Automating A New Generation Statistical Computing Environment (SCE)

Callum Charalambides & Megan Harries

22-APR-2024

veramed
End-to-End Biometrics

Table of Contents

What is a Statistical Computing Environment?

Why is it important to modernise SCEs?

Performance

Version Control

Automation Workflows

Security

Future Proofing

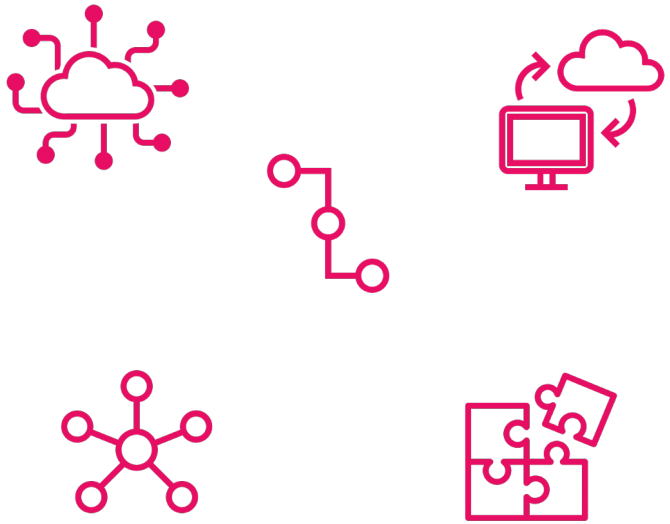
Enhanced Visualisations and Reporting

Ease of Onboarding

What did the client want?

Upgrade from UNIX SCE

Bespoke functionality

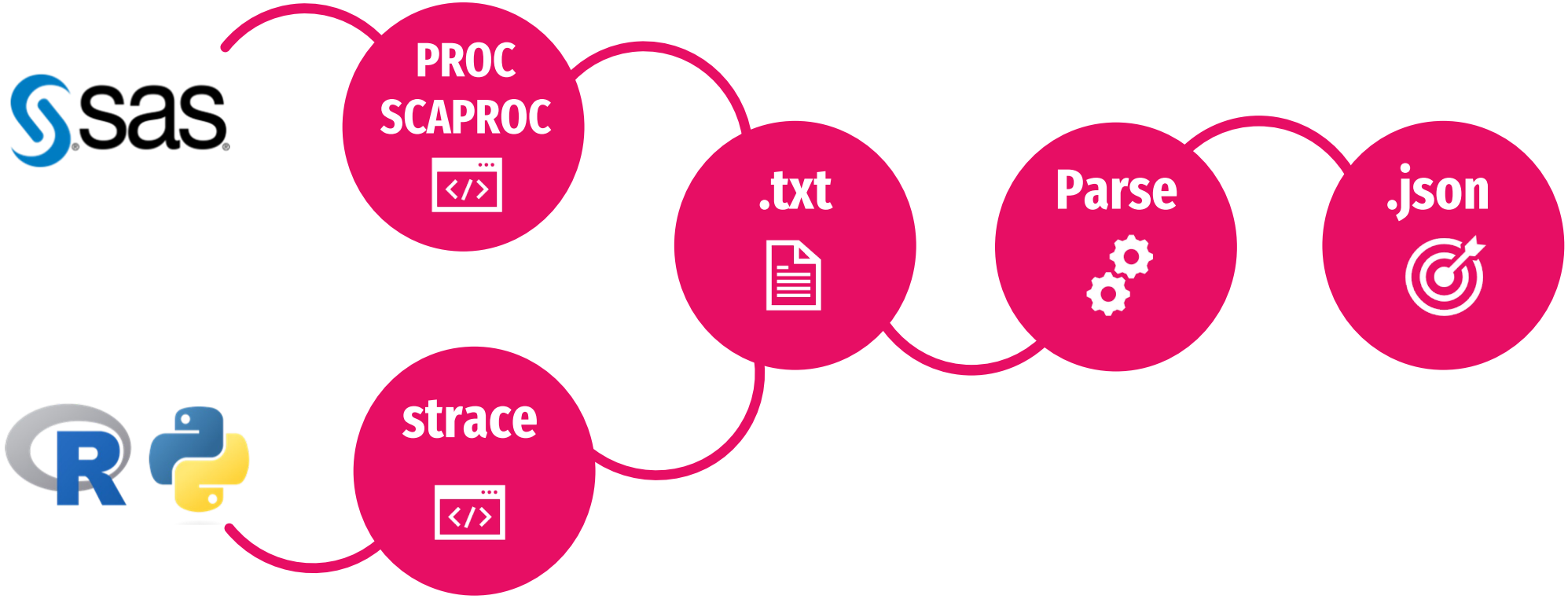


Dependency Automation

- Only run programs when required
- Automate the extraction of program dependencies
- Identify these ‘out-of-date’ programs
- Log of why it was re-run
- Create a linked dependency tree of all programs/data in a study

How We Did It

Generating Dependencies



PROC SCAPROC

```
/* JOBSPLIT: JOBSTARTTIME 06NOV2023:11:07:35.26 */
/* JOBSPLIT: TASKSTARTTIME 06NOV2023:11:07:35.26 */
/* JOBSPLIT: FILE INPUT /mnt/code/SAS/sas/import_sas7bdat.sas */
/* JOBSPLIT: DATASET INPUT SEQ DOMINO.SAS_TEST_DATA_EN.DATA */
/* JOBSPLIT: LIBNAME DOMINO "/mnt/data/Workflow_Manager" */
/* JOBSPLIT: DATASET INPUT SEQ DOMINO.SAS_TEST_DATA_FR.DATA */
/* JOBSPLIT: LIBNAME DOMINO "/mnt/data/Workflow_Manager" */
/* JOBSPLIT: DATASET INPUT SEQ DOMINO.SAS_TEST_DATA_DE.DATA */
/* JOBSPLIT: LIBNAME DOMINO "/mnt/data/Workflow_Manager" */
/* JOBSPLIT: DATASET OUTPUT SEQ WORK.TEST.DATA */
/* JOBSPLIT: LIBNAME WORK V9 '/tmp/SAS_workB8EF00000B0B_run-6548a1b6e4f8d3679715b927-dbgnb' *
/* JOBSPLIT: FILE OUTPUT <standard output> */
/* JOBSPLIT: ELAPSED 92  */
/* JOBSPLIT: SYSSCP LIN X64 */
/* JOBSPLIT: PROCNAME DATASTEP */
/* JOBSPLIT: STEP SOURCE FOLLOWS */
/* Test Case SAS02 */
/* Imported SAS7BDAT files are included as imports  */

libname Domino "/mnt/data/Workflow_Manager";

data test;
        merge Domino.sas_test_data_en
                  Domino.sas_test_data_fr
                  Domino.sas_test_data_de;
        by num;
run;

/* JOBSPLIT: JOBENDTIME 06NOV2023:11:07:35.36 */
/* JOBSPLIT: END */
```

```
{
        "version": "v1",
        "program": "/mnt/code/SAS/sas/import_sas7bdat.sas",
        "imports": [
                {
                        "name": "SAS_TEST_DATA_EN",
                        "path": "/mnt/data/Workflow_Manager/sas_test_data_en.sas7bdat",
                        "difftype": "CONTENT"
                },
                {
                        "name": "SAS_TEST_DATA_FR",
                        "path": "/mnt/data/Workflow_Manager/sas_test_data_fr.sas7bdat",
                        "difftype": "CONTENT"
                },
                {
                        "name": "SAS_TEST_DATA_DE",
                        "path": "/mnt/data/Workflow_Manager/sas_test_data_de.sas7bdat",
                        "difftype": "CONTENT"
                }
        ],
        "exports": []
}
```

Strace

```
openat(AT_FDCWD, "../r/export_txt.R", O_RDONLY) = 3</mnt/code/R/r/export_txt.R>
openat(AT_FDCWD, "/mnt/data/Workflow_Manager/r_14_first.txt", O_RDONLY) = 4</mnt/data/Workflow_Manager/r_14_first.txt>
openat(AT_FDCWD, "/mnt/data/Workflow_Manager/r_14_first.txt", O_WRONLY|O_CREAT|O_TRUNC, 0666) = 4</mnt/data/Workflow_Manager/r_14_first.txt>
write(4</mnt/data/Workflow_Manager/r_14_first.txt>, "Hello\nWorld1\n", 13) = 13
openat(AT_FDCWD, "/mnt/data/Workflow_Manager/r_14_second.txt", O_WRONLY|O_CREAT|O_TRUNC, 0666) = 4</mnt/data/Workflow_Manager/r_14_second.txt>
write(4</mnt/data/Workflow_Manager/r_14_second.txt>, "Hello", 5) = 5
write(4</mnt/data/Workflow_Manager/r_14_second.txt>, "\n", 1) = 1
write(4</mnt/data/Workflow_Manager/r_14_second.txt>, "World2", 6) = 6
```


```
{
    "version": "v1",
    "program": "/mnt/code/R/r/export_txt.R",
    "imports": [],
    "exports": [
        {
            "name": "r_14_first.txt",
            "path": "/mnt/data/Workflow_Manager/r_14_first.txt"
        },
        {
            "name": "r_14_second.txt",
            "path": "/mnt/data/Workflow_Manager/r_14_second.txt"
        }
    ]
}
```

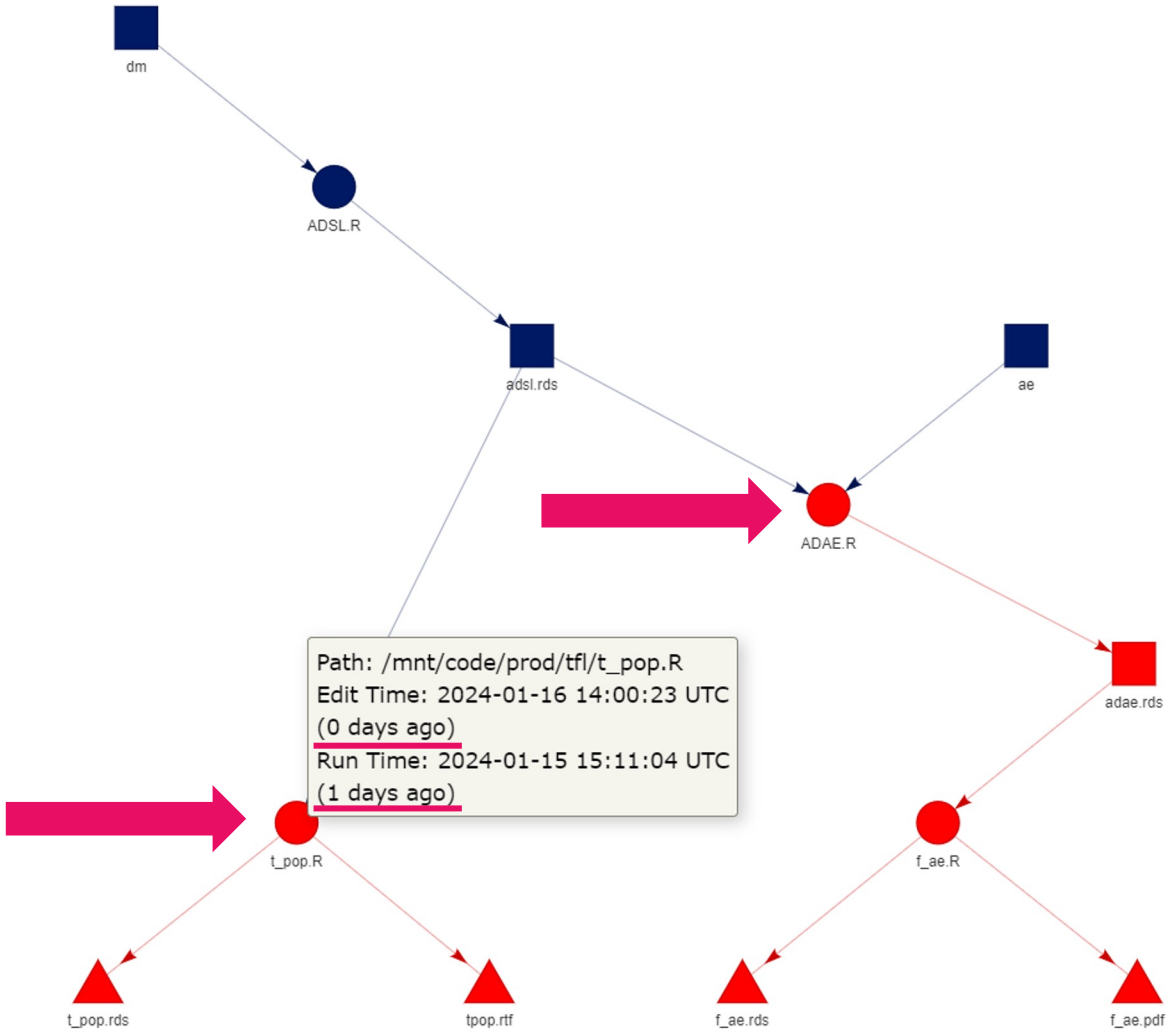

JSON

```
{
  "version": "v1",
  "program": "/mnt/code/prod/adam/ADAE.R",
  "imports": [
    {
      "name": "ae",
      "path": "/mnt/imported/data/SDTM/ae.sas7bdat",
      "difftype": "CONTENT"
    },
    {
      "name": "adsl.rds",
      "path": "/mnt/data/ADAM/adsl.rds",
      "difftype": "CONTENT"
    }
  ],
  "exports": [
    {
      "name": "adae.rds",
      "path": "/mnt/data/ADAM/adae.rds"
    }
  ]
}
```

Scheduling Jobs


```
# /mnt/code/prod/adam/ADAE.R
rule mntcodeprodadamADAER:
    input:
        "/mnt/code/prod/adam/ADAE.R",
        "/mnt/imported/data/SDTM/ae.sas7bdat",
        "/mnt/data/ADAM/adsl.rds"
    output:
        "/mnt/data/ADAM/adae.rds"
    script:
        "/mnt/imported/code/Workflow_Manager/blocking_job.py"
```

Dependency Visualisation



Parsers Testing & Validation

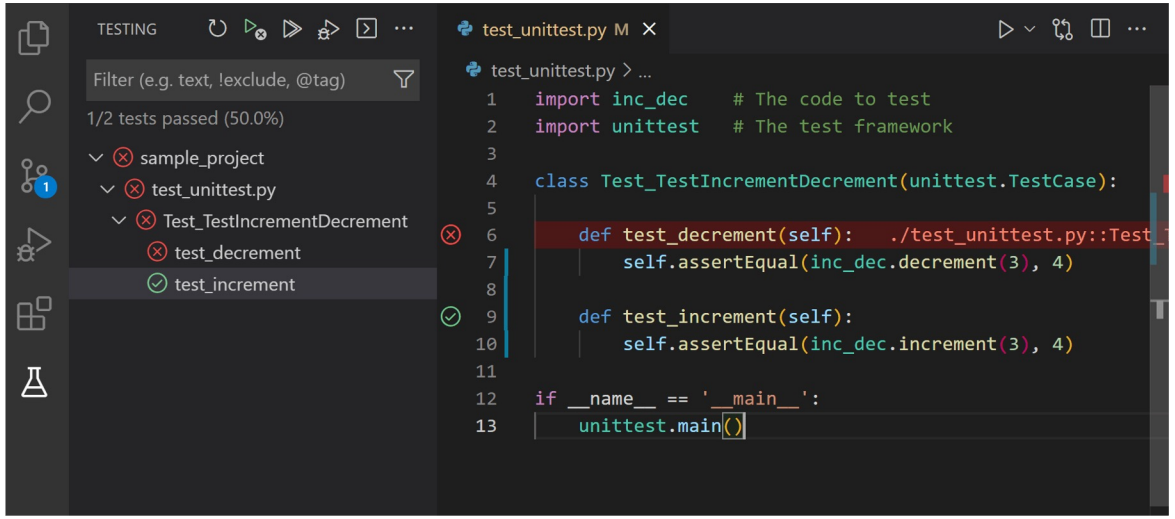
Ensure that for each language, the **inputs create the expected outputs**

Validate that the software works as expected

Unit test **one thing at a time** to isolate any potential failures

Example

Different ways to save an output in SAS are all recognised in the respective JSON file

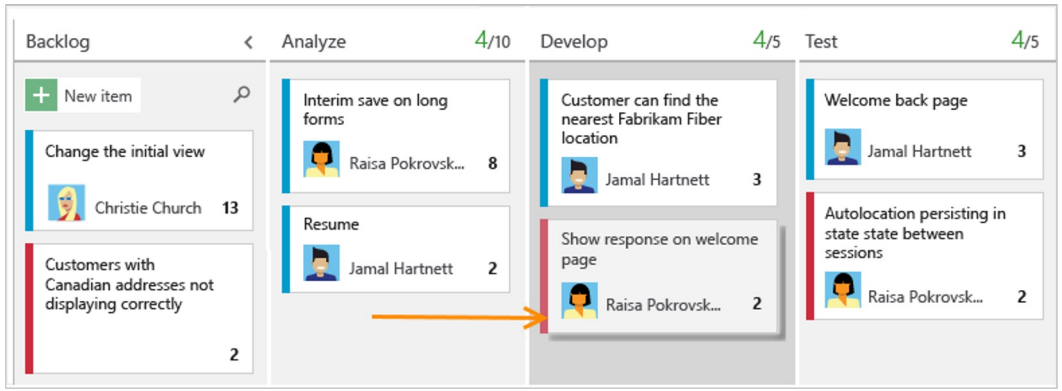


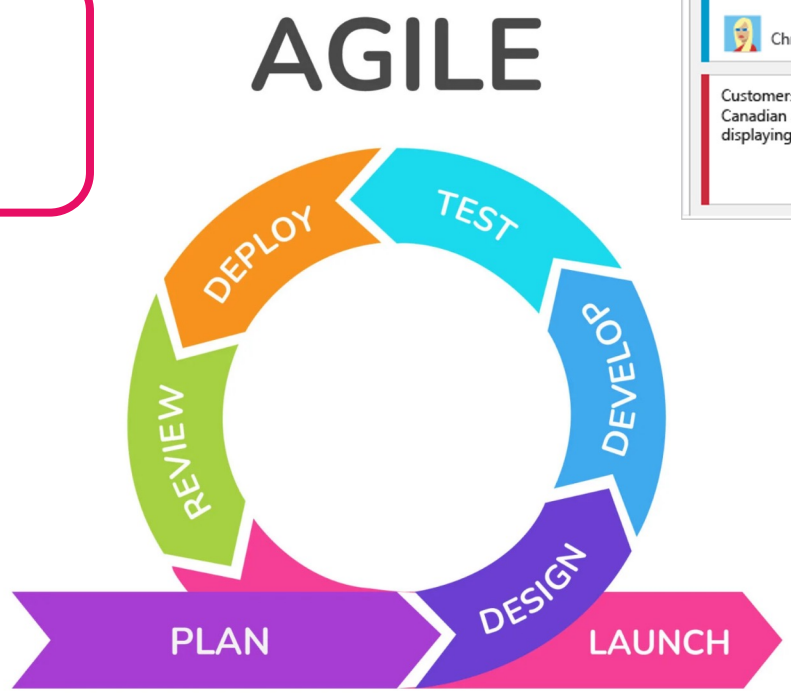
Reflection

Project Management

User Stories

As a …
I want …
So that …





2 week sprints

Azure DevOps

DSU

Sprint planning

PI planning

Retrospectives

Scrum Master

Lessons Learned

The importance of having end users in a development team

Benefits of AGILE

Software development / DevOps skills

Great way to experience independent learning at work

Similarities between QC study programming

Thank you!

Questions?

callum.charalambides@veramed.co.uk
megan.harries@veramed.co.uk