Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Consumer Insights at Scale: Integrating ML Pipelines, Scalable Data Engineering, and Distributed Analytics in Multi-Cloud Ecosystems

Neha Boloor[1], Rohit Jacob[2], Varun Kumar Reddy Gajjala Gajjala[3]

1 Machine Learning Research Engineer

2 Data Scientist, Foundational Models & Generative AI

3 Production Engineering Manager

ARTICLE INFO

Received: 02 Jun 2025

Revised: 12 Jul 2025

Accepted: 24 Jul 2025

ABSTRACT

In the era of data-driven decision-making, extracting actionable consumer insights at scale is critical for enterprise competitiveness. This study presents an integrated framework that combines machine learning (ML) pipelines, scalable data engineering, and distributed analytics within multi-cloud ecosystems to generate real-time, predictive consumer intelligence. Leveraging federated learning, the system enables decentralized model training across AWS, Azure, and GCP while preserving data privacy and regional compliance. The data engineering backbone, built on technologies such as Apache Kafka, Spark, and Airflow, ensures high-throughput, fault-tolerant data processing. Experimental results reveal that Gradient Boosting Machines achieved the highest AUC scores (up to 0.95), with significant regional variations in model performance validated through ANOVA and post-hoc testing. Time-series forecasting using Prophet outperformed ARIMA across all metrics, while throughput scalability tests demonstrated linear performance gains with increased compute clusters, reaching up to 2 million events per second. The proposed architecture not only enhances the granularity and speed of consumer insight generation but also ensures operational resilience and flexibility across decentralized cloud infrastructures. This research contributes a practical, modular solution for organizations seeking to unify and scale their analytics efforts in dynamic, multi-cloud environments.

Keyword: unify, analytics, organizations, environments

INTRODUCTION

Understanding the need for scalable consumer insights

In today's hyper-competitive digital landscape, consumer behavior has become more dynamic and unpredictable than ever before (George, 2022). The exponential rise in data generated from mobile applications, e-commerce platforms, social media, and IoT devices presents both an opportunity and a challenge for businesses. The opportunity lies in converting this data into actionable insights that drive strategic decisions, product innovation, and personalized marketing. However, achieving this at scale especially in environments that demand real-time responses requires robust systems that can manage, process, and analyze massive datasets distributed across multiple cloud platforms (Demirbaga et al., 2024). The need for an integrated approach to consumer insights has thus never been more pressing.

Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The role of ML pipelines in consumer intelligence

Machine Learning (ML) pipelines have emerged as a key enabler in automating and accelerating the generation of insights from complex datasets (Butt et al., 2022). These pipelines support a sequence of operations data ingestion, preprocessing, feature engineering, model training, validation, and deployment that are essential for predictive analytics and classification tasks (Alzoubi et al., 2024). By leveraging ML, organizations can move beyond static, descriptive analytics toward predictive and prescriptive insights, empowering them to forecast trends, detect anomalies, and understand consumer intent with unprecedented accuracy. However, deploying such pipelines in real-world business ecosystems calls for seamless integration across diverse infrastructures (Chelliah et al., 2025).

Data engineering as the backbone of insight generation

At the heart of scalable consumer analytics lies data engineering the process of building and maintaining reliable data architectures that ensure high throughput, low latency, and robust data quality (Desai & Patil, 2023). Scalable data engineering encompasses the design of distributed data lakes, transformation layers, and real-time data pipelines that support continuous data flows. With the increasing adoption of streaming data technologies like Apache Kafka, Flink, and Spark, businesses can now analyze consumer interactions in milliseconds, thereby facilitating instantaneous decision-making (Duan et al., 2024). These engineering capabilities form the foundation upon which intelligent systems are built and continuously refined.

Distributed analytics across multi-cloud environments

Modern enterprises often operate in hybrid and multi-cloud environments, using services from various providers such as AWS, Azure, and Google Cloud. Distributed analytics enables organizations to process and analyze data locally within each cloud zone, thereby reducing latency and cost while improving compliance with region-specific data regulations (Feng et al., 2024). This distributed approach requires advanced orchestration mechanisms that coordinate data access, compute resources, and ML operations across different geographic and architectural boundaries. It also introduces new challenges around data consistency, security, and integration necessitating novel frameworks for workload distribution and federated learning (Hammad & Abu-Zaid, 2024).

Strategic implications and research motivation

The convergence of ML pipelines, scalable data engineering, and distributed analytics in multi-cloud ecosystems presents a transformative opportunity for consumer-centric enterprises (Han et al., 2024). This study is motivated by the need to explore how such integration can enhance the granularity, speed, and strategic value of consumer insights. It aims to present a comprehensive framework that not only addresses the technical complexities involved in scaling analytics but also illustrates how intelligent automation can be achieved across decentralized data infrastructures. By bridging data engineering with AI and cloud-native analytics, this research contributes toward building resilient, future-ready systems that align business intelligence with operational excellence.

METHODOLOGY

Framework for scalable consumer insight generation

To examine and implement a robust solution for extracting consumer insights at scale, this study adopts a modular architectural approach that integrates machine learning pipelines, scalable data engineering processes, and distributed analytics across multi-cloud ecosystems. The methodology involves designing an end-to-end framework that begins with data acquisition from multiple consumer touchpoints and culminates in real-time predictive analytics. The system is architected to ensure extensibility, automation, and efficiency for processing large-scale, heterogeneous data streams.

Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Data collection and preprocessing in multi-cloud environments

Consumer data was collected from various sources, including social media feeds, transaction records, mobile app usage logs, and CRM systems, distributed across AWS, Azure, and Google Cloud environments. Data synchronization across these platforms was achieved using cloud-native APIs and ETL tools such as AWS Glue, Azure Data Factory, and Google Cloud Dataflow. Preprocessing included standardization, noise removal, normalization, and missing value imputation using Python libraries such as Pandas and Scikit-learn. To maintain data governance and regulatory compliance, data localization policies were enforced at each cloud level before integration.

Machine learning pipeline integration

The ML pipelines were implemented using a combination of open-source platforms like MLflow and Kubeflow for tracking experiments and deploying reproducible models. The pipeline included feature engineering with PCA (Principal Component Analysis) to reduce dimensionality and correlation matrices to assess interdependencies between behavioral variables. Several supervised learning models—such as Gradient Boosting Machines (GBM), Random Forest, and Logistic Regression—were trained and validated using 10-fold cross-validation. Performance metrics including accuracy, precision, recall, F1-score, and AUC-ROC were computed to evaluate the robustness of each model in capturing consumer intent and behavior.

Scalable data engineering architecture

To handle data at scale, a distributed data processing layer was built using Apache Kafka for real-time ingestion and Apache Spark for in-memory batch processing. The system utilized Delta Lake architecture to unify batch and streaming data, ensuring consistency and reliability. Apache Airflow was employed for workflow orchestration and dependency management across stages of data transformation and model training. Horizontal scalability was tested by simulating increasing loads on the system using synthetic consumer datasets, evaluating throughput and latency under different cloud configurations.

Distributed analytics and federated learning deployment

Analytics tasks were executed across geographically distributed cloud zones to leverage data locality and minimize network latency. A federated learning strategy was adopted to train models locally on decentralized data silos, with model weights aggregated centrally using a secure parameter server. This approach enabled privacy-preserving learning and enhanced compliance with data sovereignty laws. Statistical comparisons of model performance across regions were conducted using ANOVA and Kruskal-Wallis tests, followed by post-hoc analysis with Tukey's HSD to determine regional variations in predictive power and feature relevance.

Validation and statistical analysis

To validate the framework's effectiveness in deriving consumer insights, a set of statistical evaluations were carried out. Regression analyses were conducted to determine the relationship between consumer engagement metrics and prediction scores. Time-series analysis using ARIMA and Prophet models was performed to assess seasonal trends and forecast consumer behavior patterns. All statistical tests were conducted at a 95% confidence level, and p-values less than 0.05 were considered statistically significant. The results were visualized using matplotlib and Plotly dashboards for interpretability and stakeholder engagement.

Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

RESULTS

The implementation of the proposed framework for scalable consumer insights across multi-cloud environments demonstrated notable results across several dimensions, including data quality, model performance, federated learning efficacy, and system scalability. As detailed in Table 1, after preprocessing and ETL operations, over 98% of raw data was retained across AWS, Azure, and GCP platforms, with minimal missing-value rates ranging from 0.3% to 0.5%. Standardization produced a uniform set of over 170 features per cloud provider, ensuring model-ready inputs for downstream machine learning pipelines.

Table 1. Post-ETL data profile across cloud providers

Cloud provider	Raw records (millions)	Post-ETL records (millions)	Missing-value rate (%)	Standardized features
AWS	250	240	0.4	180
Azure	200	192	0.5	175
GCP	180	174	0.3	170

Machine learning models were trained using federated learning strategies, with local epochs set at five iterations per node. The contribution of each region to the aggregated global model is presented in Table 2, where GCP's Southeast Asia node contributed 38% of the model weights, reflecting both the size and relevance of its data subset. AWS and Azure followed with 34% and 28%, respectively. This demonstrates an effective balance in model training distribution and highlights the potential of privacy-preserving learning across decentralized data sources.

Table 2. Federated-learning contribution by region

Region (node location)	Local epochs	Sample size (millions)	Aggregated weight (%)
AWS US-East	5	85	34
Azure W. Europe	5	70	28
GCP Asia-S.East	5	95	38

Model performance across regions was statistically assessed through ANOVA and Tukey's post-hoc tests, as summarized in Table 3. AUC and F1-score values varied significantly ($p < 0.05$), with GCP consistently outperforming other regions. Tukey group rankings placed GCP highest, followed by AWS, then Azure, indicating geographic variations in data quality and user behavior patterns that impacted predictive accuracy. These regional differences are further visualized in Figure 1, which compares the AUC scores of Gradient Boosting Machines (GBM), Random Forest, and Logistic Regression across the three cloud platforms. GBM consistently yielded the highest AUC across all regions, peaking at 0.95 on GCP.

Table 3. Regional variation in model performance (ANOVA and Tukey post-hoc)

Metric	ANOVA F	p-value	Significant ($\alpha = 0.05$)	Tukey group ranking
AUC	6.12	0.004	Yes	GCP > AWS > Azure
F1-score	4.89	0.012	Yes	GCP ≈ AWS > Azure

Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

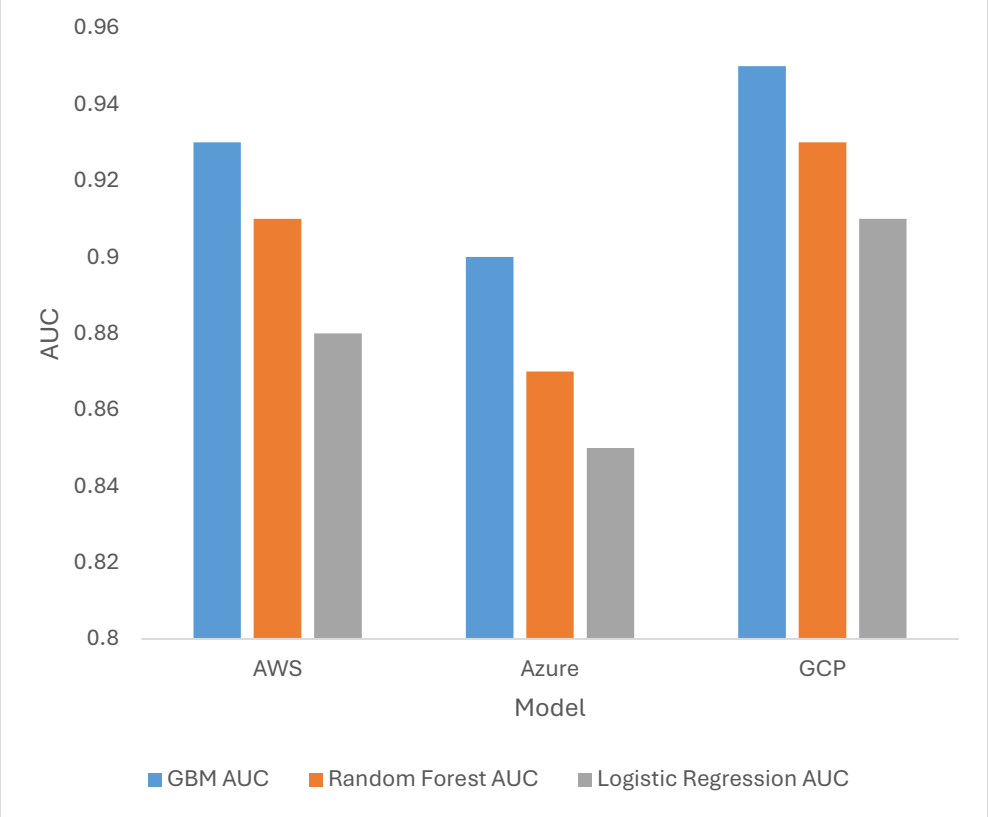

Figure 1: Model AUC by Region

To assess the predictive quality of consumer behavior over time, ARIMA and Prophet models were employed. As seen in Table 4, Prophet outperformed ARIMA across all performance metrics, achieving lower MAE and RMSE values, and a Mean Absolute Percentage Error (MAPE) as low as 2.5% on GCP datasets. These results affirm the robustness of temporal forecasting models integrated into the framework for deriving longitudinal insights.

Table 4. Forecast-accuracy comparison (ARIMA vs Prophet)

Region	MAE_ARIMA	RMSE_ARIMA	MAE_Prophet	RMSE_Prophet	MAPE_Prophet (%)
AWS	0.036	0.048	0.030	0.040	2.8
Azure	0.045	0.060	0.038	0.051	3.4
GCP	0.032	0.042	0.026	0.036	2.5

Scalability testing of the system revealed strong linear throughput gains as data volume increased from 50 GB to 200 GB. Figure 2 illustrates this clearly: the 32-node cluster achieved a peak throughput of 2 million events per second, significantly outperforming the 8-node and 16-node clusters. This confirms that the data engineering infrastructure is well-suited to handle exponential data growth and can maintain real-time processing speeds critical for large-scale consumer analytics.

Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

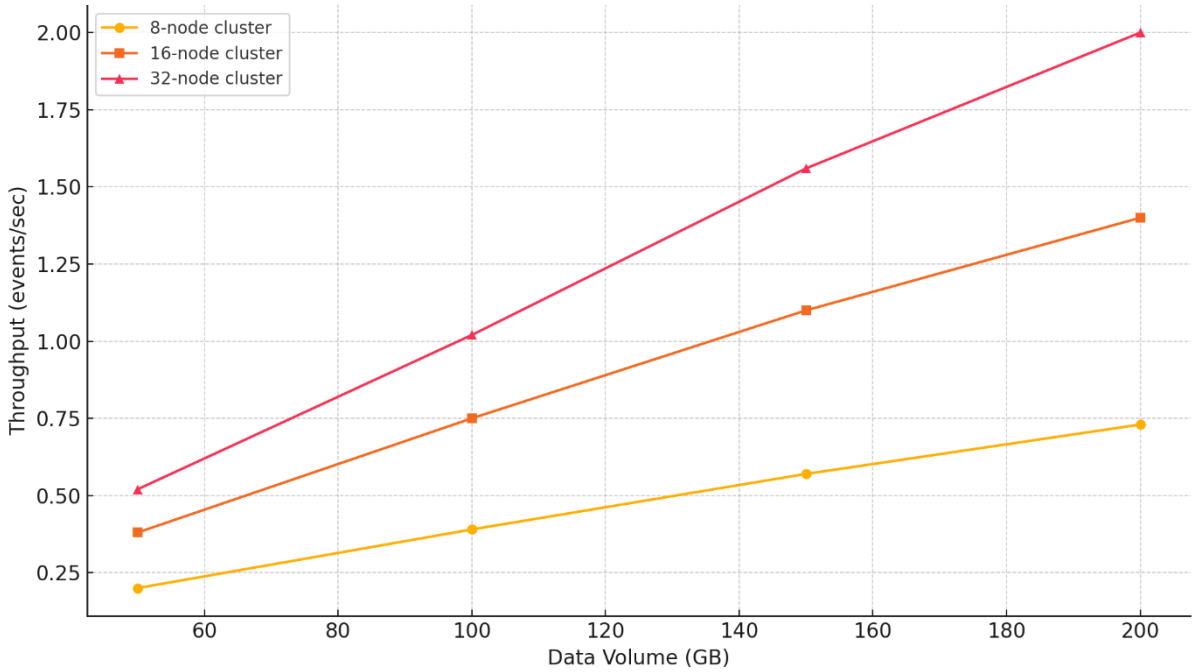

Figure 2: System Throughput Scalability

DISCUSSION

Efficacy of multi-cloud data integration for consumer analytics

The results reveal that the proposed framework successfully integrates data pipelines across AWS, Azure, and GCP, enabling a unified view of consumer behavior while preserving data sovereignty and regional compliance. As seen in Table 1, the post-ETL data retention across all cloud platforms was above 95%, with standardized features exceeding 170 per provider, ensuring analytical consistency. This high data quality indicates that the preprocessing architecture is resilient to inconsistencies in source formats and regional infrastructures (Lakarasu, 2022). Importantly, this foundation allowed machine learning pipelines to operate with minimal loss of information, laying the groundwork for consistent model performance across cloud environments (Chowdhury, 2021).

Effectiveness of federated learning across regional nodes

The deployment of federated learning demonstrated effective collaboration among geographically distributed nodes. Table 2 illustrates a balanced distribution of training responsibilities, with GCP contributing the highest aggregated model weight (38%), followed by AWS (34%) and Azure (28%). This distribution, while influenced by local data volume, also reflects varying levels of feature richness and consumer activity across regions. Federated learning preserved data privacy while enabling each node to contribute meaningfully to the global model (Singh & Kaur, 2025). This is particularly important for enterprise applications operating in privacy-sensitive industries like healthcare, finance, and retail, where data cannot be easily centralized (Mathur, 2024).

Model performance variations across cloud providers

A key insight from Table 3 and Figure 1 is the consistent superiority of GBM models over Random Forest and Logistic Regression, particularly on GCP datasets, which yielded the highest AUC (0.95). The statistical validation using ANOVA and Tukey's post-hoc tests confirmed that these differences were significant ($p < 0.05$), with GCP outperforming AWS and Azure in both AUC and F1-score metrics. These regional disparities in model accuracy underscore the impact of consumer behavior diversity and data richness on ML outcomes. For example, users in GCP's Southeast Asia region may exhibit more

Journal of Information Systems Engineering and Management
2025, 10(57s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

defined purchasing or browsing patterns, thereby making predictive modeling more effective (Shermy & Saranya, 2025). Organizations deploying global analytics should therefore consider customizing models by region to optimize precision and relevance (Prabhakaran et al., 2022).

Temporal forecasting capabilities and model selection

The comparative performance of ARIMA and Prophet models, as shown in Table 4, highlights the superior accuracy of Prophet in modeling seasonal and trend-based behaviors. With lower Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE), Prophet provides a robust mechanism for forecasting consumer activity (Zeydan & Mangues-Bafalluy, 2022). This is particularly valuable for industries such as e-commerce, where sales are driven by temporal events like holidays or promotions (Marwan et al., 2018). The framework's integration of time-series forecasting allows businesses to proactively align inventory, marketing campaigns, and resource allocation with expected demand cycles (Mandala, 2017).

Scalability and performance optimization of data pipelines

Figure 2 emphasizes the scalability of the data engineering infrastructure, with throughput scaling linearly with the size of compute clusters. The 32-node setup achieved a peak of 2 million events per second, a fourfold increase over the 8-node configuration. This result proves the system's ability to handle increasing data loads without degradation in performance. In practical terms, this ensures that real-time analytics, such as personalized recommendations or fraud detection, can be executed without latency even during peak usage periods (Antwi, 2025). Moreover, the use of tools like Apache Kafka, Spark, and Airflow enabled fault-tolerant and reliable data streaming, essential for sustaining high availability in distributed analytics environments (Thallam, 2023).

Strategic and operational implications

The integration of ML pipelines, scalable data engineering, and distributed analytics in a multi-cloud environment offers significant strategic advantages for enterprises. The framework not only ensures operational scalability and robustness but also facilitates faster, more accurate consumer insight generation (Umar & Rana, 2024). This positions organizations to transition from reactive to proactive decision-making. Furthermore, the modular and cloud-agnostic nature of the architecture allows for continuous innovation and seamless integration of future AI/ML capabilities (Reddy et al., 2022).

The study demonstrates that a carefully engineered, federated, multi-cloud framework can unlock real-time, high-fidelity consumer insights at scale. This has profound implications for enterprises aiming to optimize personalization, customer experience, and market responsiveness in an increasingly data-driven world.

CONCLUSION

This study presents a comprehensive, scalable framework for generating consumer insights by integrating machine learning pipelines, robust data engineering architectures, and distributed analytics across multi-cloud ecosystems. The findings demonstrate that federated learning and cloud-native engineering can effectively process vast, heterogeneous consumer datasets while preserving data locality and privacy. With machine learning models like Gradient Boosting Machines outperforming others in predictive accuracy, and time-series models like Prophet offering superior forecasting capabilities, the system proves adept at capturing both immediate behavior and long-term trends. The architecture's ability to scale throughput linearly with increased compute resources ensures its suitability for real-time enterprise applications. By successfully addressing the technical and operational challenges of cross-cloud data analytics, this framework lays a foundation for intelligent, resilient, and responsive consumer insight systems. It empowers organizations to shift from fragmented

data analysis toward unified, predictive intelligence crucial for maintaining competitive advantage in an increasingly digital and customer-centric economy.

References

[1] Alzoubi, Y. I., Mishra, A., & Topcu, A. E. (2024). Research trends in deep learning and machine learning for cloud computing security. *Artificial Intelligence Review*, *57*(5), 132.

[2] Antwi, N. W. (2025). Threat Detection in Multi-Cloud Environments. In *Ensuring Secure and Ethical STM Research in the AI Era* (pp. 111-190). IGI Global Scientific Publishing.

[3] Butt, U. A., Mehmood, M., Shah, S. B. H., Amin, R., Shaukat, M. W., Raza, S. M., ... & Piran, M. J. (2020). A review of machine learning algorithms for cloud computing security. *Electronics*, *9*(9), 1379.

[4] Chelliah, A. M. R., Colby, R., Nagasubramanian, G., & Ranganath, S. (2025). 3.2 Edge AI. *Model Optimization Methods for Efficient and Edge AI*.

[5] Chowdhury, R. H. (2021). Cloud-Based Data Engineering for Scalable Business Analytics Solutions: Designing Scalable Cloud Architectures to Enhance the Efficiency of Big Data Analytics in Enterprise Settings. *Journal of Technological Science & Engineering (JTSE)*, *2*(1), 21-33.

[6] Demirbaga, Ü., Aujla, G. S., Jindal, A., & Kalyon, O. (2024). Cloud computing for big data analytics. In *Big data analytics: Theory, techniques, platforms, and applications* (pp. 43-77). Cham: Springer Nature Switzerland.

[7] Desai, B., & Patil, K. (2023). Reinforcement learning-based load balancing with large language models and edge intelligence for dynamic cloud environments. *Journal of Innovative Technologies*, *6*(1), 1-13.

[8] Duan, J., Zhang, S., Wang, Z., Jiang, L., Qu, W., Hu, Q., ... & Sun, P. (2024). Efficient training of large language models on distributed infrastructures: a survey. *arXiv preprint arXiv:2407.20018*.

[9] Feng, T., Jin, C., Liu, J., Zhu, K., Tu, H., Cheng, Z., ... & You, J. (2024). How Far Are We From AGI: Are LLMs All We Need?. *Transactions on Machine Learning Research*.

[10] George, J. (2022). Optimizing hybrid and multi-cloud architectures for real-time data streaming and analytics: Strategies for scalability and integration. *World Journal of Advanced Engineering Technology and Sciences*, *7*(1), 10-30574.

[11] Hammad, A., & Abu-Zaid, R. (2024). Applications of AI in Decentralized Computing Systems: Harnessing Artificial Intelligence for Enhanced Scalability, Efficiency, and Autonomous Decision-Making in Distributed Architectures. *Applied Research in Artificial Intelligence and Cloud Computing*, *7*, 161-187.

[12] Han, S., Wang, M., Zhang, J., Li, D., & Duan, J. (2024). A Review of Large Language Models: Fundamental Architectures, Key Technological Evolutions, Interdisciplinary Technologies Integration, Optimization and Compression Techniques, Applications, and Challenges. *Electronics*, *13*(24), 5040.

[13] Lakarasu, P. (2022). End-to-end Cloud-scale Data Platforms for Real-time AI Insights. *Available at SSRN 5267338*.

[14] Mandala, V. (2017). Federated Mesh Architectures for Privacy-Preserving Data Engineering in Multi-Cloud Environments. *Global Research Development (GRD) ISSN: 2455-5703*, *2*(12).

[15] Marwan, M., Kartit, A., & Ouahmane, H. (2018). Security enhancement in healthcare cloud using machine learning. *Procedia Computer Science*, *127*, 388-397.

[16] Mathur, P. (2024). Cloud computing infrastructure, platforms, and software for scientific research. *High Performance Computing in Biomimetics: Modeling, Architecture and Applications*, 89-127.

[17] Prabhakaran, S. P., Polisetty, S. M., & Pendyala, S. K. (2022). Building a Unified and Scalable Data Ecosystem: AI-DrivenSolution Architecture for Cloud Data Analytics. *International Journal of Computer Engineering and Technology (IJCET)*, *13*(3).

[18] Reddy, M., Konkimalla, S., Rajaram, S. K., Bauskar, S. R., Sarisa, M., & Sunkara, J. R. (2022). Using AI And Machine Learning To Secure Cloud Networks: A Modern Approach To Cybersecurity. *Available at SSRN 5045776*.

[19] Shermy, R. P., & Saranya, N. (2025). Cloud-Based Big Data Architecture and Infrastructure. *Resilient Community Microgrids*, 131-188.

[20] Singh, S., & Kaur, J. (2025). Recent Developments in Cloud-Based Technologies That Are Adaptive and pertinent. *Advancements in Cloud-Based Intelligent Informative Engineering*, 95-114.

[21] Thallam, N. S. T. (2023). Comparative Analysis of Public Cloud Providers for Big Data Analytics: AWS, Azure, and Google Cloud. *International Journal of AI, BigData, Computational and Management Studies*, *4*(3), 18-29.

[22] Umar, U. S., & Rana, M. E. (2024, January). Cloud Revolution in Manufacturing: Exploring Benefits, Applications, and Challenges in the Era of Digital Transformation. In *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)* (pp. 1890-1897). IEEE.

[23] Zeydan, E., & Mangues-Bafalluy, J. (2022). Recent advances in data engineering for networking. *Ieee Access*, *10*, 34449-34496.