rapidscale

AI Armor for financial services

New strategies to protect data on AWS

aws PARTNER Premier Tier Services

aws PARTNER Security Services Competency

aws PARTNER AI Services Competency

- Agentic AI Consulting Services
- Generative AI Consulting Services

rapidscale.com

Financial services are more vulnerable than ever to sophisticated fraud schemes. Traditional fraud detection methods often struggle to keep pace with the rapidly evolving tactics of fraudsters.

Modern tools rely on advanced technologies to fight fraud, and Artificial Intelligence (AI) and Machine Learning (ML) are fast becoming the go-to players in this battle. We'll explore how AI and ML are transforming fraud detection in financial services, enhancing security, and ensuring a safer financial landscape.

This eBook describes how RapidScale and Amazon Web Services (AWS) help financial institutions build cost-effective, secure, and globally scalable fraud detection systems to act as AI Armor in securing customer data.

Building a robust fraud detection system

To build a resilient fraud detection system, financial institutions must establish a strong infrastructure foundation. Foundational components like AWS Control Tower with AWS Organizations act as an armored core, securing access to key components of the AWS ecosystem. By setting up robust account security using Service Control Policies, Virtual Private Cloud (VPC), Security Groups, Network Access Control Lists (NACLs), and Multi-Factor Authentication, financial companies can secure their core infrastructure, enabling safe access to application workloads and other AWS services. You can read more about how AWS Control Tower helps with account and security management in the cloud in this RapidScale blog: Reducing the Cost of Managing Multiple AWS Accounts Using AWS Control Tower.

Data security

Securing data is at the heart of building any robust IT solution. Fraud detection systems face higher threat levels, thus requiring additional AWS security services like AWS IAM, AWS KMS, AWS GuardDuty, AWS CloudFront, AWS Certificate Manager, AWS CloudWatch, and AWS CloudTrail. RapidScale deploys AWS services using Infrastructure as Code (IaC), a core building block of our DevOps and SecOps practices, as IaC removes the human error element for deployment consistency. By implementing a layered security solution starting at the perimeter of any AI solution, RapidScale reduces the blast radius if a breach is detected. All endpoints need to be encrypted; the perimeter defense starts with a CloudFront + WAF in front of AWS API Gateway as ingress for traffic flowing to downstream services. Each subsequent layer requires data encryption in-transit and at-rest. Encryption keys are stored in vaulted key-stores on AWS KMS with strict rules-based access control (RBAC), allowing only authorized user access to encryption keys and vaults.

Data lake and machine learning

Behind every great AI solution is good data. RapidScale and AWS are experts at helping financial services companies build a bedrock. Creating a high-quality data lake on AWS involves several critical steps to ensure data is clean, well-organized, secure, and easily accessible. Here are some fundamental techniques to help you build a robust and feature-rich data lake on AWS:

Architecture

By dividing your Data Lake architecture into data zones, companies can significantly increase the efficiencies of how their data is ingested from disparate sources, cleansed, and then curated. Zone definitions are critical to ensuring that data coming in is from credible sources and is processed in the most efficient way possible. Sample zone definitions can be as follows:

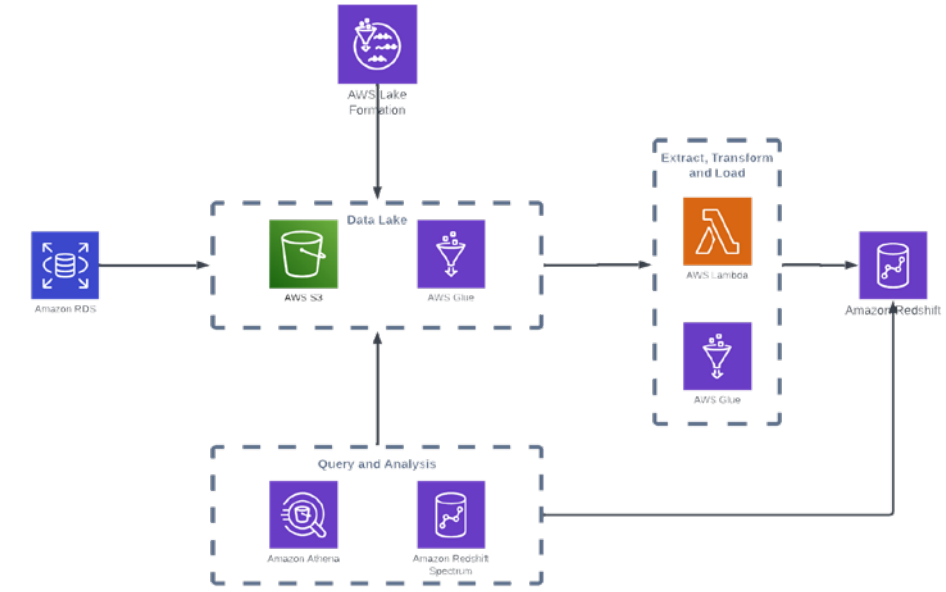
1. Raw Zone: This is where unprocessed data is stored.
2. Cleansed Zone: Data that has been cleaned and is ready for pre-processing.
3. Curated Zone: Stored data that has been processed through an ETL process.

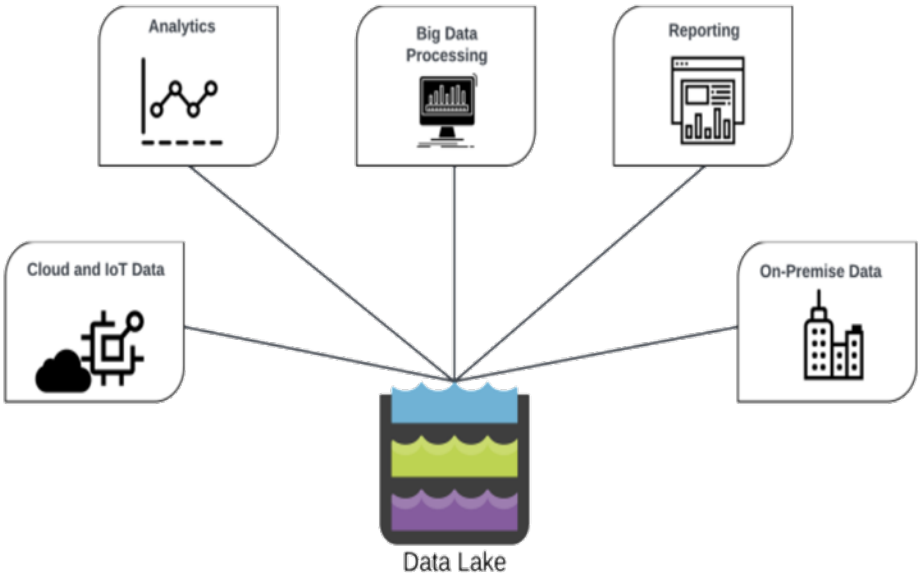
Data Lake foundations

RapidScale architects and deploys AWS serverless technologies wherever appropriate. By deploying AWS serverless technologies, RapidScale can implement cost-effective, highly secure, and scalable infrastructure for building a data lake. AWS services typically deployed in building a robust solution include:

1. **AWS Lake Formation:** A key component of the RapidScale solution is to simplify the setup, management, and security of your data lake. By using AWS Lake Formation, RapidScale can streamline the ingestion of data from multiple sources, define meta-data to properly catalog raw data which ultimately leads to better data governance and documenting data lineage for auditing purposes. AWS Data Lake Formation is an invaluable tool for cost optimization, performance tuning and preparing data for downstream processing with AWS Glue and Redshift Spectrum.
2. **Amazon S3:** As an object storage solution, S3 can store raw data files, flat files, JSON, XML, and other unstructured data files before processing. S3 is typically used as a staging point for all Raw Zone data.
3. **AWS Glue:** A highly scalable solution, AWS Glue is deployed in the Cleansed Zone. Glue is used for cataloging, ETL jobs, and schema management. AWS Glue is an efficient way of ingesting data from external sources, and both AWS Glue and AWS Data Pipeline can be used for scheduled batch processing jobs.
4. **Amazon Athena:** Athena can be utilized in both Cleansed and Curated zones. RapidScale uses AWS Athena to streamline data processing workflows, enhance data quality, and enable sophisticated analysis and reporting capabilities.
5. **Amazon Redshift Spectrum:** To perform data validation and exploratory data analysis tasks, RapidScale deploys AWS Redshift Spectrum in both the cleansed and curated zones for customers with large datasets. Redshift Spectrum can process exabytes of data in the curated zone by utilizing both "Aggregated and Materialized views" and the use of Columnar formats. RapidScale deploys Redshift Spectrum to query data in S3 without loading it into Redshift.

Building blocks of a Data Lake:





Understanding the fraud landscape

Modern financial services organizations must manage various types of fraud, including unauthorized use of customer data, identity theft, credit card fraud, insider trading, and money laundering. These are not just organizational and legal challenges but colossal technical challenges to defend against. The complexity of the fraud detection system makes it nearly impossible to perform these tasks manually or with a simple rules-based detection system.

Machine Learning and Artificial Intelligence Technologies: Technologies such as AWS SageMaker and AWS Bedrock are key components of all fraud detection systems that RapidScale deploys for customers in the financial services industry. As discussed earlier, the first step to training Large Language Models (LLMs) is having good data. Now that we have a Data Lake with curated data, we can start building the infrastructure and framework for training and tuning LLMs needed for a robust AI and ML ecosystem.

Selecting the Right LLM: Before training LLMs, selecting the right model is crucial. The following table compares different pre-trained AWS Bedrock LLMs available on AWS Marketplace:

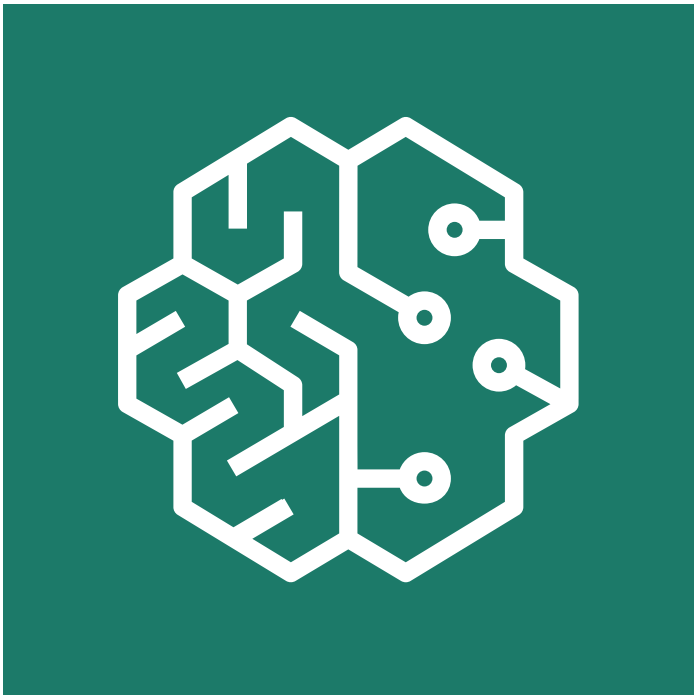
Foundation Model	Model Version	Max Capacity	Distinct Features & Languages	Supported Use Cases & Applications	Pricing
Amazon Titan	Titan Text Premier	32K tokens	Designed to deliver superior performance across a wide range of enterprise applications. This model is optimized for integration with Agents and Knowledge Bases for Amazon Bedrock, making it an ideal option for building interactive generative AI applications that can use your APIs and interact with your data.	RAG, agents, chat, chain of thought, open-ended text generation, brainstorming, summarization, code generation, table creation, data formatting, paraphrasing, rewriting, extraction, and Q&A	$0.0008 input, $0.0016 output per 1,000 tokens
	Titan Text Express	8K tokens	High-performance text model, 100+ languages	Diverse text-related tasks in content creation, classification, and open-ended Q&A, applicable in education and content marketing	$0.0008 input, $0.0016 output per 1,000 tokens
	Titan Text Lite	4K tokens	Cost-effective text generation, English	Efficient for summarization, copywriting in marketing and corporate communications	$0.0003 input, $0.0004 output per 1,000 tokens
	Titan Text Embeddings	8K tokens	Text translation to numerical representations, 25+ languages	Semantic similarity, clustering for data analysis, knowledge management	$0.0001 per 1,000 tokens

Amazon Titan	Titan Multimodal Embeddings	128 tokens, 25 MB images	Multimodal (text and image) search, English	Accurate and contextually relevant search, recommendation experiences in e-commerce, and digital media	$0.0008 per 1,000 tokens; $0.00006 per image
	Titan Image Generator	77 tokens, 25 MB images	High-quality image generation using text prompts, English	Image creation and editing for advertising, e-commerce, and entertainment with natural language prompts	512×512: $0.008, 1024×1024: $0.01 per image
Anthropic Claude	Claude 3: Opus	up to 200,000 tokens	Features: Fastest and most compact, designed for quick responses and handling simple queries efficiently. Languages: Supports multiple languages, optimized for fast and affordable operations.	Use cases: Task automation, research and development, advanced strategic analysis, and handling large-scale operations. Additionally, Opus is suitable for tasks requiring high accuracy and nuanced understanding, including solving complex problems and generating code.	$15 per million input tokens and $75 per million output tokens.
	Claude 3: Sonnet	up to 200,000 tokens	Features: Balances intelligence and speed, suitable for enterprise workloads with strong performance. Languages: Multilingual support, ideal for data processing and sales tasks.	Use cases: Data processing, sales tasks, code generation, and other time-saving operations. Claude 3: Sonnet is Ideal for enterprises needing efficient and scalable AI solutions.	$3 per million input tokens and $15 per million output tokens.
	Claude 3: Haiku	up to 200,000 tokens	Features: Fastest and most compact, designed for quick responses and handling simple queries efficiently. Languages: Supports multiple languages, optimized for fast and affordable operations.	Use cases: Customer interactions, content moderation, and cost-saving operations. Haiku is also suitable for tasks needing near-instant responsiveness	$0.25 per million input tokens and $1.25 per million output tokens.
Meta Llama 3	Llama 3.1 with 8B, 70B, and 405B parameters	Llama 3.1 models have enhanced context lengths of up to 128,000 tokens.	. The Llama 3 models support eight languages: English, French, German, Hindi, Italian, Portuguese, Spanish, and Thai. . Incorporates iterative post-training procedures with supervised fine-tuning and direct preference optimization for high-quality synthetic data. . Llama Guard 3 and Prompt Guard components help detect and prevent malicious activities and content violations.	Llama 3 is designed for high scalability and can be integrated into various systems. In addition to scalability, Llama 3 is suitable for natural language processing (NLP) tasks, real-time data analysis, and enhancing AI capabilities in platforms like WhatsApp and meta.ai. Llama 3 is also useful for cybersecurity applications, including detecting cyberattacks and preventing malicious code distribution.	Meta's business model for Llama does not focus on selling access to the AI models, as they are open-sourced. The use of Llama 3 models incurs costs related to cloud services and infrastructure from AWS

RapidScale experts work with our financial services clients to perform detailed discovery and analysis of all fraud detection use cases and available data before choosing the right LLM model to best fit the needs of the solution we architect and deploy on AWS SageMaker and AWS Bedrock.

AWS SageMaker

Released in November 2017, AWS SageMaker was one of the first Machine Learning SaaS offerings on the market. SageMaker offers developers and data scientists the ability to train custom machine learning models to best fit their individual use cases. These models can be used for embedded systems or edge devices, providing the flexibility to train models on their own data. Based on TensorFlow and Apache MXNet, AWS SageMaker provides several interfaces for developers to integrate with. By connecting to other AWS services like AWS DynamoDB and AWS Batch Processing, AWS SageMaker provides a level of abstraction for any use case requiring custom language models.

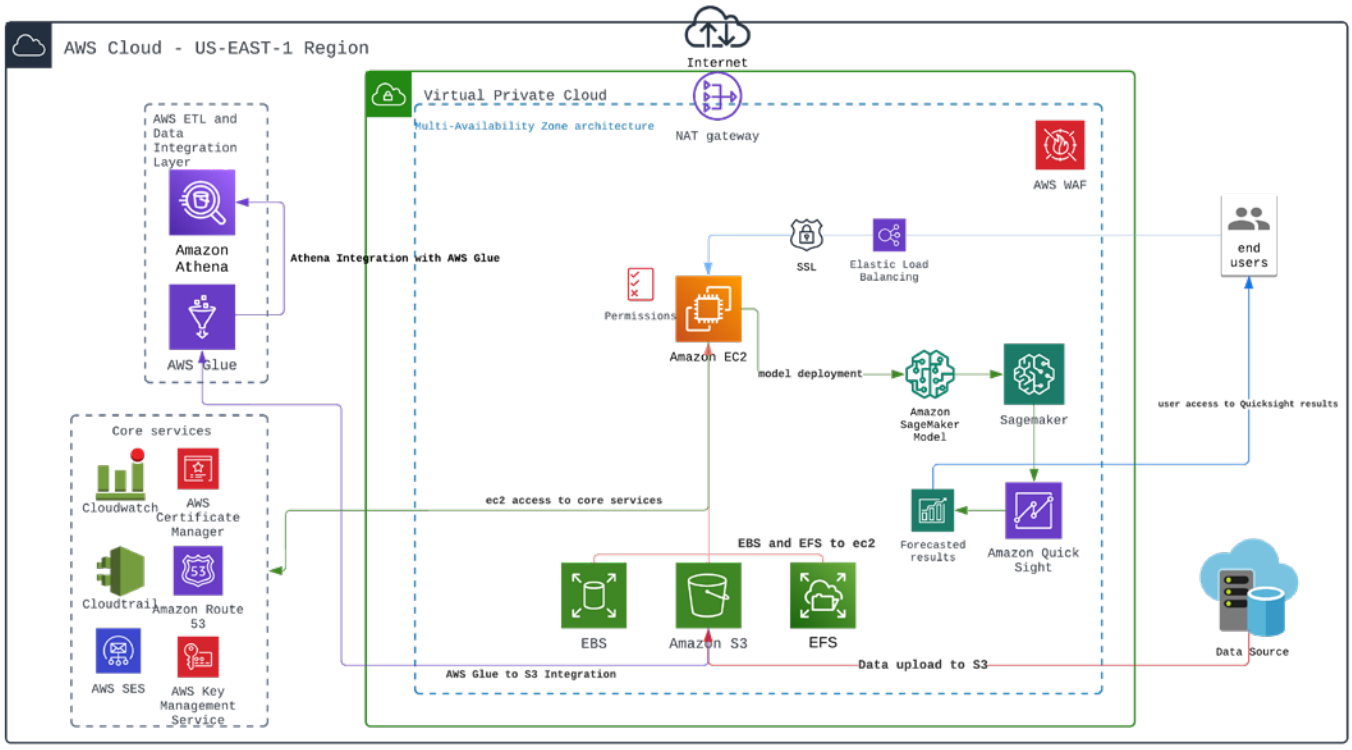


Interacting with AWS SageMaker: Using AWS SageMaker Studio tools enables developers easy access to the platform. Here's a simplified guide for a non-production environment:

1. Create an AWS SageMaker domain using the AWS SageMaker console.
   - The automated quick setup creates the user profile and IAM execution roles for accessing SageMaker services.
   - Next, create an Amazon DataZone domain, which creates a corresponding AWS SageMaker domain with AWS Glue/AWS Redshift databases for ETL processes.
2. When using SageMaker Studio, a Jupyter server is started in the background.
   - Once finished, the next screen takes you to SageMaker Jumpstart, your go-to place for SageMaker content, how-to guides, ML model selection, etc.
3. Users can experiment with their data using pre-built solutions. Automation for certain pre-built solutions can spin up other AWS services, so be aware of usage charges.

Sample architecture for anomaly detection:

The following RapidScale sample architecture illustrates the end-to-end design of an anomaly detection system using AWS SageMaker. Data ingestion from disparate sources is performed, data is sanitized using ETL services that include AWS Glue and AWS Athena to build a Data Lake. Once data is ingested and curated, custom machine learning models running on AWS SageMaker detect specific trigger elements, identify the anomaly, and present the findings to a cloud-native business intelligence solution running on AWS Quicksight.



AWS Bedrock

AWS Bedrock is the latest in Generative AI Services from AWS. Bedrock is a fully managed service providing pre-trained foundational models users can use to enable their applications with AI capabilities. RapidScale enables access to private customer datasets using Bedrock endpoints. Data privacy can be achieved by accessing private data stores within customer VPCs. Once proper access is enabled, pre-trained foundational models create original content. RapidScale uses a Retrieval-Augmented Generation (RAG) framework to further enhance the accuracy of results produced by accessing authoritative sources for exactness.

Enhancing Fraud Detection with AI and ML: Enhancing fraud detection systems using AI and ML is central to how RapidScale helps customers in the financial industry. RapidScale uses a layered approach to implementing ML and AI in fraud detection use cases, ultimately forming a formidable defense against fraud. Some techniques used include, but are not limited to:

- **Pattern Recognition and Anomaly Detection:** By using the pattern recognition capabilities of AI and ML and analyzing historical transaction data, fraud detection solutions establish "normal" behavior for individual accounts. Deviations from this learned norm can be flagged as potential fraud attempts.
- **Real-Time Analysis:** Curating an AI solution that processes and analyzes data in real-time helps detect and intercept fraudulent activities as they occur, minimizing financial impact and reducing response time.
- **Behavioral Biometrics:** Analyzing unique patterns of human behavior, such as typing speed, mouse movements, and the way a person holds their smartphone, helps in creating a behavioral profile for each user. These profiles can be used to indicate possible fraudulent activity.
- **Natural Language Processing:** By analyzing text data contained within emails, chat messages, and social media posts, an AI solution can detect social engineering attacks. This is achieved by defining specific vectors in language patterns, implemented using AWS SageMaker or AWS Bedrock.
- **Predictive Analysis:** Predictive analytics involves using historical data to predict future outcomes. Continuous learning from data improves accuracy over time, enabling fraud detection solutions to stay ahead of emerging threats.

Building a fraud detection system

RapidScale utilizes many AWS services in conjunction with AWS Bedrock to build a robust Generative AI solution for fraud detection. By implementing serverless technologies like AWS Glue, AWS Athena, AWS SageMaker, and AWS Bedrock, RapidScale provides not only technical benefits to fight fraud but also financial incentives such as optimizing cost and scaling policies to prevent waste.

Improving Accuracy:

- **Data Quality:** Ensuring a clean dataset is crucial for building an accurate fraud detection system. Define outliers correctly, eliminate missing values, and curate data by removing noise.
- **Model Selection:** Selecting accurate models like Anthropic Claude 3 is critical. These models are ideal for fraud detection use cases due to their language understanding, reasoning, and task-solving capabilities.
- **Model Deployment:** RapidScale relies on best practices for deploying models, choosing those known for handling false positives effectively. Ensemble techniques like bootstrap aggregation are used to reduce variance and increase accuracy.

Scalability and cost efficiencies

AWS SageMaker and AWS Bedrock are both managed serverless technologies, with differences in their underlying worker nodes. AWS SageMaker allows training models across multiple GPUs and utilizes the cost efficiencies of spot EC2 instances for all batch jobs. On the other hand, AWS Bedrock's infrastructure is fully serverless.

Enhanced security

Network isolation for both AWS SageMaker and AWS Bedrock is performed using VPCs. Architecting with a security-first model, RapidScale employs supporting technologies like AWS PrivateLink, AWS IAM, and AWS KMS to build a fortress around AI and ML workloads in the cloud.

AI and ML rely heavily on good data. Building a well-curated Data Lake is a vital building block for creating a well-architected AI/ML solution that works with precision and accuracy. Once a clean Data Lake is created, choosing the right ML model is the next critical step before training it for an AI solution. Large Language Models like AWS Titan, Anthropic Claude 3, and Meta Llama 3 are among the many machine learning models available on AWS Marketplace that RapidScale uses to implement robust, secure, and scalable AI & ML solutions.

AWS SageMaker and AWS Bedrock offer robust scalability and security features tailored for any fraud detection use case. SageMaker is ideal for end-to-end machine learning workflows, providing extensive support for training, tuning, and deploying models. Bedrock, on the other hand, focuses on generative AI applications, predictive analysis, and prompt engineering use cases, offering seamless access to foundational models and managing the complexities of infrastructure scaling and security.

The journey to building a formidable fraud detection system presents many challenges. RapidScale eliminates these challenges by leveraging the AWS ecosystem to build, scale, and secure machine learning and AI applications.

With RapidScale and AWS, you can build a resilient, secure, and scalable fraud detection system that leverages the power of AI and ML.

JOIN OUR AI/ML ACCELERATION WORKSHOP!

Learn how to harness the full potential of AI and ML for your fraud detection initiatives. Our hands-on workshop will guide you through the process of setting up advanced fraud detection systems, from data lake creation to model deployment.

Secure your spot today and transform your fraud detection capabilities.

REGISTER NOW

rapidscale

ABOUT RAPIDSCALE

RapidScale helps customers migrate, run, and operate mission-critical workloads on AWS with security, scalability, and efficiency baked in. RapidScale's Cloud Reliability Platform combines world-class engineering talent, policy-as-code, and integrated tooling to enable customers to confidently meet compliance regulations, security requirements, cost control, and high availability. Together with our team of dedicated certified engineers and decades of IT management experience, we ensure our customers' success across every stage of the Cloud Adoption Framework.

Speak to a RapidScale cloud expert today!

www.rapidscale.com | meet@rapidscale.com

aws PARTNER Premier Tier Services

aws PARTNER Security Services Competency