RESEARCH AND POLICY ANALYST · EI DATA CONVERSION

Proven against SAS

The new process had to give the same results as the SAS programs it was built to replace. Every step was tested against the SAS output, and the final outputs were compared value for value: every value was identical.

Explanatory diagram of the testing, drawn without data

Each step → the final outputs → identical

THE LADDER OF TESTS

EVERY STEP, EACH RAW INPUT

Tested before moving on

Every step, on each raw input, compared with the SAS output before the next began.

THEN, THE FINAL OUTPUTS

Compared as written

Nothing rounded or tidied first.
Blanks compared as blanks.

THE RESULT

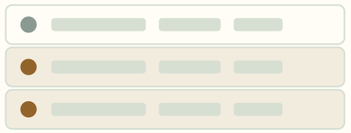
Every value identical

Every value in the final outputs matched the SAS output.

ONE DIFFERENCE, SETTLED · DRAWN WITHOUT DATA

1 Two rows share a key

ROWS, AS READ



Two rows carry the same key, shown by the same marker. Only one of them can stay.

2 SAS keeps the first row

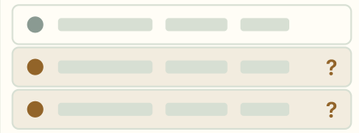
ROWS, AS READ



SAS reads the rows in order and keeps the first one it meets.

3 No “first” in parallel

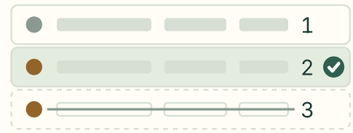
ROWS, AS READ



The new engine splits the work across many cores, so either row may be kept. The test reports the difference.

4 The fix: number the rows

ROWS, AS READ ORDER



Each row is numbered as it is read, and the lowest number is kept. The test passes.

I decided the rule; the agent wrote it; the test proved it.

The testing, simplified. The example is drawn without data: the rows are blank, and the numbers only show the order the rows are read in.

01 / EVERY STEP

Tested before moving on

Each step was compared with the SAS output, on each raw input, before the next one began. An error was found close to where it was made.

02 / STRICT

No tolerance at the end

The final outputs were compared as written: nothing rounded or tidied first, and blanks compared as blanks.

03 / LITERAL

Match first, speed later

The code was made to behave as SAS does, then made faster. The tests were run again after every change.

> **In more detail.** SAS applies rules that are easy to miss: how missing values sort and compare, the order rows are read in, how a value is carried from one row to the next, and which row is kept when two share a key. A database engine that works in parallel does none of this by default. Each rule was reproduced on purpose, then proven by the tests.

An explanatory reconstruction of how the results were checked. The example is drawn without data.

Thomas Eapen
nowlabs.net/research-policy-analyst

Drawn without data. The diagram is an explanatory reconstruction. Tools: AI coding agents, for building and reviewing.

Contact me to discuss this work.