



Case Study

AI-POWERED DOCUMENT INTELLIGENCE FOR AIRWORKS

At a Glance

AirWorks, an aviation and MRO (Maintenance, Repair & Overhaul) organization, partnered with AceCloud to build a secure, AI-powered document query platform on AWS. The new system lets employees search and ask natural-language questions across thousands of pages of aviation, MRO, finance, safety, and operational documentation and receive relevant, citation-backed answers in seconds while keeping sensitive records within a governed, auditable environment.

The platform combines Retrieval-Augmented Generation (RAG) with a production-grade cloud foundation, including automated deployments, continuous monitoring, layered security controls, and a tested disaster-recovery plan, making the AI capability as operationally reliable as it is useful.

- P95 response time below 8 seconds for natural-language queries across the document library.
- 99.5% monthly availability objective, continuously tracked through automated synthetic monitoring.

- New documents made searchable within 5 minutes of upload through an automated ingestion pipeline.
- AI responses include source citations, allowing users to verify answers against the underlying documents.
- Zero open high- or critical-severity security findings at go-live, verified against a CIS-aligned security baseline.
- Automated blue/green releases designed to avoid planned deployment downtime, with failed changes rolled back automatically.

The Client

AirWorks operates in the aviation and MRO space, managing large volumes of technical, financial, safety, and operational documentation. Much of this material is sensitive by nature, including financial records, safety-critical MRO documentation, and compliance-related content, which makes both fast access and strict control over who can see what equally important.

As the organization grew, so did its library of aviation manuals, MRO records, financial reports, and safety documentation, along with the number of employees and stakeholders who needed fast, reliable answers drawn from that material.

The Challenge

Finding the right information inside a large, fast-growing document library was a manual, time-consuming process. Employees needed to know which document to open and where to look inside it, and there was no way to ask a plain-language question and get a direct, sourced answer.

This was made harder by the sensitivity of the content. Financial and MRO records could not simply be made broadly searchable without strict access control, strong authentication, and a complete audit trail of who accessed what and when.

Any AI-powered search layer also needed to be designed around AirWorks' own documents, rather than relying on unsupported model knowledge. Source citations were essential so employees could verify generated answers before using the information in operational, safety-related or compliance-related contexts.

AirWorks needed a solution that combined the speed and convenience of conversational AI search with the rigor of an auditable, access-controlled enterprise system.

The Solution

ceCloud, delivering as an AWS partner under Real Time Data Services, designed and deployed a serverless, event-driven Retrieval-Augmented Generation (RAG) platform purpose-built for AirWorks' document library.

Documents are uploaded to a dedicated Amazon S3 bucket, which triggers an automated ingestion pipeline: a Lambda function reads each document, splits it into chunks, and generates vector embeddings using Amazon Bedrock's Titan Embedding model. Those embeddings are indexed in Amazon OpenSearch Service, giving the platform fast, semantic (meaning-based, not just keyword) search across the entire document library.

On the query side, users authenticate through Amazon Cognito with multi-factor authentication before submitting a natural-language question. The platform converts the question into an embedding, retrieves the most relevant document chunks from OpenSearch, and passes that retrieved context, together with the original question, to Amazon Bedrock's Claude 3 Sonnet model, which generates a grounded, citation-backed answer. Every query and response is logged to Amazon DynamoDB and exported to Amazon S3 for long-term, auditable retention.

A lightweight analytics layer sits on top of this. Amazon Athena makes the query logs available for SQL-based compliance reporting, while Amazon QuickSight dashboards give AirWorks stakeholders visibility into usage volume, response quality and cost without needing to access the underlying infrastructure.

- **A Governed, Well-Architected Cloud Foundation**

The entire workload runs inside a dedicated production VPC spanning two Availability Zones, with public and private subnets cleanly separating the internet-facing application tier from backend data services. Every AWS service in the pipeline is reached through a tightly scoped, least-privilege IAM role, so a function can only touch the specific resources it needs.

Deployments are fully automated and reviewed. Application changes move through a CI/CD pipeline that builds, tests, and stages each release before anything reaches production. The web application tier uses a blue/green deployment model: a parallel environment is stood up, validated against a full suite of automated smoke tests, and only then does live traffic shift over, with the previous version kept warm in case a fast rollback is needed. The AI query and ingestion functions use a similar approach, shifting traffic gradually to new versions and automatically reverting if monitoring detects a problem during the rollout, with no manual intervention required.

- **Continuous Monitoring and Rapid Incident Response**

A central operations dashboard tracks query latency, error rates, search cluster health, database throttling, authentication activity, and infrastructure utilization in real time. A prioritized alerting system automatically flags issues ranging from elevated response times and search cluster outages to spikes in failed sign-in attempts that could indicate credential stuffing. Each alert is routed to the appropriate response playbook, with clear escalation paths and defined response-time targets based on severity.

Every layer of the platform is instrumented. Application logs, infrastructure logs and a complete API audit trail are retained according to a defined retention schedule. Anomaly detection covering both system behavior and cloud spend flags unusual patterns before they become incidents.

- **Continuous Monitoring and Rapid Incident Response**

The platform is built on a CIS-aligned security baseline: account-wide audit logging, continuous configuration compliance checks, threat detection, enforced multi-factor authentication, and encryption at rest and in transit across every data store. No component of the pipeline is exposed to the public

internet beyond the application's own front door, and human users never hold long-lived AWS credentials.

On top of that baseline, the platform adds controls specific to an AI-driven document system, including strict input validation to defend against prompt-injection attempts, a model configuration that grounds every response strictly in retrieved AirWorks documents rather than the model's general knowledge and full logging of every AI model invocation, including input and output, for audit and quality review.

- **Resilience and Disaster Recovery**

The workload is classified and protected as a business-critical (Tier 2) system, with a defined recovery time objective of 4 hours and a recovery point objective of 1 hour. Backups run continuously and automatically: point-in-time recovery and daily snapshots for the database layer, versioning and cross-region replication for stored documents and audit logs, and regular snapshots for the search index and application servers. The disaster recovery plan is tested on a regular cycle rather than left untested until it's needed.

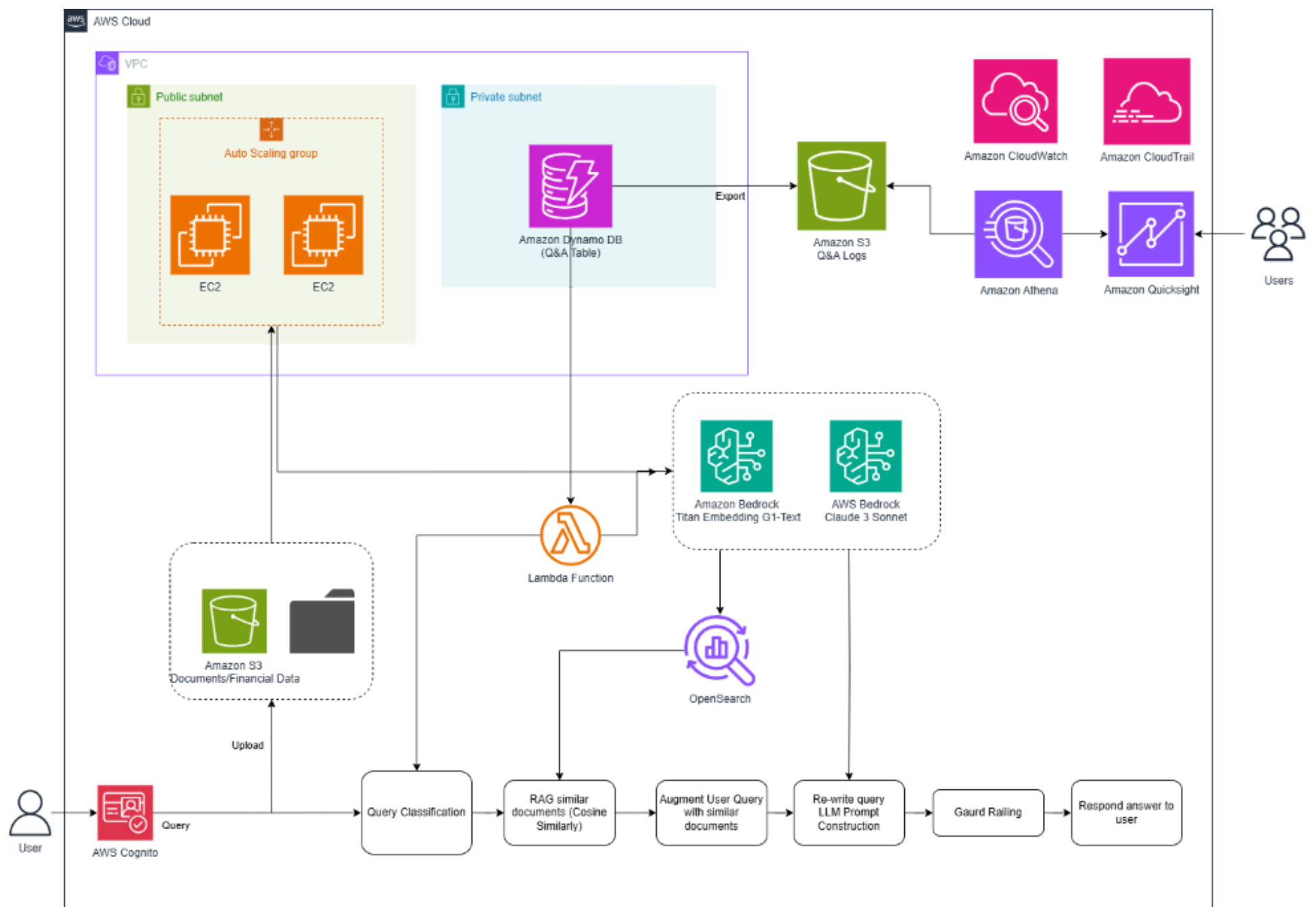
- **Responsible AI by Design**

Because the platform touches safety, financial, and compliance-sensitive material, responsible AI practices are built into the architecture rather than bolted on afterward. Retrieval-Augmented Generation was chosen specifically over fine-tuning an AI model on AirWorks' data, so proprietary information stays inside AirWorks-controlled storage rather than being absorbed into a model. Every response carries source citations so a human can verify it against the original document. Automated safeguards screen for potential exposure of personal information and inappropriate content. The system is explicitly designed to inform, not decide, with every safety-critical or MRO decision remaining with AirWorks personnel.

A simple feedback loop supports continuous improvement. Users can rate individual responses, and those ratings feed into regular quality reviews so any decline in answer quality can be identified and addressed quickly.

Architectural Diagram

The diagram below shows the full platform: document ingestion and embedding on the left, the conversational query flow along the bottom, and the monitoring, analytics, and audit layer on the right.



Results and Benefits

- **Faster Answers:** employees get direct, sourced answers to natural-language questions instead of manually searching through documents.
- **Trustworthy AI:** Every answer is grounded in AirWorks' own documentation and comes with citations, so users can verify what the AI tells them.
- **Enterprise-Grade Security:** Multi-factor authentication, encryption, least-privilege access, and a full audit trail protect sensitive financial and MRO records.
- **Operational Resilience:** Automated blue/green deployments and Lambda traffic-shifting mean new features and fixes ship without risking downtime, with automatic rollback if something goes wrong.
- **Proactive Monitoring:** A prioritized alerting system and central dashboard catch issues, from latency spikes to security anomalies, before they affect users.
- **Business Continuity:** Consistent resource tagging and automated cost-anomaly detection keep cloud spend transparent and predictable.
- **Cost Visibility:** A tested disaster-recovery plan and continuous, cross-region backups protect against data loss and extended outages.
- **Responsible AI Governance:** Citation-backed, grounded responses, human oversight, and a continuous feedback loop keep the AI system accountable.

AWS Services Used

The platform draws on a broad, purpose-fit set of AWS services across ingestion, AI, application, security, and operations:



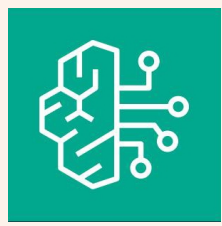
Amazon S3



**AWS
Lambda**



**Amazon
OpenSearch Service**



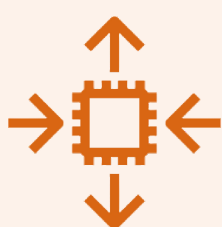
**Amazon Bedrock
(Titan Embedding &
Claude 3 Sonnet)**



**Amazon
DynamoDB**



**Amazon
Cognito**



**Amazon EC2
(Auto Scaling)**



**Amazon Athena and
Amazon QuickSight**



**Amazon CloudWatch
and AWS CloudTrail**



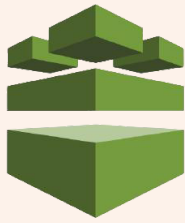
**Amazon
VPC**



AWS IAM



**AWS CodePipeline,
CodeBuild, CodeDeploy,
and CodeCommit**



AWS Config, AWS Security Hub, and Amazon GuardDuty



AWS Backup



AWS Secrets Manager and AWS Systems Manager Parameter Store

Conclusion

By combining Retrieval-Augmented Generation with a disciplined, well-architected AWS foundation, AceCloud gave AirWorks a fast, auditable AI document assistant that employees can use with confidence without compromising the security and reliability required for sensitive aviation, MRO and financial data.

With automated deployments, continuous monitoring, tested disaster recovery and responsible AI governance built in from the start, AirWorks now has a platform that can keep pace with its growing document library and expanding user base safely and sustainably.



Contact Us

Are you ready to cut down your cloud spending? Contact AceCloud Experts today to get a customized quote.



awspartner@acecloud.ai



<https://acecloud.ai/aws/>

Our Location

**2069 Oneill Dr, Bethel
Park, Pennsylvania,
15102-6602**

**2637 E Atlantic Blvd
#23519 Pompano Beach,
FL 33062, Florida**

**Ace Tower, 809-A, Udyog
Vihar, Phase 5, Gurugram
122016, Haryana, India**

About Us



Years of Exp

15+



Data Center

10+



Awards

100+

AceCloud is a leading provider of end-to-end cloud computing solutions to global organizations across industries at scale. It offers a full spectrum of cloud services, including Public Cloud, Application Hosting, AWS Services, Managed Security Services, and Hosted Virtual Desktop Solutions.

AceCloud is a brand of Real Time Data Services (RTDS) group of companies – a leading provider of information technology capabilities specializing in Cloud Computing and Cloud Telephony. RTDS Group has an employee base of 600+ across India, the US, and the UK, supporting over 20,000+ customers and IT infrastructure in 10+ data centers spanning the globe. It offers industry-leading technological solutions that help customers streamline their operations and enhance efficiency.

As an advanced consulting and certified AWS Well-Architected Partner, AceCloud recently achieved the prestigious AWS Storage Competency and is also recognized as an Amazon RDS Service Delivery Partner. These achievements underscore its deep expertise in AWS technologies and commitment to delivering high-quality cloud solutions.



- Public Sector
- Amazon RDS Delivery
- Storage Services Competency
- Well-Architected Partner Program