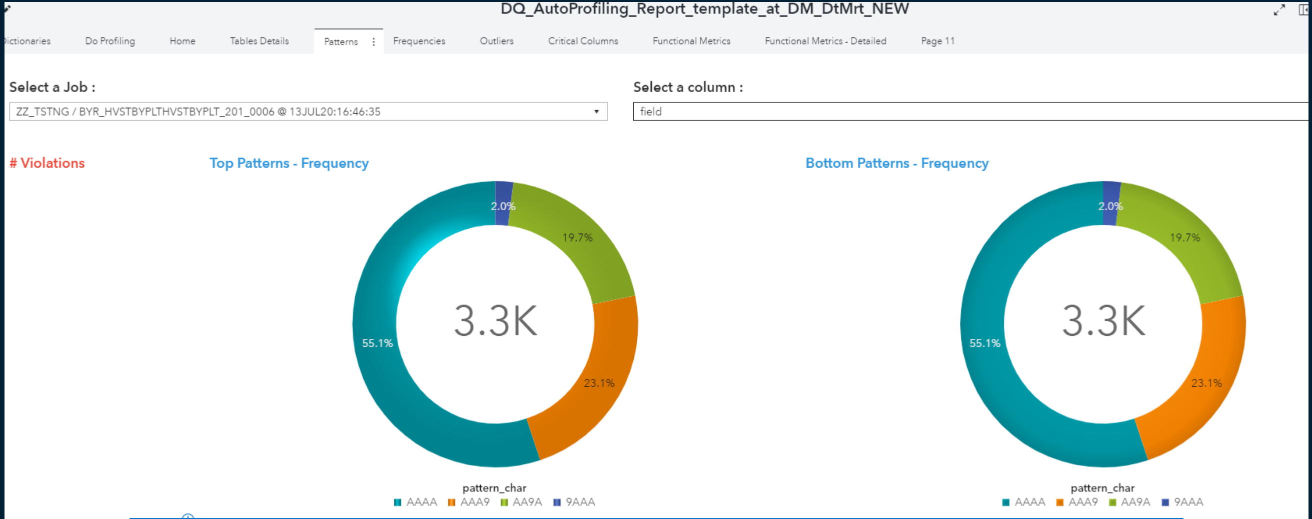
Number of unique : 881.02857143

Defaults

Number of blanks : 0

Number of pattern violation : (missing)

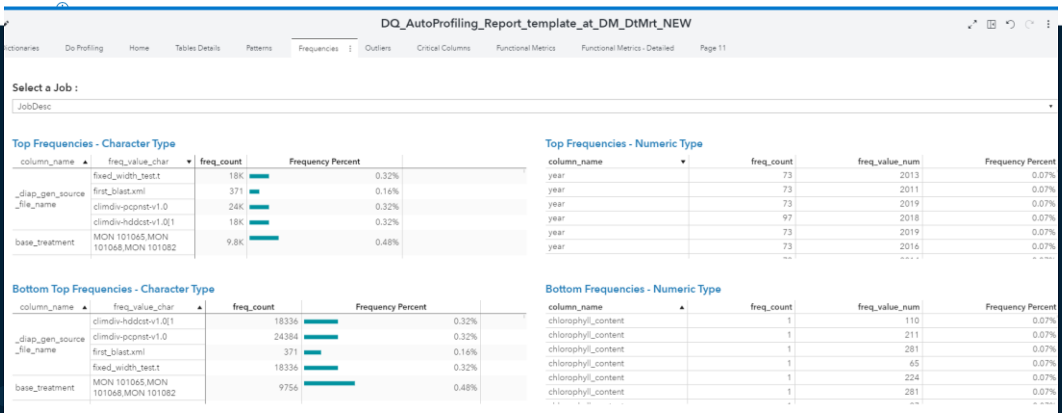
Integration with Profiling Solution (AutoProfiling from SAS Europe)



DQ_AutoProfiling_Report_template_at_DM_DtMrt_NEW

Select a Job : ZZ_TSTNG / BYR_HVSTBYPLTHVSTBYPLT_201_0006 @ 13JUL20:16:46:35

Select a column : field



DQ_AutoProfiling_Report_template_at_DM_DtMrt_NEW

Select a Job : JobDesc

Top Frequencies - Character Type

column_name ▲	freq_value_char ▼	freq_count	Frequency Percent
_diap_gen_source_file_name	fixed_width_test.t	18K	0.32%
	first_blast.xml	371	0.16%
	climdiv-pcpnst-v1.0	24K	0.32%
	climdiv-hddcst-v1.0[1	18K	0.32%
base_treatment	MON 101065,MON 101068,MON 101082	9.8K	0.48%

Top Frequencies - Numeric Type

column_name ▼	freq_count	freq_value_num	Frequency Percent
year	73	2013	0.07%
year	73	2011	0.07%
year	73	2019	0.07%
year	97	2018	0.07%
year	73	2019	0.07%
year	73	2016	0.07%

Bottom Top Frequencies - Character Type

column_name ▲	freq_value_char ▲	freq_count	Frequency Percent
_diap_gen_source_file_name	climdiv-hddcst-v1.0[1	18336	0.32%
	climdiv-pcpnst-v1.0	24384	0.32%
	first_blast.xml	371	0.16%
	fixed_width_test.t	18336	0.32%
base_treatment	MON 101065,MON 101068,MON 101082	9756	0.48%

Bottom Frequencies - Numeric Type

column_name ▲	freq_count	freq_value_num	Frequency Percent
chlorophyll_content	1	110	0.07%
chlorophyll_content	1	211	0.07%
chlorophyll_content	1	281	0.07%
chlorophyll_content	1	65	0.07%
chlorophyll_content	1	224	0.07%
chlorophyll_content	1	281	0.07%

Demo…

Questions?