Supporting Complex Deployments of Next Generation AI Analytics in AWS

aws partner network

Executive Summary

Vertical Relevance partnered with AI Squared to support one of its marquee customers, a leading investment advisor and asset manager, in standing up a custom configuration of AI Squared's predictive analytics application and platform within AWS. This custom deployment of AI Squared featured tight integration with the end customers AWS architectural deployment and security patterns, ensuring seamless consumption within the organization.

Custom Multi-Service Proxy Deployment Configuration

In an effort to align with the best practices and architectural standards of the client, our team implemented a sophisticated solution that involved using NGINX as a reverse proxy to front multiple application services. This strategic approach allowed for a single web endpoint to expose various services, streamlining access in accordance with the organization's requirements for simplified infrastructure and enhanced security. By deploying each application service within its own Amazon ECS (Elastic Container Service) container, we leveraged the robustness of managed container orchestration to facilitate ease of deployment and scalability. NGINX was meticulously configured to route incoming traffic to the appropriate service based on the request path, ensuring efficient and secure access to the diverse services through a unified interface. This architecture not only minimized the public exposure of internal services but also upheld the organization's standards for security, efficiency, and manageability, demonstrating our commitment to tailoring solutions that meet and exceed client expectations.

Authentication Integration

The application's integration with Okta to utilize the customer's Identity Provider (IDP), Azure AD, as the standard authentication source, was expanded to include the supporting Chrome extension, ensuring a cohesive and secure user authentication experience across both platforms. This enhancement enabled the seamless use of Azure AD for authentication, providing users with secure single sign-on (SSO) capabilities not only within the web application but also when interacting with the Chrome extension. By leveraging Okta's advanced identity and access management features, we established a unified authorization flow that maintained strict adherence to the organization's security protocols and identity management policies. This strategic implementation across the application and its Chrome extension counterpart ensured that users could access services efficiently and securely, regardless of the platform, thereby upholding the organization's standards for security, user experience, and access control. Through this comprehensive approach, our team delivered a solution that not only met the client's requirements for secure and streamlined access but also reinforced the application's security posture by integrating robust and scalable identity management capabilities.

About AI Squared

ai squared®

AI Squared facilitates the integration of AI and ML into business applications, offering a drag-and-drop solution that minimizes code and IT dependency. Its technology significantly reduces neural network resource requirements without affecting performance, targeting data scientists and engineers struggling with ML deployment. The platform provides easy model integration into web applications, streamlining decision-making and innovation. Originating from the creators of BeyondML, AI Squared enhances scalability and collaboration in internal ML initiatives.

Automated Performance & Load Testing Capabilities

To ensure the newly integrated application architecture could withstand the demands of the customer's environment, we deployed the Smart Capacity Finder, an intelligent load testing tool. Developed with Locust and enhanced by Python customizations, this tool automates the process of identifying an application's capacity threshold. It begins testing with a minimal load, gradually increasing it in an exponential manner until failures are detected, thereby pinpointing the system's maximum capacity within a configurable tolerance. This methodical approach allows for precise adjustments based on real-time performance feedback, effectively simulating various user load scenarios to assess the resilience of the application.

The deployment of the Smart Capacity Finder in the AWS cloud was facilitated through a CDK script, simplifying the setup process by automating the infrastructure requirements and ensuring seamless integration with the existing environment. This strategic implementation of the Smart Capacity Finder not only validated the application's performance under load but also reinforced our commitment to delivering robust and reliable solutions.

Enhanced Logging & Monitoring Capabilities

In response to the client's operational and data security requirements, we enhanced the logging and monitoring capabilities of their application and Chrome extension. Given the constraints on data collection, particularly for uptime and service availability, we developed a tailored logging framework that captures essential system logs without collecting customer-specific data, thus adhering to strict privacy policies. This improvement enables precise monitoring of the application's performance, supporting Recovery Point Objectives (RPOs) and Recovery Time Objectives (RTOs), while ensuring compliance with data protection standards. This streamlined approach to logging and monitoring not only ensures the application's reliability and functionality but also aligns with the client's need for operational integrity and data security, showcasing our commitment to delivering solutions that meet complex requirements.

Conclusion

In conclusion, our collaboration with AI Squared focused on deploying and optimizing their application for a flagship customer, involving the strategic implementation of NGINX reverse proxy, secure authentication via Okta and Azure AD, advanced load testing with the Smart Capacity Finder, and enhanced logging and monitoring. These efforts significantly improved performance, security, and manageability, ensuring the application met the rigorous standards and specific needs of the customer. Our comprehensive support played a pivotal role in optimizing the application's resilience and operational efficiency, underlining our commitment to empowering AI Squared in achieving success with their flagship deployment.

About Vertical Application

Vertical Relevance is a Financial Services focused (Wealth Management, Asset Management, Banking, Insurance) consulting firm helping with the design & delivery of effective transformation programs across people, process, & systems. With 10+ years of AWS & 20+ years of Financial Services experience, we understand the business needs & build solutions to meet sales, marketing & compliance goals.

VR VERTICAL RELEVANCE