Testing Big Data Systems: A Multivocal Review and LLM-Driven Support Tool

Icaro S. de Oliveira
Master's Student at PPGCC
Ceará State University (UECE)
Fortaleza, Brazil
icaro.santos@aluno.uece.br

Ismayle S. Santos
Advisor
Ceará State University (UECE)
Fortaleza, Brazil
ismayle.santos@uece.br

Rossana M. C. Andrade
Co-advisor
Federal University of Ceará (UFC)
Fortaleza, Brazil
rossana@ufc.br

I. Context

Big Data refers to the vast volumes of data generated at high velocity and in a wide variety of formats. The defining characteristics of Big Data are often encapsulated by the "Five Vs": Volume, Velocity, Variety, Veracity, and Value [3][5]. Gao et al. (2016)[4] and Arshad et al.(2023)[3] emphasize that Big Data quality is affected by challenges like user consent, data ownership, and source limitations, which can lead to ethical concerns and inaccurate results.These quality issues directly affect the reliability of decisions derived from data-intensive systems.

This research builds upon a systematic literature review published at SBQS 2024, which mapped methods, tools, and best practices for testing Big Data systems. Additionally, we contributed to an experience report and survey published at ICEIS 2025, focusing on testing practices in a real-world governmental Big Data platform. To broaden this foundation, we are currently conducting a snowballing study based on two established secondary studies and performing grey literature mining (LinkedIn, Medium, Dev.to, Stack Overflow). Using LDAvis, we compare the themes emerging from grey and formal sources, aiming to consolidate these insights into a multivocal literature review suitable for journal publication.

Looking ahead, this project proposes the development of a support tool for data quality assurance, powered by Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG). The tool will guide users through a structured checklist to define quality rules, translate them into a domain-specific language (DSL), and automatically generate executable code. These rules will then be executed in a Spark-based cluster, enabling scalable testing for Big Data systems.

1) Research Questions: RQ1: What are the methods and techniques for Big Data testing?

RQ2: What tools and methods for the testing community are most discussed?

RQ3: How can an LLM+RAG, checklist-driven approach reliably translate data-quality requirements into a DSL and executable Spark tests?

II. Aims

This research aims to (i) develop a multivocal literature review on testing Big Data systems by combining academic and grey sources. (ii) design a support tool that assists in defining data quality rules through a checklist-driven approach powered by Large Language Models (LLMs).

III. Method

This research combines multiple methods across two main phases. First, we conducted a systematic literature review (SBQS 2024) and are currently extending it through snowballing and grey literature mining (e.g., LinkedIn, Dev.to), using LDAvis for comparative topic modeling. These efforts will be consolidated into a multivocal literature review [1].

Second, we adopt a Design Science Research (DSR) approach to develop a support tool for data quality testing. The tool leverages LLMs + RAG to generate quality rules from a checklist, translate them into a DSL, and execute the resulting code on a Spark cluster for validation at scale [2].

IV. Expected Results

We are extending a prior systematic review through snowballing (47 papers) and mining grey literature from online sources (3301 posts). These data will support a multivocal literature review combining academic and practitioner insights. Based on this foundation, we follow a Design Science Research (DSR) approach to develop a support tool. The tool uses LLMs enhanced with RAG to guide the creation of data quality rules.

A. Next Steps

The next steps include finalizing the multivocal literature review, developing the support tool using the DSR approach, and evaluating it through a case study on a real Big Data platform.

References

[1] Wohlin, C. (2014, May). Guidelines for snowballing in systematic literature studies and a replication in software engineering.

[2] Per Runeson and Martin Höst. 2009. Guidelines for conducting and reporting case study research in software engineering. Empirical software engineering.

[3] Arshad, I., Alsamhi, S. H., & Afzal, W. (2023). Big Data testing techniques: taxonomy, challenges and future trends. arXiv preprint arXiv:2111.02853.

[4] Gao, J., Xie, C., & Tao, C. (2016, March). Big data validation and quality assurance–issues, challenges, and needs. In 2016 IEEE symposium on service-oriented system engineering (SOSE).

[5] Daase, C., Staegemann, D., & Turowski, K. (2024). Overcoming the Complexity of Quality Assurance for Big Data Systems: An Examination of Testing Methods.