Clearscale

Case Study

Solving the Challenge of Large-Scale Document Workflows

Clearscale implemented an automated OCR pipeline on AWS that improved efficiency, resilience, and scalability.

Client Profile

Industry: SaaS / Content Management

Technology: Migration and Modernization

Overview

A technology enterprise needed a way to manage and transform large volumes of documents in its SaaS platform. Clearscale designed and implemented an event-driven OCR pipeline on AWS that automated multi-stage processing, delivered scalability, and ensured resilience. The solution not only improved efficiency but also enhanced product value by making documents searchable and laying the foundation for future document intelligence.

Meet Our Hero:

This global software provider helps enterprises manage and access their content at scale. As customer needs grew, the organization struggled with orchestrating a multi-stage process for handling massive volumes of documents.

The lack of automation slowed operations, while limited resilience created risks around performance and customer experience. The company needed a cloud-native solution that could reliably process diverse file types, make documents searchable, and deliver results at scale

The Goal

- Reliably orchestrate an OCR pipeline on AWS
- Handle massive volumes of documents with automation
- Make documents searchable to increase usability
- Ensure resilience with fault-tolerant workflows
- Enhance the end-user experience

The Challenge

Challenge 01

Needed to handle diverse file types in document processing

Challenge 02

Multi-stage workflows were complex and hard to manage

Challenge 03

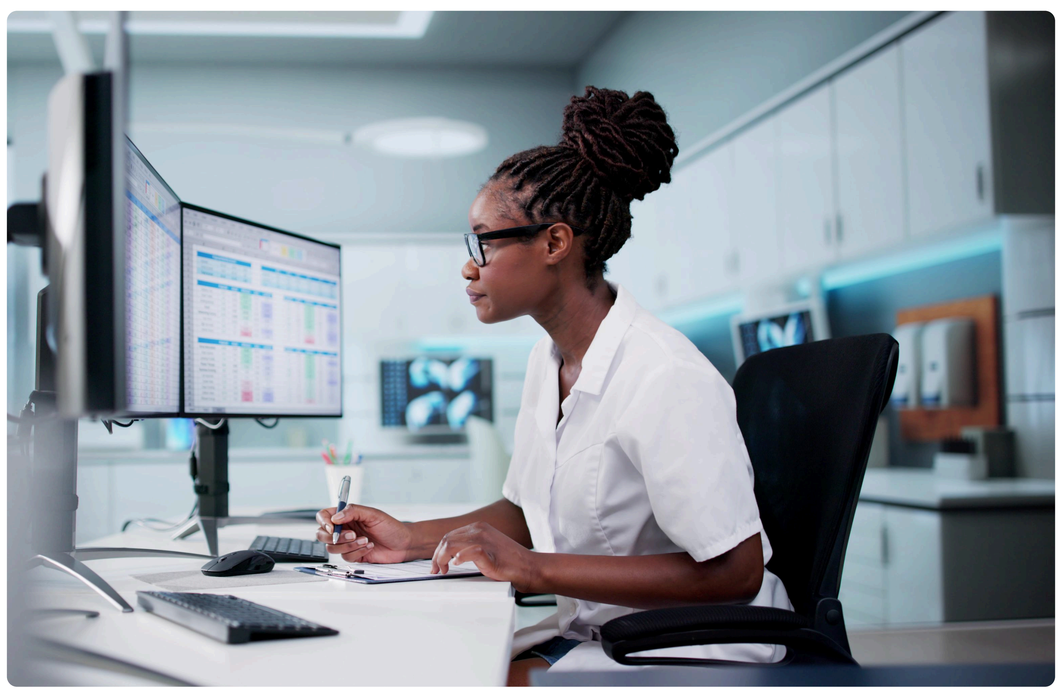
Scalability challenges limited the ability to handle large volumes

Challenge 04

State management was difficult in a decoupled architecture

Challenge 05

End-user experience suffered due to delays and inefficiencies



The Solution

Clearscale implemented a fully automated OCR pipeline on AWS using decoupled microservices.

Step 01: Event-Driven Architecture

- Built the pipeline around AWS services for orchestration and resilience
- Used SQS queues, S3 buckets, and DynamoDB counters to manage state

Step 02: Multi-Stage Processing

- Applied Apache Tika for text extraction
- Leveraged Tesseract for optical character recognition (OCR)
- Added a text overlay service to enrich documents
- Assembled files through a dedicated service to complete the pipeline

Step 03: Automation and Resilience

- Deployed services on Amazon ECS for container orchestration
- Implemented error handling and DLQs to ensure fault tolerance
- Used IAM and AppConfig for secure, configurable operations

Outcomes & Impact

Searchable documents, increasing the platform’s product value

Massive scalability, handling large volumes of files efficiently

High resilience, with DLQs and automated recovery in place

Improved efficiency, streamlining multi-stage document processing

Future-ready architecture, providing a foundation for advanced document intelligence features

Turn Cloud Chaos Into Clear Results On AWS

Clearscale helps technology enterprises break free from cloud chaos and experience clear results on AWSIf your workflows are holding back efficiency and innovation, let’s talk.

Talk to an Engineer

aws PARTNER Premier Tier Services