Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 5 Issue 3 (2025)

Generative AI Meets Data Engineering: Automating Code, Query Generation, and Data Insights in Large scale Enterprises

Kushvanth Chowdary Nagabhyru[1], Majjari Venkata Kesava Kumar[2]
[1]Senior Data Engineer ORCID 0009-0004-7175-7024
[2]Assistant Professor, EEE department JNTU KALIKIRI Andhra Pradesh, India
keshavakumar.eee@jntua.ac.in

Abstract:
Generative AI for Large-scale Enterprise Data Engi- neering: Automating Code and Query Generation, Data Insights, and Analytics. In a data-driven economy, organizations across verticals show a growing demand for ingesting terabytes (or petabytes) of data frequently and performing rapid analytics on the data. However, a sizable effort is required to transform the data into a standardized, query-optimized format. Data scientists have long wished to automate code and query generation for their projects. The advent of generative AI holds the promise to enable such automation. Major technology companies have incorporated generative AI modules within their products. For instance, publicly available large language models (LLMs) can take natural language text as input and generate code in Python, Java, and SQL. OpenAI publicly hosts a jacuzzi to invoke these LLMs, and the open-source community has built similar services. These tools have gained immense popularity, with over 100 million users in just one year. This paper discusses the integration of generative AI tools into large-scale enterprise data engineering workflows for code and query generation, data insights, and analytics.

Keyword: Generative AI, Enterprise Data Engineering, Code Generation, Query Generation, Data Insights, Data An- alytics, Large Language Models, Natural Language Processing, Python, Java, SQL, Data Standardization, Query Optimization, Automation, Data Transformation, Open-Source Services, Cloud- Hosted Models, Enterprise Workflows, AI Integration, Data- Driven Economy.

1. Introduction

Generative AI broadly refers to techniques that enable machines to produce new content such as images created by DALL·E or art created by Midjourney. It is also used to author computer code as demonstrated by GitHub Copilot. At the forefront of technological advancement, generative AI is poised to redefine workflows through certain types of automation. In Big Data, it promises to simplify the gen- eration of complex queries required to extract specific data insights. In data engineering, the technology is advancing toward generating and automating code. The discussion begins by outlining the current state of generative AI applications in data engineering, then examines role-specific scenarios for data infrastructure engineers, followed by related challenges and issues. The traditional approach to data engineering in- volves building pipelines and preparing data for downstream consumption. The two main goals of the activity are to manage large amounts of infrastructure and to explore and code data pipelines. Large enterprises have taken concrete steps in introducing automation in the first category by optimizing the necessary infrastructure—serverless setups, cost/size optimiza- tions, access management—and making the setup self-service so teams can directly spin reliable pipelines and data marts. The second, more creative and exploratory part is still done by data infrastructure engineers. Supported by spreadsheets and email interaction, it is becoming increasingly complex, slow, and error-prone.

A. Overview of Generative AI and Its Significance in Data Engineering

Generative AI, also known as foundation or large-language AI models in 2024, are deep-learning models trained on both labeled and unlabeled data. Trained in a self-supervised manner using unlabeled data, Generative AI excels at complex tasks such as text generation, speech production, and language translation. In the domain of enterprise data engineering and data analytics, Generative AI has evolved into a powerful en- abler. Recent developments in open-source-release foundation models and cloud-offering AI capabilities unlock enterprise data. Leveraging natural-language programming capabilities, Generative AI automates code generation, query creation, data analysis, and visualization.

I. The Role of Generative AI in Modern Enterprises

In the current landscape, large-scale enterprises generate massive volumes of data across diverse domains. This infor- mation unlocks vast potential for valuable insights through advanced analytics. As the quantity and complexity of data escalate, data engineers dedicate significant effort to organiz- ing, processing, maintaining, and managing it. By automating data engineering workflows, enterprises can reduce costs, save time, and increase efficiency. Generative AI can automate the generation of code, queries, and analytics for large-scale en- terprises. Conversational AI agents—built on the transformer architecture—are trained on large-scale corpora and excel at

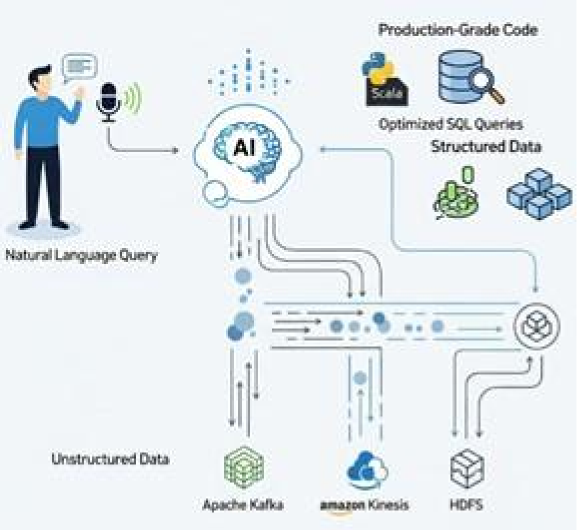


Fig. 1. Conversational AI for Data Engineering Automation

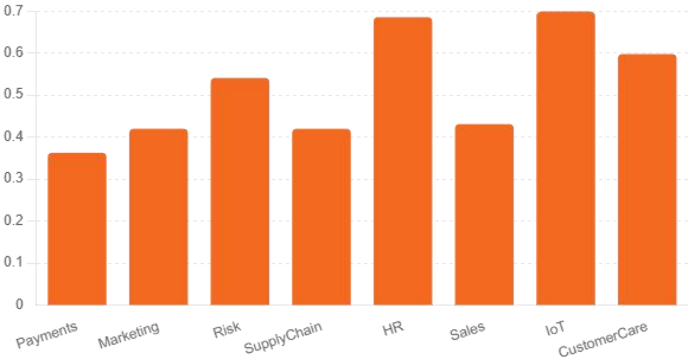


Fig. 2. Code Automation Efficiency by Team

- **Code Automation Efficiency**

$Timesaved$

$$= T_{-T} \quad , \quad CAE = \frac{Tman - Tauto}{} = 1 - \frac{Tauto}{} .$$

$manautoTman$

$Tman$

context understanding and query generation. These capabil- ities allow the agents to generate production-grade code in programming languages such as Python and Scala. Data engi- neering code contains intrinsic domain-specific details. When companies employ domain-specific data engineering frame- works or architectures, the generated code must incorporate specific keywords and methods relevant to those frameworks. The code typically connects to external sources of structured or unstructured data—for example, Amazon Kinesis, Apache Kafka, or HDFS—and processes the data with frameworks like Apache Spark. Conversational AI agents can generate semantic table descriptions of large-scale enterprise tabular data, extract a business glossary from tables and columns, and create optimized SQL queries from natural-language queries. Additionally, they can analyze data distributions and highlight anomalies or outliers.

A. The Impact of Generative AI on Data Engineering Work- flows

Today's ecosystem of generative AI automation tools has become indispensable for many enterprise-scale data engineer- ing workflows. Its value proposition is striking: Any required code—whether procedural or declarative—can be generated based on plain-English explanations, making automation ac- cessible to every member of the team. Scalability is not a concern either; the model learns from your body of data to deliver its best effort regardless of project size. Furthermore, a natural-language query approach offers a familiar, accessi- ble medium for data consumption. Business team members need only articulate their questions in conversational language without delving into technical details or consulting with a data engineer. Finally, given the enormous information vol- ume managed on big-data platforms, generative AI becomes instrumental in identifying useful sets of correlations, outliers, and patterns, as well as designing insightful visualizations to illustrate these findings.

II. AUTOMATING CODE GENERATION

Numerous generative AI–based methods automate code generation in data engineering. State-of-the-art natural lan- guage processing models parse human-language data re- quests and create corresponding ingest, transform, or integrate pipelines. This automation streamlines data engineers' work- flows, enabling scalability at enterprise intensity. Real-world examples demonstrate that enterprises can deploy simple data engineering functions quickly and efficiently. For instance, a large enterprise with more than 46,000 employees automated the creation and management of more than 800 database scripts, procedures, and jobs that support critical functions. Key performance indicators were tracked and analyzed for each generated code segment.

A. Techniques for Code Automation

Enterprise-scale data engineering workflows involve the de- velopment and testing of diverse code artifacts across various platforms, including SQL, Python, Informatica Data Factory, shell scripting, and XML. Generative AI now enables the automated generation of development and test code, even for complex use cases within large enterprises. It can use re- quirements and test case descriptions to automate the creation of data engineering code and generate test cases, test data, and test automation scripts, thereby significantly accelerating the development and testing processes. In the domain of data query generation, natural language processing (NLP) techniques are gaining momentum. Enterprises can lever- age NLP to automatically generate data quality, monitoring, and validation queries—key activities within data engineering workflows. Moreover, by harnessing NLP, organizations can generate data and analytics insights derived from information embedded in big data. A series of use cases have been developed to demonstrate the efficacy and enterprise readiness of generative AI in automating code and query generation, as well as in delivering robust data insights—all crucial for large-scale data engineering. These examples underscore that data engineering contributes to each of these facets. Finally, an examination of the challenges associated with implementing generative AI for data engineering within modern enterprises illuminates considerations related to ethics and governance.

- Query Optimization Gain (QOG)

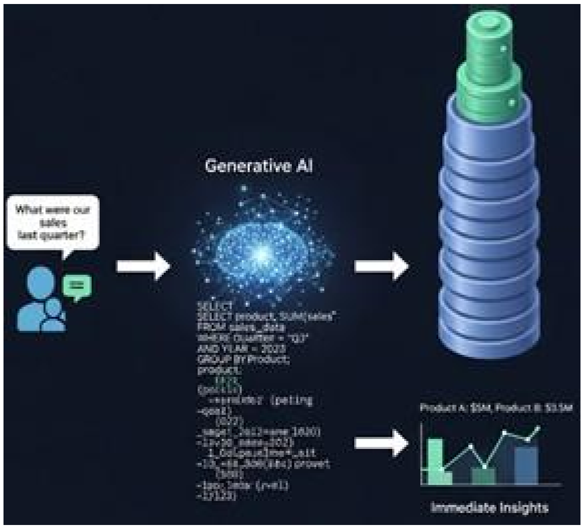
$QOG = L0L0 - L1 = 1 - L0L1.$ (2)

B. Case Studies of Automated Code Generation

The use of generative AI to automate code generation in data engineering workflows is a relatively new field where few concrete examples have yet emerged. SQLærn.ai's recent blog offered an initial proof of concept that generates not only SQL queries but also the connecting code for diverse databases—a capability essential in data engineering use cases and illustrative given SQLærn's limited two-day development period. Similarly, Supabase's ChatSpot chatbot has been in- tegrated with OpenAI models to convert English requests directly into SQL queries, supporting data exploration through natural language queries. Product support is also expanding; for instance, Databricks SQL's chatbot features a ”query generation” function that translates natural language prompts into queries. Companies in data management are actively exploring these capabilities: IBM reckons that generative AI will ”enable automation through every stage of AI lifecycle and across many personas including DataOps. Support for automation: AI-assisted data pipeline development, including code generation”. IBM's DataOps automations now encompass data pipeline development and batch pipeline scheduling encompassing multiple execution engines and clouds. Okera has produced a prototype that enhances data governance by automatically generating PySpark code from English prompts. The feature sets offered by Large Language Models (LLMs) for data engineering continue to evolve rapidly.

Query Generation with Generative AI

Natural language processing is increasingly popular, en- abling the generation of queries or search inputs for graphical data visualization tools and pattern detection during query execution. Thesemethods are also better capable of summa- rizing the results of the executed queries and generating answers for automation and human consumption. Additionally, higher layers in data analytics pipelines are incorporating such technologies, easing data engineering tasks while offering a descriptive and predictive view. However, summarized results are more prone to inconsistencies when the underlying data quality is poor or lacks timeliness, granularity, or correctness. Prompt engineering has emerged as an art form, where care- fully designed prompts with real-time context and examples can significantly improve the quality of responses. Holistic ecosystem development is becoming critical, as regulation and ethics need to be introduced at every level—from data collec- tion and movement to data engineering and data engineering automation—to optimize the quality, usefulness, and inter- pretability of the results. Enterprises typically maintain large


Generative AI
What were our sales last quarter?
Product A: $5M, Product B: $3.5M
Immediate Insights

```
SELECT
SELECT product, SUM(sales"
FROM sales_data
WHERE Quarter = "Q3"
AND YEAR = 2023
GROUP BY Product;
product;
```

Fig. 3. Natural Language to SQL with Generative AI

production workloads and face scalability challenges in using generative artificial intelligence, as data volumes grow rapidly without corresponding increases in data-engineering arms. Planning ahead is therefore vital to ensuring that enterprise- level deployments can be successfully scaled.

A. *Natural Language Processing in Query Generation*

Natural language processing (NLP) techniques allow natural language queries to be automatically translated into structured queries that retrieve the appropriate data. Enterprise-scale datasets often involve queries dozens or hundreds of lines long. Manually generating such queries is a highly skilled and error-prone task, creating a bottleneck in the enterprise data engineering workflow. Automating query generation with generative AI allows users with a simple natural language question or statement to retrieve exact sets of results from an enterprise dataset. Large-scale enterprises generate petabytes of data every day, and those volumes only grow with time. Generative AI techniques can help organizations draw immedi- ate insights from their information troves, across these massive datasets. From White House budget documents to tweets, to YouTube comments, generative AI is helping organizations search, summarize, and visualize their data like never before.

B. *Performance Metrics for Generated Queries*

Metrics are an integral part of every system, providing meaningful information on the system's performance. The primary objective of an evaluation metric is to represent how well the system has performed and analyze each component. To begin with, language models today have the capability to perform tasks in multiple languages. As a result, during query generation it is possible that the generated queries can be in any language, which is why it must be guaranteed that the generated query is in the required language. For this purpose, the Query Language Accuracy metric can be used. This metric captures the number of generated queries in the required language, followed by normalization. The test dataset uses the desired language for the queries like Czech, German, Italian, Arabic, or Japanese. Another important contribution to the test is Query Translation Accuracy—whether the translation from question to query is correct or not. Since the existing translation accuracy metric does not consider query structure, a new accuracy metric is proposed that determines if the structure of the generated and target queries is the same. In cross-domain evaluation, it has been observed that the performance of a model varies with different languages. Cross- domain is evaluated by finding translation accuracy on an out-of-domain database, Spider, using the Exact Set Match metric. The Exact Set Match metric measures the correct prediction of the set of tables and columns used across databases. Its performance enhances with increase in training data. Further, Natural Language to Query helps in minimizing the involvement of users in many areas. It enables users to ask questions in any of the existing natural languages and get answers from the required database. On running a SQL query, the reply can be either the requested data from the database or error messages in cases such as a malformed query or query with invalid elements. When a particular query raises an error, there is a possibility that the error is generated because of the grammar of the query itself or due to selection of inappropriate elements in the query. The Query Syntax Correctness metric helps to figure that out.

DATA INSIGHTS AND ANALYTICS

Data analysis is a key aspect of business investments; a thorough analysis will provide valuable transparency into the activities of the business. AI enables businesses to extract insights from massive datasets that put human analysis beyond reach, and it is now possible to automate these insights. Transforming an insight into a visual representation requires a natural language query; combining these processes makes it possible to automatically generate analytics visuals or dashboards from data. Automation can also generate ML model candidates, transforming a business problem statement into a model for evaluation. Organizational databases are typically utilized for transactional input/output while centralized data repositories are utilized for business intelligence and analysis. A pipeline can be created to automatically process data as it gets generated and execute analytics on the data in the reposi- tory. Answering questions against such a data-repository—the proverbial search for the needle in the haystack—is often performed utilizing a query system that makes use of particular design elements to allow for optimization of execution. The datastore must support fast analytical execution over billions of calls by the users. Modern users look for an easy interface; a natural language–based interface satisfies this need while generating the query automatically. Analytics is a never-ending cycle—users want to keep performing different analyses on their data—and an easy, natural language–based interface assists greatly in this desire. The key aspects of the datastore's operational efficiency are many, but execution time, fault tolerance, and cost all require particular attention. All these

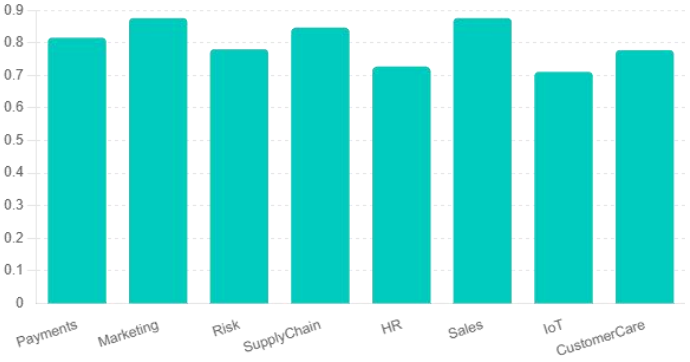


Fig. 4. Data Insight Precision by Team

aspects warrant analysis of their individual trade-offs when crafting the query system. When considering scalability, the system must be optimized to execute not only few queries over lot of data, but also, the ability to execute many queries on fast changing datasets.

A. *Leveraging AI for Data Insights*

Generative AI engines can also be employed to generate insights from the data in datasets. The Data Engineer can craft an appropriate template query which depicts the kind of insight sought from the Data. The AI engine then uses the Natural Language Processing (NLP) model to generate a suitable query that may require one or more operations such as joins, sub-selects, unions, and aggregation functions. The generated query is run on the dataset, and the results are fed to the Data Engineer. Based on the Data Engineer's feedback, the AI engine iteratively analyses the previous results and may invoke other operations like ranking, DOSKEY, or comparable analysis to achieve the desired output. As enterprises scale up, the corresponding growth in the amount of data generated is so vast and varied that identifying commercial decision-making use cases becomes almost impossible. These enterprises may utilize generative AI to discover potential use cases from the available datasets. A Data Quality Analyst may formulate suitable queries, which are then processed through an NLP model to perform various operations, including joins, sub- selects, unions, and ranking functions. The resulting data is analysed, and the generative AI engine suggests possible use cases and relevant analysis techniques that the enterprise can apply to the datasets.

- Data Insight Precision (DIP)

$$DIP = TP + FPTP. \tag{3}$$

B. *Visualizing Insights from Large Datasets*

Insights from large datasets often contain valuable nuggets of information, which may include answering specific queries, demonstrating a trend, detecting anomalies, providing compar- ative information, and so forth. A textual-based explanation typically lacks simplicity in conveying complex facts or narra- tives. Leveraging generative AI to visualize these insights can

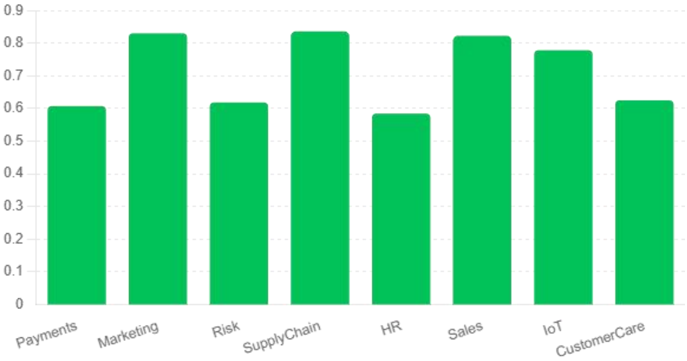


Fig. 5. Enterprise Automation Index by Team

make the explanation far more interesting and simpler to grasp. Indeed, services such as PowerBI, Tableau, and Dall-E offer the capability to generate such visual explanations. The crucial distinction when using generative AI for data engineering lies in its productivity and scalability for enterprise data engineer- ing workflows compared to non-data-engineering scenarios. For instance, consider a virtual assistant that automatically creates a PowerBI dashboard and reports for a dataset that contains weekly UAE cargo volume at a container terminal, including loading and discharge weight for each shipment.

Challenges in Implementing Generative AI

Generative AI unlocks automation and innovation for tech- nology in large enterprises, helping machines generate code, queries, and data insights. However, implementing generative AI introduces a new set of challenges. These include data quality and integrity, scalability, data governance, data privacy, and regulatory compliance. Data quality and integrity is one challenge. Generative AI applications require large volumes of good, clean, and well-structured data to help the AI model develop an understanding of the domain. If these aspects are not taken into consideration, the output generated by these applications can be misleading and biased. In complex IT environments with multiple data sources, data quality across systems needs to be assessed and ensured before adopting a generative AI-first approach. Without proper data governance and integration processes, generative applications will not scale seamlessly as the enterprise matures.

- Enterprise Automation Index (EAI)

foundation requires that the automated pipelines do not lead to error propagation and contamination, thereby guaranteeing the delivered data's quality and reliability. Therefore, data quality remains a central challenge, even for large-scale enterprises. The next-generation, AI-powered data engineering transforma- tion unlocks new ways of eliciting insights and analytics from all the often unpopular yet still necessary tedious and boring activities involved in data engineering. For example, natural language processing (NLP) techniques can now be leveraged to generate data engineering code and queries with little input from experts. Automated workflow and pipeline generation makes it possible to schedule data jobs automatically and even add quality gates at each step in the pipeline. Scaling data engineering with AI can thus reduce manual work, technical debt, and data shenanigans.

—

B. *Scalability Concerns in Large Enterprises*

Generative AI's recent advances empower data engineering with capabilities such as code generation, query generation, and data insights. These three levels of automation target some of the costliest and most time-consuming data engineering workflows. However, recent work tends to focus on scenarios where data is aggregated in smaller data repositories or lakes. In contrast, large-scale enterprises often handle billions of transactions daily, accompanied by complex transaction de- pendencies and intricate data quality rules. Accordingly, their data engineering operations must process large volumes of data with rigorous data quality and integrity checks. Data quality rules span multiple domains, including schema validation, data distribution validation, completeness, consistency, and accuracy. While data quality and data integrity involve rig- orous data checks, data privacy requires personal or sensitive information in data stores to be encrypted or masked. Despite their importance, existing generative AI-autonomous data en- gineering systems overlook these crucial enterprise concerns. Ignoring them not only impairs data quality and integrity but also risks breaching regulations like GDPR, HIPAA, and CCPA.

III. ETHICS AND GOVERNANCE IN AI AUTOMATION

The large-scale automation of data engineering workflows with generative AI however is not without challenges. Enter- prises wanting to use generative AI in their data engineering

$EAI = (C^{\alpha}Q^{\beta}A^{\gamma})\frac{1}{\quad} = \exp \frac{\alpha \ln C + \beta \ln Q + \gamma \ln A}{\quad}$. workflows need to carefully think about ethics and governance.

$\alpha+\beta+\gamma$

A. *Data Quality and Integrity Issues*

B. $\alpha + \beta + \gamma$

The quality of the underlying structured and unstructured data sets is crucial for the quality of the generated output. Assessing the quality of data in the context of an automation use case is another major consideration. Data quality, data

Data quality and integrity represent two of the most critical concerns when automating data engineering at scale. Inac- curate, inconsistent, or incomplete data results in unreliable outputs, undermining stakeholders' confidence. However, en- suring data quality extends beyond the enterprise boundary, as data obtained from third parties, although useful at times, often suffers from questionable integrity. Building a corporate data integrity, and data scalability considerations have a critical role to play. There may be interesting restrictions in terms of who can use generative AI for data product development, data scientist augmentation, data analytics, and the automation of other data engineering use cases. The ethics and governance considerations therefore have an important role to play in addition to the technical considerations related to building a successful generative AI solution. Currently, generative AI is at the forefront of technology discussions in practically every aspect of society. It will continue to expand into all domains and

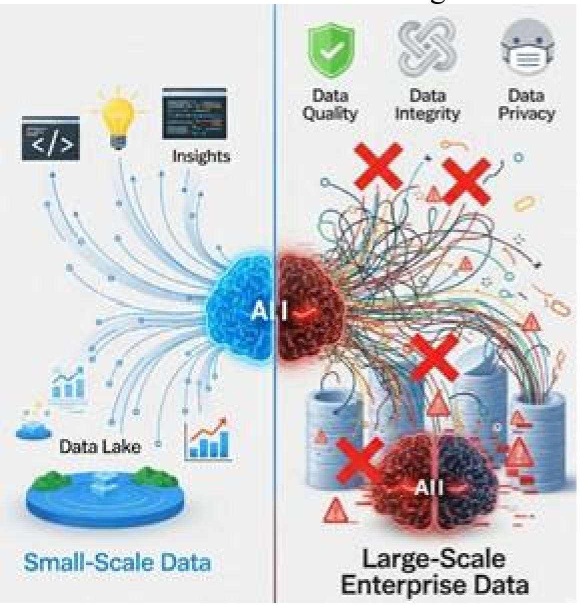


Fig. 6. Caption

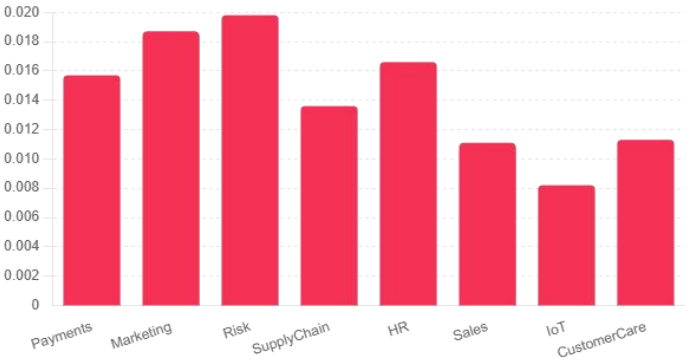


Fig. 7. Generative Model Utilization by Team

industries within the next decade. Along with that expansion, the use of generated content can lead to concerns related to copyright, regulation, and data privacy. Enterprises should take a cautious approach and put in place strong governance frameworks and guardrails. Organizations must adhere to internal and external policies to address the potential risks associated with content generation. On the generative AI technology front, there is active ongoing research in content attribution to enhance ethical and responsible usage.

- Generative Model Utilization (GMU)

$$Acceptedtimeperday = 3600p \cdot a \cdot s \text{ hours}, \quad (5)$$

$$GMU = Hpas/3600 \in [0, 1].$$

A. *Ethical Considerations in AI Deployment*

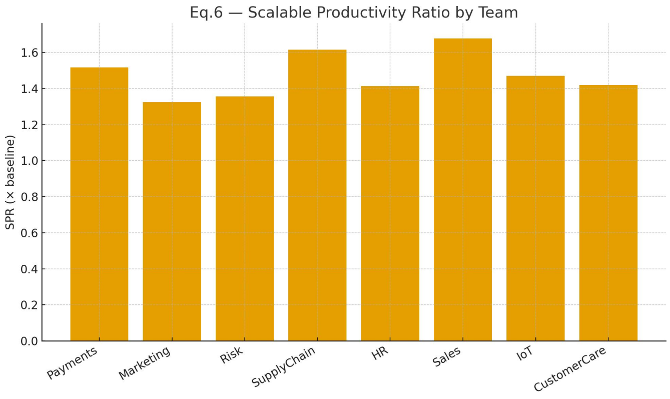
AI continues to reshape the way enterprises develop auto- mated data engineering workflows. Now capable of creating their own code and SQL queries, such algorithms can also be utilized to derive valuable insights from vast amounts of data. However, the extensive use of such powerful tools raises important ethical questions, necessitating that organi- zations strike the right balance to support developers effec- tively without compromising ethical standards. Recent studies demonstrate that generative model–based prompt engines en- hance efficiency and scalability in automated data engineering frameworks. By leveraging large language models powered with natural language processing, enterprises can now design business metrics and generate queries alongside Python or Scala code for data ingestion, data engineering, and data quality. Generative AI–based analytics engines tap into the inherent patterns and connections among data points to extract insights, which can be further visualized using tools like Neo4J or Amazon Quicksight. Despite this advanced utilization, ethical aspects remain crucial across all applications. Issues surrounding data quality and integrity during ingestion and quality validation not only impact the correctness of insights but also raise compliance concerns. Consequently, enterprises must operate these systems with the utmost consideration of governance, compliance, and privacy requirements.

B. *Regulatory Compliance and Data Privacy*

Data engineering in large-scale enterprises often requires processing sensitive and regulated information, making it imperative to maintain compliancy with predefined rules. Gen- erative AI automation must therefore orchestrate in accordance with these industry-specific regulations, security procedures, and data governance policies. Although generative AI enables creation of code and queries at speed, it remains susceptible to data quality and integrity issues, particularly when han- dling vast datasets. Regulatory guidelines introduce constraints regarding data utilization and management, especially when private or confidential records are concerned. Additionally, legal frameworks may impose limitations on using generative AI for automating tasks that handle such information. These considerations contribute to the complexity of deploying gen- erative AI at enterprise scale.

IV. FUTURE TRENDS IN GENERATIVE AI AND DATA ENGINEERING

Emerging technologies such as chatGPT/Bard-like applica- tions will further advance the current capabilities and enable new use cases. In the coming years, Generative AI will act as a business analyst on top of data engineering workflows for various enterprises at scale. Few examples are given below: Create code on the fly: — Build Spark pipelines based on a business description in plain English. — Modify existing pipelines based on additional changes. — Generate test cases to validate existing pipelines. — Generate GitHub issues based on sample logs. Generate queries from data, schema, and queries in natural language. This can be powered using



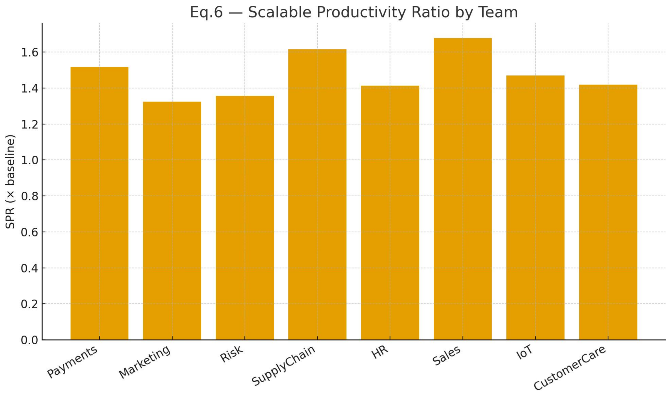


Fig. 8. Scalable Productivity Ratio by Team

traditional Natural Language Processing techniques. Provide data insights from traditional big data analytic systems in enterprises and visualize these insights. Enterprises can benefit from such solutions if the generated insights are contextual to their business.

A. Emerging Technologies in AI

Emerging typed-nature-inspired generative Alpha Tensor Algorithm (Alpha Tensor Algorithms) is designed to decom- pose a plethora of matrix multiplication forms, ranging from the classical to the contemporary. This algorithm identifies the most efficient algorithms for computing specific matrix prod- ucts. Additionally, Text-to-Image Diffusion Models constitute a significant advancement in AI-driven image generation. A Multi-agent chat tool has also been introduced, facilitating conversations between human agents and bots. The conver- gence of cognitive science and AI is creating models that simulate human yet genuinely intelligent trade-offs, shaping human–AI collaborations. A secondary literature reveals that a large proportion of foundation-model-related papers employ reinforcement learning with human feedback. Despite their remarkable adaptations to human traits, the latest models still do not possess third-order human theory of mind.

- Scalable Productivity Ratio (SPR)

$$SPR = \tau_0 \, \tau_1 \qquad (6)$$

B. *Predictions for the Next Decade*

Generative AI tools have been shaping the modern world for some time now. Recent developments across NLP, generative AI, and large language models have pushed AI capabilities into a new phase. The potential of new technologies such as OpenAI's GPT 4 or Google Bard, unaffiliated with this author, is undeniable. The next decade will see further ex- traordinary advances in the fields of AI and data science, but current data quality and governance concerns will remain. The computing power of large language models means that enabling creative and repeated interaction between humans and AI will become easier and cheaper. AI can already be used to generate code, spreadsheet functions, and queries from

Fig. 9. Scalable Productivity Ratio by Team

unstructured natural language input. It can also be used to analyze large quantities of data and provide natural language summaries of the results, freeing human intelligence for more strategic work. These foundations pave the way for highly creative applications and research. Generative AI is often viewed as an unnatural and mechanical process; however, the lack of practical nuance in some recent applications of AI mirrors certain dystopian portrayals. Data science is already an intensely creative process, driven by creativity as much as by computational power. Data engineering is a particularly important component of the data processing pipeline in large- scale organizations. Data format, filtering, and governance are critical. Code generation not only rapidly removes time- consuming coding tasks but also provides recommen dations and guidance. Query generation dramatically increases acces- sibility in complex data environments; NLP-enabled analytics allows those with limited expertise to derive insights from the data simply by interacting with the AI.

CONCLUSION

Generative AI for code and query generation, data insights, and analysis is revolutionizing data engineering workloads at scale—especially within enterprise-scale, big-data envi- ronments. Today's automation tools span everything from Jupyterhub or Zeppelin notebook query assistance (eg, Chat- GPT, Tableau Ask Data) to enterprise-scale data pipelines for code and query generation or code debugging. The goal is to further automate workloads, eliminate manual effort, and accelerate time to data and innovation. However, execution re- mains contingent on high-quality, battle-tested historical data, comprehensive governance and controls, and scalable enter- prise processes and methods. Automated-environment data and model governance, auditing, explainability, and observability are critical for success: Enterprises must ensure AI systems are governed, ethical, and compliant. When these strategic, ethical, and operational considerations are addressed, Genera- tive AI becomes an indispensable assistant in navigating vast information landscapes.

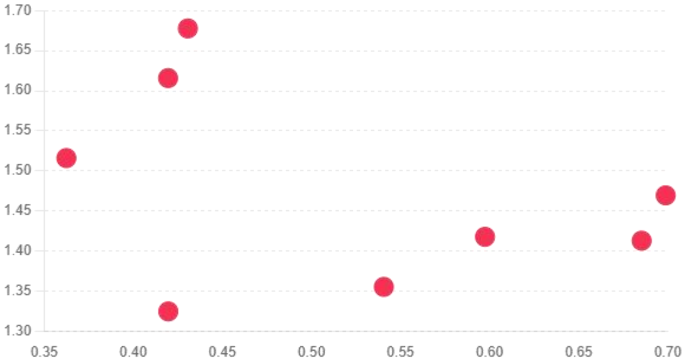


Fig. 10. Relationship Code Automation Efficiency vs Scalable Productivity Ratio

A. Final Thoughts and Key Takeaways

Generative AI, such as OpenAI's ChatGPT and Codex, is becoming an important productivity tool for Cloud Enterprise Data Engineering at scale. Key applications include engineer- ing automation of code and query generation, implementation of data quality and integrity checks, basic data profiling and analytics, and analytics automation for Data Insight and Data Storytelling. These technologies truly move Data Engineering beyond the concept of Big Data, into the real world of En- terprise Data with Volume, Variety, and Velocity at scale. En- terprises are incorporating natural language processing (NLP) techniques to automatically generate queries for insight on large volumes of data. Generative AI can also help automate many manual and repetitive data engineering tasks, including writing code for data ingestion and transformation workflows, and creating data quality and integrity checks. As with any new technology, there are challenges around data quality, data integrity, governance, and management. As these tools become further embedded into Data Engineering at scale, ongoing effort in these areas will help minimize risk and maximize benefit for any organization.

References

1. G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol.
3. 2. Oxford: Clarendon, 1892, pp.68–73.
4. I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
5. K. Elissa, "Title of paper if known," unpublished.
6. R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.
7. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
8. M. Young, The Technical Writer's Handbook. Mill Valley, CA: Univer- sity Science, 1989.