A REVIEW OF PAIR PROGRAMMING (GEN-AI) TOOLS IN DATA ENGINEERING

Prathik Reddy Chintala*1, Vishnu Vardhan Reddy Chilukoori*2,
Sadha Shiva Reddy Chilukoori*3

*1Graduate Student, Business School, University Of Colorado, Denver, CO, USA.
*2Lead Data Engineer, Amazon.com Services LLC, Seattle, WA, USA.
*3Lead Data Engineer, Meta Platforms Inc, Menlo Park, CA, USA.

DOI: https://www.doi.org/10.56726/IRJMETS53795

ABSTRACT

Pair programming, a collaborative coding practice, has been widely adopted in software development, but its application in Data Engineering is still emerging. We will review the benefits of pair programming in Data Engineering, particularly with the integration of Generative AI (Gen-AI) tools. We discuss how pair programming enhances code quality, reduces errors, and improves knowledge sharing among data engineers. Furthermore, we examine how Gen-AI assistants, like Meta-AI, Github co-pilot, and Google’s Gemini in Colab, can augment pair programming by providing real-time suggestions, automating routine tasks, and facilitating more efficient problem-solving. Our review reveals several advantages of combining pair programming with Gen-AI in Data Engineering, such as improved code quality, enhanced collaboration, and more.

Keywords: Pair-Programming, Data Engineering, Generative AI, Collaboration, Productivity, Code Quality, Problem-Solving, Data Management, Data Security.

I. INTRODUCTION

Data Engineering is a critical component of modern data science, involving designing, developing, and maintaining large-scale data systems. Data engineers face increasing challenges in managing and processing data efficiently as data volumes and complexities grow. Pair programming, a collaborative coding practice, has been widely adopted in software development to improve code quality, reduce errors, and enhance knowledge sharing. Generative AI (Gen-AI) tools have emerged, offering real-time suggestions, automation, and problem-solving capabilities. We will highlight the benefits of combining pair programming with Gen-AI tools in Data Engineering and some downsides.

II. METHODOLOGY

We conducted a simple coding challenge where we tried to build SQL code for calculating standard metrics with a combination of filters, aggregations, and engines. We wrote the code samples without using AI tools and repeated the same with the Gen-AI tools to investigate the impact of pair programming with Gen-AI tools on data engineering projects. Our sample consisted of 5 queries on Presto, Hive, and Spark SQL dialects using the TPCH data. These queries included steps such as casting timestamps, flattening maps, and array objects, as these tasks are implemented quite differently in the engines, resulting in many syntax errors.

III. KEY BENEFITS

Our experience suggests that combining pair programming with Gen-AI tools can significantly improve productivity, collaboration, and innovation in Data Engineering projects. By embracing this collaborative approach, data engineers can unlock new possibilities in data processing, analytics, and insights. Using pair programming with Gen-AI tools, we were approximately 50% faster in writing the SQL snippets and made fewer syntax errors. Moreover, we spent less time formatting the code, improving our focus on building the logic behind the metric. We have summarized our discoveries below:

● **Improved code quality**: Pair programming using Gen-AI tools leads to more robust, efficient, and maintainable code.
● **Enhanced collaboration**: Real-time feedback and suggestions from Gen-AI assistants facilitate more effective communication and knowledge sharing among data engineers.
Increased productivity: Gen-AI tools automate routine tasks, allowing data engineers to focus on complex problems and reducing overall project duration.

Faster problem-solving: Pair programming with Gen-AI enables data engineers to tackle complex challenges more efficiently, leading to faster solution development.

Better data management: Gen-AI assistants help data engineers optimize data processing, storage, and retrieval, leading to improved data quality and reduced data debt.

Improved data security: Pair programming with Gen-AI tools enhances data security by identifying potential vulnerabilities and suggesting security patches.

IV. CONCLUSION

Pair programming with Gen-AI tools offers a powerful combination for Data Engineering projects, enhancing code quality, collaboration, productivity, and problem-solving efficiency. As data volumes and complexities grow, embracing this collaborative approach can help data engineers unlock new possibilities in data processing, analytics, and insights. Data engineers can improve data management, security, and overall project success by integrating pair programming with Gen-AI tools.

V. REFERENCES


