LEVERAGING GENAI TECHNIQUES FOR OPTIMISATION IN AI-DRIVEN ETL PIPELINE DEVELOPMENT

Name: Mukund Kulkarni
Designation: Senior Engineer.
Affiliation- Ernst & Young US.
Location- Dallas, Texas, 75068, USA.
Email- mukundkut@gmail.com

Abstract **-** This paper focuses on adopting GenAI Techniques to optimise AI-driven ETL Pipeline Development. Generative AI, adds flexibility and intelligence to the process of ETL. By linking manual data pipelines with the ability of AI, companies unlock new paths to optimise and automate how data flows. The ETL method extracts data from multiple sources, transforms it into the correct format, and loads it into a database or the data warehouse. This method enables companies to arrange their data for reporting, analysing, and decision-making. GenAI-driven pipelines refer to core issues in data management. Facilities that empower companies with faster insights, better decision-making, and interventions for data-intensive and complex circumstances.

Keywords: Artificial Intelligence, generative AI, ETL pipeline, Flexibility, Adaptability.

I. INTRODUCTION

A. Background to the Study

GenAI, or generative AI, is a type of artificial intelligence that develops content including music, images, text, audio, and even video content. GenAI can modify the way of traditional data management [1]. This applies high-level AI frameworks such as foundation models to learn structures, patterns, and functions from input data. An ETL pipeline is a method of developing a set of phases to extract, transform, and load data from multiple sources into one storage system.

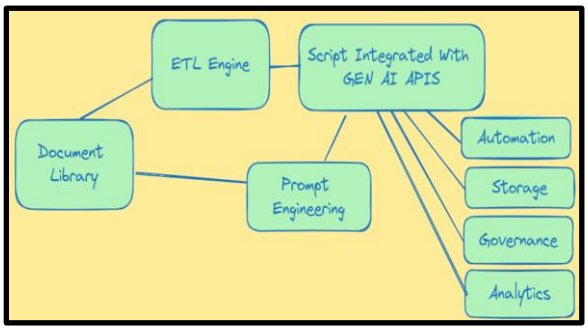

Figure 1: Transforming ETL with GenAI [2]

Figure 1 shows the transformation of ETL with GenAI integrated with automation, storage, governance, and analytics. However, ETL methods observe disruptions because of data errors, schema changes, and system limitations. In this context, GenAI comes to adding flexibility and intelligence to the ETL method. "ETL data pipeline" is working and running on Apache Spark [3]. By linking traditional data pipelines with the potentialities of AI, companies discover new paths for optimising and automating data flows.

B. Overview

GenAI has been approached as a developmental instrument to optimise AI-based "Extract, Transform, Load" pipelines. ETL pipelines are created to perform an aspect of data preferred to ELT models [4]. Models of GenAI enable precision and automation in data generation, transformation, and the application process of the data, effectively decreasing manual mediation. For example, BloombergGPT was created by Bloomberg, which is a wide language model created to process financial information, allowing valuations that

doi: 10.18178/ijiee.2024.14.4.5

increase decision-making [5]. Additionally, Databricks integrates GenAI to improve the application of data strategy enabling companies such as Shell to apple large set datasets while decreasing flaws and developing performance levels. Through "Self-Updating Data Pipelines," "Self-Healing Capabilities," "Handling Logical Changes in Data," and "Automating Data Cleanup and Harmonisation" GenAI improves the process of ETL.

C. Problem Statement

The ETL is core to the pipeline of data, specifying data is accurately generated, transformed, and loaded for its further use. Traditional ETL methods are commonly error-prone, time-consuming, and lack scalability, reducing their performance in managing wide data streams. ETL pipelines have been identified as an abstract demonstration of end-to-end data [6]. The application of Generative AI creates emerging interventions; however, many companies observe issues with its application because of the threats to resource allocation, personalisation, and maintaining the integrity of data in the model. As an outcome, in the recent ETL modelling, a lack of decision-making and real-time automation abilities hampers data-driven innovation. Referring to these issues is major to improving the efficiency of data and authorising scalable interventions.

D. Objectives

The primary objectives of this research are 1. To create techniques through GenAI to automate ETL methods, enhancing efficiency and decreasing manual intervention. 2. To apply GenAI-based interventions to enable real-time ETL flows, enabling companies to analyse and process data streams effectively. 3. To highlight challenges companies, face while integrating GenAI to technique optimise AI-Driven ETL Pipeline Development. 4. To propose scalable ETL methods through GenAI to optimise human resources and computational while maintaining performance in high-demand instances. Therefore, this objective aims to discover practical implications and potential facilities for applying GenAI into ETL Pipeline Development, ultimately upgrading its reliability and efficiency.

E. Scope and Significance

This research discovers how GenAI transforms ETL pipeline development, error reduction, upgrading automation, and initiates scalability in data processing. This research shows the application of AI to refer to incompetence's in manual ETL methods and their contribution to initiating real-time decision-making. Additionally, practically, companies such as Shell, Databricks, and others demonstrate the ability of GenAI to develop governance of data and manage wide databases. These integrations are effective around multiple sectors such as healthcare, finance, energy, and others, where data-driven strategies are major. Therefore, the particularity of this research lies in improving operational performance, driving innovation, and decreasing costs leading to trillions in economic value internationally.

II. LITERATURE REVIEW

A. GenAI Techniques to optimise AI-Driven ETL Pipeline

The ETL process takes out data from multiple sources, transforms it into correct outlooks, and loads it into a data warehouse or database. ETL process based on GenAI is created with the help of "data engineering strategy" [7]. This method enables companies to arrange their data to be analysis-ready, reporting, and decision-making. However, apart from its major role, manual ETL faces multiple issues, including, data quality issues, schema changes, error handling, and scalability concerns. With these difficulties and the volume of data, companies require more contemporary and flexible methods, and here GenAI comes in

doi: 10.18178/ijiee.2024.14.4.5

with interventions to go beyond manual ETL developments.

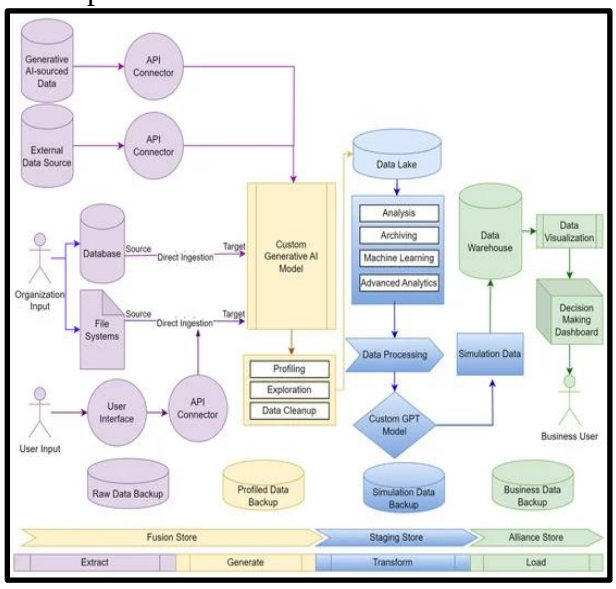

Figure 2: EGTL model high-level solution architecture

[7]

Figure 2 shows the levels of the EGTL model by data flow in the EGTL process. In this regard, real-time processing improves efficiency, anomaly detection, adaptability, schema matching, and scalability are included as techniques. These processes increase speed, decrease charges, and enable actionable valuations from emerging data streams. For example, companies like Shell imposed GenAI through Databricks to aerodynamic data governance and its application, whereas BloombergGPT integrates GenAI to interpret economic datasets of their company. Furthermore, "Self-Updating Data Pipelines" as an AI-enabled method highlights modifications in schemas and data outlines and immediately adjusts code in the pipeline to lodge them [8]. This removes the requirement for updates and manual coding methods whenever new data files are launched or formats are modified.

B. Real-Time Data Processing

By specifying scalable and accurate processing of live data, companies create a competitive edge for themselves, decrease operational challenges, and improve customer satisfaction rates. With artificial intelligence anticipated to develop critical financial value, real-time GenAI-driven ETL workflows are essential for efficiency and innovation. Thus, AI in the data pipeline shows a major potential to account for data preparation and create real-time insights [9]. Manual ETL processes commonly fail to manage data demands in real-time, creating delays in decision-making and insights creation. GenAI refers to this with the help of cleansing, extraction, and data transformation, through ML frameworks able to approach and anticipate emerging patterns of data.

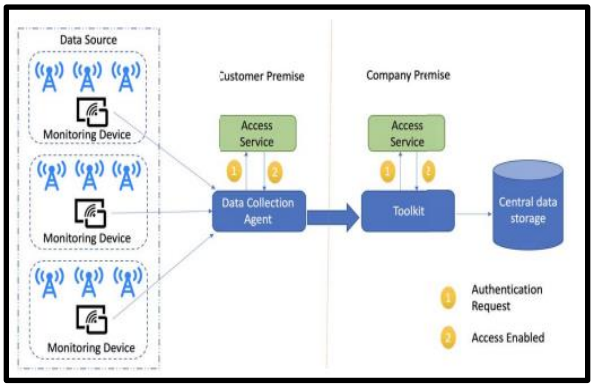

Figure 3: Data collection process

[6]

Figure 3 shows the data collection process by monitoring the data source and device. GenAI effectively identifies and makes corrections to inconsistencies in data without human interactions. Thus, the significance of this research lies in decreasing inactivity, enabling companies to react to changes in the market instantly.

C. GenAI Challenges

Applying GenAI techniques to optimise AI-driven ETL pipeline development shows multiple disruptions for business. ***The intensity of resources and costs*** refers to interventions through GenAI that need major funding in infrastructure, including GPU and could computing resources, SMEs find it problematic to assign these resources. For instance, Snowflake faced problems with

doi: 10.18178/ijiee.2024.14.4.5

extreme cloud expenses when scaling their ETL pipelines regarding real-time analytics. ***Compliance*** refers to the management of sensitive data, specifically in banking, which creates obedience issues. For example, the approach of HSBS related to AI to detect fraud needed strict conformance to the norms of data protection including, GDPR, which decreased application because of the concerns related to delicate customers' information. However, GenAI-based companies automate complex data analysis to anticipate their market trend [10]. ***Application with legacy systems*** refers that, most of the companies depend on manual systems that are not capable of AI-based interventions. For example, manufacturing companies faced problems in applying GenAI-based ETL with legacy ERP methods, needing enlarging overhauls of the system.

Workforce Readiness and skill gap address that organisations often lack skilled employees in AI and ETL processes. Companies applying Databricks highlight disruptions due to their data engineers' required training on GenAI models. Furthermore, ***Bias*** in generative AI frameworks leads to inaccurate data insights and transformations. AI-based ETL processes loudening differences in regional sales because of biased training information. Referring to these issues need skilled personnel, strategic investments, and strong frameworks for compliance and data governance.

D. Scalable ETL frameworks

Proposing scalable ETL models or frameworks with GenAI enables companies to manage high-demand and complex environments in data effectively while optimising both human and computational resources. Scalable ETL pipelines allow companies to impose ML frameworks at their scale [10]. ETL frameworks have been leveraging the abilities of GenAI to anticipate uncertainties of workflow, automate repetitive activities, and effectively allocate resources, specifying continually in performance also at the time of data surges. GenAI improves the allocation of computational resources with the help of "predictive analysis." This enables systems to proactively "on-premise infrastructure" or "scale cloud." For example, Michelangelo's domain from Uber, integrates artificial intelligence to generate extreme ride data, maintaining low discontinuation at the time of rush hours.

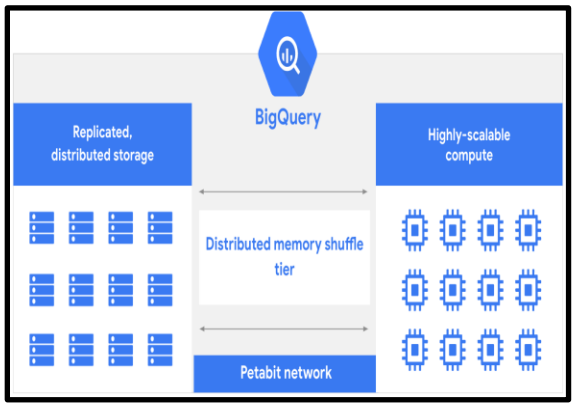

Figure 4: Outline of BigQuery [12]

In this regard, "***Google BigQuery***" applies GenAI frameworks to automate data transformations and also optimising query implementation plans. It is created for wide data analytics with applied ML-APIs. "Google BigQuery" produces a server-free design a real-time stream process of data, decreasing inactivity and statistical overhead. For example, Spotify has been using BigQuery to manage its user activities, specifying real-time evaluations while scaling for a wide base of customers. After that, ***Fivetran*** applies to integrate automation in the ETL methods with the help of GenAI for error correction, schema generation, and replication of data around several domains [13]. Fivetran fastens the setup process of ETL with less configuration and its pre-developed connectors, enabling companies to scale it effectively. For instance, Canva has been using Fivetran to modify its marketing

doi: 10.18178/ijiee.2024.14.4.5

information, shortening data reporting and analytics.

Furthermore, applying Kubernetes as containerisation tools with generative AI enables AI-based ETL pipelines towards effective resource allocation depending on the demand of the workplace, further developing scalability.

III. METHODOLOGY

A. Research Design

In order to identify the role of "GenAI Techniques for Optimisation in AI-Driven ETL Pipeline Development" applied both quantitative and qualitative research methods. The research onion is a model that guides researchers create research methodology and research design. It is an instrument that guides the researcher through the decisions that are required to create when working on research methodology for research purpose.

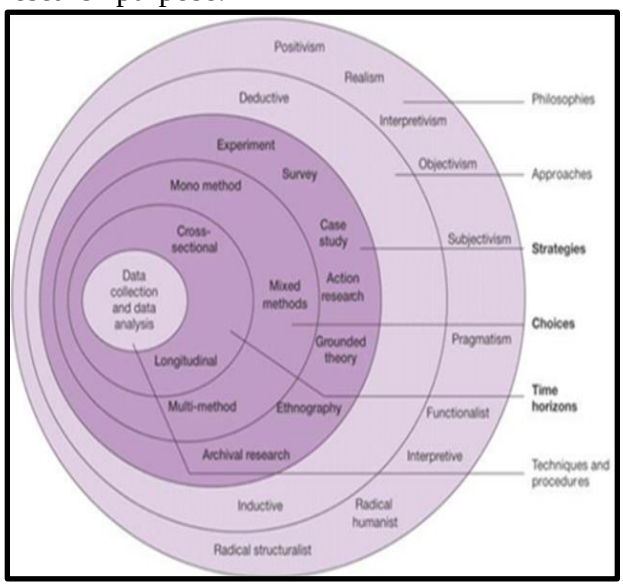

Figure 5: Research Onion Model

[14]

This figure has shown an outline of the research onion with data collection and analysis techniques. Hence, to analyse "Optimisation in AI-Driven ETL Pipeline Development" the researcher has utilised a ***multi-method*** including secondary qualitative and quantitative research methods. Additionally, in this study "interpretivism philosophy" has been applied to investigate this topic. This has been chosen to refer to major research designs that include multiple operational decisions depending on identifying suitable steps to achieve research objectives. After that, "an inductive approach" has been selected here as it enables this study to do the investigation with creativity and intuition to interpret theoretical descriptive evaluation depending on observed real-life scenarios. Moreover, "an explanatory research design" has also been chosen here, as this study approached both qualitative and quantitative information from real-life insights depending on the given context.

B. Data Collection

Data sources were quantitative and qualitative, and the investigation was conducted by collecting journals and articles from multiple authors and information taken from trustworthy websites to highlight how GenAI Techniques optimised AI-driven ETL Pipeline Development. Journals and articles are used particularly for qualitative methods and graphs, charts, and statistical information to do quantitative analysis. Moreover, this research depends on case study examples, documented reports, measures of organisational performance, and initiatives to compare and contrast the effectiveness of GenAI on AI-Driven ETL Pipeline Development.

C. Case Studies Examples

Case Study 1: Amazon Web Services

Amazon Web Services has created an AI-enabled ETL domain referred to as, "AWS Glue." This initiates automation in its ETL process. By using techniques of GenAI, including ML algorithms, and "natural language processing,", AWS Glue interpreted unorganised information, and transformed and cleaned the data [4]. This decreases the requirement for manual generation of data to identify insights of the

doi: 10.18178/ijiee.2024.14.4.5

customers in sectors such as finance, healthcare, manufacturing, retail, and others.

Case Study 2: Netflix

Netflix applies AI-based ETL pipelines to operate its wide range of data generated by its customer base. With the help of frameworks in ML, Netflix has been optimising methods in ETL specifying infestation, process, and loading of data, decreasing its traditional intervention and developing the performance of the pipeline [14]. This process improves the real-time suggestion system, generating customers with customised content seamlessly.

Case Study 3: Spotify

Spotify applies an artificial intelligence-based system as a part of its ETL pipeline to track and manage its user activities. With the help of GenAI, this company automates the transformation and observation of multiple data types, including, streaming patterns [15]. This revolution specifies the accuracy of data, decreasing flaws in its "recommendation engine." This upgrade improves the performance of data processing, assisting Spotify in creating customised suggestions seamlessly for its customers.

IV. RESULTS

A. Data presentation

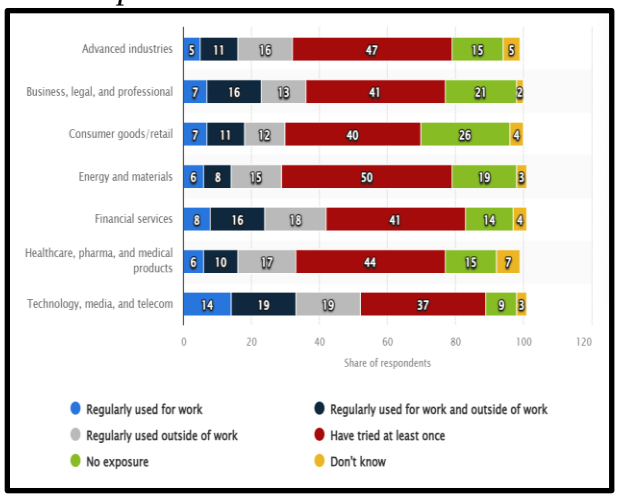

Figure 6: Generative AI tool in industries [16]

Figure 6 has highlighted that most of the people in various sectors have tried to impose generative AI. Customers and the retail group were the largest sectors that had no such exposure to GenAI instruments. Around 45% of companies have observed twice the employee performance rate because of the GAI, with 59% citing the accuracy of the surge and 54% observing faster time to the market. 51% over, 1,000 marketers Salesforce are already using or experimenting with GenAI in their workplace [16]. 85% of companies refer to the raised engagement of their users and 80% increased customer satisfaction rate because of generative AI. In this regard, around 75% of generative AI users search to automate activity to work and apply GenAI to communicate at work. However, 73% think that generative AI launches new security threats [16].

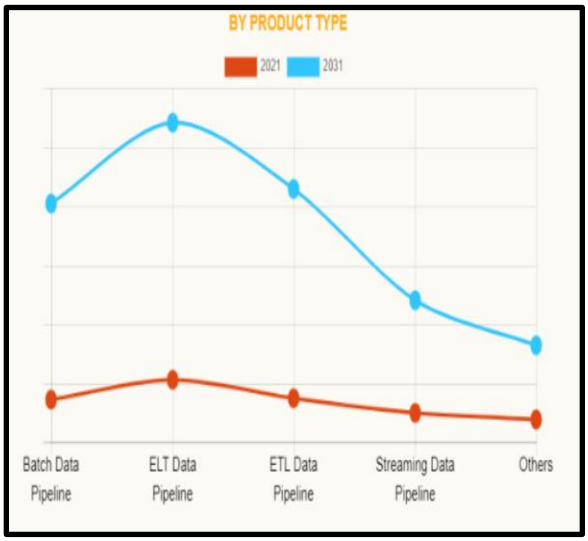

Figure 7: Data Pipeline Market [18]

Data quality pipelines create potentialities such as regular standardisation of new customer names. Real-time client refers verification would be observed as an element of the data pipeline at the time of a credit application. GenAI-based ETL pipeline initiates improvements ranging between 30–50%, in the operations of data handling [18]. Automation with the help of GenAI decreases flaws by 60%, specifying quality outcomes and improving decision-making.

doi: 10.18178/ijiee.2024.14.4.5

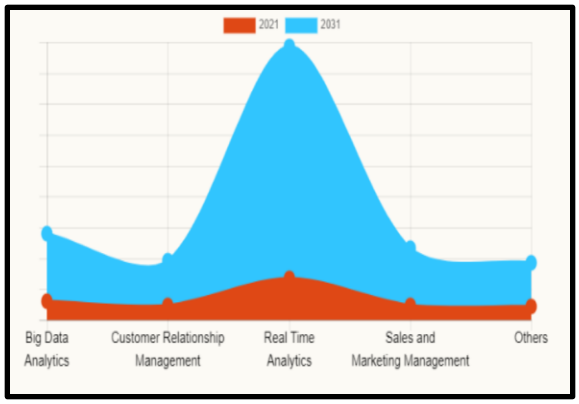

Figure 9: Data Pipeline Market by Application Area

[18]

Figure 9 has highlighted, the market of data pipeline by its application area and real-time analytics has most of the market share, at the time of a forecast duration. GenAI such as LLMs or "large language models" outlines structures from unorganised data particularly. For instance, applying LLMs such as, "Databricks Meta-Llama" enables automatic elicitation, division, and sentiment analysis, decreasing work pressure, conducting analysis of "free-from data," and processing time effectively [18]. These processes have been major in areas such as the process of insurance claims when unorganised charges claim notes are altered into actionable evaluations.

B. Findings

Advancements show the developmental effect of generative AI in offering scalability, engineering, and raised accuracy in the ETL methods [16]. From automating the cleanup of data to performing analyst-level activities, GenAI supports the creation process of data effectively and seamlessly [18]. While there are still issues, around security and privacy concerns, facilities in secure AI deployment are making it possible to harness the ability of AI without disrupting the integrity of data.

C. Case study outcomes

Company	ETL Optimisation with GenAI	Analysis
AWS	Division of data based on NLP and data cleaning via AWS Glue [4].	Better customer insights, seamless preparation of data, and decreased ETL charges.
Netflix	Observation of automated oddity while transforming the data [14].	Better accuracy in content suggestion and decreased downtime in the pipeline.
Spotify	GenAI-enabled transformation of data regarding user activity information [15].	Improved recommendation engine and generation of customised user experiences.

Table 1: Data Presentation of findings from the case study

D. Comparative analysis

Attributes	Traditional ETL	EGTL	Development rate
Accuracy in data quality	80%	93%	+16.25%
Processing Speed	100	100 to 130	+30%
Throughput	2000	2000 to 2600	+30%
Automation	Around 60	100% automat	+40%

doi: 10.18178/ijiee.2024.14.4.5

	automations	ion [6]	
Latency	15	9	(40%)
Application with Data Mesh	50%	90% to 95%	More than 90%

Table 2: GenAI in ETL Optimisation

Table 2 has highlighted major data points deduced from both manual ETL development and EGTL frameworks with the help of GenAI. For example, generative AI develops data quality with the help of advanced data cleaning, data profiling, and augmentation in steps such as, "Generate" decreasing limitations and anomalies in the data extraction process. Additionally, in terms of scalability, GenAI-based pipelines apply with contemporary designs including, scaling and data mesh around confederate processes.

V. DISCUSSION

A. Interpretation of results

Real-time ETL workflow through GenAI supports anomaly detection, ongoing data monitoring, and production of rapid insights of data which are major for companies working with time-sensitive information such as logistics and healthcare. As an outcome, GenAI has been optimising AI-driven ETL pipelines by automating its transformation, and data preparation, decreasing its traditional effort, and application processes, and developing data accuracy [14]. Further, ETL pipelines are used to improve data accuracy, decrease manual intervention, and accelerate the workflow of data integration. Generative AI highlights errors in data pipelines by developing "self-healing pipelines" that treat, detect, and resolve errors without the recruitment of manual efforts. The findings also refer to the faster extraction and analysis of data that decreases latency, initiating real-time data insights which are major for decision-making processes.

B. Practical Implications

ETL pipelines through GenAI improve data quality, automate data profiling, and encourage data processing, allowing more credible and efficient data management. By applying anomaly detection, real-time validation, and anomaly observation, GenAI decreases inactivity and fallacy rates [20]. Adaptability of this validates contemporary architectures including data mesh, anticipating scalable, and segregated processes. These facilities authorised companies with on-time insights, streamlined operations, and fostered innovation in data-intense ecosystems.

C. Challenges and Limitations

Despite these present facilities, this work has its limitations. A quantitative study shows a few company examples, however, lacks statistical analysis through survey or interview data [19]. This certain limitation decreases the generalisation of the obtained research outcomes.

D. Recommendations

Applying primary quantitative and secondary quantitative methods. Furthermore, improving data governance, scalability and obedience of ethical AI in segregated outlines stays critical for upgrading AI-based pipelines [20].

VI. CONCLUSION AND FUTURE WORK

In conclusion, generative AI is forming how data pipelines and ETL processes operate. Artificial intelligence is launching adaptability and particularity in data management by making data pipelines self-healing, self-updating, data matching, and data aggregation. By balancing resource efficiency and automation technology, scalable GenAI-based ETL models authorised companies to process wide data volumes seamlessly, improving their

doi: 10.18178/ijiee.2024.14.4.5

operational performance in high-demand contexts.

Hence, future work concentrates on empirically testing GenAI-based ETL frameworks around sectors to support achievements from the performance. Creating standardised measures to cultivate GenAI and discovering its application in technologies such as edge AI and quantum computing are encouraging domains.

VII. REFERENCES

[1] Aydın, Ö. and Karaarslan, E., 2023. Is ChatGPT leading generative AI? What is beyond expectations?. *Academic Platform Journal of Engineering and Smart Systems*, *11*(3), pp.118-134.

[2] Linkedin.com, (2024). *Transforming ETL Processes with Generative AI: A Revolution in Data Management*, Available at: https://www.linkedin.com/pulse/transforming-etl-processes-generative-ai-revolution-data-anjani-kumar-wpndc (Accessed on: 16 December 2024).

[3] Batmaci, G., 2022. Etl Data Pipelines Configurations in Spark.

[4] Mbata, A., Sripada, Y. and Zhong, M., 2024. A Survey of Pipeline Tools for Data Engineering. *arXiv preprint arXiv:2406.08335*.

[5] Bloomberg.com, (2023). *Introducing BloombergGPT, Bloomberg's 50-billion parameter large language model, purpose-built from scratch for finance*, Available at: https://www.bloomberg.com/company/press/bloomberggpt-50-billion-parameter-llm-tuned-finance/ (Accessed on: 16 December 2024).

[6] Raj, A., Bosch, J., Olsson, H.H. and Wang, T.J., 2020, August. Modelling data pipelines. In *2020 46th Euromicro conference on software engineering and advanced applications (SEAA)* (pp. 13-20). IEEE.

[7] Vesjolijs, A., 2024. The E (G) TL Model: A Novel Approach for Efficient Data Handling and Extraction in Multivariate Systems. *Applied System Innovation*, *7*(5), p.92.

[8] Linkedin.com, (2024). *How Generative AI is Revolutionizing Data Pipelines and ETL*, Available at: https://www.linkedin.com/pulse/how-generative-ai-revolutionizing-data-pipelines-etl-094nc#:~:text=This%20is%20where%20generative%20AI,and%20optimizing%20how%20data%20flows. (Accessed on: 16 December 2024).

[9] Pattyam, S.P., 2021. Data Engineering for Business Intelligence: Techniques for ETL, Data Integration, and Real-Time Reporting. *Hong Kong Journal of AI and Medicine*, *1*(2), pp.1-54.

[10] Banu, A., 2024. Harnessing GenAI for Advanced Data Analytics in Real-Time Streaming with Google Cloud.

[11] Yadav, H., 2024. Scalable ETL pipelines for aggregating and manipulating IoT data for customer analytics and machine learning. *International Journal of Creative Research In Computer Technology and Design*, *6*(6), pp.1-30.

[12] Cloud.google.com, (2024). *BigQuery overview*, Available at: https://cloud.google.com/bigquery/docs/introduction (Accessed on: 16 December 2024).

[13] Go.fivetran.com, (2024). *Hundreds of data sources at your fingertips*, Available at: https://go.fivetran.com/demo?_bt=696470892629&_bk=fivetran&_bm=e&_bn=g&_bg=161419549856&campaignid=21181153400&device=c&utm_term=&gad_source=1&gclid=Cj0KCQiAvP-6BhDyARIsAJ3uv7bGjScqQkg-lqovhpPE5oiZVGdhiNFXy4TC40HG2VVVaDqTc_mFWDMaAopxEALw_wcB (Accessed on: 16 December 2024).

doi: 10.18178/ijiee.2024.14.4.5

[14] Saunders, M., Lewis, P. and Thornhill, A., (2009). Research methods for business student. 5th ed.Harlow: Pearson

[14] Netflixtechblog.com, (2022). *Ready-to-go sample data pipelines with Dataflow*, Available at: https://netflixtechblog.com/ready-to-go-sample-data-pipelines-with-dataflow-17440a9e141d (Accessed on: 16 December 2024).

[15] Medium.com, (2023). *Simple ETL Pipeline with Python: Spotify*, Available at: https://medium.com/@mervegunak/simple-etl-pipeline-with-python-spotify-b0bc4bbf8890 (Accessed on: 16 December 2024).

[17] Salesforce.com, (2024). *Top Generative AI Statistics for 2024*, Available at: https://www.salesforce.com/news/stories/generative-ai-statistics/#:~:text=Marketers%20believe%20generative%20AI%20will,experimenting%20with%20it%20at%20work. (Accessed on: 16 December 2024).

[18] Cloudaeon.co.uk, (2024). *Optimising Your ETL Pipelines*, Available at: https://cloudaeon.co.uk/innovation-lab/etl-optimization-1114/ (Accessed on: 16 December 2024).

[19] Chintale, P.: DevOps Design Pattern: Implementing DevOps Best Practices forSecure and Reliable CI/CD Pipeline (English Edition). BPB Publications, 2023.

[20] Chintale, P.: DevOps Design Pattern: Implementing DevOps Best Practices forSecure and Reliable CI/CD Pipeline (English Edition). BPB Publications, 2023.

[21] Nookala, G., Gade, K.R., Dulam, N. and Thumburu, S.K.R., 2020. Automating ETL Processes in Modern Cloud Data Warehouses Using AI. *MZ Computing Journal*, *1*(2).

[22] P. Chintale, R. K. Malviya, N. B. Merla, P. P. G. Chinna, G. Desaboyina and T. A. R. Sure, "Levy Flight Osprey Optimization Algorithm for Task Scheduling in Cloud Computing," 2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS), Hassan, India, 2024, pp. 1-5, doi: 10.1109/IACIS61494.2024.10721633.

[23] Awiti, J., Vaisman, A.A. and Zimányi, E., 2020. Design and implementation of ETL processes using BPMN and relational algebra. *Data & Knowledge Engineering*, *129*, p.101837.

doi: 10.18178/ijiee.2024.14.4.5