IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SERVERLESS DATA LAKE FOR REAL-TIME FINANCIAL ANALYTICS USING AWS

[1]Velagapudi Drakshayani, [2]K S Venkata Kushal, [3]M Teja Vardhan, [4]T Sri Saranya Reddy

[1]UG Scholar, [2]UG Scholar, [3]UG Scholar, [4]UG Scholar

[1]Department of CSE,

[1]Koneru Lakshmaiah Education Foundation, Guntur, India

Abstract: The rapid growth of financial data has created a strong demand for scalable and efficient systems for data storage, processing, and analysis. Traditional data processing approaches often rely on complex infrastructure, leading to high operational costs and limited scalability. To address these challenges, this paper proposes a serverless data lake architecture for real-time financial analytics using cloud-based services.

The proposed system leverages Amazon S3 for scalable data storage, AWS Glue for automated schema discovery and data cataloging, Amazon Athena for serverless querying, and Amazon QuickSight for interactive data visualization. By eliminating the need for infrastructure management, the architecture simplifies deployment while ensuring high scalability, flexibility, and cost efficiency.

A financial dataset is utilized to evaluate the system's performance and analytical capabilities, including trend analysis, volume analysis, and price comparison. The results demonstrate that serverless data lake architectures provide an effective and practical solution for modern financial analytics, enabling faster insights with reduced operational overhead. The architecture also supports dynamic scaling based on workload demands, making it suitable for varying data volumes.

Keywords: Serverless Architecture, Data Lake, Financial Analytics, Cloud Computing, AWS, Data Visualization.

I. INTRODUCTION

With the rapid growth of financial markets and digital transactions, an enormous volume of data is generated continuously from sources such as stock exchanges, trading platforms, and financial institutions. Efficient processing and analysis of this data are critical for deriving insights, detecting trends, and supporting decision-making processes. Traditional data processing systems primarily rely on on-premise infrastructure, which often involves high setup costs, limited scalability, and significant maintenance efforts. As data volume and velocity increase, these systems struggle to meet performance and flexibility requirements.

Cloud computing has emerged as a transformative solution by providing scalable and on-demand resources over the internet. In particular, serverless architecture has gained attention due to its ability to eliminate the need for infrastructure management. In a serverless model, computing resources are automatically provisioned and scaled based on workload demands, allowing developers to focus solely on application logic. This paradigm not only reduces operational complexity but also improves cost efficiency through a pay-as-you-go model.

In the context of big data analytics, serverless data lake architectures have become increasingly popular. A data lake enables the storage of large volumes of structured, semi-structured, and

unstructured data in its native format, supporting diverse analytical workloads. By integrating serverless computing with cloud storage and query services, it is possible to build flexible and efficient data processing pipelines without the need for dedicated infrastructure.

This paper focuses on the design and implementation of a serverless data lake architecture for financial analytics. The proposed system facilitates efficient data ingestion, storage, cataloging, and querying of financial datasets. It aims to improve scalability, reduce operational overhead, and enable faster analytical processing compared to traditional approaches.

II. Literature Survey

Previous research in big data analytics has extensively utilized distributed processing frameworks such as Hadoop and Spark. These frameworks enable parallel processing of large-scale datasets across clustered environments, significantly improving computational efficiency. However, they often require substantial infrastructure setup, cluster management, and continuous monitoring, which increases operational complexity and cost. Additionally, performance tuning and resource allocation in such systems demand specialized expertise.

With the evolution of cloud computing, researchers have explored managed and serverless architectures to overcome these limitations. Serverless computing platforms allow developers to execute code without provisioning or managing servers, thereby simplifying deployment and scaling. Services such as function-as-a-service (FaaS) automatically allocate resources based on workload demand, improving efficiency and reducing idle resource costs.

Recent studies have introduced the concept of serverless data lakes, where large volumes of structured, semi-structured, and unstructured data are stored in cloud object storage systems. These architectures integrate services for data ingestion, cataloging, and querying, enabling users to perform analytics directly on stored data without complex ETL pipelines. Metadata management tools and automated schema inference techniques further enhance data accessibility and usability.

Compared to traditional big data systems, serverless data lake architectures offer improved scalability, flexibility, and cost-effectiveness. They support pay-as-you-go pricing models and eliminate the need for dedicated infrastructure maintenance. Despite these advantages, several challenges persist, including query latency, cold start overheads, and efficient metadata handling for large datasets.

Therefore, there is a need for optimized solutions that leverage serverless technologies while addressing these performance and management challenges. This work focuses on designing an efficient serverless data.

III. Methodology

A. System Architecture

The proposed system is designed using a serverless architecture that integrates multiple cloud-based services to enable efficient financial data processing and analytics. The architecture follows a modular pipeline approach, consisting of data ingestion, storage, processing, and visualization stages.

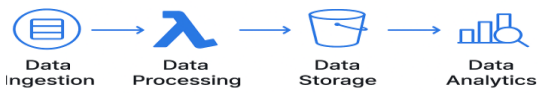
Amazon S3 serves as the primary data lake, providing highly scalable and durable storage for financial datasets in various formats such as CSV and JSON. Data is ingested into S3 from external sources and organized into structured buckets for efficient access and management.

AWS Glue is employed for data cataloging and schema management. Glue crawlers automatically scan the stored datasets, infer their structure, and generate metadata tables in the data catalog. This eliminates the need for manual schema definition and ensures consistency across the system.

For query processing, Amazon Athena is utilized to perform serverless, SQL-based queries directly on the data stored in S3. This approach removes the need for data movement or dedicated database systems, thereby reducing latency and operational overhead.

Finally, Amazon QuickSight is used to create interactive dashboards and visualizations. It connects directly to Athena, allowing users to perform real-time analysis and generate insights such as trend analysis, volume patterns, and price comparisons.

The overall architecture ensures scalability, cost efficiency, and ease of deployment by eliminating infrastructure management while supporting large-scale financial data analytics.



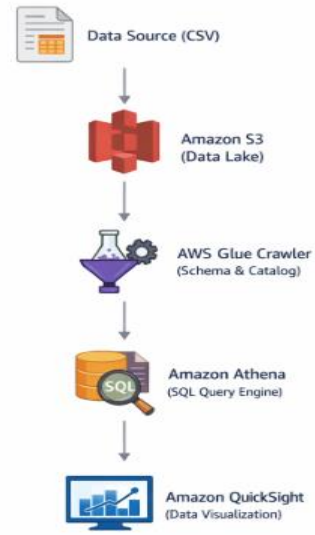


Fig. 1: Serverless Data Lake Architecture

B. Workflow

The system workflow consists of multiple stages that enable efficient processing and analysis of financial data in a serverless environment. The workflow begins with the ingestion of financial datasets in CSV format, which are uploaded to Amazon S3. S3 acts as the central data lake, providing scalable and durable storage for raw data.

Once the data is stored, AWS Glue is used for data preprocessing and cataloging. A Glue crawler automatically scans the dataset, identifies its structure, and generates a corresponding schema in the data catalog. This process ensures that the data is organized and readily accessible for querying without manual intervention.

In the next stage, Amazon Athena is utilized to perform SQL-based queries directly on the data stored in S3. This enables efficient data analysis without requiring data movement or dedicated database infrastructure. Users can execute complex queries to extract meaningful insights such as trends, patterns, and comparisons.

Finally, Amazon QuickSight is used for data visualization. It connects to Athena to generate interactive dashboards and graphical representations of the analyzed data. These visualizations help users interpret financial trends, volume distributions, and price variations effectively.

The overall workflow ensures a seamless, automated, and scalable data processing pipeline, minimizing operational complexity while enabling real-time analytics.

I. IMPLEMENTATION

The implementation of the proposed system was carried out using a serverless architecture provided by Amazon Web Services. The objective was to build a scalable and efficient pipeline for financial data analysis without managing any physical infrastructure.

A. Data Upload and Storage

The financial dataset in CSV format was initially uploaded to Amazon S3, which serves as the data lake. The storage layer provides high durability and scalability, allowing large volumes of structured data to be stored efficiently.

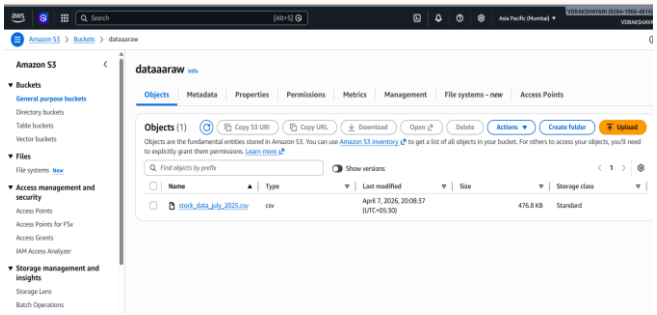


Fig. 1: Dataset in S3

B. Data Cataloging using Glue

To enable querying, the dataset was processed using AWS Glue. A crawler was configured to scan the data stored in S3 and automatically infer the schema. The crawler generated a metadata table in the data catalog, making the dataset ready for querying.

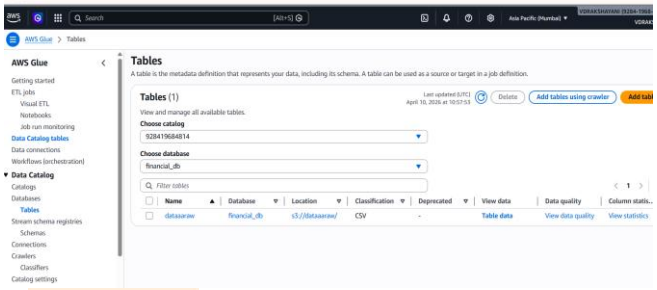


Fig. 2: Glue crawler

C. Query Execution using Athena

After cataloging, Amazon Athena was used to perform SQL-based queries directly on the data stored in S3. This eliminated the need for data movement and enabled fast and efficient analysis. Queries such as data retrieval and aggregation were executed to validate the dataset.

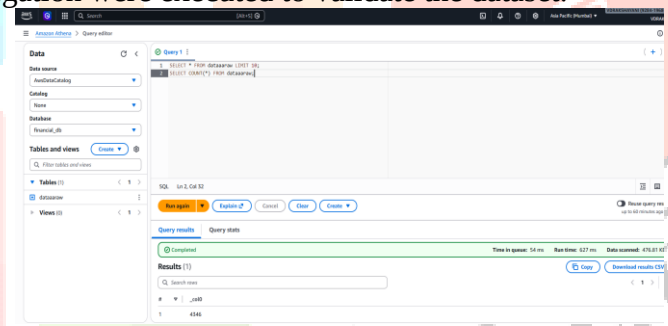


Fig. 3: Athena Query

D. Data Visualization using QuickSight

For visual analysis, Amazon QuickSight was integrated with Athena. The dataset was imported into QuickSight and used to create interactive dashboards. Various visualizations such as line charts, bar charts, and KPI indicators were developed to analyze trends, trading volume, and price variations.

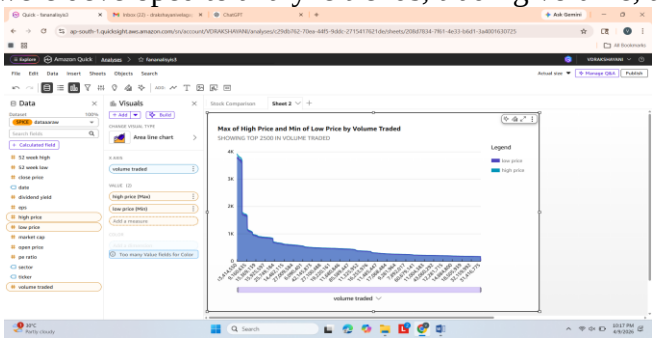


Fig. 4: QuickSight Dashboard

I. Results And Analysis

The dataset contains attributes such as date, open price, high price, low price, close price, and volume. Several visualizations were created:

A. Price Trend Analysis

A line chart was used to visualize the closing price over time, showing fluctuations and trends in the stock market.

B. Volume Analysis

A bar chart was used to analyze trading volume across different dates, indicating market activity levels.

C. High vs Low Comparison

A line chart comparing high and low prices provided insights into market volatility.

D. Key Metrics

A KPI visualization was used to calculate average closing price, providing a quick summary of the dataset.

These visualizations demonstrate how financial data can be analyzed effectively using serverless tools.

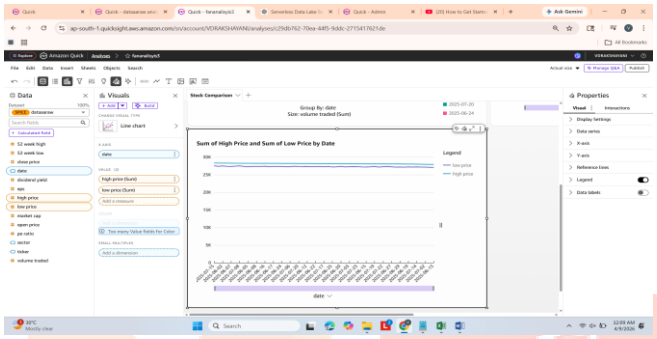


Fig. 5: Comparison of High and Low Prices over Time

II. Conclusion

This paper presented a serverless data lake architecture for financial analytics using cloud-based services. The proposed system successfully demonstrated how financial data can be stored, processed, and visualized without the need for dedicated infrastructure. By integrating Amazon S3, AWS Glue, Amazon Athena, and Amazon QuickSight, the architecture enables efficient data handling and real-time analytical capabilities.

The experimental results, supported by multiple visualizations, highlight trends, trading activity, and price fluctuations within the dataset. The system proved to be scalable, flexible, and cost-efficient, making it suitable for modern financial data analysis applications.

Overall, the study confirms that serverless architectures provide a practical and effective solution for large-scale financial analytics. Future enhancements may include real-time data streaming and predictive analytics to further improve decision-making capabilities.

III. References

[1] Amazon Web Services, *"Overview of Amazon Web Services,"* 2023. [Online]. Available: https://aws.amazon.com/what-is-aws/
Accessed: Apr. 2026.

[2] Amazon S3 Documentation, *"Amazon Simple Storage Service (S3) – Scalable Object Storage in the Cloud,"* Amazon Web Services, 2023. [Online]. Available: https://docs.aws.amazon.com/s3/
Accessed: Apr. 2026.

[3] AWS Glue Documentation, *"AWS Glue Developer Guide: Data Catalog and ETL Services,"* Amazon Web Services, 2023. [Online]. Available: https://docs.aws.amazon.com/glue/
Accessed: Apr. 2026.

[4] Amazon Athena Documentation, *"Amazon Athena: Interactive Query Service,"* Amazon Web Services, 2023. [Online]. Available: https://docs.aws.amazon.com/athena/ Accessed: Apr. 2026.

[5] Amazon QuickSight Documentation, *"Amazon QuickSight User Guide: Business Intelligence and Data Visualization,"* Amazon Web Services, 2023. [Online]. Available: https://docs.aws.amazon.com/quicksight/ Accessed: Apr. 2026.

[6] Kaggle, *"Stock Market Dataset,"* [Online]. Available: https://www.kaggle.com/ Accessed: Apr. 2026.

[7] J. Dean and S. Ghemawat, *"MapReduce: Simplified Data Processing on Large Clusters,"* IEEE Symposium on Operating Systems Design and Implementation, 2004.

[8] T. White, *"Hadoop: The Definitive Guide,"* 4th ed., O'Reilly Media, 2015.

[9] M. Zaharia et al., *"Apache Spark: A Unified Engine for Big Data Processing,"* Communications of the ACM, vol. 59, no. 11, pp. 56–65, 2016.

[10] P. Mell and T. Grance, *"The NIST Definition of Cloud Computing,"* National Institute of Standards and Technology (NIST), 2011.