GRAAHO

CASE STUDY

Scalable Real-Time Recommendation Engine on AWS

Executive summary

Customer: ALGOREC
Partner: Graaho Technologies LLC
Industry: Technology / SaaS
Location: United States

Challenge: ALGOREC needed to build a hyper personalized recommendation engine capable of delivering real-time, context aware product suggestions while managing fluctuating traffic loads and maintaining cost efficiency.

Solution: Graaho Technologies deployed ALGOREC, an ML-powered recommendation platform built on Amazon SageMaker and scalable AWS infrastructure, enabling intelligent real-time predictions with automated model training and orchestration.

Results:

- 20% increase in sales conversion
- 30% boost in user engagement metrics
- 40% reduction in inference latency
- 25% decrease in operational costs
- Enhanced scalability supporting 10x traffic spikes
- Automated ML lifecycle management
- Enterprise grade security and compliance

Business Challenges

ALGOREC faced several critical challenges in building & scaling their personalization platform:

Operational Challenges

- **Performance:** Need for sub second recommendation delivery during user sessions
- **Scalability Constraints**: Inability to handle traffic spikes during peak shopping periods
- **High Infrastructure Costs:** Traditional ML infrastructure required significant capital investment and maintenance
- **ML Operations:** Time-consuming model training, deployment, and monitoring processes
- **Data Processing Bottlenecks:** Difficulty processing large volumes of behavioural data for model training periods

AlgoRec

About the Customer

ALGOREC is a digital personalization platform that delivers intelligent recommendations to enhance user experiences across various digital touchpoints. As demand for hyper personalized interactions grew, ALGOREC recognized the need to evolve their recommendation engine with advanced machine learning capabilities and scalable infrastructure to meet increasing user expectations and business requirements.

GRAAHO

About the AWS Partner

Graaho Technologies LLC is an AWS Advanced Tier Services Partner specializing in machine learning solutions, data analytics, and cloud-native application development. With deep expertise in Amazon SageMaker, ML-Ops automation, and scalable data pipelines, Graaho helps enterprises transform customer experiences through intelligent, production ready ML platforms. The company focuses on delivering end-to-end solutions that combine cutting-edge machine learning with enterprise grade reliability, security, and operational excellence.

Technical Requirements

- **ML Excellence:** Build sophisticated recommendation algorithms based on collaborative filtering, content based filtering, and deep learning
- **Real-Time Inference**: Deliver personalized recommendations with minimal latency
- **Automated ML Lifecycle:** Implement continuous model training, evaluation, and deployment
- **Data Pipeline Efficiency:** Process streaming and batch data for feature engineering and model training
- **Cost Optimization:** Achieve predictable, usage-based pricing without over-provisioning

Strategic Goals

- Increase customer engagement and sales conversion through better personalization
- Reduce time-to-market for new recommendation features
- Scale infrastructure efficiently to support business growth
- Maintain competitive advantage through continuous ML innovation
- Ensure data security and regulatory compliance

Solution Overview

Graaho Technologies designed and implemented **ALGOREC**, a comprehensive ML-powered recommendation platform leveraging AWS's machine learning and data services.

Architecture Highlights

Machine Learning & AI:

- **Amazon SageMaker:** End-to-end ML platform for model development, training, and deployment
- **SageMaker Endpoints:** Real-time inference with auto-scaling capabilities
- **SageMaker Pipelines:** Automated ML workflows for continuous model improvement
- **SageMaker Feature Store**: Centralized repository for feature engineering and reuse

Data Processing & Storage:

- **Amazon S3:** Scalable data lake for raw and processed data
- **Amazon RDS:** Relational database for user profiles and transactional data
- **AWS Glue:** Serverless ETL for data transformation and preparation
- **Amazon Kinesis:** Real-time data streaming for behavioral events

Compute & API Layer:

- **Amazon EC2:** Flexible compute for Apache Airflow orchestration
- **AWS Lambda:** Serverless functions for event-driven processing
- **Amazon API Gateway:** Secure, scalable API endpoints for recommendation requests
- **Amazon SQS:** Message queuing for asynchronous processing

GRAAHO

Security & Monitoring:

- **Amazon Cognito:** User authentication and authorization
- **AWS IAM:** Fine-grained access control for ML resources
- **Amazon CloudWatch:** Monitoring, logging, and alerting
- **AWS KMS:** Encryption key management for data protection

Orchestration & Automation:

- **Apache Airflow on EC2:** For scheduled model training
- **AWS Step Functions:** Serverless workflow coordination
- **AWS Cloud-Formation:** IaC for consistent deployments

Key Features Implemented

Real-Time Recommendation Engine:

- Context-aware predictions based on user behaviour, preferences, and session data
- Multi-model ensemble approach for improved accuracy
- A/B testing framework for continuous optimization
- Personalized ranking algorithms for diverse content types

Automated ML Pipeline:

- Scheduled data ingestion from multiple sources
- Automated feature engineering and transformation
- Continuous model training with hyper parameter optimization
- Auto model evaluation and deployment with approval gates
- Model monitoring and drift detection

Scalable Infrastructure:

- Auto-scaling SageMaker endpoints based on traffic patterns
- Serverless components for cost-efficient processing
- Multi-region deployment for low-latency global access
- Caching layer for frequently requested recommendations

Implementation Process

Phase 1: Discovery & Design (Weeks 1-3)

- Assessed the existing recommendation system
- Analysed user behaviour data and business requirements
- Designed an AWS Well-Architected–aligned architecture
- Created proof of concept with Amazon SageMaker
- Defined success metrics and KPIs Bedrock SageMaker

Phase 2: Data Infrastructure & Pipeline (Weeks 4-7)

- Established data lake on Amazon S3 with organized data zones
- Implemented real-time data streaming with Amazon Kinesis
- Built ETL pipelines and engineering workflows using AWS Glue
- Established data quality validation processes

Phase 3: ML Model Development & Training (Weeks 8-12)

- Developed collaborative filtering and content-based models
- Built deep learning models for sequential recommendations
- Created ensemble models combining multiple techniques
- Performed hyper parameter tuning and baseline validation

Phase 4: Real-Time Inference Infrastructure (Weeks 13-15)

- Deployed SageMaker endpoints with auto-scaling configuration
- Built API layer with Amazon API Gateway
- Implemented caching and fallback mechanisms
- Conducted load testing for 10x expected traffic

Phase 5: ML Operations Automation (Weeks 16-18)

- Configured Apache Airflow for workflow orchestration
- Implemented SageMaker Pipelines for automated retraining
- Established model monitoring and drift detection
- Built CI/CD pipelines and CloudWatch dashboards

Phase 6: Testing, Optimization & Launch (Weeks 19-20)

- Conducted A/B testing framework validation
- Performed security penetration testing
- Optimized inference latency and cost efficiency
- Executed phased rollout and knowledge transfer

Results & Benefits

Performance Improvements

Inference Latency Reduction:

- Achieved 40% reduction in recommendation response times
- Average latency reduced from 250ms to 150ms
- 95th percentile latency under 300ms
- Consistent performance during peak traffic periods

Scalability Achievement:

- Successfully handling 10x traffic spikes during promotional events
- Auto-scaling responds to demand within 2 minutes
- Support for millions of daily recommendation requests
- No performance degradation during scaling events

Model Accuracy Improvements:

- 15% improvement in recommendation relevance metrics
- 25% increase in click-through rates
- 18% improvement in add-to-cart conversion
- Improved recommendation diversity reducing filter bubble effects

Business Impact

20% Increase in Sales:

- Direct revenue attribution to personalized recommendations
- Higher average order values through intelligent cross-selling
- Improved conversion rates across customer segments
- Enhanced customer lifetime value

30% Boost in User Engagement:

- Increased time spent on platform
- Higher page views per session
- Improved return visitor rates
- Enhanced user satisfaction scores

Customer Experience Enhancement:

- More relevant product discoveries
- Reduced search friction
- Personalized shopping journeys
- Improved mobile experience

Cost Optimization

20% Increase in Sales:

- Direct revenue attribution to personalized recommendations
- Higher average order values through intelligent cross-selling
- Improved conversion rates across customer segments
- Enhanced customer lifetime value

30% Boost in User Engagement:

- Increased time spent on platform
- Higher page views per session
- Improved return visitor rates
- Enhanced user satisfaction scores

Customer Experience Enhancement:

- More relevant product discoveries
- Reduced search friction
- Personalized shopping journeys
- Improved mobile experience

Operational Excellence

Automated ML Lifecycle:

- Model retraining frequency increased from monthly to weekly
- Deployment time reduced from days to hours
- Auto quality gates prevent poor models from reaching production
- Continuous monitoring ensures model health

Development Velocity:

- 50% faster time-to-market for new recommendation features
- Simplified experimentation with SageMaker notebooks
- Reusable ML pipelines accelerate development
- Collaborative environment for data scientists

Reliability & Availability:

- 99.95% uptime for recommendation API
- Automated failover and recovery mechanisms
- Multi-AZ deployment for high availability
- Graceful degradation during service disruptions

GRAAHO

Technical Deep Dive

Machine Learning Architecture

Model Development Pipeline:

- Raw Data → Data Pre-processing → Feature Engineering → Model Training → Model Evaluation → Model Selection → Deployment → Real-Time Inference

ML Models Implemented:

- Implemented collaborative and content-based recommendation models
- Developed hybrid approaches combining multiple techniques
- Applied deep learning models for advanced pattern recognition

Feature Engineering:

- User behavioural features (click history, purchase history, browsing patterns)
- Item features (category, attributes, popularity, ratings)
- Contextual features (time of day, device type, user location)
- Temporal features (seasonality, trends, recency)

Data Pipeline Architecture

Data Ingestion:

- Real-time event streaming from application servers
- Batch data imports from external sources
- Data validation and quality checks
- Deduplication and data cleaning

Data Processing:

- Apache Airflow orchestrating complex workflows
- Scheduled data transformations and aggregations
- Feature computation and storage in SageMaker Feature Store
- Data partitioning for efficient querying

Data Storage:

- Amazon S3 data lake with Parquet format for efficient storage
- Amazon RDS for transactional data
- ElastiCache for frequently accessed features
- Lifecycle policies for cost optimization

Real-Time Inference System

- **Inference Architecture:** User Request → API Gateway → Load Balancer → SageMaker Endpoints → Feature Retrieval → Model Inference → Response Caching → User Respons

AWS Services Used

AI & Machine Learning:

- Amazon Bedrock (AgentCore)
- Amazon SageMaker Autopilot
- Amazon SageMaker Feature Store

Compute:

- Amazon EC2
- AWS Lambda

Database & Storage

- Amazon S3
- Amazon RDS
- Amazon ElastiCache

Analytics & Data Processing

- Apache Airflow on EC2
- Amazon Athena

Networking & Content Delivery

- Amazon API Gateway
- Amazon VPC
- Amazon CloudFront

Developer Tools

- AWS CodePipeline
- AWS CodeBuild
- AWS CodeDeploy
- AWS CloudFormation

Security, Identity & Compliance

- AWS IAM
- Amazon Cognito
- AWS KMS
- AWS CloudTrail
- AWS Secrets Manager

Management & Governance

- Amazon CloudWatch
- AWS Systems Manager
- AWS Config

GRAAHO

Technical Deep Dive

Performance Optimization:

- Model endpoint auto-scaling based on request volume
- Request batching for improved throughput
- Response caching for frequently requested recommendations
- Asynchronous processing for non-critical requests using SQS

Monitoring & Observability:

- Real-time performance metrics in CloudWatch
- Model performance tracking and drift detection
- Request latency and error rate monitoring
- Custom dashboards for operational visibility

DevOps & Automation

- **CI/CD Pipeline:** Code Commit → CodeBuild (Test) → Security Scanning → CodeDeploy (Staging) → Integration Testing → Approval Gate → CodeDeploy (Production) → Monitoring

Infrastructure as Code:

- Used AWS CloudFormation with version-controlled infrastructure templates
- Automated environment provisioning processes
- Ensured consistent configurations across all environments

Model Deployment:

- Automated model evaluation, validation and rollback on performance degradation
- Canary deployments for safe model updates
- A/B testing framework for model comparison

Security Implementation

Data Protection:

- Encryption at rest using AWS KMS and in transit using TLS 1.3
- Secure credential management with AWS Secrets Manager
- Regular backup and disaster recovery testing

Access Control:

- Implemented least-privilege IAM policies and role-based access control
- Enforced multi-factor authentication for sensitive operations
- Enabled comprehensive audit logging for access and changes

Compliance & Governance:

- Enabled CloudTrail logging and AWS Config for audit and compliance tracking
- Conducted regular security assessments and penetration testing
- Maintained comprehensive compliance documentation and reporting

Customer Testimonial

— CTO, ALGOREC

"Graaho Technologies has revolutionized our recommendation capabilities with their expertly designed ML platform on AWS. The combination of Amazon SageMaker's powerful ML tools and their deep expertise in MLOps has delivered exceptional results. We've seen a 20% increase in sales and 30% boost in engagement while reducing our operational costs by 25%. The automated model training and deployment pipelines have transformed how quickly we can innovate and respond to changing customer preferences. The platform's scalability has been remarkable—handling our Black Friday traffic without any issues. This partnership has been instrumental in establishing our competitive advantage in the e-commerce personalization space."

GRAAHO

Best Practices Implemented

AWS Well-Architected Framework

Operational Excellence:

- Infrastructure as Code for all resources
- Automated deployment pipelines with comprehensive testing
- Comprehensive monitoring and alerting with CloudWatch
- Runbook automation for common operational scenarios
- Regular operational reviews and optimization

Security:

- Defense in depth with multiple security layers
- Principle of least privilege for all access controls
- Automated security scanning in CI/CD pipelines
- Regular security assessments and vulnerability management
- Encryption for data at rest and in transit

Reliability:

- Implemented multi-AZ deployments with automated backup and recovery
- Applied circuit breakers, retry logic, and health checks for resilience
- Enabled automated failover and chaos engineering for reliability testing

Performance Efficiency:

- Right-sized compute instances based on workload analysis
- Auto-scaling for dynamic resource allocation
- Efficient data storage with appropriate formats and partitioning
- Caching strategies for frequently accessed data
- Regular performance optimization and tuning

Cost Optimization:

- Used Reserved and Spot Instances for cost-efficient compute utilization
- Automated resource clean-up and lifecycle management
- Conducted regular cost optimization reviews and recommendations
- Applied data storage lifecycle policies for efficient storage management

Sustainability:

- Efficient resource utilization with auto-scaling
- Spot instances reduce carbon footprint
- Serverless architecture minimizes idle resources
- Optimized model inference for energy efficiency

Lessons Learned

Technical Insights:

- Optimized SageMaker instances and endpoints for cost and performance
- Built robust feature stores for faster, consistent model development
- Implemented multi-tier caching to cut inference costs
- Monitored models continuously to prevent drift

Process Improvements:

- Automated retraining with validation to keep recommendations relevant
- Phased A/B rollout validated changes before full deployment
- Regular cross-team syncs ensured alignment
- Early documentation accelerated knowledge transfer

Partnership Success Factors:

- Defined KPIs upfront for clear success measurement
- Two-week sprints with demos maintained momentum
- Followed AWS Well-Architected Framework to reduce technical debt
- Conducted hands-on training for team independence

Future Roadmap

Graaho Technologies and ALGOREC are planning several enhancements to further leverage AWS capabilities:

Short-term (3-6 months)

- Implement Amazon Personalize for additional recommendation algorithms
- Add real-time feature computation using Amazon Kinesis Data Analytics
- Expand A/B testing framework with multi-armed bandit algorithms
- Integrate Amazon QuickSight for business intelligence dashboards

Medium-term (6-12 months)

- Deploy deep learning models using SageMaker with GPU instances
- Implement explainable AI features for recommendation transparency
- Add computer vision capabilities using Amazon Rekognition for image-based recommendations
- Expand to additional markets with multi-region deployment

Long-term (12+ months)

- Explore reinforcement learning for dynamic recommendation optimization
- Implement federated learning for privacy-preserving personalization
- Develop industry-specific recommendation templates
- Create AWS Marketplace offering for rapid customer on boarding

Contact Information

To learn more about this solution:

Graaho Technologies LLC

❑ Website: www.graaho.com

❑ Email: aws@graaho.com

❑ AWS Partner Profile: Graaho Technologies LLC

AWS Partner Team

❑ For partnership inquiries: +1 703 936 9360

Additional Resources

❑ Amazon SageMaker Documentation
❑ ML Lens - AWS Well-Architected Framework
❑ MLOps Best Practices on AWS
❑ AWS Architecture Center - ML Solutions
❑ Amazon SageMaker Examples Repository

Document Version: 1.0

Last Updated: 17-02-2026

Classification: Public

Conclusion

The partnership between Graaho Technologies LLC and ALGOREC demonstrates how AWS's machine learning and analytics services can transform digital personalization at scale. By leveraging Amazon SageMaker, serverless architectures, and comprehensive DevOps automation, Graaho delivered a production-grade ML platform that exceeds performance expectations while significantly reducing costs.

The solution's success—characterized by 20% sales increase, 30% engagement boost, 40% latency reduction, and 25% cost savings—showcases how modern cloud-native ML architectures deliver both technical excellence and measurable business value. The fully automated MLOps pipelines enable continuous innovation, allowing ALGOREC to maintain its competitive edge in the rapidly evolving personalization landscape.

This case study exemplifies AWS's commitment to empowering partners and customers to build intelligent, scalable, and cost-effective solutions that drive digital transformation and enhance customer experiences through data-driven personalization.