DuploCloud

BLACKBIRD.AI

Narrative Intelligence at Production Scale — 396 AI Containers Running Securely on AWS

How Blackbird.AI deployed a production-grade AI platform in under 90 days using DuploCloud's DevSecOps automation on AWS.

396

AI containers in production

17

Isolated tenants

<90

Days to production

Zero

Critical incidents

At Glance

The foundation behind Blackbird.AI's production AI platform

Challenge

Running GPU-accelerated AI at government-grade security — with a lean engineering team

Solution

DuploCloud DevSecOps platform + dedicated engineering across a GPU-accelerated AI stack on AWS

Architecture Overview

Blackbird.AI — AWS Gen AI platform architecture via DuploCloud

Results

A production-grade AI platform — fast, secure, and built to scale

At a Glance

The foundation behind Blackbird.AI's production AI platform

COMPANY

Blackbird.AI

INDUSTRY

AI / Narrative Intelligence & Security

CLOUD

US East, US West (3 accounts)

ZERO CRITICAL INCIDENTS

Since production launch (June 2025)

CUSTOMER SINCE

October 2024

DUPLOCLOUD SERVICES

DevSecOps Platform; EKS Cluster Management; Advanced Observability Suite (AOS); Dedicated Engineering; Standard Support

Production AI platform live in under 90 days — March to June 2025

396 containers across 17 security-isolated tenants with zero critical incidents

EKS upgraded from v1.32 > v1.33 with AL2023 AMIs — zero downtime in production

GPU-accelerated AI inference workloads (CUDA) managed on dedicated tenant infrastructure

Vanta compliance posture maintained alongside AI workload deployment

DuploCloud Advanced Observability Suite (AOS) deployed to staging — production rollout in progress

The Challenge

Running GPU-accelerated AI at government-grade security — with a lean engineering team

Blackbird.AI defends truth at scale — their platform analyzes billions of signals across social media, news, and communications to detect disinformation campaigns and influence operations in real time. Delivering this mission requires running complex, GPU-accelerated AI inference pipelines, large-language model serving, and graph-analytics workloads across multiple AWS environments simultaneously, while maintaining the strict data isolation and compliance posture demanded by government and enterprise security customers.

With a lean engineering team responsible for both advancing core AI models and keeping infrastructure healthy, Blackbird.AI needed a platform that could absorb the operational weight of production AWS management — multi-account governance, EKS lifecycle, GPU node management, secrets handling, and continuous compliance — without diverting engineers from their core mission.

Blackbird.AI needed an infrastructure partner that could stand up a production-grade, GPU-accelerated AI platform on AWS — fast — while enforcing the security isolation and compliance posture required to serve government and enterprise intelligence customers.

The Solution

DuploCloud DevSecOps platform + dedicated engineering across a GPU-accelerated AI stack on AWS

Blackbird.AI's platform is powered by **RAV3N Risk LLM** — a proprietary multimodal, multilingual model purpose-built for narrative intelligence, threat detection, and risk analysis. Unlike general-purpose LLMs, RAV3N is engineered to process and reason across multilingual content at scale — analyzing text, images, audio, and video from global intelligence sources to identify coordinated disinformation campaigns, influence operations, and emerging reputational threats. The platform ingests signals across social media, news, web, forums, dark web, images, audio, and video through a multi-source intelligence pipeline — with GPU-accelerated processing on AWS enabling real-time analysis across billions of signals.

The GenAI architecture is built around retrieval-augmented contextual analysis — grounding AI-driven risk scoring and threat detection in real-time external content rather than relying on static model knowledge. Semantic retrieval and narrative clustering identify coordinated inauthentic behavior and emerging reputational risks across sources, while RAG-based reasoning combines live ingested signals with AI analysis to surface actionable intelligence.
The **Constellation** platform serves as the custom orchestration layer — coordinating automated reporting, analyst investigation workflows, and executive intelligence dashboards with no reliance on third-party orchestration frameworks such as LangChain or LlamaIndex.

The business outcomes span the full enterprise risk and intelligence lifecycle: faster identification of misinformation, disinformation, and coordinated narrative attacks; improved executive situational awareness through automated intelligence summaries and risk prioritization; reduced manual analyst workload via AI-assisted monitoring, investigation, and reporting workflows; better brand and reputational protection through earlier detection and response capabilities; and enhanced risk management across cyber, social, geopolitical, and information-threat environments. DuploCloud provides the fully managed AWS infrastructure that makes all of this possible — handling EKS lifecycle, GPU node management, multi-tenant isolation, and compliance controls so Blackbird.AI's engineering team can stay focused on advancing RAV3N and the Constellation platform.

AMAZON EKS | EC2 GPU INSTANCES | AWS SECRETS MANAGER | AMAZON ELASTICACHE | AMAZON RDS

AMAZON S3 | AMAZON CLOUDWATCH | AWS LAMBDA | AWS INSPECTOR | AMAZON SQS

AWS IAM | AMAZON ECR

GPU-ACCELERATED AI WORKLOADS

DuploCloud orchestrates CUDA-accelerated GPU nodes on Amazon EKS for embedding generation and AI inference. Dedicated tenant isolation (prd-graph) ensures model-level security separation and resource governance for sensitive AI workloads processing intelligence data.

SECURE AI DEPLOYMENT PIPELINE

GitHub Actions and Terraform workflows run through DuploCloud JIT credentials, enabling automated CI/CD for AI model deployment across environments — without exposing long-lived AWS credentials or secrets to the pipeline at any stage.

MULTI-TENANT ISOLATION

17 isolated DuploCloud tenants across dev, staging, and production provide separate networking, IAM boundaries, and resource quotas — critical for AI workloads processing sensitive government and enterprise intelligence data.

SECRETS & CREDENTIAL MANAGEMENT

AWS Secrets Manager integration via DuploCloud secures AI model API keys, database credentials, and third-party service tokens — with automated rotation and audit trails meeting enterprise compliance requirements alongside Okta SSO.

OBSERVABILITY FOR AI WORKLOADS

OpenTelemetry, CloudWatch, and Prometheus are configured out of the box, giving the team full visibility into AI inference latency, GPU utilization, pod health, and cost per tenant — without manual instrumentation overhead.

ADVANCED OBSERVABILITY SUITE (AOS)

DuploCloud's AOS delivers natural-language Kubernetes management, automated health monitoring, and intelligent anomaly detection — reducing the operational burden on Blackbird.AI's lean engineering team as the AI platform scales.

Architecture Overview

Blackbird.AI — AWS Gen AI platform architecture via DuploCloud

DATA INGESTION

Social Media APIs | News & Web | Forums & Dark Web | Image / Audio / Video | Amazon SQS

AWS Lambda Preprocessors

AI PROCESSING

RAV3N Risk LLM (Proprietary) | Multimodal / Multilingual Inference | Semantic Retrieval & RAG

Narrative Clustering Engine | Constellation (Orchestration) | Graph Analytics Engine | EC2 GPU Nodes (CUDA)

EKS TENANTS (17 TOTAL)

prd-graph (GPU / AI) | prd-es (Search) | prd-vault (Secrets) | Staging & Dev Clusters | 13 Additional Tenants

DATA LAYER

Amazon S3 (Artifacts) | Amazon RDS | Amazon ElastiCache | AWS Secrets Manager

SECURITY, COMPLIANCE & OBSERVABILITY

DuploCloud DevSecOps Platform | Amazon CloudWatch | OpenTelemetry + Prometheus | AWS Inspector

Vanta (Compliance) | DuploCloud AOS

DEVELOPER EXPERIENCE & CI/CD

GitHub Actions + Terraform | DuploCloud JIT Credentials | Okta SSO | Retool Dashboards | Amazon ECR

The Results

A production-grade AI platform — fast, secure, and built to scale

Production AI Platform in Under 90 Days

From infrastructure setup in October 2024 to pre-production in March 2025 and full production cutover in June 2025, Blackbird.AI launched a production-grade AI platform with 396 containers running across 17 isolated tenants — well within typical enterprise timelines for a build of this complexity.

<90 Days

Zero-Downtime EKS Upgrades Across All Environments

Both non-production and production EKS clusters were upgraded from Kubernetes v1.32 to v1.33 with AL2023 AMIs during coordinated maintenance windows. DuploCloud engineers provided real-time support throughout, ensuring AI workloads experienced no service disruption.

0 Downtime

Zero Critical Incidents Since Platform Launch

With DuploCloud's automated health monitoring, proactive alerting, and same-day engineering response, Blackbird.AI's production AI platform has maintained zero critical incidents since go-live — meeting the availability demands of government and enterprise security customers.

Zero SEV1/2

Compliance-Ready Infrastructure for Enterprise AI

DuploCloud's built-in security controls, Vanta integration, and automated policy enforcement give Blackbird.AI the compliance posture required to serve security-sensitive customers — without dedicating engineering cycles to manual audit preparation or infrastructure hardening.

Vanta ✓

AWS Gen AI Competency Highlights

DuploCloud's proven capability across the four dimensions of AWS Gen AI Competency — demonstrated through Blackbird.AI's production AI infrastructure

RAV3N Risk LLM — Proprietary Multimodal Narrative Intelligence

- RAV3N Risk LLM is a proprietary multimodal, multilingual model for threat detection, disinformation analysis, and narrative risk scoring.
- Ingests signals across social media, news, web, forums, dark web, image, audio, and video for comprehensive coverage.
- Constellation platform orchestrates automated reporting, analyst workflows, and executive intelligence dashboards.
- CUDA-accelerated GPU nodes on Amazon EKS (prd-graph tenant) power real-time RAV3N inference at production scale.
- DuploCloud maintains environment parity across dev, staging, and production for reliable model promotion.

Data Privacy & Security for Sensitive AI Workloads

- 17 tenant-isolated environments prevent cross-contamination of intelligence data across customer workloads.
- AWS Secrets Manager handles secure rotation of model API keys and service credentials.
- DuploCloud JIT credentials + Okta SSO eliminate long-lived access tokens from CI/CD pipelines.
- Vanta + AWS Inspector provide continuous compliance automation and vulnerability management.

Scalability, Optimization & Responsible AI Infrastructure

- Zero-downtime EKS upgrades (v1.32 > v1.33) with pod disruption budgets protecting live AI workloads.
- ElastiCache provides low-latency caching for AI inference results — reducing repeat GPU compute costs.
- OpenTelemetry + CloudWatch provide full observability into inference latency, GPU utilization, and drift.
- DuploCloud AOS delivers automated anomaly detection — giving early warning of AI workload behavior deviations.