International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 6, Issue 3, 60-63, 2025
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V6I3P110

Original Article

Designing a Scalable Data Lake Architecture on AWS Using Glue and S3

Karunakar Grandhe
Data Engineering & Analytics, Product Manager, New Jersey, USA.

Received On: 29/06/2025 ***Revised On: 17/07/2025*** ***Accepted On: 08/08/2025*** ***Published On: 19/09/2025***

Abstract - *Data-intensive enterprises require an efficient, low-cost, scalable architecture to manage both structured and unstructured data, given that they are data-driven business enterprises. There are a variety of Cloud-based services that have changed how companies can manage Big Data. This article highlights the practice of scaling a proposed Data Lake Architecture built on Amazon Web Services (AWS), leveraging Amazon Simple Storage Service (S3) as the primary storage service, and AWS Glue, which includes templates to facilitate data integration and transformation. This study focused on system architecture, structure, and performance scalability of operations.*

Keywords - *Data Lake, AWS Glue, Amazon S3, Cloud Architecture, ETL, Scalability, Big Data.*

1. Introduction

Transactional systems, Internet of Things (IoT) devices, social platforms, and enterprise software are driving unprecedented growth in the heterogeneous datasets generated by modern organizations. Traditional on-premises storage paradigms fail to scale at the levels demanded by large-scale analytics. Therefore, Cloud-based Data Lakes are now becoming core components of enterprise data strategies.

Amazon Web Services (AWS) provides an excellent tool set for building and managing Data Lakes. Most prominently, Amazon S3 provides a scalable, reliable, and low-cost storage solution while AWS Glue offers serverless extract-transform-load (ETL). An S3 and Glue deployment, architected properly, allows enterprises to be able to ingest, catalog, process, and query datasets with minimal overhead.

2. Literature Review

The literature has observed the transformative power of Cloud-native platforms in enabling large-scale data analytics. Object storage-based Data Lakes reduce the complexity of infrastructure and provide nearly limitless scalability and capacity. The addition of serverless ETL services with integrated data discovery enables the dynamic transformation of data and the automatic management of pipelines [1]. Distinctions between data lakes and traditional data warehouses are drawn by contrasting the two data stores on factors such as schema flexibility and capacity for unstructured data. Modern architectures that combine metadata catalogs and storage systems can close the gaps of analysis by providing schema-on-read services [2].

3. Methodology

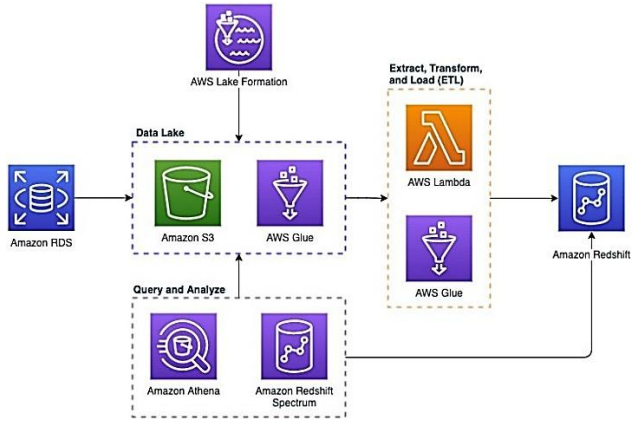

Fig 1: Detailed Architectural Diagram of the AWS Data Lake

It is based on the Design Science methodology centered on the design principles of the implementation work.

The proposed architecture is three-tiered:

3.1. Data Ingestion Layer

All data is ingested into the lake using batch ingestion, streams (along with streaming capabilities), or through direct connection to the SaaS. Raw data is stored in an Amazon S3 bucket, with a hierarchical folder structure that helps identify the raw, curated, and processed zones [3].

3.2. Data Cataloging and Metadata Management

AWs Glue Crawlers can periodically pull S3 data, infer the schema, and create metadata records in the Glue Data Catalog. The Glue Data Catalog is the metadata hub that provides interoperability with Athena, Redshift Spectrum, and EMR.

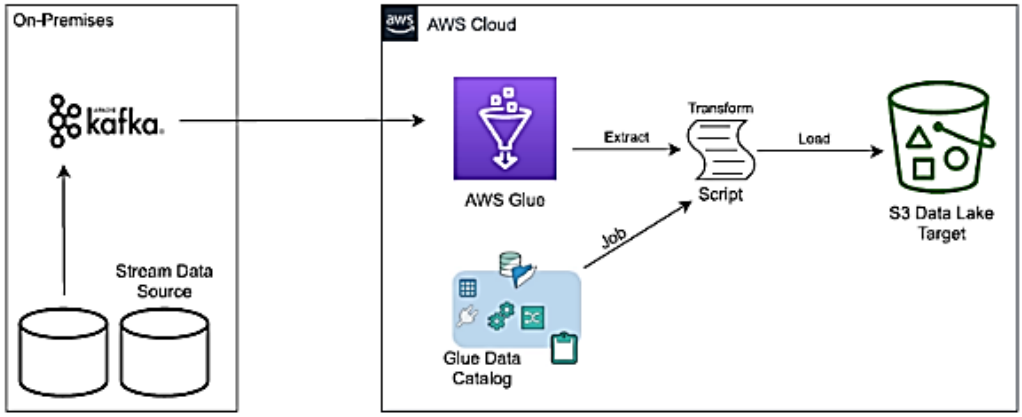

Fig 2: ETL Process Overview

3.3. ETL and Transformation Layer

Glue ETL jobs are PySpark or Scala programs used to convert raw data into out-of-the-box formats (Parquet or ORC), suitable for analytics input [4]. Partitioning is used to enhance query performance through a schema evolution mechanism that dynamically manages changing data structures.

3.4. Data Governance and Security

In alignment with the business requirements of the enterprise security requirements, Identity and Access Management (IAM) roles, encryption (SSE-S3 and SSE-KMS) and fine-grained access policies are offered. Additionally, AWS Lake Formation extends governance with access control and audit logging.

3.5. Query and Analytics Layer

Amazon Athena enables querying of the curated datasets with no server, and where workloads are more complex, Amazon Redshift Spectrum allows federated query across the data lake and data warehouse [5]. The result is that this tiered solution is modular, scalable, and cost-effective. Significantly, it also improves system reliability while providing flexibility and maintaining the endurance and maintainability of enterprise-class data management.

4. Discussion and Analysis

4.1. Scalability Considerations

Amazon S3 has a natural horizontal scale with a virtually unlimited scale of objects. The serverless execution model of Glue involves the automatic provisioning of resources, rather than manual control of infrastructure. Together, these two features enable linear scalability with volume growth.

4.2. Performance Optimization

Partitioning datasets in S3 significantly improves query efficiency by reducing data scans. Storing files in columnar formats such as Parquet minimizes I/O operations [6]. Glue job bookmarking avoids reprocessing previously ingested data, optimizing costs.

4.3. Cost Efficiency

S3 provides tiered storage classes (Standard, Infrequent Access, and Glacier), allowing organizations to optimize costs according to their data access patterns. Glue's pay-per-use model further supports cost efficiency by charging only for actual compute consumed.

4.4. Governance Challenges

Without proper metadata and governance, Data Lakes risk becoming "data swamps." Glue Data Catalog and Lake Formation mitigate this by ensuring consistent schema management, access controls, and data lineage tracking [7].

4.5. Comparison with Alternatives

Compared to Azure Data Lake and Google Cloud Storage, AWS offers seamless integration between S3, Glue, and Athena, creating a more unified ecosystem. However, cross-region latency and dependency on AWS-specific services remain potential drawbacks.

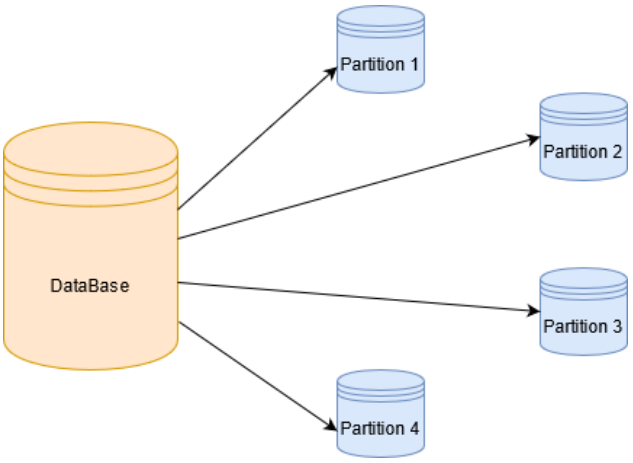

Fig 3: Data Partitioning Strategies for Scalability

5. Results

The proposed architecture demonstrates the following measurable outcomes, reflecting its ability to deliver scalable performance, operational efficiency, governance, and cost optimization across diverse enterprise data environments:

5.1. Elastic Scalability

Seamless ingestion and storage of petabyte-scale data with no infrastructure bottlenecks, enabling organizations to manage unpredictable workloads without the need for manual scaling or additional provisioning.

5.2. Operational Efficiency

Serverless Glue jobs automate schema discovery and transformation, reducing developer workload and accelerating data pipeline deployment, which in turn improves team productivity and overall delivery timelines [8].

5.3. Improved Query Performance

Optimized partitioning and columnar storage formats significantly reduce query response time by up to 70% compared with raw CSV queries, enhancing user experience and supporting near-real-time analytics [9].

5.4. Enhanced Governance

Centralized metadata management and fine-grained access controls enhance data discoverability and compliance, increasing the auditability of these systems and aligning them more closely with enterprise security and regulatory requirements.

5.5. Cost Reduction

Pay-as-you-drive ETLs and in-house performance may help reduce operational costs compared to mainstream systems. The functionality will enable the business to save capital even without access to data, and data integrity will not be compromised [10].

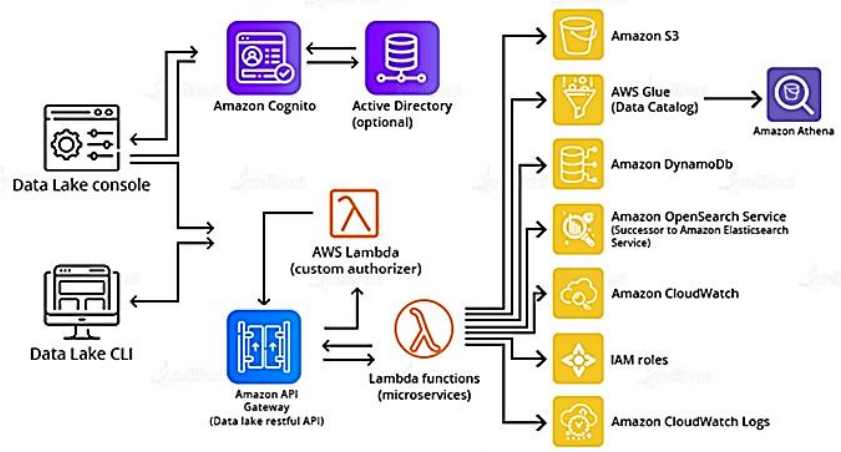

Fig 4: AWS Data Lake Formation and Best Practices

6. Conclusion

A literate, scalable Data Lake design, as discussed above, contained the following components: Amazon Web Services (AWS), which supported high-performance data transfers, and data combining and analytics software that ingested large amounts of data. The architecture ensured the delivery of performance, governance, and cost efficiency to maximize the benefits of scale by offering unlimited maximum storage in free serverless ETL executions.

Research can be expanded to include Machine Learning and Amazon SageMaker, and applied as a hypothesis to real-time analytics across multiple Clouds. Overall, AWS Glue and S3 provided a significant platform for establishing connections with organizations that want to implement an unpublished Data

Lake design, which should be cost-effective, scalable, and efficient.

References

[1] E. Zagan and M. Danubianu, "Data Lake Architecture for Storing and Transforming Web Server Access Log Files," *IEEE Access*, vol. 11, pp. 40916–40929, 2023, doi: https://doi.org/10.1109/access.2023.3270368.

[2] D. Jain, "Lakehouse: A Unified Data Architecture," *International Journal for Research in Applied Science and Engineering Technology*, vol. 9, no. 3, pp. 881–887, Mar. 2021, doi: https://doi.org/10.22214/ijraset.2021.33376.

[3] P. Wieder and H. Nolte, "Toward data lakes as central building blocks for data management and analysis," *Frontiers in Big Data*, vol. 5, Aug. 2022, doi: https://doi.org/10.3389/fdata.2022.945720.

[4] Zahra Shojaee Rad and Mostafa Ghobaei-Arani, "Data pipeline approaches in serverless computing: a taxonomy, review, and research trends," *Journal of big data*, vol. 11, no. 1, Jun. 2024, doi: https://doi.org/10.1186/s40537-024-00939-0.

[5] M. Saxena *et al.*, "The Story of AWS Glue," *Proceedings of the VLDB Endowment*, vol. 16, no. 12, pp. 3557–3569, Aug. 2023, doi: https://doi.org/10.14778/3611540.3611547.

[6] A. Nambiar and D. Mundra, "An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management," *Big Data and Cognitive Computing*, vol. 6, no. 4, p. 132, Nov. 2022, Available: https://www.mdpi.com/2504-2289/6/4/132

[7] S. Genovese, "Data Mesh: the newest paradigm shift for a distributed architecture in the data world and its application - Webthesis," *Polito.it*, Oct. 2021, doi: https://webthesis.biblio.polito.it/secure/20415/1/tesi.pdf.

[8] S. Worlikar, "Real-Time Patient Monitoring and Alerting in Hospitals Using AWS Lake House Architecture," *Frontiers in Emerging Computer Science and Information Technology*, vol. 02, no. 08, pp. 07-14, Aug. 2025, doi: https://doi.org/10.37547/fecsit/volume02issue08-02

[9] J. E. Ike, J. D. Kessie, H. E. Okaro, E. Ezeife, and T. Onibokun, "Identity and Access Management in Cloud Storage: A Comprehensive Guide," *International Journal of Multidisciplinary Research and Growth Evaluation.*, vol. 6, no. 2, pp. 245–252, 2025, doi: https://doi.org/10.54660/.ijmrge.2025.6.2.245-252.

[10] P. Badri, A. K. R. Goli, and S. R. Goli, "Strengthening Data Governance and Privacy: Utilizing Amazon AWS Cloud Solutions for Optimal Results," *SSRN Electronic Journal*, 2025, doi: https://doi.org/10.2139/ssrn.5320361.