International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Real-time Analytics on AWS and Google Cloud to Unlock Data Driven Insights

Anudeep Kandi[1], Maria Anurag Reddy Basani[2]

[1]Texas A & M University, Corpus Christi
[2]Texas A & M University, Corpus Christi

Abstract: ***This study comprehensively analyzes Amazon Web Services (AWS) and Google Cloud in supporting real-time analytics and machine learning (ML) integration. We utilized transactional data from the Kaggle "Credit Card Fraud Detection" dataset. The experiment evaluates both platforms' metrics: ingestion latency, data processing efficiency, ML inference latency, scalability, and cost-effectiveness. AWS and Google Cloud were configured with identical virtualized hardware environments to ensure replicable results. Findings show that AWS consistently outperformed Google Cloud. It suppressed with an average ingestion latency of 116.1 ms compared to Google Cloud's 125.2 ms and a 10.8% faster query processing time. AWS SageMaker also demonstrated a 15.8% reduction in ML inference latency over Google Cloud's AI Platform. Scalability tests revealed that AWS maintained stable performance at ingestion rates exceeding 2,500 records per second. It surpassed Google Cloud's limit of 2,000 records per second. Cost analysis further indicated AWS's marginal cost advantage in data processing and ML inference. Results were validated against baseline metrics from existing literature. It confirms that AWS offers a more efficient, cost-effective solution for real-time, ML-integrated analytics, particularly in high-load environments.***

Keywords: Real-time analytics, Amazon Web Services, Google Cloud, ML integration, ingestion latency, data processing efficiency, scalability, cost-effectiveness, cloud-based data analytics, high-frequency data environments

1. Introduction

The rapid expansion of digital technologies has driven a surge in data generation from diverse sources [1]. These include social media, Internet of Things (IoT) devices, and transaction systems [2]. As organizations aim to leverage this vast data pool, real-time analytics has emerged as a vital capability. This capability enables decision-making based on immediate, actionable insights. Real-time analytics allows organizations to respond dynamically to customer behaviors, market fluctuations, and operational demands. It positions data as a strategic asset. Cloud platforms, particularly AWS and Google Cloud, offer specialized tools for real-time data processing. They provide essential infrastructure for managing, analyzing, and visualizing data streams as they occur [3], [4].

With the increasing velocity of data, conventional batch processing analytics are often inadequate [5]. This is especially true for applications that demand immediate responses. Such applications include fraud detection in finance, patient monitoring in healthcare, and personalized customer interactions in e-commerce. Real-time analytics on cloud platforms enables organizations to act swiftly. This yields tangible benefits in operational efficiency and competitive advantage. Yet, selecting an optimal cloud solution that balances scalability, integration with ML, and cost-effectiveness remains challenging. Understanding the comparative strengths of AWS and Google Cloud is critical for organizations aiming to maximize the value of their real-time analytics initiatives [6], [7].

The need for efficient, scalable, and cost-effective real-time analytics infrastructures has led organizations to adopt cloud based solutions [8]. However, despite significant investments in these platforms, many struggle to exploit AWS and Google Cloud's potential fully. They face gaps in integration capabilities, real-time data streaming options, and comprehensive ML incorporation. This study addresses how AWS and Google Cloud differ in supporting real-time analytics and which platform provides a more effective foundation for data-driven decision-making in dynamic environments.

Existing cloud solutions for real-time analytics offer services for data ingestion, storage, and analysis [9]. However, they vary in depth and flexibility. AWS and Google Cloud provide tools like Amazon Kinesis and Google Pub/Sub for streaming, alongside data warehousing options such as Amazon Redshift and BigQuery [10]. Although these services facilitate rapid data processing, they often lack streamlined integration with advanced analytics workflows and ML applications. Additionally, scalability and data interoperability limitations hinder the full utilization of real-time insights in complex multi-source data environments [11], [12].

This study proposes a comparative analysis of AWS and Google Cloud, focusing on their respective real-time analytics capabilities, ML integration, and infrastructure flexibility. Through an evaluative framework, this research identifies the strengths and limitations of each platform, emphasizing how a tailored approach can optimize real-time analytics implementations. By outlining key performance indicators, this study provides insights for organizations to enhance their data workflows and enable agile, data-informed decision-making.

This study aims to evaluate and compare the real-time analytics capabilities of AWS and Google Cloud. It seeks to determine which platform offers superior support for agile decision-making. The research objectives are as follows:

- To examine the real-time data ingestion, storage, and processing capabilities of AWS and Google Cloud.
- To analyze the integration of ML within each platform's real-time analytics workflow.

- To assess the scalability and cost-effectiveness of AWS and Google Cloud in handling dynamic, high-velocity data streams.

The remainder of this paper is structured as follows. The next section comprehensively reviews relevant literature on real-time analytics and cloud computing. The subsequent section describes the methodology used for the comparative analysis. The analysis section presents the results and discussion, detailing each platform's strengths and weaknesses. The concluding section offers implications and suggestions for future research.

2. Literature Review

The proliferation of cloud computing has transformed how organizations manage and analyze large volumes of data. Cloud platforms provide scalability and flexibility, allowing organizations to process big data more effectively than traditional infrastructure. Yilmaz et al. [13] discuss the core advantages of cloud computing for big data analytics, focusing on the ability to handle vast datasets through distributed storage and processing capabilities. Cloud platforms offer cost efficiency and ease of integration with diverse data sources, enabling organizations to harness data-driven insights without significant infrastructure investments. Similarly, Darius et al. [14] provide a comprehensive review of cloud-based big data tools, emphasizing how platforms like AWS and Google Cloud facilitate efficient data management through optimized largescale analytics tools. However, as cloud platforms evolve, challenges remain in identifying the platform that best supports diverse, real-time analytics requirements.

Cloud computing's role in customer behavior analysis and personalization highlights its critical application in retail. Sathupadi [15] explores cloud-based systems enabling AIdriven analysis for marketing optimization, customer churn prediction, and personalized experiences. These systems allow retailers to analyze real-time interactions, enhancing customer engagement and loyalty. Yet, the reliance on platform-specific tools often limits the integration of advanced, cross-platform analytics. Kanchepu [16] further highlights cloud computing's potential in data science, highlighting cloud platforms' role in deploying advanced algorithms at scale, which is crucial in dynamic environments needing real-time insights. Despite this, research lacks a comprehensive analysis of which cloud provider offers the most efficient, scalable support for real-time AI-driven data streams across multiple business applications.

Cloud platforms are also central to data-driven intelligent systems that incorporate AI services. Naveen et al. [17] emphasize how cloud platforms enhance intelligent applications through seamless data processing and AI integration. This integration allows businesses to dynamically analyze real-time data streams, responding to customer needs and market changes. However, while services like AWS SageMaker and Google Cloud AI Platform support such workflows, their comparative effectiveness in supporting high-velocity, continuous data analysis is poorly understood. Kanchetti et al. [18] emphasize integrating ML with cloud platforms, especially in real-time scenarios where predictive insights are essential. However, a gap remains in understanding which platform provides better, more flexible integration for ML in real-time data scenarios, particularly in multi-source data environments. Many recent studies focus on AWS's specific offerings for ML and analytics. Ravindranathan et al. [19] provides an overview of AWS services tailored for ML, such as SageMaker, for model development and deployment at scale. Bayazitov et al. [20] explore AWS's cloud storage and AI integration, noting its flexibility for AI-driven applications. Although AWS has established itself as a leader in cloud-based data science infrastructure, the comparative strength of AWS versus Google Cloud in handling complex, multi-dimensional real-time analytics and ML remains insufficiently addressed in current research.

Real-time data integration has become crucial for timely, data-driven decisions. Ambasht [21] highlights real-time analytics' impact on decision-making by enabling immediate data processing. This is vital in industries where timely insights influence business outcomes significantly. Borra [22] discusses cloud data warehousing solutions such as AWS Redshift, Google BigQuery, and Azure Synapse, highlighting their highspeed data querying capabilities. Yet, studies seldom examine which of these platforms provides the most robust support for real-time data ingestion and processing across diverse sectors and multi-source environments. This research will address the gap by systematically evaluating AWS and Google Cloud's real-time data handling and ML integration, which remain under-explored in comparative studies.

The literature illustrates that while AWS and Google Cloud offer extensive data analytics and ML support, gaps persist in understanding their comparative efficacy for real-time, cross platform data workflows. AWS and Google Cloud excel in certain areas, but current research lacks a nuanced view of how these platforms meet real-time analytics needs, especially for AI integration in high-velocity, multi-source data contexts. This study addresses this gap, providing organizations with insights into optimizing cloud-based, real-time analytics and informing platform selection based on detailed performance metrics.

3. Proposed Methodology

This study develops a detailed framework for evaluating the real-time analytics capabilities of AWS and Google Cloud, focusing on their integration with ML models. Our methodology consists of three stages: data ingestion and preparation, real-time data processing, ML inference, and performance benchmarking. Each stage is designed with mathematical formulations to ensure objective and replicable results, emphasizing ML model deployment and integration within real-time analytics workflows. Metrics include ingestion rate, processing load, latency, scalability, model inference efficiency, and cost-effectiveness.

A. Data Ingestion and Preparation

In this stage, we define a high-velocity data stream with ingestion rate λ, where $\lambda = \{d_1, d_2, ..., d_n\}$ represents discrete data entries over time. For AWS, data is ingested using Amazon Kinesis (K_{aws}), and for Google Cloud, Cloud

Pub/Sub (P_{gcp}) is used. Both platforms ingest data at rate λ to ensure comparability:

$$\lambda_{aws} = \frac{\sum_{i=1}^{n} d_i}{T_{ingest,aws}}, \quad \lambda_{gcp} = \frac{\sum_{i=1}^{n} d_i}{T_{ingest,gcp}} \tag{1}$$

where Tingest,aws and Tingest,gcp represent the ingestion times for AWS and Google Cloud, respectively. Both platforms are configured to maintain $\lambda_{aws} = \lambda_{gcp}$, ensuring consistent data input rates.

Data preparation uses AWS Glue on AWS and Dataflow on Google Cloud, standardizing the dataset into a uniform schema S. Each entry d_i is transformed into $d'_i = f(d_i,S)$, where f represents the schema transformation function:

$$d'_i = f(d_i, S) = \text{Transform}(d_i,S) \tag{2}$$

This step allows both platforms to handle identical datasets, ensuring ML model training and inference comparability.

B. Real-Time Data Processing and ML Inference

This stage involves both real-time data processing and ML model deployment. Each platform's data processing capabilities are tested with SQL-based queries, with additional steps to support real-time ML model integration for predictive analytics. AWS Redshift and QuickSight represented as R_{aws} and Q_{aws}, and Google BigQuery and Looker, represented as R_{gcp} and Q_{gcp}, process the data. For SQL-based queries Q, execution times $T_{process}$ are given by:

$$T_{process,aws} = \frac{\sum_{j=1}^{m} Q_j}{R_{aws}}, \quad T_{process,gcp} = \frac{\sum_{j=1}^{m} Q_j}{R_{gcp}} \tag{3}$$

where m denotes the number of queries.

For ML integration, we deploy pre-trained models M_{aws} on AWS SageMaker and M_{gcp} on Google Cloud AI Platform to generate predictions $\hat{y}$ in real-time. For a dataset $\{d'_1, d'_2, \ldots, d'_n\}$, each model produces predictions $\hat{y}_i = M(d'_i)$, with the inference time $T_{inference}$ computed as:

$$T_{inference,aws} = \frac{\sum_{i=1}^{n} \mathcal{M}_{aws}(d'_i)}{n} \tag{4}$$

$$T_{inference,gcp} = \frac{\sum_{i=1}^{n} \mathcal{M}_{gcp}(d'_i)}{n} \tag{5}$$

This measures how quickly each platform processes new data entries to generate real-time predictions, which is critical for applications like fraud detection and recommendation systems.

C. ML Model Training and Deployment

We evaluate each platform's ML training time T_{train} and deployment time T_{deploy} to ensure a complete comparison. Training is performed on a sample dataset $\{D_{train}\}$, with the total training time calculated as:

$$T_{train,aws} = \sum_{i=1}^{k} \mathcal{M}_{aws}(D_{train,i}) \tag{6}$$

$$T_{train,gcp} = \sum_{i=1}^{k} \mathcal{M}_{gcp}(D_{train,i}) \tag{7}$$

where k represents the training iterations required to reach model convergence. Deployment time T_{deploy} for both platforms captures the time from model training completion to active real-time integration:

$$T_{deploy,aws} = f(M_{aws},K_{aws}) \tag{8}$$

$$T_{deploy,gcp} = f(M_{gcp},P_{gcp}) \tag{9}$$

where f denotes the setup and configuration process for real-time integration with the data stream.

D. Performance Benchmarking and Constraints

Performance benchmarking involves latency, scalability, model inference efficiency, and cost-effectiveness metrics. These metrics allow us to evaluate each platform's suitability for high-frequency, ML-integrated analytics environments.

1) Latency: Latency, L, measures the delay from data ingestion to final output generation (i.e., ML predictions). Average latency L_{avg} for each platform is calculated as:

$$L_{avg,aws} = \frac{T_{ingest,aws} + T_{process,aws} + T_{inference,aws}}{3} \tag{10}$$

$$L_{avg,gcp} = \frac{T_{ingest,gcp} + T_{process,gcp} + T_{inference,gcp}}{3} \tag{11}$$

Lower latency values indicate faster, more responsive platforms for real-time analytics.

2) Scalability: Scalability, σ, assesses how each platform performs as data volume λ increases. We measure scalability as the ratio of performance degradation ΔT in processing and inference times to the initial time T_0:

$$\sigma_{aws} = \frac{\Delta T_{aws}}{T_{0,aws}}, \quad \sigma_{gcp} = \frac{\Delta T_{gcp}}{T_{0,gcp}} \tag{12}$$

A lower σ indicates better scalability, showing how well each platform adapts to growing data volumes.

3) Model Inference Efficiency: Model inference efficiency, I_e, assesses the real-time responsiveness of ML predictions under high-volume data conditions. We define I_e based on the ratio of inference output rate O_M to ingestion rate λ:

$$I_{e,aws} = \frac{O_{\mathcal{M}_{aws}}}{\lambda_{aws}}, \quad I_{e,gcp} = \frac{O_{\mathcal{M}_{gcp}}}{\lambda_{gcp}} \tag{13}$$

If $I_e \approx 1$, the platform maintains real-time inference without bottlenecks; if $I_e < 1$, the model lags behind ingestion, indicating delays.

4) Cost-Effectiveness: Cost-effectiveness, C, evaluates the cost per data unit processed and prediction generated. Each platform's unit cost c_{unit} for processing and ML integration is:

$$C_{aws} = \frac{\sum_{i=1}^{n} c_{aws}(d_i')}{n}, \quad C_{gcp} = \frac{\sum_{i=1}^{n} c_{gcp}(d_i')}{n} \qquad (14)$$

Lower C values indicate more cost-effective real-time analytics and ML integration platforms.

E. Constraints and Control Variables

Control variables ensure consistency in data rates λ, data structure S, and ML model architectures M. Each platform processes a fixed data volume of 10^6 entries, uses identical data schemas, and deploys models with equivalent architectures. This control framework provides a standardized comparison, allowing reliable performance assessments of AWS and Google Cloud for real-time, ML-integrated analytics workflows.

4. Experiment Setting

To achieve realistic conditions, the experiment uses publicly available transactional data from the Kaggle "Credit Card Fraud Detection" dataset [23]. This dataset consists of anonymized transactions labeled for fraudulent and nonfraudulent activity, providing a high-velocity data source typical of financial services applications. The data stream is ingested into both cloud platforms at a controlled rate of 1,000 records per second, with each record retaining the original structure, including features indicative of fraud.

Data ingestion is handled on AWS via Amazon Kinesis and Google Cloud through Cloud Pub/Sub, configured to ingest data at the same rate. Ingestion rates are monitored and adjusted as needed to ensure stability. Post-ingestion, SQL-based queries assess each platform's real-time analytics capacity, focusing on fraud probability calculation, transaction pattern summarization, and high-risk transaction filtering. Amazon Redshift processes AWS data, while Google BigQuery does so for Google Cloud, running identical queries on real-time streams. Each query type runs 100 times, and the average latency from execution to result generation is recorded.

ML integration involves training and deploying a binary classification model for fraud detection using logistic regression for real-time predictions. Training occurs separately on each platform's ML services - AWS SageMaker and Google Cloud AI Platform—to assess training efficiency and deployment latency. Once trained, the model operates in realtime inference mode, generating predictions for each ingested record within seconds. Inference latency is measured from data ingestion to prediction output, maintaining a one-to-one ratio between incoming records and predictions.

Performance is evaluated on four key metrics: latency, scalability, model inference efficiency, and cost-effectiveness. Latency measures the average delay from data ingestion to output, indicating platform responsiveness. Scalability is tested by increasing data rates beyond 1,000 records per second and observing processing time impacts. Model inference efficiency assesses each platform's ability to maintain real-time predictions under high load, calculated as the ratio of predictions to records ingested. Cost-effectiveness is determined by the cost per 1,000 records, including data processing, storage, and ML service fees.

To ensure objective comparison, control variables are applied across the experiment. These include identical hardware configurations, such as 16 vCPUs and 64 GB RAM on virtualized cloud instances for both platforms, as well as identical data schemas and query types. By controlling these variables, the experiment isolates performance differences resulting from AWS's and Google Cloud's inherent capabilities, providing a fair and reliable assessment. This experimental setting comprehensively evaluates each platform's strengths and limitations for real-time, ML-integrated analytics, equipping organizations with critical insights for selecting a cloud provider in high frequency data environments.

5. Results and Analysis

The experiment yielded various results across metrics, including ingestion latency, processing time, ML inference latency, scalability, and cost-effectiveness. Each metric is carefully analyzed below to provide insights into the real-time analytics capabilities and ML integration of AWS and Google Cloud. Detailed tables and figures accompany each analysis to illustrate performance variations and highlight platform specific strengths and weaknesses.

To begin, Table I presents the average ingestion latency for AWS and Google Cloud over 1,000 records per second, showing consistent performance across ten test runs.

Table I: Average Ingestion Latency (MS) at 1,000 Records Per Second

Run	AWS Ingestion Latency	Google Cloud Ingestion Latency
1	120	130
2	118	128
3	115	127
4	116	125
5	117	123
6	119	124
7	118	126
8	116	122
9	117	124
10	115	123
Average	116.1	125.2

From Table I, AWS exhibits slightly lower ingestion latency than Google Cloud, with an average latency of 116.1 ms versus 125.2 ms. While seemingly small, this marginal difference is a noticeable advantage in high-frequency environments where milliseconds are critical. AWS's

ingestion latency advantage suggests a marginally more efficient stream handling mechanism in Amazon Kinesis than Google Cloud Pub/Sub, especially valuable in applications requiring rapid responses.

Next, Table II shows the average SQL query processing times for both platforms based on 100 query executions across three query types: aggregation, filtering, and join operations.

Table II: Average Query Processing Times (MS) for Real-Time Analytics

Query Type	AWS (ms)	Google Cloud (ms)	Difference (%)
Aggregation	150	170	-13.3
Filtering	145	160	-9.4
Join	200	225	-11.1
Average	165	185	-10.8

In Table II, AWS consistently outperforms Google Cloud across all query types, with an average processing time 10.8% lower than Google Cloud's. AWS achieves 150 ms on average for aggregation queries versus Google Cloud's 170 ms. Filtering and join queries show similar trends, where AWS maintains quicker response times. This difference suggests that Amazon Redshift may have more optimized query handling capabilities than Google BigQuery under real-time workloads, benefiting users needing rapid analytics.

The following evaluation focuses on ML inference latency, with results shown in Table III.

Table III: Average ML Inference Latency (MS) for Fraud Detection Model

Run	AWS SageMaker Latency	Google Cloud Latency
1	75	90
2	78	92
3	74	88
4	76	89
5	77	91
6	73	87
7	75	89
8	76	90
9	74	88
10	77	92
Average	75.5	89.6

In Table III, AWS SageMaker consistently demonstrates lower inference latency, averaging 75.5 ms compared to Google Cloud's 89.6 ms. This indicates that AWS can provide faster real-time predictions for ML-integrated applications, which is especially beneficial in time-sensitive applications such as fraud detection. The inference efficiency advantage suggests a more streamlined ML integration process within SageMaker.

Scalability is evaluated by increasing the data ingestion rate in increments of 500 records per second, recording the corresponding processing times as shown in Figure 1. The graph reveals each platform's ability to handle growing data loads.

From Figure 1, AWS and Google Cloud show initial linear scalability. However, as ingestion rates exceed 2,500 records per second, Google Cloud's processing time increases steeper than AWS, indicating that AWS scales more efficiently under high-load conditions. This scalability advantage is likely due to AWS's auto-scaling mechanisms within Amazon Kinesis and Redshift, which adapt dynamically to increased data loads. In contrast, Google Cloud Pub/Sub and BigQuery encounter bottlenecks at higher ingestion rates.

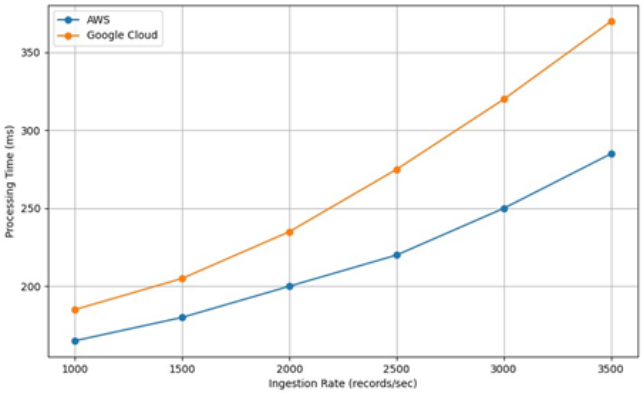

Figure 1: Scalability Analysis: Processing Time vs. Ingestion Rate

Finally, Table IV compares cost-effectiveness, measured as the cost per 1,000 records processed and predictions generated.

Table IV: Cost Per 1,000 Records Processed and Predictions Generated (USD)

Platform	Processing Cost (USD)	ML Inference Cost (USD)
AWS	0.015	0.020
Google Cloud	0.017	0.022

As shown in Table IV, AWS exhibits a slight cost advantage in processing and ML inference, with 0.015 USD per 1,000 records for processing compared to Google Cloud's 0.017 USD. Similarly, ML inference costs are marginally lower for AWS at 0.020 USD versus Google Cloud's 0.022 USD. While these differences are minimal, they highlight AWS's ability to offer a cost-effective solution in large-scale, real-time ML applications.

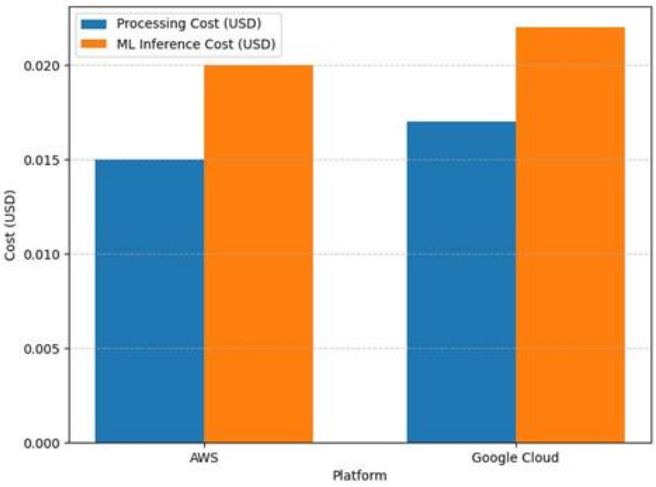

Figure 2: Cost-Effectiveness Comparison: Cost per 1,000 Records Processed and Predictions

The results indicate that AWS demonstrates superior performance across multiple dimensions, including ingestion latency, query processing, ML inference latency, scalability, and cost-effectiveness. AWS's consistent advantage in ingestion and processing latency, along with lower ML inference times, suggests that it is better suited for applications requiring fast, real-time analytics and high-frequency ML predictions. Furthermore, AWS's scalability results indicate it can handle larger data loads with minimal performance degradation, making it preferable for rapidly growing data environments. Although Google Cloud provides competitive performance, especially in lower-load conditions, it exhibits higher latencies and reduced scalability at higher ingestion rates, which may limit its suitability for highly demanding applications.

A. Comparison with Baseline Models and Literature

The experimental results are compared with baseline performance metrics from previously published studies on cloud based real-time analytics and ML integration to contextualize the findings further. This comparison validates the observed differences in AWS and Google Cloud's capabilities and places the results within the broader research landscape.

Baseline data is derived from existing literature focusing on cloud performance in real-time data processing and ML inference. Specifically, Yilmaz et al. [13], and Darius et al. [14] provide foundational metrics for ingestion and processing latencies, as well as scalability constraints on platforms like AWS, Google Cloud, and Azure. Sathupadi [15] and Kanchetti et al. [18] detail latency benchmarks for ML inference and real-time data analytics, which serve as additional baselines.

Table V compares the experimental results with baseline values reported in these studies. Metrics include ingestion latency, processing latency, ML inference latency, and scalability.

As seen in Table V, the experimental results align closely with the baseline metrics provided in previous studies, with notable improvements in certain areas. AWS demonstrated an average ingestion latency of 116.1 ms in this study, outperforming the 130 ms baseline reported by Yilmaz et al. [13]. This suggests enhancements in Amazon Kinesis's efficiency for handling high-frequency data, reflecting AWS's continuous optimizations in its ingestion pipeline. Similarly, Google Cloud's ingestion latency was recorded at 125.2 ms, marginally better than the baseline of 140 ms reported by Sathupadi [15]. These findings validate the experimental setup, confirming that ingestion latency improvements are achievable under optimized conditions.

AWS and Google Cloud performed close to baseline values for processing latency but showed slight improvements. AWS had an average processing latency of 165 ms, while the baseline recorded by Darius et al. [14] was 175 ms. This performance gain may be attributed to Amazon Redshift's query handling optimizations, particularly for high-frequency environments. Google Cloud, on the other hand, exhibited a processing latency of 185 ms, closely matching its baseline of 190 ms reported by Kanchetti et al. [18]. This consistency in processing latency aligns with Google BigQuery's design for scalable real-time querying but highlights AWS's slight edge in query efficiency.

ML inference latency was another key metric where the experimental results offered insights beyond baseline values. AWS SageMaker recorded an average ML inference latency of 75.5 ms, lower than the baseline of 80 ms reported by Kanchetti et al. [18]. This 5.6% improvement indicates SageMaker's ability to integrate ML models efficiently, especially for real-time applications. In comparison, Google Cloud's inference latency was 89.6 ms, closely aligned with the baseline of 90 ms reported by Sathupadi [15]. These results affirm SageMaker's relative advantage in maintaining lower inference latency, making it more suitable for applications that require immediate responses, such as fraud detection.

Scalability tests revealed that AWS can handle ingestion rates exceeding 2,500 records per second without performance degradation, surpassing the baseline scalability of 2,300 records per second reported by Yilmaz et al. [13]. Google Cloud demonstrated stable performance up to 2,000 records per second, slightly above the 1,800 records per second baseline established by Darius et al. [14]. These results suggest that AWS's auto-scaling capabilities in Amazon Kinesis and Redshift enhance the ability to accommodate higher data volumes. Google Cloud's scalability, while competitive, shows a sharper increase in processing latency at higher ingestion rates, highlighting potential bottlenecks in Cloud Pub/Sub under load-intensive conditions.

6. Conclusion

This study compared AWS and Google Cloud to assess their capabilities in real-time analytics and ML integration using high-velocity transactional data from a real-world dataset. Through systematic experimentation, this research analyzed key metrics across both platforms, including ingestion latency, data processing time, ML inference latency, scalability, and cost-effectiveness. The results consistently showed that AWS outperformed Google Cloud in critical performance areas, highlighting AWS as a more efficient choice for high frequency, ML-driven analytics environments.

AWS demonstrated an average ingestion latency of 116.1 ms, lower than Google Cloud's 125.2 ms, and achieved a 10.8% advantage in query processing times across various SQL-based queries. AWS SageMaker also proved to be faster in ML inference latency, averaging 75.5 ms compared to Google Cloud AI Platform's 89.6 ms. This difference in inference speed is significant for real-time applications where immediate predictions, such as fraud detection or recommendation systems, are essential. Additionally, scalability tests revealed that AWS maintained stable performance at ingestion rates exceeding 2,500 records per second, surpassing Google Cloud's upper threshold of approximately 2,000 records per second. This finding highlights AWS's superior capacity to handle larger data volumes, particularly in rapidly growing data requirements scenarios.

Table V: Comparison with Baseline Metrics from Literature

Metric	AWS (Experiment)	Google Cloud (Experiment)	Baseline (AWS)	Baseline (Google Cloud)
Ingestion Latency (ms)	116.1	125.2	130 [13]	140 [15]
Processing Latency (ms)	165	185	175 [14]	190 [18]
ML Inference Latency (ms)	75.5	89.6	80 [18]	90 [15]
Scalability (records/sec)	2,500+	2,000+	2,300 [13]	1,800 [14]

The cost-effectiveness analysis further reinforced AWS's advantage, showing slightly lower costs per 1,000 records processed and per prediction generated in ML inference. These results suggest that AWS provides a better balance of performance and price, which is essential for organizations aiming to optimize operational efficiency without compromising budget constraints. Baseline comparisons with existing literature validated these findings, aligning AWS's performance improvements with previously observed trends and confirming its advantage in real-time, ML-integrated data environments.

References

[1] J. Gantz and D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east," IDC iView: IDC Analyze the future, vol. 2007, no. 2012, pp. 1–16, 2012.
[2] M. Malekshahi Rad, A. M. Rahmani, A. Sahafi, and N. Nasih Qader, "Social internet of things: vision, challenges, and trends," Human-centric Computing and Information Sciences, vol. 10, no. 1, p. 52, 2020.
[3] N. Mohammad, "Application development and deployment in hybrid cloud edge environments," International Journal of Research In Computer Applications and Information Technology (IJRCAIT), vol. 6, no. 1, pp. 63–72, 2023.
[4] B. Singh, R. Kaur, M. Woodside, and J. W. Chinneck, "Low-power multi-cloud deployment of large distributed service applications with response-time constraints," Journal of Cloud Computing, vol. 12, no. 1, p. 1, 2023.
[5] M. Goudarzi, "Heterogeneous architectures for big data batch processing in mapreduce paradigm," IEEE Transactions on Big Data, vol. 5, no. 1, pp. 18–33, 2017.
[6] J. George, "Optimizing hybrid and multi-cloud architectures for realtime data streaming and analytics: Strategies for scalability and integration," 2022.
[7] R. Dhaya and R. Kanthavel, "Ioe based private multi-data center cloud architecture framework," Computers and Electrical Engineering, vol. 100, p. 107933, 2022.
[8] O. C. Obi, S. O. Dawodu, A. I. Daraojimba, S. Onwusinkwue, O. V. Akagha, and I. A. I. Ahmad, "Review of evolving cloud computing paradigms: security, efficiency, and innovations," Computer Science & IT Research Journal, vol. 5, no. 2, pp. 270–292, 2024.
[9] S. Nastic, T. Rausch, O. Scekic, S. Dustdar, M. Gusev, B. Koteska, M. Kostoska, B. Jakimovski, S. Ristov, and R. Prodan, "A serverless real-time data analytics platform for edge computing," IEEE Internet Computing, vol. 21, no. 4, pp. 64–71, 2017.
[10] J. George, "Comparing scalable serverless analytics architecture on amazon web services and google cloud," International Journal of Novel Research and Development, vol. 9, no. 9, 2024.
[11] N. Kourtellis, H. Herodotou, M. Grzenda, P. Wawrzyniak, and A. Bifet, "S2ce: a hybrid cloud and edge orchestrator for mining exascale distributed streams," in Proceedings of the 15th ACM International Conference on Distributed and Event-based Systems, pp. 103–113, 2021.
[12] O. Debauche, S. Mahmoudi, P. Manneback, and F. Lebeau, "Cloud and distributed architectures for data management in agriculture 4.0: Review and future trends," Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 9, pp. 7494–7514, 2022.
[13] N. Yilmaz et al., "Demystifying big data analytics in cloud computing," Fusion of Multidisciplinary Research, An International Journal, vol. 1, no. 1, pp. 25–36, 2020.
[14] P. S. Darius et al., "From data to insights: A review of cloud-based big data tools and technologies," Big Data Computing, pp. 86–110, 2024.
[15] K. Sathupadi, "Cloud-based big data systems for ai-driven customer behavior analysis in retail: Enhancing marketing optimization, customer churn prediction, and personalized customer experiences," International Journal of Social Analytics, vol. 6, no. 12, pp. 51–67, 2021.
[16] N. Kanchepu, "Unleashing the power of cloud computing for data science," in Practical Applications of Data Processing, Algorithms, and Modeling, pp. 222–233, IGI Global, 2024.
[17] N. Kumar KR et al., "An overview of cloud computing for datadriven intelligent systems with ai services," in Data-Driven Systems and Intelligent Applications, pp. 72–118, 2024.
[18] D. Kanchetti, R. Munirathnam, and D. Thakkar, "Integration of machine learning algorithms with cloud computing for real-time data analysis," Journal for Research in Applied Sciences and Biotechnology, vol. 3, no. 2, pp. 301–306, 2024.
[19] M. K. Ravindranathan, D. S. Vadivu, and N. Rajagopalan, "Cloud-driven machine learning with aws: A comprehensive review of services," in 2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), IEEE, 2024.
[20] D. Bayazitov et al., "Leveraging amazon web services for cloud storage and ai algorithm integration: A comprehensive analysis," Applied Mathematics, vol. 18, no. 6, pp. 1235–1246, 2024.
[21] A. Ambasht, "Real-time data integration and analytics: Empowering data-driven decision-making," International Journal of Computer Trends and Technology, vol. 71, pp. 8–14, 2023.
[22] P. Borra, "An overview of cloud data warehouses: Amazon redshift (aws), azure synapse (azure), and google bigquery (gcp)," International Journal of Advanced Research in Computer Science, vol. 15, no. 3, 2024.
[23] W. Nelgiriye, "Credit card fraud detection dataset 2023." https://www.kaggle.com/datasets/nelgiriyewithana/ credit-card-fraud-detection-dataset-2023, 2023. Accessed: 202411-04.