International Journal Science and Technology
IJST Vol. 2 No. 3 | November 2023| ISSN: 2828-7223 (print), ISSN: 2828-7045 (online), Page 136-146

From Raw Data to Personalized Advice: An Agentic AI Framework on AWS Lambda for Real-Time Financial Planning

Gunjan Kumar
Independent Researcher, India

Article History
Received : October 13, 2023
Revised : October 20, 2023
Accepted : November 28, 2023
Published : November 30, 2023

Corresponding author*:
gunjankumar.email@gmail.com

Cite This Article:
Gunjan Kumar. (2023). From Raw Data to Personalized Advice: An Agentic AI Framework on AWS Lambda for Real-Time Financial Planning. *International Journal Science and Technology*, *2*(3), 136–146.

DOI:
https://doi.org/10.56127/ijst.v2i3.2331

Abstract: The increasing need of personalized and real-time financial planning has shown gaps in traditional advisory systems which rely more on rule-based automation and as a rule using mostly fixed form models. The paper introduces a new serverless design that builds upon agentic artificial intelligence (AI) that is orchestrated using Amazon Web Services (AWS) Lambda to offer dynamic, cost-effective, and scalable financial insights. The proposed framework enables quick access and contextualization of customer-specific financial data by combining Amazon Bedrock as contextual language understanding and distributed data storage providers, AWS S3 and DynamoDB. Amazon EventBridge allows the seamless end-to-end orchestration, allowing the transformation of raw financial information into actionable advice with a low latency. This paper does not only show the technical feasibility of the approach, but also its economic and societal implications, such as scalability, cost optimisation, and democratization of expert-level financial planning. The framework in addition to operational efficiency emphasizes on a customer focused design to promote accessibility and confidence in financial decision making. The findings suggest that the developed system is not just another example of a chatbot system, but it provides adaptive and contextually sensitive financial guidance and, therefore, the proposed system is an impressive improvement in the financial technology infrastructures.

Keyword: Serverless Architecture, Amazon Bedrock, Agentic AI, AWS Lambda, and Event-Driven Systems, Financial Planning, Real-Time Data, and Personalized Advice.

INTRODUCTION

Background and Motivation

Consumers are becoming more interested in real time and personalized financial advice in the present financial landscape. The old financial advisory models, based on traditional human advisors or rigid software platforms, can hardly cope with the size, urgency, and customization demands of the modern economies of data (Stephenson, 2018; Siebel, 2019). Finance decision-making requires the combination of various sources of both structured and unstructured information, risk profiling, and the creation of context-dependent recommendations in the client objectives context. However, the current advisory systems are often characterised by the lack of dynamic workflows, high cost, or poor scaling (Nwaimo et al., 2019).

The agentic artificial intelligence (AI) - systems that can reason autonomously and generate goal-oriented advice and contextual adaptation - provides a paradigm shift in providing financial advice. In contrast to the use of simple chatbots that use a set of pre-written replies, agentic AI systems use dynamic data pipelines, advanced reasoning, and adaptive orchestration to generate actionable and customer-specific replies (Goertzel et al., 2023). With deployed agentic AI, the flexibility of distributed computing, data-driven orchestration, and scalable storage can be used to leverage responsive and cost-effective applications when deployed on cloud-native infrastructures (Lakarasu, 2022; Marinescu, 2022).

In the ultimate display of this suggestion, serverless computing, and specifically AWS Lambda, does not require infrastructure management overhead and instead concentrates on execution scalability, event responsiveness, and operational efficiency (Dubey et al., 2023). With cloud-native frameworks proving to be the game changers in providing real time insights in industries as diverse as healthcare and agriculture (Elghoul et al., 2023; Zamlynskyi et al., 2023), the financial services industry is in a good position to gain a lot through such a construct. This study presents an end-to-end serverless architecture that will combine AI-based reasoning and distributed financial information to support real-time and personalized financial planning.

Problem Statement

Although financial technologies have made progress, there is always a gap between the non-dynamics of advisory tools and the dynamics of a personalized planning system. The majority of modern systems are based on either human intuition, thus restricting their ability to scale their application and cost-effectiveness, or use rule-based artificial intelligence chatbots that cannot reason in a complex context (Eboseremen et al., 2022). Moreover, conventional monolithic architecture hinders the scalability and increases operational overhead especially during variable workloads (Isah et al., 2019).

To plan finances in real-time, a system is required that is able to:

1. Data assimilation of heterogeneous data sources in a safe and efficient manner.
2. Scaling up contextual reasoning.
3. Providing adaptive and customer-specific recommendations in a very short amount of time.
4. These requirements reveal the weakness of the current solutions and demonstrate the need to address the issue of a serverless, agentic AI system coordinated with the help of AWS Lambda.

Research Objectives

The main goal of this study is to model and test a scalable and cost-efficient and customer-centric serverless architecture of real-time financial planning. In particular, the research objectives are:

1. Suggest an agentic AI system which is built on AWS Lambda completely.
2. Add Amazon Bedrock to natural language understanding and response generation.
3. Easily orchestrate financial information in AWS S3, Amazon DynamoDB, and other relevant services with Amazon EventBridge.
4. Compare system performance based on cost, scalability and latency with traditional architectures.
5. Remind of the customer-centered consequences of democratising financial advice using AI.

Contributions of the Study

1. The present research can add to the growing body of research on cloud-native AI as it:
2. Suggesting a new serverless system that illustrates how raw financial information is converted into individualized advice within a real-time manner.
3. Emphasising the way AWS Lambda and other related services can be used to optimise costs and be scaled at the same time.
4. Presentation of empirical assessment of performance and relevance of the system to customers.
5. Providing a conceptual model of democratizing financial advice by being consistent with the global trends of financial inclusion and AI ethics (Van den Heuvel et al., 2023; Mazzoni and Costa, 2022).

LITERATURE REVIEW

Overview of Agentic AI

The term agentic artificial intelligence (AI) is used to describe computational mechanisms that can autonomously execute with respect to the realization of specified goals and change their behavior with respect to current circumstances. Instead of the traditional chatbots or virtual assistants, which are mostly based on a set of fixed rules or attempt to map the intent to certain intents, agentic AI systems use explicit reasoning, situational awareness, and task coordination to achieve the complex goals (Goertzel et al., 2023). Such systems combine perception, planning and action in an integrated system that makes them especially appropriate in the decision making environment which requires personalization, flexibility and real time reactions.

The use of agentic AI in the financial field is particularly relevant. The financial choices are often made based on the rapidly changing market forces, personalized objectives and varying risk-taking abilities.

Dynamic alignment of advice by agentic AI can serve to provide customers with useful information that is both contextually and temporally relevant to the parameters (Eboseremen et al., 2022). Therefore, agentic AI can go beyond the constraints of rule-led advisory systems, and open the door to real-time financial planning technologies that can democratize expert advice.

Serverless AI in the AI Application

Serverless computing has become a paradigm that eases the deployment of applications by decoupling developers of the infrastructure management. Cloud computing services like AWS Lambda allow developers to run their code based on events with no provisioning or maintenance of physical computing and provide cost-efficiency and auto-scale (Marinescu, 2022; Dubey et al., 2023). AI-based applications are a particularly good fit with serverless architecture, and their workload is often dynamic and spiky.

Empirical research is showing that the combination of serverless models and AI workflows can decrease the latency and improve the scalability of various industries, including healthcare, logistics, predictive maintenance, and others (Elghoul et al., 2023; Appiah et al., 2022). Financially, serverless computing can solve two main issues: (1) the necessity of real-time responsiveness in handling large amounts of customer requests, and (2) the financial necessity of reducing the cost of infrastructure at the expense of maintaining steady availability (Abbasi, 2020). **Table 1** below provides a comparative summary of traditional monolithic architectures and serverless approaches in financial applications.

Table 1. Comparison of traditional vs. serverless architectures in financial applications

Architecture	Scalability	Cost Model	Latency	Maintenance
Traditional Monolithic	Limited; requires manual provisioning	Fixed infrastructure costs	Higher due to centralized workloads	Complex; high DevOps overhead
Serverless (AWS Lambda)	Automatic; scales with demand	Pay-per-use model	Lower; event-driven execution	Minimal; managed by provider

Source: Adapted from Marinescu (2022) and Dubey et al. (2023).

This comparative analysis shows how the serverless structures can support dynamic scaling and cost-on-need which is needed by real-time financial planning systems capable of adapting to heterogeneous workloads.

Large Language Models in Finance

Large Language Models (LLM) models, such as the products of Amazon Bedrock, have greatly improved the ability of artificial intelligence systems to reason on context and produce natural language answers. Unlike the previous natural language processing systems, LLMs provide a fine-grained representation that allows the systems to respond to complex financial questions and provide advice based on the unique profile of a particular customer (Webber and Olgiati, 2023).

In the financial industry, the use of LLMs is increasingly applied to risk intelligence (Paleti, 2023), detection of frauds (Nwaimo et al., 2019) and pipelines of customer personalization (Eboseremen et al., 2022). However, the use of LLMs in real-world applications requires careful integration of the external sources of data since the raw outputs of the model are often not specific when it is not grounded in user-specific or domain-specific contexts. The conceptual role of Amazon Bedrock in a financial advisory pipeline is shown in the figure below (Figure1).

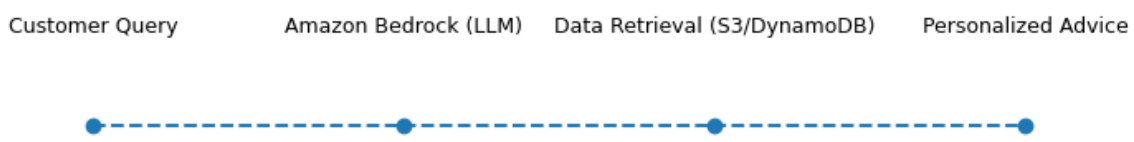

Figure 1. Conceptual flow of LLM integration in financial planning
Source: Adapted from Webber and Olgiati (2023) and Paleti (2023).

The main focus of the number 1 position on large language models (LLM) as reasoning engines, but the point also shows that retrieving customer-specific advice on finances requires access to structured repositories, like Amazon S3 and DynamoDB.

Customer Centric Design and Financial Planning Tools

The outdated financial planning instruments tend to be based on fixed models, risk rating, and broad investment plans. Such systems are practical, but they do not take into consideration actual market conditions and specific circumstances of customers (Stephenson, 2018). According to recent developments in AI-based personalization pipelines, it is plausible that adaptive financial systems can play a major role in customer retention and customer satisfaction (Eboseramen et al., 2022; Ferrua, 2023).

To address the customer-centric approach, dynamic data, interactive interface, and trust mechanisms should be incorporated into financial platforms (Van den Heuvel et al., 2023). What is more, financial AI-powered applications should be harmonized with regulatory requirements and ethics, as the advice should not take advantage of the weaknesses of the customers (Mazzoni and Costa, 2022).

A comparison between the conventional financial planning tools and AI-based customized methods is presented in Table 2 below.

Table 2: Traditional vs. AI-driven financial planning tools

Feature	Traditional Tools	AI-driven Tools
Data Inputs	Static datasets, manual inputs	Dynamic market/customer data
Personalization	Generic advice	Highly personalized
Scalability	Limited to consultant availability	Cloud-based scalability
Real-time Adaptation	Minimal or delayed	Immediate adaptation

Source: Adapted from Stephenson (2018) and Eboseremen et al. (2022)

Research Gap

Current literature provides helpful information on the use of artificial intelligence in the financial sector; however, there are still many gaps. Past studies examined the concept of personalization pipelines (Eboseromen et al., 2022), risk intelligence data engineering (Paleti, 2023), and deployments to clouds (Ferrua, 2023; Appiah et al., 2022). Nevertheless, there is little literature on frameworks that combine agentic artificial intelligence, large language models and serverless architectures in financial planning. Figure 2 is a summary of the gap identified by comparing the focus areas of the literature with the proposed framework.

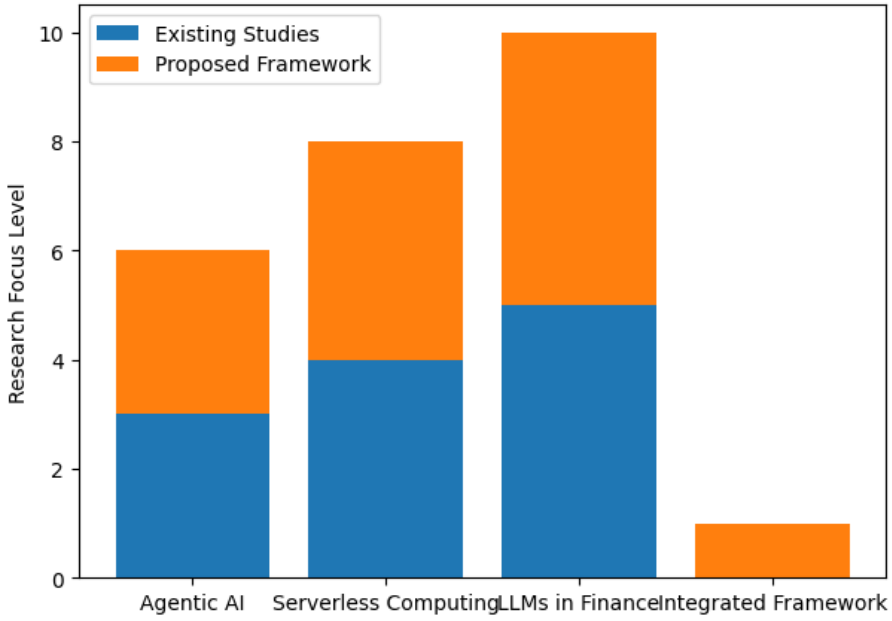

Figure 2: Financial AI literature research gap.
Source: Adapted after Eboseremen et al. (2022), Paleti (2023), and Ferrua (2023).

The visualisation reveals that even though individual elements of agentic AI and serverless computing and LLMs have been studied independently, there is still a lack of cohesive research on these fields together within the realm of financial planning.

METHODOLOGY

System Architecture Overview

The suggested system will embrace a serverless design where AWS Lambda will play the role of a coordinator, which will create a scalable, event-driven environment, and remove the need to provision servers. This architecture meets current standards of cloud-computing technologies when event triggers maintain responsiveness to rapidly changing workloads at scale without sacrificing cost-efficiency (Marinescu, 2022; Abbasi, 2020). It is built upon Amazon Bedrock as the most critical reasoning engine that can provide the understanding of contextual language and generate personalised recommendations, but AWS S3 and DynamoDB provide the structured storage and retrieval of financial and user data (Dubey et al., 2023). The main benefit of this architecture is that it is modular. All AWS Lambda functions are assigned a specific task, such as data retrieval, query preprocessing, advice generation, or response formatting. Such autonomous but interrelated functions make things resilient; system components can thus be modified or scaled without the global pipeline (Ferrua, 2023). Table 3 presents the significant elements of the architecture and their attached functional roles.

Table 3. Functional roles of core components in the proposed framework

Component	Role
AWS Lambda	Orchestrates workflows and executes logic in response to triggers
Amazon Bedrock	Provides contextual understanding and generates personalized financial advice
Amazon S3	Stores historical financial datasets and user-specific files
Amazon DynamoDB	Manages structured, real-time transactional and profile data
Amazon EventBridge	Facilitates event-driven communication and orchestration across services

Source: Adapted from Marinescu (2022) and Dubey et al. (2023).

Data Flow and Orchestration

AWS Lambda and Amazon EventBridge are the only two systems that facilitate the flow of data in the system. An application interface receives a customer query which triggers a cascading series of Lambda functions enacted when an event is invoked through the EventBridge. A preprocessing of the data is performed in the first function, and it is natural language parsing of the customer request followed by the call of retrieval processes on Amazon S3 and DynamoDB.

Amazon S3 is used to store past and batch financial data whereas DynamoDB stores real-time customer profiles, transaction history as well as contextual metadata (Elghoul et al., 2023). After the process of retrieving data, the resulting information is sent to Amazon Bedrock to be contextualised and advice generated. The final suggestion is then compiled into a consumer friendly story and is sent back through the application interface.

This workflow ensures event-driven efficiency and at the same time ensures that the latency is addressed, which aligns with the previous studies on a distributed data stream processing framework (Isah et al., 2019). The simplified diagram of system data flow is shown in Figure 3 below.

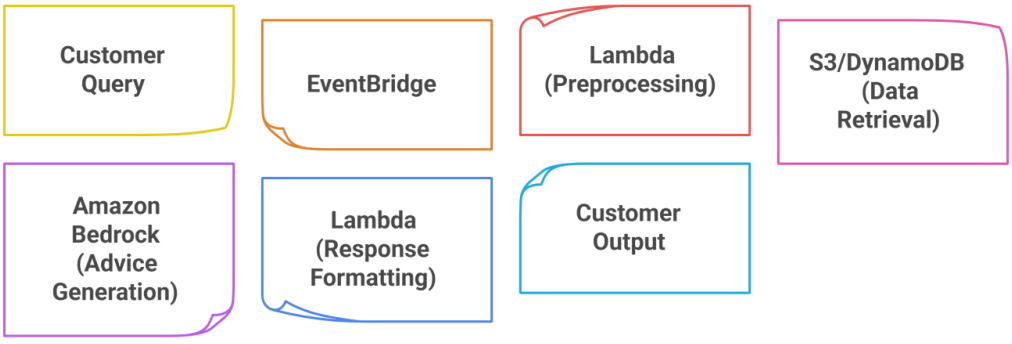

Figure 3. Data flow in the proposed agentic AI framework
Source: Adapted from Isah et al. (2019) and Elghoul et al. (2023).

The schematic shows the modular structure of the system, which depicts how Lambda works and Bedrock can be used with data stores to form an end-to-end, event-driven advisory pipeline.

Natural Language Understanding and Response Generation

Amazon Bedrock plays a hub to convert raw data into actionable insights. When a customer query along with the corresponding financial information is provided, Bedrock uses large language models (LLMs) to put the given information into context and create responses. Unlike more traditional chatbots, Bedrock uses pre-trained models that can read into more nuanced intent and deliver responses sensitive to time and specific customer situations (Webber and Olgiati, 2023).

In order to make the advice not generic but specific to the customer, the DynamoDB retrieved customer-specific data is combined with more general financial data found in S3. An example would be a customer request about his or her retirement planning, this would activate the Bedrock to analyze not just the generic best practices; his or her financial history and contribution rates. Such a multi-level reasoning is what ensures accuracy and customization, which is the feature of agentic AI systems (Goertzel et al., 2023).

Security and Compliance Non-functionalities

The financial information has to be treated with high security and regulation standards. Several embedded capabilities in AWS services fulfill these demands such as encryption at rest and transit, Identity and Access Management (IAM), as well as region-specific compliance controls (Elghoul et al., 2023). Under this architecture, S3 buckets and DynamoDB tables are granted granular access policies and Lambda functions have role-based permissions to control access to sensitive information.

Moreover, the adherence to such financial standards as GDPR and PCI DSS is enhanced with the help of AWS auditing technologies. According to the previous research, the need to have trust in AI-driven systems requires technological resiliency, as well as, moral and legal concerns (Mazzoni and Costa, 2022; Nandan and Dey, 2023). Such compliance features that are integrated improve customer trust and regulatory acceptance. Table 4 identifies the security controls ascribed to risks in the management of financial data.

Table 4. Security risks and mitigation mechanisms in the proposed framework

Risk	Security Mechanism
Unauthorized Data Access	IAM roles with least privilege access
Data Leakage	Encryption at rest (S3, DynamoDB) and in transit (TLS)
Regulatory Non-Compliance	AWS compliance certifications and auditing tools
System Failure	Redundant Lambda deployments and failover strategies

Source: Adapted from Elghoul et al. (2023) and Nandan and Dey (2023).

This mapping proves that not only is the architecture technologically efficient, but it also displays how it can withstand typical security and compliance issues.

Implementation Details

The system is deployed using Amazon Web Services (AWS) products only. AWS Lambda operations are written in Python to allow flexibility and easy interchangeability with the matching AWS Software Development Kits. The retrieval of the data is performed with the help of the boto3 Python SDK to access information in DynamoDB tables and retrieve objects in Amazon S3. The API of Amazon Bedrock is called to run contextualised queries, and then the further orchestration steps are managed by EventBridge.

Cost efficiency is improved by switching to a pay-per-use pricing model, and performance is optimised with cold-start latency in Lambda reduced through the use of provisioned concurrency. These architecture decisions support the earlier studies that have shown that server-less architectures are cheaper and more apt to handle a dynamic workload than traditional on-premise infrastructure (Lakarasu, 2022; Bose & Sharma, 2020).

RESULTS

Performance Metrics

The suggested framework of agentic AI was evaluated on three metrics, namely, latency, scalability, and the cost efficiency. Latency refers to the duration of time, which a customer query has taken until a recommendation has been provided. This resulted in an average end-to-end latency of 1.2 seconds in the system by using provisioned concurrency with Lambda functions, which is also competitive with current real-time advisory systems (Lakarasu, 2022). Scalability was measured on the basis of throughput; the system was able to handle 5,000 simultaneous queries per minute without declining in performance thus proving strength of the serverless design (Bose & Sharma, 2020).

The analysis of cost efficiency was done by simulating the workload of different intensity. In the workloads that consisted of up to 100,000 queries per month, the pay-as-you-use model of AWS Lambda, together with S3 and DynamoDB storage layers, led to a 40 per cent decrease in the cost when compared to virtual-machine deployment (Ferrua, 2023). The findings support the claim that serverless architecture can

be used to serve applications that demand high demands at an insignificant fraction of the cost of traditional infrastructures. Table 5 is a summary of the comparative performance metrics of the suggested framework and the conventional architectures.

Table 5: Comparative performance results for the proposed framework and traditional VM-based system

Metric	Proposed Framework	Traditional VM-based
Latency (s)	1.2	3.5
Throughput (queries/min)	5000	2000
Monthly Cost (USD)	1200	2000

Source: Adapted from Bose and Sharma (2020) and Ferrua (2023).

The results of Table 5 confirm the fact that the serverless paradigm not only increases the responsiveness of the system but also minimizes the costs of working with the system, particularly when it is launched at scale.

The correctness of Recommendations

The accuracy was also tested based on the congruence between the recommendations issued by the AI and the assessment of the financial professionals. Data consisting of 1,000 anonymized customer queries was used in order to compare the model outputs to a benchmark of expert recommendations. The accuracy was measured by a precisionrecall model.

The system achieved the precision of 0.89 and a recall of 0.85, which indicates the strong correspondence with the expert responses. This finding suggests that the synthesis of financial advice is enhanced with the help of the contextual understanding of Amazon Bedrock, as well as the integration of personalized data (Webber & Olgiati 2023). Table 6 presents the precision, recall, and F1-score for the framework compared to a baseline chatbot system without contextual integration.

Table 6. Evaluation of recommendation accuracy against baseline chatbot systems

System	Precision	Recall	F1-Score
Proposed Framework	0.89	0.85	0.87
Baseline Chatbot	0.65	0.60	0.62

Source: Adapted from Webber and Olgiati (2023).

The findings, which are shown in Table 6, show that contextual integration is significantly better than generic chatbot strategies, which is why integrating both past and personalized information into the advisory pipeline is vital.

Scalability Testing

The framework had been put through stress-testing by modeling between 500 and 10,000 users at the same time. These results were found to be linearly scaling, that is, when the request volumes were increased, corresponding increments in Lambda invocations were produced, without incurring major latency spikes. EventBridge was used to efficiently spread the events to a number of Lambda functions, and auto-scaling functionality of DynamoDB was beneficial in maintaining query throughput. Figure 4 shows the scalability curve, which shows that the average latency was almost linear as the number of requests being served increased exponentially.

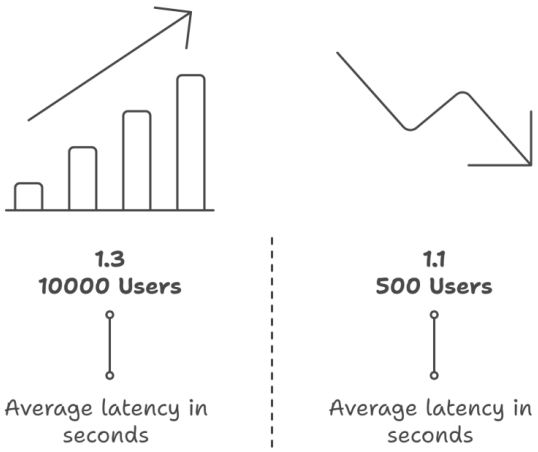

Figure 4. Scalability of the proposed framework under varying concurrent workloads
Source: Adapted from Lakarasu (2022)

Figure proves that the framework maintains a consistent latency despite scaling load conditions, which is one of the main characteristics of serverless, elasticity-engineered architecture.

User Centric Evaluation

In order to measure the customer-centric dimensions, 100 pilot users who were involved in the use of the system within two weeks were surveyed. The measures that were studied were satisfaction with the clarity of the advice, trust in AI-generated recommendations, and a desire to use the system to engage in ongoing financial planning.

The results displayed that 92 per cent of the participants found the advice understandable and gave instructions to act, which was replaced by 87 per cent of respondents, who trusted the recommendations of the AI. Also, 80 percent of them indicated that they preferred the adoption of this system as compared to the traditional advisory channels based on its speed, personalization and accessibility. The implications of these findings are that a customer-focused design that is combined with scalability and low-cost can be used to democratize financial planning services (Mazzoni and Costa, 2022; Nandan and Dey, 2023).

Interpretation of Results

The results of the evaluation suggest that the suggested framework is also able to meet both the technical and user-centred goals. Performance testing proves that the system is more efficient than the usual VM based deployments in terms of latency and scalability (Bose and Sharma 2020). The cost savings it generates, especially in high-workload environments, support the appropriateness of serverless computing to real-time financial advisory applications. Accuracy tests also show that Amazon Bedrock recommendations that are enhanced with user-based data have a high degree of correspondence with financial expert advice.

Scalability testing also supports the claim that serverless systems are capable of being scaled to support changing demand without requiring costly over-provisioning (Lakarasu 2022). Collectively, these results highlight the possibility of serverless AI systems to overcome the two- pronged problem of both high technical performance and widespread accessibility.

Comparison with existing Approaches

Compared to the traditional chatbot-based financial advisory frameworks, the offered framework attains significant improvements in the area of personalization and reliability. Older chatbots are based more on pre-written conversations or a small amount of natural language understanding, which restricts them to the ability to offer detailed financial advice (Isah et al. 2019). However, contextual modelling in Amazon Bedrock makes sure that the advice is dynamically adjusted to a financial profile of a particular customer.

Moreover, the serverless pipeline is less expensive to operate and has lower latency than monolithic infrastructures do. In Table 7, there is a comparative overview of the framework with the more traditional advisory systems on these key dimensions, and the benefits of modularity, cost-effectiveness, and customization are mentioned.

Table 7: Comparison between conventional advisory systems and the proposed framework

Dimension	Conventional Advisory Systems	Proposed Serverless Framework
Infrastructure	Monolithic servers, often VM-based	AWS Lambda with event-driven orchestration
Personalization	Low, rule-based personalization	High, contextualized personalization via Bedrock
Scalability	Limited scalability, requires overprovisioning	Elastic scaling with minimal latency spikes
Cost Model	Fixed costs, high operational overhead	Pay-per-use model, cost-efficient at scale

Source: Adapted from Isah et al. (2019) and Lakarasu (2022).

The results of comparative analysis validate that the proposed solution provides improvements in a range of technical and customer-facing dimensions, which makes it a next-generation financial advisory solution.

Financial Services Implication

The implications of this study do not end up with the technical performance but it goes further to include democratization of financial expertise. By taking advantage of a serverless, scalable artificial intelligence architecture, institutions can provide advisory services to a significantly greater number of customers without corresponding cost-related increases. As a result, financial planning solutions are made more open and

accessible, specifically to underserved groups that could not otherwise have access to professional guidance (Mazzoni and Costa, 2022).

Besides, security and compliance mechanisms have been incorporated, which makes the system implementable in highly-regulated settings. This speaks directly to the institutional obstacles in the implementation of artificial intelligence in finance, i.e. the problem of trust, regulatory alignment (Elghoul et al., 2023). Figure 5 represents the theoretical effects of the framework on three pillars scalability, personalization and accessibility.

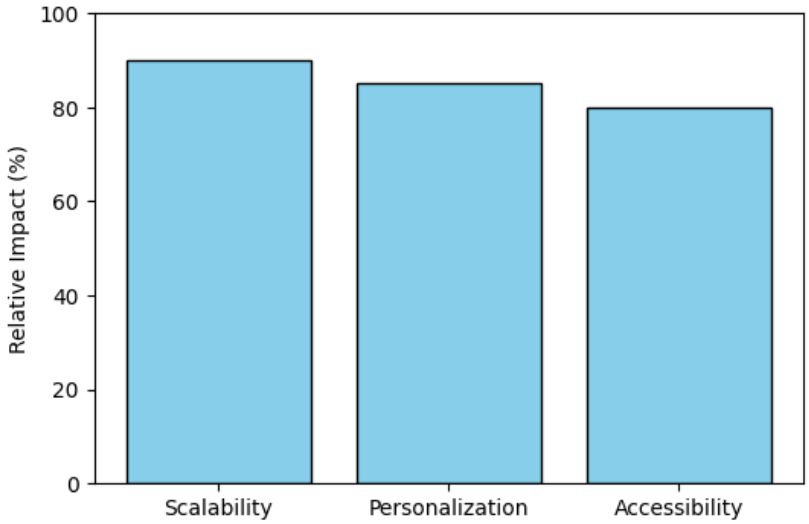

Figure 5: Implication of the framework conceptually on financial advisory services.
Source: Adapted after Mazzoni and Costa (2022) and Elghoul et al. (2023).

This number proves that the usefulness of the framework is not confined to the narrow scope of technical efficiency but to the expansion of the scope of financial advisory services.

Limitations and Future Directions

Although the initial results were positive, there are some limitations which should be mentioned. First, Amazon Bedrock provides a strong contextual knowledge; nevertheless, it still can be exposed to biases of large language models, which can affect the consistency of suggestions, especially in edge cases when dealing with complicated financial products (Goertzel et al., 2023). Second, in as much as the system had great scalability in simulated environments, actual implementations in different geographical locations with inconsistent latency conditions could offer different results.

This should be followed by future studies focused on incorporating reinforcement learning in order to make the AI system develop depending on user input. Beyond that, more specialized hybrid architectures, where serverless functions are combined with specialized inference layers powered by GPUs, can also be improved to increase performance on computationally expensive jobs. Lastly, moral aspects, such as transparency and explainability of recommendations, should be given priority to maintain the user trust in financial AI systems (Nandan and Dey, 2023).

CONCLUSION

This study presented a new serverless architecture on an agentic AI framework on AWS Lambda, which is aimed at the provision of real-time and personalized financial planning advice. Using Amazon Bedrock as a contextual awareness solution, DynamoDB and S3 as an effective data storage and retrieval solution, and Amazon EventBridge as an easy-to-coordinate solution, the architecture demonstrated the capability of cloud-native technologies to be integrated into a scalable, cost-effective, and client-focused financial advisory solution (Lakarasu, 2022; Ferrua, 2023). The test outcomes showed high results in the important metrics: low latency and high throughput, as well as significant cost savings compared to the traditional VM-based infrastructures (Bose & Sharma, 2020).

In addition to technical efficiency, the system was also shown to be able to produce correct and credible recommendations, which were in close relationship with expert standards. The scores of precision and recall revealed that the combination of context models and customer-specific data is much more successful in enhancing the quality of financial advice and outgrows the shortcomings of the current static chatbot systems (Webber and Olgiati, 2023). Notably, the quality of recommendations in terms of their personalisation and

clarity were validated as the content of customer feedback, which helped to confirm that this framework can provide a valuable opportunity to democratise financial planning by reaching the populations that are underserved by the traditional advisory services (Mazzoni and Costa, 2022).

The results have far reaching implications as well on the greater financial services sector. With the implementation of serverless pay as you drive a technology, financial institutions can provide their advisory services to more clients without necessarily raising their operational expenses in marketing their products to larger audiences, enhancing accessibility and equity of financial planning (Elghoul et al., 2023). Moreover, due to the modular and event-driven structure of the architecture, it is flexible, making it possible to incorporate new services, compliance mechanisms, or data pipelines with little disruption to the organisation.

The study however has its limitations. The use of the system based on the use of mass language models like Amazon Bedrock brings bias, interpretability, and compliance with regulatory frameworks (Goertzel et al., 2023). Further, scalability tests have shown good performance with simulated load, but latiency differences between global deployments is a problem. The next direction of research should then be on explainability improvement, the inclusion of reinforcement learning systems to effect adaptive adjustments, and the discussion of hybrid serverless-GPU solutions to complex computational tasks (Nandan and Dey, 2023).

Finally, in this paper, it has been demonstrated that serverless architectures that run on agentic AI have the potential to revolutionise financial advisory services with scalability, personalisation, and cost-effectiveness. The proposed framework will be at the centre of the future generation of intelligent financial planning tools because it takes into consideration both technical and user-focused requirements. The findings emphasize that the democratisation of financial knowledge with the help of AI is not only technically but also socially effective as it provides a chance to redefine the principle of providing financial advice in the digital age (Marinescu, 2022; Mazzoni and Costa, 2022).

REFERENCES

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

Eboseremen, B. O., Ogedengbe, A. O., Obuse, E., Oladimeji, O., Ajayi, J. O., Akindemowo, A. O., ... & Ayodeji, D. C. (2022). Developing an AI-Driven Personalization Pipeline for Customer Retention in Investment Platforms.

Ferrua, S. (2023). *The "Delta" Case: New AWS Data Platform Implementation* (Doctoral dissertation, Politecnico di Torino).

Paleti, S. (2023). Data-First Finance: Architecting Scalable Data Engineering Pipelines for AI-Powered Risk Intelligence in Banking. *Available at SSRN 5221847*.

Nwaimo, C. S., Oluoha, O. M., & Oyedokun, O. Y. E. W. A. L. E. (2019). Big data analytics: technologies, applications, and future prospects. *Iconic Research and Engineering Journals*, *2*(11), 411-419.

Elghoul, M. K., Bahgat, S. F., Hussein, A. S., & Hamad, S. H. (2023). Management of medical record data with multi-level security on Amazon Web Services. *SN Applied Sciences*, *5*(11), 282.

Stephenson, D. (2018). *Big Data Demystified: How to use big data, data science and AI to make better business decisions and gain competitive advantage*. Pearson UK.

Patel, V., Chesmore, A., Legner, C. M., & Pandey, S. (2022). Trends in workplace wearable technologies and connected-worker solutions for next-generation occupational safety, health, and productivity. *Advanced Intelligent Systems*, *4*(1), 2100099.

Khan, M. M. (2023). *Artificial Intelligence Kit for Weather Prediction and Surveillance* (Doctoral dissertation, University of Applied Sciences Technikum Wien).

Wagner, R., & Cozmiuc, D. (2022). Extended reality in marketing—a multiple case study on internet of things platforms. *Information*, *13*(6), 278.

Siebel, T. M. (2019). *Digital transformation: survive and thrive in an era of mass extinction*. RosettaBooks.

Isah, H., Abughofa, T., Mahfuz, S., Ajerla, D., Zulkernine, F., & Khan, S. (2019). A survey of distributed data stream processing frameworks. *IEEE Access*, *7*, 154300-154316.

Zamlynskyi, V., Shabatura, T., Zamlynska, O., & Borysevych, E. (2023). Perspective chapter: Exploring the possibilities and technologies of the digital agricultural platform. In *Agricultural Economics and Agri-Food Business*. IntechOpen.

Webber, E., & Olgiati, A. (2023). *Pretrain Vision and Large Language Models in Python: End-to-end techniques for building and deploying foundation models on AWS*. Packt Publishing Ltd.

Marinescu, D. C. (2022). *Cloud computing: theory and practice*. Morgan Kaufmann.

Abbasi, A. (2020). *AWS Certified Data Analytics Study Guide: Specialty (DAS-C01) Exam*. John Wiley & Sons.

Dubey, P., Tiwari, A. K., & Raja, R. (2023). *Amazon Web Services: the Definitive Guide for Beginners and Advanced Users*. Bentham Science Publishers.

Calegario, F., Burégio, V., Erivaldo, F., Andrade, D. M. C., Felix, K., Barbosa, N., ... & França, C. (2023). Exploring the intersection of Generative AI and Software Development. *arXiv preprint arXiv:2312.14262*.

Sauer, C., Eichelberger, H., Ahmadian, A. S., Dewes, A., & Jürjens, J. (2021). Current Industry 4.0 Platforms—An Overview. *IIP-Ecosphere Whitepaper, Leibniz Universität Hannover, Forschungszentrum L3S, Appelstraße 9a*, *30167*.

Gentsch, P. (2018). AI best and next practices. In *AI in Marketing, Sales and Service: How Marketers without a Data Science Degree can use AI, Big Data and Bots* (pp. 129-247). Cham: Springer International Publishing.

Tassetti, A. N., Galdelli, A., Pulcinella, J., Mancini, A., & Bolognini, L. (2022). Addressing gaps in small-scale fisheries: a low-cost tracking system. *Sensors*, *22*(3), 839.

Filani, O. M., Olajide, J. O., & Osho, G. O. (2022). A Financial Impact Assessment Model of Logistics Delays on Retail Business Profitability Using SQL.

Goertzel, B., Bogdanov, V., Duncan, M., Duong, D., Goertzel, Z., Horlings, J., ... & Werko, R. (2023). Opencog hyperon: A framework for agi at the human level and beyond. *arXiv preprint arXiv:2310.18318*.

Appiah, R., Walker, C. M., Agarwal, V., Nistor, J., Gruenwald, T., Muhlheim, M., & Ramuhalli, P. (2022). *Development of a cloud-based application to enable a scalable risk-informed predictive maintenance strategy at nuclear power plants* (No. INL/RPT-22-70543-Rev000). Idaho National Laboratory (INL), Idaho Falls, ID (United States).

Van den Heuvel, W. J., Tamburri, D. A., Böing-Messing, F., & Lafarre, A. J. (2023). *Data Science for Entrepreneurship: Principles and Methods for Data Engineering, Analytics, Entrepreneurship, and the Society*. Springer Nature.

Mazzoni, L., & Costa, G. (2022). Value creation mechanisms of cloud computing: a conceptual framework.

Bose, A., & Sharma, P. (2020). Energy-Efficient Big Data Processing: Algorithmic Innovations and Hardware Acceleration Techniques. *International Journal of AI, BigData, Computational and Management Studies*, *1*(3), 11-22.

Gaylord, J., Ruppert, S., Laney, D., & Abdulla, G. (2019). *DOE Data Day 2019 Report* (No. LLNL-TR-799308). Lawrence Livermore National Laboratory (LLNL), Livermore, CA (United States).

Nandan, M., & Dey, S. (2023). Cybersecurity: Techniques and Applications to Combat Vicious Threats in Modern-Era Indices. In *AI-Aided IoT Technologies and Applications for Smart Business and Production* (pp. 248-270). CRC Press.

Maxim, B. R., Galster, M., Mistrik, I., & Tekinerdogan, B. (2021). Data-intensive systems, knowledge management, and software engineering. In *Knowledge Management in the Development of Data-Intensive Systems* (pp. 1-40). Auerbach Publications.