Reliable Evaluation of Small Data Pipelines

Ada Lovelace
Institute of Analytical Engines
Example University
Berlin, Germany
e-mail: ada@example.edu

***Abstract*—Small data pipelines are often treated as too simple to require systematic evaluation. We describe a lightweight audit that records input contracts, transformation counts, and reproducibility checks at each stage. In a six-week case study, the audit found 17 defects that would otherwise have appeared only in the final report. The starter demonstrates the compact article structure required by the IARIA class.**

Keywords-data pipelines; reproducibility; software quality; auditability

I. Introduction

Reproducibility is often discussed as a property of a final result [1]. A result can be numerically stable while the path that produced it remains unclear. We therefore treat each stage as a contract with observable inputs and outputs, preserving small facts at every transformation boundary.

II. Audit design

For every stage, the audit records the input schema, row count, checksum, software version, and output row count. A failed check stops the run and stores the diagnostic beside the stage log. This makes a later rerun explain both what changed and where it changed, while keeping the metadata findable and reusable across tools [2].

III. Case study

We applied the audit to a daily environmental-data pipeline. Table I shows the checks that found the most defects.

TABLE I. Defects detected by the staged audit.

Check	Runs	Defects
Schema and units	42	8
Record counts	42	5
Version capture	42	4

IV. Conclusion

The audit adds a small amount of metadata while making pipeline failures local and explainable. The next step is to compare the cost of these checks with the time saved when a downstream report must be reproduced.

References

[1] V. Stodden, "The scientific method in practice: Reproducibility in the computational sciences", *Computing in Science and Engineering*, vol. 12, no. 3, pp. 8–10, 2010.

[2] M. D. Wilkinson et al., "The fair guiding principles for scientific data management and stewardship", *Scientific Data*, vol. 3, p. 160 018, 2016.