



Chapter 3

Generative AI as an Accelerator for Data Engineering


N. Rajkumar

 <https://orcid.org/0000-0001-7857-9452>
*Alliance School of Advanced
Computing, Alliance University, India*

C. Viji

 <https://orcid.org/0000-0002-2759-8896>
*Alliance School of Advanced
Computing, Alliance University, India*

Balusamy Nachiappan

 <https://orcid.org/0009-0006-0951-8078>
Prologis, USA


A. Mohanraj

*Sri Eshwar College of Engineering,
India*

Judeson Antony Kovilpillai

Amity University, India

Oswalt Manoj S.

 <https://orcid.org/0000-0001-8251-2920>
*Alliance School of Advance Computing,
Alliance University, India*

ABSTRACT

The exponential growth of data in the digital era has heightened the importance of efficient data engineering practices. Traditional approaches to data ingestion, cleaning, transformation, and governance often require significant manual intervention, domain expertise, and development time, leading to bottlenecks in analytics and decision-making. By leveraging natural language prompts, contextual learning, and synthetic data generation capabilities, GenAI can facilitate code generation for data pipelines, improve data quality, enhance metadata documentation, and support intelligent decision-making in real-time environments. This chapter explores the integration of Generative AI into data engineering workflows, providing a comprehensive review of relevant technologies, conceptual frameworks, industrial

DOI: 10.4018/979-8-3373-5616-7.ch003

applications, and challenges. Several case studies highlight the tangible benefits of this synergy, including improved productivity, reduced operational costs, and higher data accuracy.

INTRODUCTION

In these days's virtual monetary device, records has emerged as a strategic asset, frequently defined because the "new oil." but, in comparison to crude oil, raw information in itself holds limited fee; it must go through a chain of strategies collectively with ingestion, cleansing, transformation, enrichment, and governance earlier than it can be converted into actionable insights. Those procedures together define the challenge of data engineering, which paperwork the backbone of modern-day analytics, commercial enterprise intelligence, and artificial intelligence structures. Over the last decades, data engineering has advanced appreciably. In the early 2000s, Extract-Transform-Load (ETL) methods dominated, usually going for walks in batch mode on relational databases. The upward push of huge data within the 2010s, supported with the aid of allotted computing frameworks like hadoop and spark, ushered in new paradigms which consist of data lakes, schema-on-examine, and actual-time streaming. Extra these days, cloud-native architectures and the concept of the information mesh have in addition decentralized and democratized information engineering responsibilities throughout organizations (duan et al., 2024; nguyen & hoang, 2024).

In spite of these technological shifts, a fundamental assignment persists making ready first-rate, integrated, and reliable information stays labour-in depth. Studies via gartner (2023) highlights that information scientists nevertheless spend an average of 60–eighty% of their challenge time on data instruction in place of analysis or modelling. This bottleneck not simplest delays insight generation however also restricts the scalability of data-driven projects.

Despite the fact that automation tools together with apache nifi, talend, and dbt have alleviated a number of the load, traditional information engineering workflows nonetheless face sizable obstacles. The art work often needs excessive guide overhead, requiring engineers to write down ETL scripts, control schema adjustments, and keep documentation. The tool environment is complex, forcing practitioners to navigate exceptional systems for ingestion, transformation, excellent tests, and storage. Talent shortages in addition compound the hassle, as data engineering calls for proficiency in a couple of programming languages, database structures, and cloud offerings—abilities which may be increasingly scarce. Scalability additionally stays a hurdle, particularly with rapidly developing data volumes and the rise of streamingIOTrecords that necessitates near actual-time processing. Furthermore,

24 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/generative-ai-as-an-accelerator-for-data-engineering/395322

Related Content

Focused Error Analysis: Examples from the Use of the SHEEP Model

Deborah J. Rosenorn-Lanngand Vaughan A. Michell (2016). *International Journal of Big Data and Analytics in Healthcare* (pp. 30-48).

www.irma-international.org/article/focused-error-analysis/171403

ICTs and Domestic Violence (DV): Exploring Intimate Partner Violence (IPV)

Bolanle A. Olaniran (2021). *International Journal of Big Data and Analytics in Healthcare* (pp. 31-44).

www.irma-international.org/article/icts-and-domestic-violence-dv/277646

A Brief Survey on Big Data in Healthcare

Ebru Aydindag Bayrakand Pinar Kirci (2020). *International Journal of Big Data and Analytics in Healthcare* (pp. 1-18).

www.irma-international.org/article/a-brief-survey-on-big-data-in-healthcare/253842

Online Tourism Destination Image by User-Generated Content: A Systematic Literature Approach

Valerie Stopp, Nelson deMatosand Célia M. Q. Ramos (2025). *Advancing Smart Tourism Through Analytics* (pp. 291-314).

www.irma-international.org/chapter/online-tourism-destination-image-by-user-generated-content/362490

Information Strategy: Implementing and Managing a Digital Strategy in a Portuguese Company

Sérgio Maravilhas, Paulo Meloand Sérgio Goes Oliveira (2019). *Handbook of Research on Expanding Business Opportunities With Information Systems and Analytics* (pp. 225-251).

www.irma-international.org/chapter/information-strategy/208566