International Journal of Core Engineering & Management
Volume-7, Issue-05, 2023 ISSN No: 2348-9510

SCALABLE REAL-TIME ANALYTICS ON AWS: LEVERAGING KINESIS, EMR, AND REDSHIFT FOR HIGH-VELOCITY DATA PROCESSING

Girish Ganachari
girish.gie@gmail.com

Abstract

The current work carries out a study of scalable real-time analytics solutions based on AWS and investigates such components as Kinesis, EMR and Redshift. As for the contemporary issues, it explores the performance and adaptability of such services to process high-speed streams of data in healthcare, finance, online trade, and similar spheres. Real-life examples show that health status tracking, fraud mitigation, and targeted promotion become possible with the help of big data technologies. Main trends regard the incorporation of digital twins, enhancement of security, energy efficiency of computations, as well as enhancement of machine learning algorithms. In this regard, the study demonstrates the ability of AWS's real-time analytics framework to meet evolving needs in data-driven societies.

Keywords: Amazon Web Service Kinesis, Amazon Web Service Electronic Medical Record, Amazon Web Service Redshift, Real-time analytics, High-velocity data, Scalable architecture, Big data processing, Cloud computing, Data streaming, Predictive analytics, IoT integration, Digital twins, Machine learning, Data security, Energy-efficient computing

I. INTRODUCTION

Technological advancement mainly in data storage has led to the massive generation of big data requiring real-time responses. Amazon Web Services has a rich set of products specifically aimed to deal with big data stream; allowing organizations to derive insights and extract value in near real-time. This paper delves into design and real-time analytics on AWS using Kinesis, EMR and Redshift.

II. AWS REAL-TIME ANALYTICS ARCHITECTURE

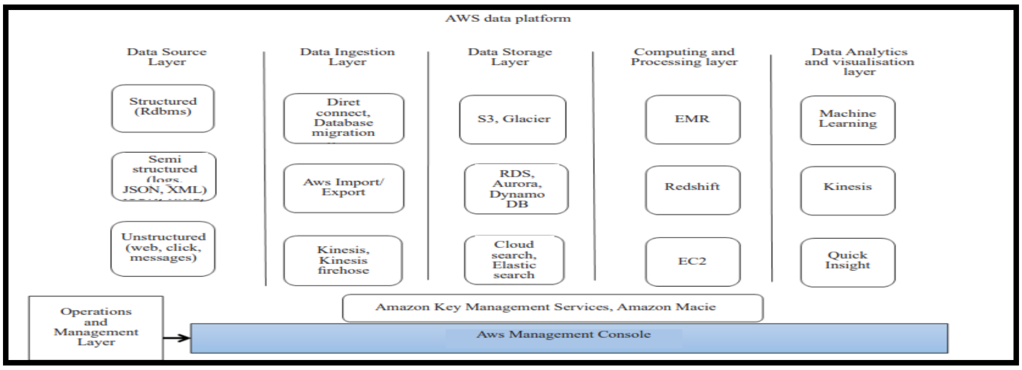
AWS data platform
Data Source Layer
Structured (Rdbms)
Semi structured (logs, JSON, XML)
Unstructured (web, click, messages)
Data Ingestion Layer
Diret connect, Database migration
Aws Import/ Export
Kinesis, Kinesis firehose
Data Storage Layer
S3, Glacier
RDS, Aurora, Dynamo DB
Cloud search, Elastic search
Computing and Processing layer
EMR
Redshift
EC2
Data Analytics and visualisation layer
Machine Learning
Kinesis
Quick Insight
Amazon Key Management Services, Amazon Macie
Operations and Management Layer
Aws Management Console

Figure 1: AWS Architecture
(Source: [26])

1. AWS Kinesis

AWS Kinesis is a service through which real-time data streaming and analysis can be done. They include Kinesis data streams for input, Kinesis data fire hose for data transfer, and Kinesis data analytics for processing.

A. Kinesis Data Streams

Kinesis Data Streams also enable the ingestion of stream of data from different sources for instance IoT devices, logs files and feeds from social media among others [12]. The data is ready for further analysis and decision-making in milliseconds which are very helpful in today's business environment [16].

B. Kinesis Data Firehose

Kinesis data firehose helps in sending streaming data to destinations such as Amazon S3, redshift, Elastic Search etc [10]. It supports the change of data structure and data aggregation before passing the data, thus improving the effectiveness of data transport and storage [21].

C. Kinesis Data Analytics

Kinesis Data Analytics applies real-time SQL processing to streaming data which allows for instantaneous decisions and actions [2]. This capability is important in applications where real-time processing is essential as in fraud detection and dynamic pricing [4].

2. AWS EMR

AWS EMR is a big data service offered on Amazon Web Services that is used to process large datasets using distributed frameworks such as Hadoop, Spark, and HBase [14]. EMR involves both batch mode processing as well as real-time processing, which gives a facility of scalability [20].

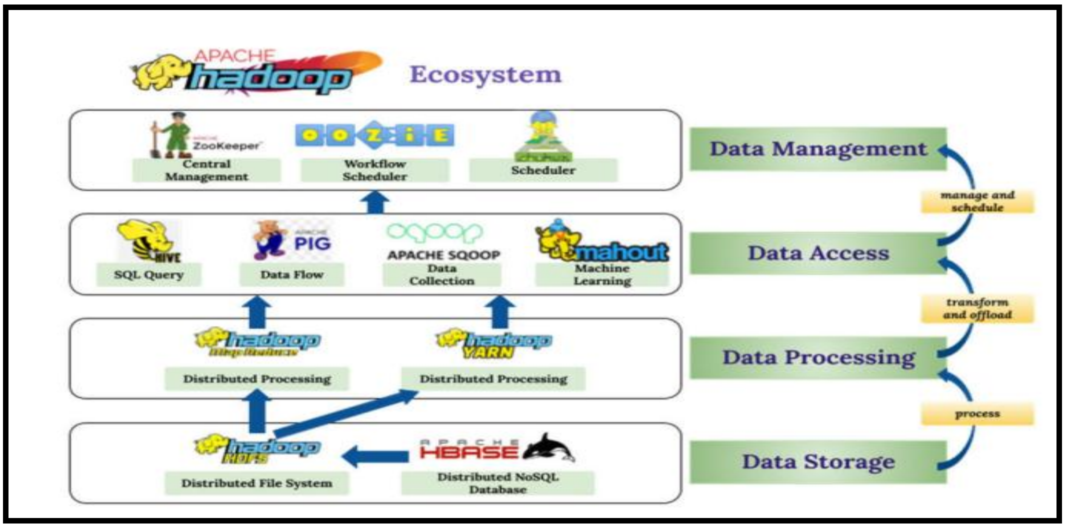

Figure 2: Apache Hadoop Ecosystem
(Source: [27])

3. AWS Redshift

Amazon Redshift is used for analytics and is a petabyte-scale data warehouse on-demand service provided by Amazon [3]. It is capable of addressing the application of queries on petabytes of data through its Massively Parallel Processing (MPP). Redshift can easily work in coordination with other AWS services and this makes it efficient, especially in real-time analytics [8]

III. IMPLEMENTATION OF SCALABLE REAL-TIME ANALYTICS

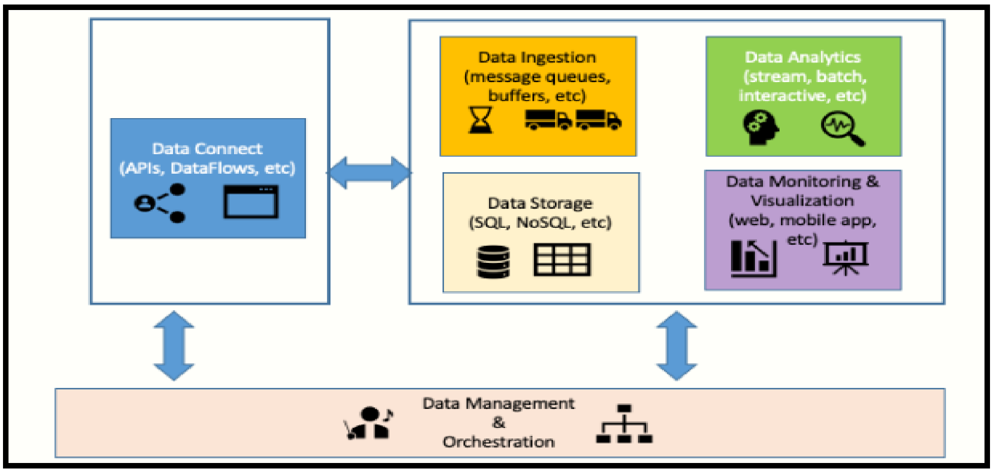

Figure 3: Implementation of Real-Time Data Analytics
(Source: [28])

1. Data Ingestion and Processing

Data ingestion occurs in Kinesis Data Streams as the first step after sources feed data into it. Kinesis Data Firehose follows this data in real-time to Amazon S3 for storage [1]. Apache Spark jobs are executed on AWS EMR that performs ETL operations on the data before it is transferred to Amazon Redshift for analysis [22].

2. Real-Time Data Analytics

Kinesis Data Analytics supports real-time future analysis and lets users issue SQL instructions to the streaming data immediately [6]. It is vital for applications such as real-time recommendation and dynamic pricing models [13].

3. Data Warehousing and Querying

Amazon Redshift is used for storing transformed data for post-processing and high-level analytical queries [9]. Spectrum also enables analytical queries to be made against data stored in S3 meaning it composes well with data lakes [18]. These two combined enable historical and real-time analysis and offer a solution to big data.

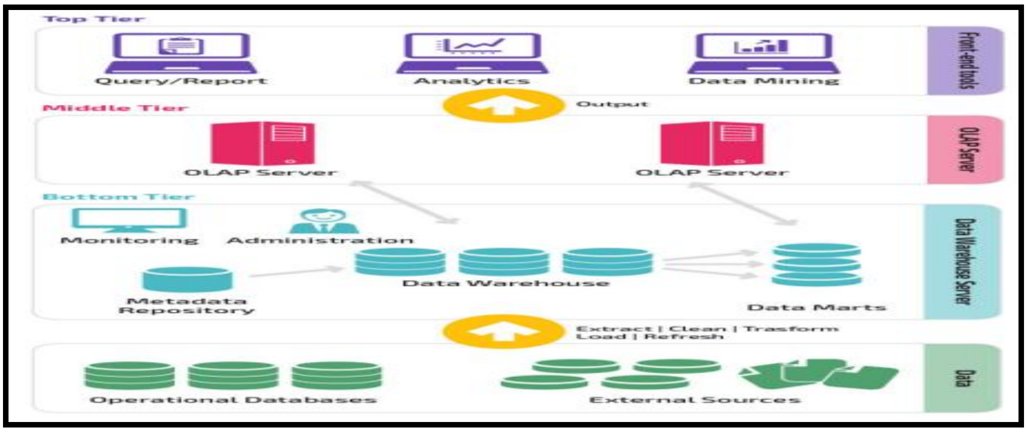

Figure 4: Data Warehousing
(Source: [29])

IV. METHODOLOGY

1. Research Design

This research falls under descriptive research type and includes secondary data collected from 25 peer-reviewed journal articles from 2018 to 2023. Thus, it combines the case studies and theoretical reviews of the advantages of cloud-based services AWS Kinesis, EMR, Redshift in the analysis of large-scale streams of information in real-time for various industries. These comprise of talking about the implementation frameworks, performance indicators and the future development.

2. Data Collection and Analysis

Data collection is achieved through a method of identifying 25 peer-reviewed journal articles from the year 2018-2023 on the implementation and performance of AWS Kinesis, EMR, AND Redshift in real-time analytics. Such analytical methods involve qualitative content analysis to discern the themes and patterns, and comparative analysis to assess the case studies in terms of applied branches as healthcare, financial sector, e-commerce, and others. It also uses cross-reference to amalgamate findings on the direction for the future such as the use of advanced machine learning, integration of digital twin, and the efficient use of energy in data processing [24].

V. RESULTS

The qualitative nature of this study reveals the thematic analysis of the research to describe strategies in the utilization of AWS Kinesis, EMR, and Redshift for real-time analysis. Among the most recognized of these themes are the following: First, scalability – a number of works focused on the specifics of the effective use of Amazon Web Services to analyse large amounts of data are worth mentioning [6]. Another equally important aspect is real-time data processing that is essential for such use cases as fraud detection and continuous patient monitoring [12]. In terms of the future plan, a critical area is the integration of novel artificial intelligence technologies which includes an improvement in machine learning integration, and the use of digital twins to improve the accuracy of analytics of predictive and procedural improvements to the business [13]. Finally, energy efficiency and security that is consistently featured as an issue, to reiterate the importance of reliable and energy efficient data processing in cloud environments [24]. All these themes as shown above indicate the effectiveness of AWS in tackling modern data analytics issues.

VI. DISCUSSION

AWS Kinesis, EMR, and Redshift are products that are used to process high-velocity data in organizations' premises. It integrates the systems in such a manner that real-time data from the IoT devices, social media, and application logs are collected and made available for real-time analysis [16]. EMR optimizes the handling of big data with the help of Hadoop and Spark and is suitable for handling ETL processes, as well as working with actual data [14] Redshift of AWS is suitable for a petabyte-scale data warehouse MPP and designed for efficient querying and integration with other AWS services for advanced analytics [8].

The usage of AWS Kinesis, EMR, and Redshift takes significant effects. In health care it improves outpatient tracking and risk modelling [15]. Thus, the foregoing findings make clear that while financial institutions obtain real-time fraud identification [10], e-commerce gets better immediate recommendation and intelligent, changing prices [15]). Machine learning and digital twins' application improves predictive modelling and asset management [13], thereby promoting sustainability and security [23].

VII. CASE STUDIES

1. Healthcare Industry

In health care, there is optimization for patient tracking in addition to forecasting on the use of AWS Kinesis, EMR, and Redshift. One healthcare provider utilized this architecture for the processing of real-time data originating from patient monitoring gadgets to boost response time and subsequently the patient's condition [1].

2. Financial Services

Financial organisations utilize AWS to perform the functions of fraud detection in real-time and risk evaluation. Such institutions can easily infer fraudulent activities with the risks and contain them when the transaction data are collected in real-time through Kinesis on EMR and stored and analysed using Redshift [10].

3. E-Commerce

Electronics trade organizes the work with data in real-time with the help of AWS services to increase comfort levels for buyers with the help of such parameters as, personalization of recommendations and the dynamics of the price. Kinesis helps in extracting the customer's interaction details and despite EMR analysing these details, Redshift is crucial in presenting the required insights to the marketing and sales branches [15].

4. Transportation Analytics

In the transport subsector, it is in real-time processing for monitoring traffic and scheduling timetables for public transport [7]. The approach of Lambda architecture can be used on the platform AWS with low cost and good performance for transport analytics [5]. For real-time data ingestion, Kinesis is used, for batch processing EMR is used, and for storage and querying the aforementioned system meaningfully tracks bus delays and ETAs Redshift is used.

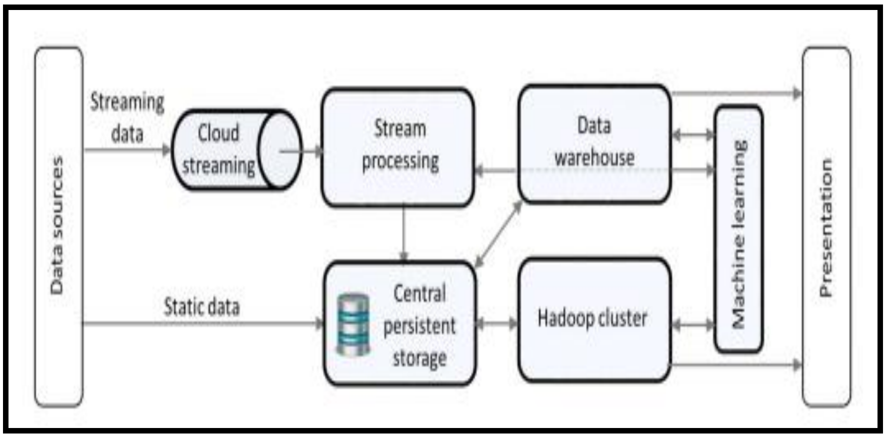

Figure 5: Lambda Architecture
(Source: [30])

5. IoT-Based Livestock Management

The incorporation of the Internet of Things in smart farming is beneficial in measuring and processing data in real-time on the health of the animals and the operations of the business [17]. The pseudo-IoT design developed for targeting a big-scale smart animal farming system was also founded on AWS services [9]. Of these tools Kinesis was used to capture the data from the Sensors in the IoT promptly and process it, EMR was used to analyse the data further and Redshift was used to store processed data.

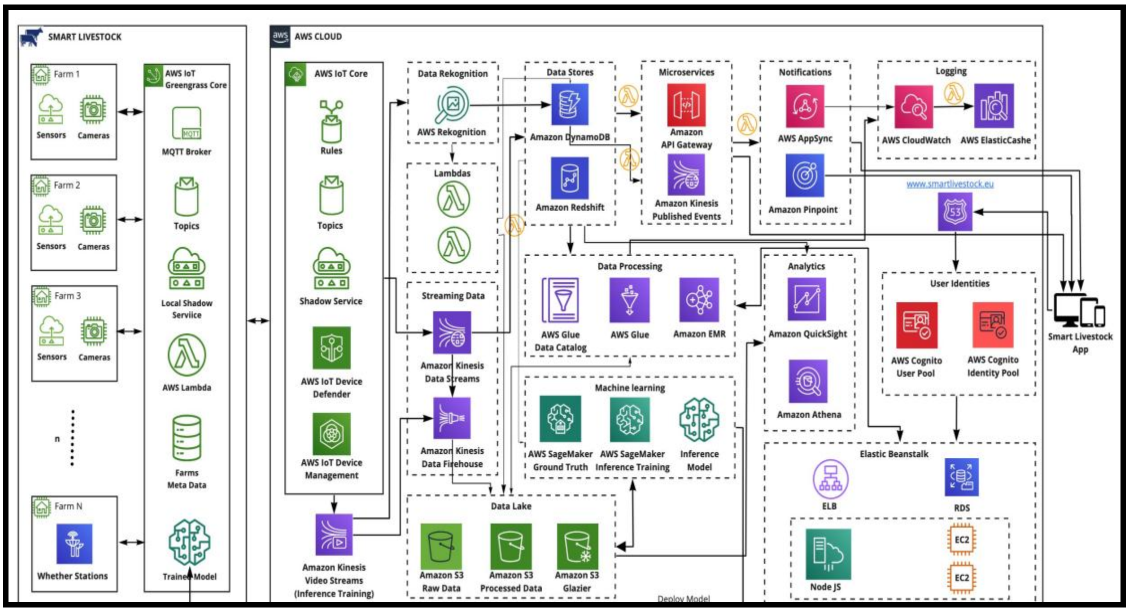

Figure 6: Architecture of Livestock Management
(Source: [31])

VIII. CHALLENGES AND FUTURE DIRECTIONS

1. Challenges

Data Integration: Data integration is also another challenge that may be by sources as well as the format of data collected. Due to the nature of diverse sources and data heterogeneity, the processes of data transformation and normalization are crucial [6].

Latency: Low latency of data processing and querying is important to accommodate real-time queries. There are often latency problems that negatively affect the use of real-time analysis [25].

Scalability: The issues of managing and scaling structures capable of supporting data volume and velocity increases are very challenging [11]. There is a need to have the right approach to the management of resources if the performance is to be sustained [24].

2. Future Directions

Enhanced Machine Learning Integration: Integrating machine learning to development systems for mass consumer markets can significantly enhance real-time analytics that may include predictive and prescriptive ones [23]. Solutions like Sage Maker presented by AWS can be provided to the analyst that can be incorporated into the pipeline for better analysis.

Edge Computing: To avoid long latencies and maximize the response time real-time analytics can be shifted to the edge. In detail, edge computing brings the computational capability closer to where data is created to minimize the time required to transfer data to centralization data centres [22].

Multi-Cloud Architectures: Getting services from multiple cloud providers is also reliable since it minimizes reliance on one provider. Multi-clouds assist an organization in partitioning the computational undertakings on various solutions that are highly proficient, and some of them, cost-efficient [19].

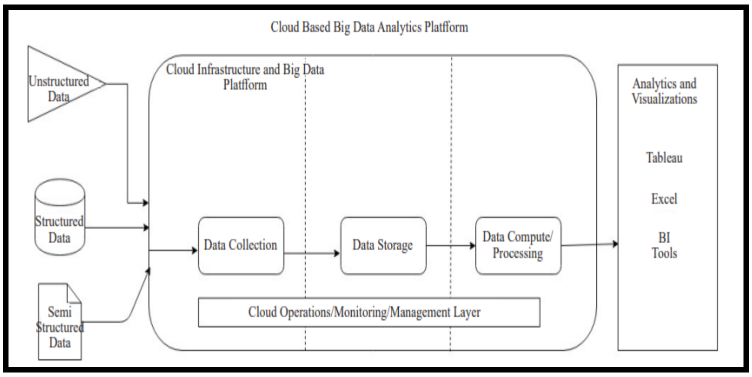
Cloud Based Big Data Analytics Plattform
Unstructured Data
Structured Data
Semi Structured Data
Cloud Infrastructure and Big Data Plattform
Data Collection
Data Storage
Data Compute/ Processing
Cloud Operations/Monitoring/Management Layer
Analytics and Visualizations
Tableau
Excel
BI Tools

Figure 7: Cloud Architecture
(Source: [26])

Real-time data integrations over digital twins: Real-time data analysis combined with the concept of Digital Twins is one of the promising future outcomes among the listed ones. Real-time data can be employed in updating the simulation models of physical systems which can be referred to as digital twins [13]. AWS Solution for real-time data integration with the virtual identity of the physical assets for better management and controls.

It is recommended to integrate advanced machine learning and edge computing with AWS Kinesis, EMR, and Redshift to further enhance real-time analytics capabilities and optimize high-velocity data processing across industries.

IX. CONCLUSION

- The integration of AWS Kinesis, EMR, and Redshift creates a resonant foundation for real-time analysis, which indicates that they have sufficient solutions to the problems of high-velocity data processing.
- This architecture enables the Real-time processing of data and decision-making related to all sectors of the economy as long as it concerns the health sector, finance, retail trade, transportation, smart manufacturing industries, etc.
- The kind of data processing needs as demonstrated in the case studies is evidence of the versatility of AWS as does the future development direction of the subject in which areas that can be enhanced include Digital twin and security, data processing energy efficiency, and Machine learning.
- In the future, with the development of technology, especially, the multi-cloud strategy and integration of advanced analytics, the characteristics of the real-time data processing system will further improve in line with the development of data-oriented environments.

REFERENCES

1. Abdel-Rahman, M. and Younis, F.A., 2022. Developing an Architecture for Scalable Analytics in a Multi-Cloud Environment for Big Data-Driven Applications. International Journal of Business Intelligence and Big Data Analytics, 5(1), pp.66-73.
2. Sharma, R.S., Mannava, P.N. and Wingreen, S.C., 2022. Reverse-engineering the design rules for cloud-based big data platforms. Cloud Computing and Data Science, pp.39-59.
3. Grzegorowski, M., Zdravevski, E., Janusz, A., Lameski, P., Apanowicz, C. and Ślęzak, D., 2021. Cost optimization for big data workloads based on dynamic scheduling and cluster-size tuning. Big Data Research, 25, p.100203.

4. Davoudian, A. and Liu, M., 2020. Big data systems: A software engineering perspective. ACM Computing Surveys (CSUR), 53(5), pp.1-39.
5. Zeydan, E. and Mangues-Bafalluy, J., 2022. Recent advances in data engineering for networking. IEEE Access, 10, pp.34449-34496.
6. Qolomany, B.M.B., 2018. Efficacy of Deep Learning in Support of Smart Services.
7. Cardoso, D.S.D., 2020. Framework for collecting and processing georeferencing data (Master's thesis, Universidade do Porto (Portugal)).
8. Enes, J., Expósito, R.R. and Touriño, J., 2020. Real-time resource scaling platform for big data workloads on serverless environments. Future Generation Computer Systems, 105, pp.361-379.
9. Dineva, K. and Atanasova, T., 2021. Design of scalable IoT architecture based on AWS for smart livestock. Animals, 11(9), p.2697.
10. Wang, X., Guo, P., Li, X., Gangopadhyay, A., Busart, C.E., Freeman, J. and Wang, J., 2023. Reproducible and portable big data analytics in the cloud. IEEE Transactions on Cloud Computing, 11(3), pp.2966-2982.
11. Jannapureddy, R., Vien, Q.T., Shah, P. and Trestian, R., 2019. An auto-scaling framework for analyzing big data in the cloud environment. Applied Sciences, 9(7), p.1417.
12. Mendhe, C.H., Henderson, N., Srivastava, G. and Mago, V., 2020. A scalable platform to collect, store, visualize, and analyze big data in real time. IEEE Transactions on Computational Social Systems, 8(1), pp.260-269.
13. Djonov, M. and Galabov, M., 2020, June. Real-time data integration AWS Infrastructure for Digital Twin. In Proceedings of the 21st International Conference on Computer Systems and Technologies (pp. 223-228).
14. Chen, W., Milosevic, Z., Rabhi, F.A. and Berry, A., 2023. Real-time analytics: Concepts, architectures and ML/AI considerations. IEEE Access.
15. Khan, A., Nawaz, U., Ulhaq, A. and Robinson, R.W., 2020. Real-time plant health assessment via implementing cloud-based scalable transfer learning on AWS DeepLens. Plos one, 15(12), p.e0243243.
16. Kipf, A., Pandey, V., Böttcher, J., Braun, L., Neumann, T. and Kemper, A., 2019. Scalable analytics on fast data. ACM Transactions on Database Systems (TODS), 44(1), pp.1-35.
17. Gupta, U. and Sharma, R., 2023. A Study of Cloud-Based Solution for Data Analytics. In Data Analytics for Internet of Things Infrastructure (pp. 145-161). Cham: Springer Nature Switzerland.
18. He, J., Chen, Y., Fu, T.Z., Long, X., Winslett, M., You, L. and Zhang, Z., 2018, July. Haas: Cloud-based real-time data analytics with heterogeneity-aware scheduling. In 2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS) (pp. 1017-1028). IEEE.
19. Khriji, S., Benbelgacem, Y., Chéour, R., Houssaini, D.E. and Kanoun, O., 2022. Design and implementation of a cloud-based event-driven architecture for real-time data processing in wireless sensor networks. The Journal of Supercomputing, 78(3), pp.3374-3401.
20. Al-Gumaei, K., Müller, A., Weskamp, J.N., Santo Longo, C., Pethig, F. and Windmann, S., 2019, September. Scalable analytics platform for machine learning in smart production systems. In 2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA) (pp. 1155-1162). IEEE.
21. Pérez-Arteaga, P.F., Castellanos, C.C., Castro, H., Correal, D., Guzmán, L.A. and Denneulin, Y., 2018. Cost comparison of lambda architecture implementations for transportation analytics using public cloud software as a service. Special Session on Software Engineering for Service and Cloud Computing, pp.855-862.

22. Kaplunovich, A. and Yesha, Y., 2018, December. Consolidating billions of Taxi rides with AWS EMR and Spark in the Cloud: Tuning, Analytics and Best Practices. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 4501-4507). IEEE.
23. Dzulhikam, D. and Rana, M.E., 2022, March. A critical review of cloud computing environment for big data analytics. In 2022 International Conference on Decision Aid Sciences and Applications (DASA) (pp. 76-81). IEEE.
24. Canizo, M., Conde, A., Charramendieta, S., Minon, R., Cid-Fuentes, R.G. and Onieva, E., 2019. Implementation of a large-scale platform for cyber-physical system real-time monitoring. IEEE Access, 7, pp.52455-52466.
25. Mehmood, E. and Anees, T., 2020. Challenges and solutions for processing real-time big data stream: a systematic literature review. IEEE Access, 8, pp.119123-119143.
26. Sharma, Ravi & Mannava, Purna & Wingreen, Stephen. (2022). Reverse-Engineering the Design Rules for Cloud-Based Big Data Platforms. Cloud Computing and Data Science. 1-21. 10.37256/ccds.3220221213.
27. Turing.com, (2024), Hadoop Ecosystem: Hadoop Tools for Crunching Big Data Problems, Available at: https://www.turing.com/kb/hadoop-ecosystem-and-hadoop-components-for-big-data-problems [Accessed on: 01.08.2024]
28. Fabric.inc, (2024), Data Orchestration is Essential for E-Commerce, Available at: https://fabric.inc/blog/developer/data-orchestration-ecommerce [Accessed on: 01.08.2024]
29. Panoply.io, (2024), Data Warehouse Architecture, Available at: https://panoply.io/data-warehouse-guide/data-warehouse-architecture-traditional-vs-cloud/ [Accessed on: 01.08.2024]
30. Geeksforgeeks.org, (2024), What is Lambda architecture, Available at: https://www.geeksforgeeks.org/what-is-lambda-architecture-system-design/ [Accessed on: 01.08.2024]
31. Dineva, Kristina & Atanasova, Tatiana. (2021). Design of Scalable IoT Architecture Based on AWS for Smart Livestock. Animals. 11. 2697. 10.3390/ani11092697.