From ETL to ELT to LLMs: Redefining Data Engineering in the Generative AI Era

Jayanth Veeramachaneni

Independent Researcher
Missouri University of Science and Technology, Rolla

***Abstract*— Generative AI is radically changing the career of data engineering, particularly through the application of generative AI, and more specifically, Large Language Models (LLMs). The following review outlines the transformation that is to be conducted in the field to convert the traditional extraction, transform and load (ETL) systems into dynamic extract, load, transform (ELT) systems and eventually to the intelligent, LLM-driven data pipelines. The given paper is dedicated to the critical analysis of the pipeline generation process, semantic transformation, and errors that are minimized and generated automatically and with the help of LLMs. A review of recent publications and current trends in the industry—mentioning the overlap of LLMs with new data architecture such as the data lakehouse, the current popularity of semantic ETL, the introduction of the term 'LLMOps', and democratizing access to data using natural language interfaces—is contained in the paper. The generative paradigm will minimize the overhead and the technical complexity to a bare minimum, and it is the paradigm change that could help organizations to come up with smarter data systems that are more dynamic and user-friendly. The review also presupposes the detailed study of the process of transforming the ideals of data engineering at the age of generative AI when the LLMs are utilized.**

***Index Terms*— ETL, ELT, Large Language Models, Generative AI**

1. Introduction

Information engineering has already changed its environment radically in the last few decades and the archaic paradigm of Extract, Transform, Load (ETL) is now substituted with a far more modern and adaptable Extract, Load, Transform (ELT) paradigm. Critical issues that have brought about the immense growth of the quantity of data are the driving force of these changes that necessitate real-time analytics. The new development of data engineering was influenced by the recent addition and integration of Large Language Models (LLMs). The resulting change brings about the new paradigm which is more automation, flexibility, and savvy in data handling. The paper argues that ELT was previously referred to as ETL and bringing the LLMs into the existing data engineering processes, where one is interested in the perception of how the inventions are revolutionizing the field at the dawn of generative AI.

2. Evolution from Traditional ETL to ELT

It has modified the DNP model as the outline of the data engineering activities that are a sequential process with the first stage being the acquisition of data out of the source systems to be transformed to useful formats prior to being introduced into the target storage or databases. It fitted very well in this model when the size of the data was very small and the transformation logic was fixed and rigid. However, the weaknesses of ETL were felt when companies started generating and consuming volumes of transformed and unstructured data. It was particularly in the transformation in the mid-way of the pipeline and the inability to adjust the schema as it was rigidly built.

ELT model uses existing architectures and it is transform and load reversal. ELT will store data, initially load data on scalable storage data lakes or lakehouse, and subsequently optimize the performance with further transformations by use of optimally configured compute engines, which will be performing parallel computing. The inversion enables more scalability to be achieved and raw data can be stored in a format that can be re-processed at some later stage in the future and is much more suited to cloud-native data systems. ELT model can facilitate accelerated cycle development and allow the information researchers and designers to test the logic of transforming cloud-based analytics models in a few days [1].

It is a huge invention in the history of data engineering. However, the shift to ELT was essential to enable the introduction of the incredible stream of human resources in the pipeline construction process, not to mention pipeline maintenance and the correction of all the errors that had been made. The engineers also dedicated their time to doing the corrections, other than attempting to lay the pipelines that would minimize the speed at which the yielding of insights would be made. The ever-increasing range of data inputs and data types has provided the impetus to adopt a new model of automation that has taken advantage of the development of Artificial Intelligence (AI) in the data pipelines in the recent past [1].

3. The Rise of Generative AI in Data Engineering

The adoption of Generative AI (and, in particular, the GPT-based) allows shifting the paradigm of the data engineering processes implementation. The models are capable of absorbing commands in natural languages, and to an even greater extent are capable of generating syntactically valid code, and the models are put in a situation where they can optimize the data workflows. They have been incorporated in the data pipelines and it is an easy way of executing the ETL and ELT processes since it will automatically carry out the schema mapping, data transformation, detecting failures, and enhancing performance.

The ability to make and streamline the pipeline scripts, which is facilitated through plain language, is the most notable input of generative AI into the presented situation, as per the customer's will. The technical skills required to apply the data engineering capabilities are also complicated and removed in order to accomplish the tasks of the LLMs that understand the instructions in natural language and convert them into runnable SQL or Python code. Such operations can be used to democratize access to and analysis of data within an organization to make sure that the business users are the participants in the pipeline programming [2], analysts are the participants, and citizen data scientists are the participants.

The latter is also supported by smart pattern recognition and applied to the code generation by non-automation LLMs. Such models can be trained on big data corpus and can detect abnormality, inefficiency, and discover the opportunities for data workflow optimization. The consequential outcome is a self-optimization of the pipelines and a mammoth task in terms of a cost reduction in the maintenance cost is also met. The intentional application of LLMs is the creation of an additional layer of intelligent data streams capable of adaptively altering its information format and usage disposition even with the factual change in data form and style of utilization without the need for a large portion of human intervention [2].

4. Enhancing Workflow Intelligence and Adaptability

The fact that generative AI is introduced to the process of data processing is linked to the problem of the work that is already in place and the task of the further acquisition of automation, and the reorganization of the working process and its optimization. The data from earlier executions of the pipelines can be used by the LLMs to offer optimized configurations that will reduce latency, throughput delays, and resource consumption. This kind of meta-analysis has potential, which will revolutionize the themes of the data ecosystem, and not the passive agents of the data.

In addition to that, these models allow dynamically updating pipelines logic by either a new information source or tracking schema. These flows will inevitably cause the failure of the classical systems until it is manually fixed with a worse downtime and data quality penalty. Generative AI already has the contextual awareness, which can rearrange transformation logic or refer to a possible breakdown [3].

Otherwise, semantic transformations in generative AI are also present and the conceptualization of the transformations is syntactically and contextually based. One of the secluded examples of tacit knowledge is the international conversion address format where one of the rule-based conventional systems was found not to be productive. What makes the LLM a necessity is the ability to retrieve various types of data that can identify such contextual deviation and, consequently, takes minimal time cleaning and normalizing various types of data [3].

Figure 1 shows the contrast between traditional ETL pipelines and modern LLM-powered workflows that integrate automation and semantic transformation.

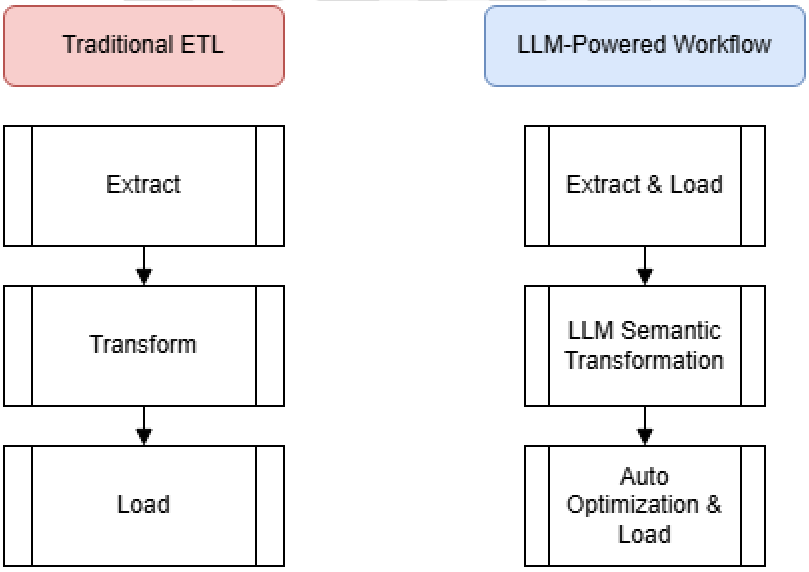

Figure 1: Conceptual Diagram of LLM-Assisted Data Engineering Workflow

(Source: Adapted and synthesized based on [1], [2], [3])

5. Intelligent ETL Automation

Among the theoretical causes, there is also an increasing number of actual jobs of developing AI in ETL. It is being related to business-level systems and they pursue GPT-based systems at their coordination levels. Having the case of AI agents, it is now feasible to suggest the most suitable data retrieving policies, suggest information quality policies, as well as run their pipeline in preliminary to determine the possible errors or bottlenecks. It has also implemented AI modules that have minimized the duration of designing and testing of the pipeline through the application of AI like IBM DataStage [4].

The semantic understanding has been added to the ETL automation. The LLMs are able to analyze documents or metadata and produce transformation scripts as opposed to rule-based transformation scripts in order to be semantically consistent with the organization. This is highly convenient in scenarios where the systems companies have been working with have little or only obsolete documentation. In generative AI, the logic of transformation is reproduced with a high level of accuracy, according to the available sample data, reports, and, therefore, can be applied to the new data platforms, without needing to be transferred, which is the traditional method of doing so [4].

As shown in Table 1, the key differentiators between traditional and LLM-powered ETL systems revolve around automation, intelligence, and adaptability.

Table 1: Comparison Between Traditional ETL and LLM-powered ETL

Feature	**Traditional ETL**	**LLM-Powered ETL**
Pipeline Design	Manual scripting	Natural language to code automation
Schema Handling	Static and rule-based	Adaptive and semantic-aware
Error Detection	Post-failure logs	Predictive anomaly detection
Documentation Requirements	High	Minimal, inferred from metadata
Maintenance Frequency	High manual involvement	Self-optimizing and auto-correcting
User Accessibility	Technical users only	Open to business users and non-technical staff

Source: Compiled from [1], [2], [3], [4], [5]

6. Emergence of Semantic ETL and LLM Synthesized Code

The generative data engineering of AI includes the semantic ETL generation. They include how requests are to be implemented by users, and logic will be produced that is semantically aligned with domain-specific requirements with the aid of the LLMs. A generative AI as a medical organization can help write an ETL script that can be used to proceed with the HIPAA, and a banking business can create a logic which will be able to detect suspicious transactions passively with the help of the AML standards [5].

Semantic ETL engines are based on the interpretative capabilities of the LLM to generate semantically and syntactically meaningful SQL and PL/SQL. These scripts are referred to as business logic constraints and they can be executed in the production environments at much higher, faster rates. This will do away with chances of miscommunication and the looping process will be quicker, and ultimately, the institutions will be streamlined and sensitive to the business demands [5].

7. The Data Lakehouse Architecture and Its Convergence with LLMs

It has come with the effect of the growing big data system, the inadequacy of the former data warehouse system that offers the data warehouse formality in its design with the dynamism and volume of the data lake. Besides possessing the capability to store structured and unstructured information on the same platform, the convergence will enable the execution of workloads of analytical and machine learning on the identical platform.

The discussed case of the data lakehouse can be further developed with the assistance of generative AI that can automatically classify the metadata and make the data more usable and adapt the solution to the search query optimization depending on the purpose of data utilization. The other application of LLMs is to transform unstructured data into semi-structured data in a manner that allows the query and analytics to be applied in real-time. A natural language processing model is employed to retrieve data from tables in PDFs, emails, or free-text fields, which can be queried in the lakehouse system and would otherwise require significant levels of pre-processing [6].

The next benefit which is associated with the usage of LLMs in creating the lakehouse is that it automatically creates the data catalogs. The application of the classical metadata administration tools is based on manual annotations and the interventions of data stewards, which are not likely to be consistent and comprehensive. Instead, field-level and record-level semantics that the LLMs are able to produce by datasets and a more detailed and machine-readable metadata layer can supplement the data governance, data lineage, and privacy enforcers [6].

The systems are utilized in the lakehouse and through this, organizations have the opportunity to use their AI-native data platform. The information can be stored and processed in these platforms and even relay the information to the users using natural language interfaces. The tradition is futuristic, in which the person can input a lot of data and does not have to type one line of SQL, and all that he or she has to do is to state his or her informational preferences in his or her own expression [6].

8. ETL Pipeline Generation through AI Modules

Even the implementation and use of ETL modules into the organization is now being considered as a major factor in the development of the LLM. Here, a collection of schemas or descriptions of a business process as it should be mounted on a computer in the form of modular ETL code was demonstrated in doctoral research in the field, when tools of generators were actually run. It is high productivity, meaning, in which parallel logic of pipeline is used in other departments or areas of application.

The advantages of the code-generating process include the reinforcement learning and the instant tunings which make sure the products generated are grammatically correct, in addition to ensuring that the products generated are contextually correct to the data structure of an organization. For example, the trained model on an organization's data warehouse could learn the naming conventions used, index strategies employed, and data constraints, to ensure that the generated code aligns with best practices and also reduces query costs [7].

Moreover, the AI-driven systems will be capable of recognizing the change and reconstructing the parts of ETL that could have been modified in case we know that we are conducting a schema change, i.e., introduction of a new attribute or ensuring dataset consolidation etc. This will result in giant time savings on data infrastructure procurement. Rather than altering the change request cycles and the manual development process, AI modules are capable of reorganizing the processes with the minimum number of human interferences and ensuring the quality of data, homogeneity of enterprise logic [7].

The multi-cloud/hybrid systems are also adding to the integration issues. The architecture and language disparity across the platforms would also be solved with the assistance of the LLMs. Data engineering can be made easier with an AI-assisted ecosystem of scale, whereby it can be created to run on a new cloud engine, e.g., BigQuery, Snowflake, or Azure Synapse. It is not binding one that reduces the vendor lock-in in addition to being a more responsive and adaptive data architecture strategy [7].

9. Personalized Instruction and Automation in Data Engineering Education

The transformational opportunities of the education and training field of the LLMs can be found in other technical aspects of the data engineering world. The individualization of the learning experiences is related to the fact that they may be used as personal tutors or teaching assistants. The availability of the course contents, the sample of the coding, feedback on the actions, and modeling the situations with the aid of the LLMs assist the larger impetus of the data engineers to survive in the future.

As a result of the mechanisms that help to balance the work of the LLM and the learning outcomes, these days, the concept of AI teaching assistants in the learning programs can be applied to the learning programs. These support staff will be working close to students and areas of weakness can be determined and incremental assistance given depending on the speed of the learner and comprehension. Not only that this type of automation could be used to facilitate the interaction between the students, but also guarantee the improved perception of this complex of issues such as the coordination of pipes, languages of data transformation, and distributed computing networks [8].

They can also be extended to additional use cases to address the gap between industry applications and what is taught in academic settings. The students will not have to rely only on the theoretical teaching under consideration, but rather they will be in a position to work with the AI systems, which are designed considering the actual problems in the life of the pipeline failures, the quality of information, and the problem of performance optimization. This introduces the added training and equips them to fit in the employment market in the dynamic information engineering [8].

10. Democratizing Data Access with Generative AI

This has been manifested as one of the largest limitations of the conservative data ecosystems since the data tools and platforms are limited to the technical stakeholders. It can act as a bottleneck and slow down the decision-making process since business users usually need the help of data engineers or data analysts to gain a specific insight or produce dashboards. This can change with the advent of natural language interfaces of data communication applying generative AI.

The integration of the LLMs within the business intelligence industry will enable the data platforms to deliver self-service data experience within the businesses. The model may be in a position to decode natural language queries to SQL queries or API calls to answer user queries. Not only does it augment productivity, but it also allows the use of data across the entire process of decision-making to the entire organizational levels. Moreover, AI agents are also able to offer the rest of the relevant datasets, concentrate on the problem of the quality of the data, and even offer a visualization based on the data presented by a user [9].

This is not limited to the query generation only. Even the generative AI will arrive at the point of assisting in the creation of the KPIs and correlating the information with the benchmarks and predicting the tendencies on the basis of the created statistical models. The additions make the data platforms quite useful and enable them to be more amenable to the changing demands of business. To be more specific, the democratization of data with the use of the LLM will result in the democratic organizational culture, where the data it contains will be a shared resource and not its own property [9].

11. Managing LLMs in Production: The Rise of LLMOps

Their dominance on the models of the production systems is therefore a tough subject to discuss since the LLMs will be better suited to be incorporated into the data engineering systems. The management of big generative models through execution, monitoring, and execution is estimated to be done by the new architecture of LLMOps (Large Language Model Operations). The LLMs do have some infrastructure requirements relative to the traditional ones, which are token requirements, trigger engineering plans, latency improvements, and memory utilization.

The practices in the data engineering of LLMOps in versioning of models, high-quality prompt curating, and model drift or hallucination checking are the practices. Such operations are necessitated in the context of ensuring that the products that are produced by AI are right, especially when the product is executable code or code of transformation which could affect the pipelines of data.

This is the other issue that LLM needs to address: security and compliance. The companies will make sure that the LLMs will not spill secret information, open loopholes, or have biased output. Among the risks that can be reduced in timely sanitization and numerous others, there are access control and differential privacy. The user comments and the performance measures are not left out either to optimize the models in production by fine-tuning the models and therefore retain the same model intact with the organizational goals [10].

Another move toward the standardization of the LLMOps will be observed to be introduced into the data platforms in the enterprise so that its use and administration can be predicted for the effective execution of the generative AI into data engineering. The standardized APIs, model registration, observability dashboards, and the generative workload support tools would be some of the long-term impacts of the LLMOps [10].

12. Conclusion

The redesign of the ETL to ELT and the subsequent intelligent data engineering, which is built on the basis of the LLM, is the reconsideration of the data processing within the enterprise. Contextual ability and semantic intelligence is very extreme in generative AI compared to rule-based systems. With the introduction of LLMs in data engineering, the level of automation, data efficiency, and operational costs in businesses are not comparable. Besides, the implementation of some disciplines like the LLMOps will be likely to contribute to the delivery of responsible and scalable implementation of such models.

Since the information will be the blood of the digital transformation, not only will it be possible to streamline the already existing pipelines to which the generative AI will be introduced, but also to think over the new imaginings of the business intelligence, the education, and design of the platforms. The data engineers of the future will not have the appearance of coders, as they will have the appearance of the creators of the AI-powered systems because they will be in a constant learning and development.

References

[1] Oluwaferanmi, J. K. A. (2025). Automating ETL Pipelines Using Artificial Intelligence: Transforming Legacy Data Integration Systems into Intelligent Data Workflows.
[2] Kanagarla, K. (2025). Data Engineering with Generative AI Automating Pipelines and Transformations Using LLMS like GPT. Available at SSRN 5107348.
[3] Khuat, Q. H. (2025). Leveraging Generative AI for Data Engineering Workflows. Journal of Computer Science and Technology Studies, 7(3), 120-140.
[4] Benedetti, A. U. (2025). Automation of ETL Pipelines in DataStage (Doctoral dissertation, Politecnico di Torino).
[5] Ananthakrishnan, V., Kondaveeti, D., & Mohammed, A. S. (2025). GenAI-Driven Semantic ETL:: Synthesizing Self-Optimizing SQL & PL/SQL. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 4(2), 29-43.
[6] Generative, A. I., & Kaushikk, R. The Data Lakehouse Revolution.
[7] Michelotti, I. (2025). Progettazione e realizzazione di moduli per la generazione di pipeline ETL= Design and Implementation of Modules for ETL Pipeline Generation (Doctoral dissertation, Politecnico di Torino).
[8] Cao, D. (2025). Instructional Alignment of Large Language Models: A Framework for Creating Personalized AI Teaching Assistants in Engineering Education (Master's thesis, University of Southern California).
[9] Laakkonen, S. (2025). Enabling self-service business intelligence in an organization with generative AI.
[10] Aryan, A. (2025). LLMOps: Managing Large Language Models in Production. " O'Reilly Media, Inc.".