Artificial Intelligence

Accelerate intelligent document processing with generative AI on AWS

by Bob Strahan, Vamsi Thilak Gudi, David Kaleko, Joe King, Mofijul Islam, Rafal Pawlaszek, Spencer Romo, and Vincil Bishop | on 22 AUG 2025 | in Amazon Bedrock Data Automation, Generative AI | Permalink | Comments | Share

Every day, organizations process millions of documents, including invoices, contracts, insurance claims, medical records, and financial statements. Despite the critical role these documents play, an estimated 80–90% of the data they contain is unstructured and largely untapped, hiding valuable insights that could transform business outcomes. Despite advances in technology, many organizations still rely on manual data entry, spending countless hours extracting information from PDFs, scanned images, and forms. This manual approach is time-consuming, error-prone, and prevents organizations from scaling their operations and responding quickly to business demands.

Although generative AI has made it easier to build proof-of-concept document processing solutions, the journey from proof of concept to production remains fraught with challenges. Organizations often find themselves rebuilding from scratch when they discover their prototype can't handle production volumes, lacks proper error handling, doesn't scale cost-effectively, or fails to meet enterprise security and compliance requirements. What works in a demo with a handful of documents often breaks down when processing thousands of documents daily in a production environment.

In this post, we introduce our open source GenAI IDP Accelerator—a tested solution that we use to help customers across industries address their document processing challenges. Automated document processing workflows accurately extract structured information from documents, reducing manual effort. We will show you how this ready-to-deploy solution can help you build those workflows with generative AI on AWS in days instead of months.

Understanding intelligent document processing

Intelligent document processing (IDP) encompasses the technologies and techniques used to extract and process data from various document types. Common IDP tasks include:

- **OCR (Optical Character Recognition)** – Converting scanned documents and images into machine-readable text
- **Document classification** – Automatically identifying document types (such as invoices, contracts, or forms)
- **Data extraction** – Pulling structured information from unstructured documents
- **Assessment** – Evaluating the quality and confidence of extracted data
- **Summarization** – Creating concise summaries of document content
- **Evaluation** – Measuring accuracy and performance against expected outcomes

These capabilities are critical across industries. In financial services, organizations use IDP to process loan applications, extract data from bank statements, and validate insurance claims. Healthcare providers rely on IDP to extract patient information from medical records, process insurance forms, and handle lab results efficiently. Manufacturing and logistics companies use IDP to process invoices and purchase orders, extract shipping information, and handle quality certificates. Government agencies use IDP to process citizen applications, extract data from tax forms, manage permits and licenses, and enforce regulatory compliance.

The generative AI revolution in IDP

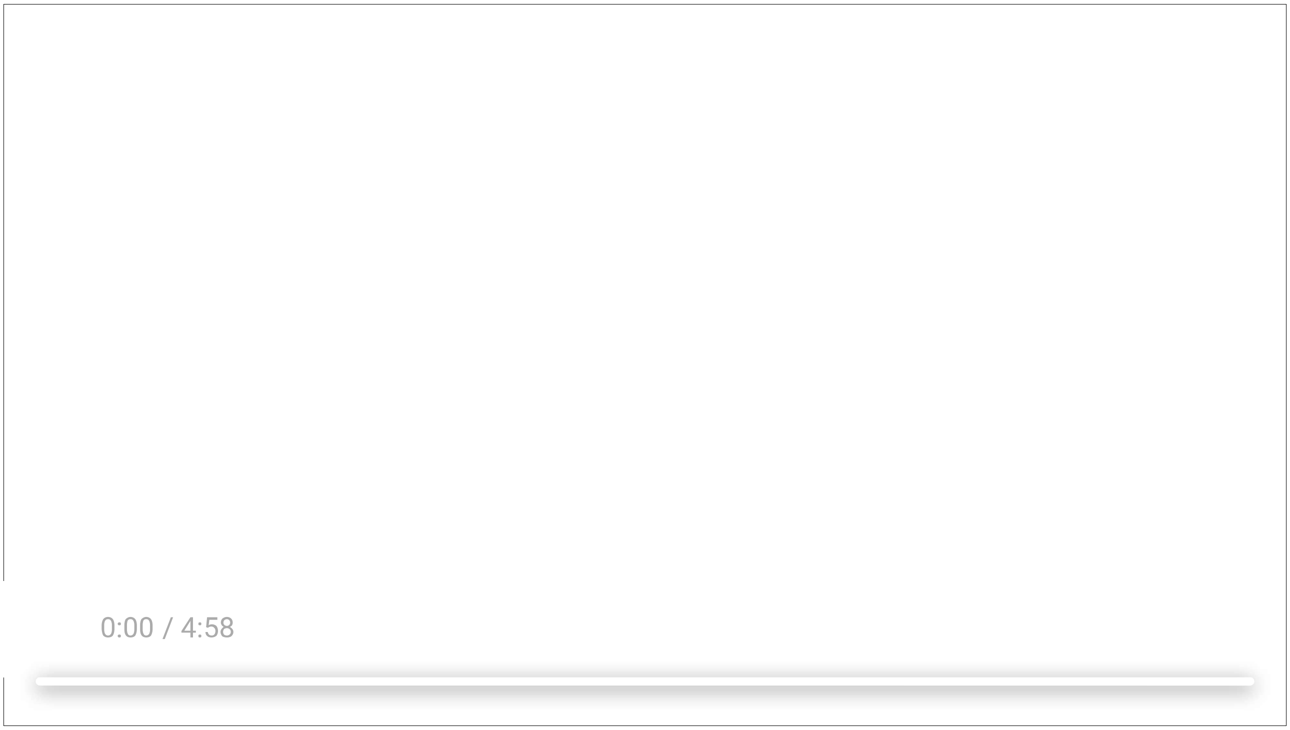
Traditional IDP solutions relied on template-based extraction, regular expressions, and classical machine learning (ML) models. Though functional, these approaches required extensive setup, struggled with document variations, and achieved limited accuracy on complex documents.

The emergence of large language models (LLMs) and generative AI has fundamentally transformed IDP capabilities. Modern AI models can understand document context, handle variations without templates, achieve near-human accuracy on complex extractions, and adapt to new document types with minimal examples. This shift from rule-based to intelligence-based processing means organizations can now process different document types with high accuracy, dramatically reducing the time and cost of implementation.

GenAI IDP Accelerator

We're excited to share the GenAI IDP Accelerator—an open source solution that transforms how organizations handle document processing by dramatically reducing manual effort and improving accuracy. This serverless foundation offers processing patterns which use Amazon Bedrock Data Automation for rich out-of-the-box document processing features, high accuracy, ease of use, and straightforward per-page pricing, Amazon Bedrock state-of-the-art foundation models (FMs) for complex documents requiring custom logic, and other AWS AI services to provide a flexible, scalable starting point for enterprises to build document automation tailored to their specific needs.

The following is a short demo of the solution in action, in this case showcasing the default Amazon Bedrock Data Automation processing pattern.



Real-world impact

The GenAI IDP Accelerator is already transforming document processing for organizations across industries.

Competiscan: Transforming marketing intelligence at scale

Competiscan, a leader in competitive marketing intelligence, faced a massive challenge: processing 35,000–45,000 marketing campaigns daily while maintaining a searchable archive of 45 million campaigns spanning 15 years.

Using the GenAI IDP Accelerator, Competiscan achieved the following:

- 85% classification and extraction accuracy across diverse marketing materials
- Increased scalability to handle 35,000–45,000 daily campaigns
- Removal of critical bottlenecks, facilitating business growth
- Production deployment in just 8 weeks from initial concept

Ricoh: Scaling document processing

Ricoh, a global leader in document management, implemented the GenAI IDP Accelerator to transform healthcare document processing for their clients. Processing over 10,000 healthcare documents monthly with potential to scale to 70,000, they needed a solution that could handle complex medical documentation with high accuracy.

The results speak for themselves:

- Savings potential of over 1,900 person-hours annually through automation
- Achieved extraction accuracy to help minimize financial penalties from processing errors

- Automated classification of grievances vs. appeals
- Created a reusable framework deployable across multiple healthcare customers
- Integrated with human-in-the-loop review for cases requiring expert validation
- Leveraged modular architecture to integrate with existing systems, enabling custom document splitting and large-scale document processing

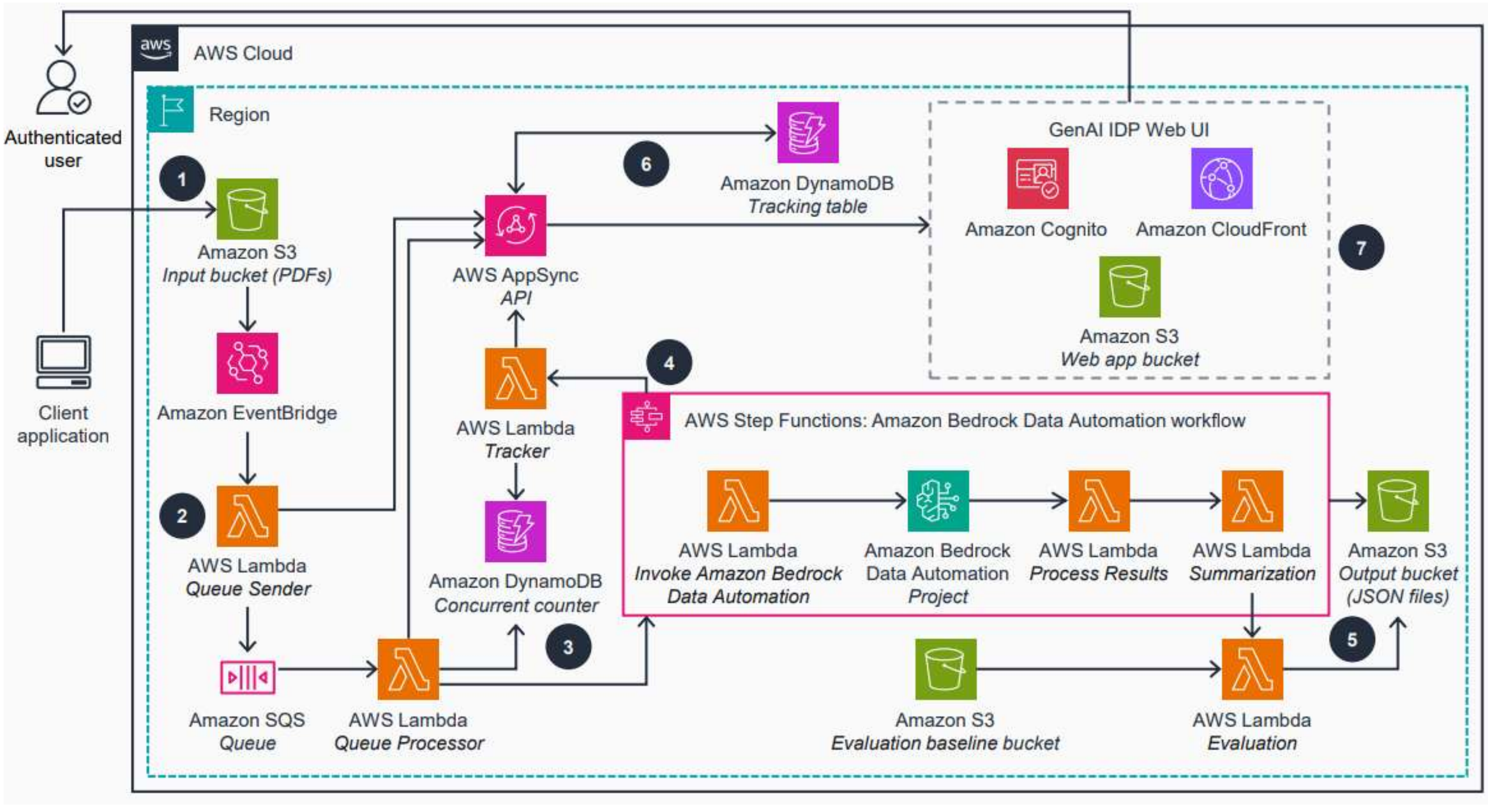
Solution overview

The GenAI IDP Accelerator is a modular, serverless solution that automatically converts unstructured documents into structured, actionable data. Built entirely on AWS services, it provides enterprise-grade scalability, security, and cost-effectiveness while requiring minimal setup and maintenance. Its configuration-driven design helps teams quickly adapt prompts, extraction templates, and validation rules for their specific document types without touching the underlying infrastructure.

The solution follows a modular pipeline that enriches documents at each stage, from OCR to classification, to extraction, to assessment, to summarization, and ending with evaluation.

You can deploy and customize each step independently, so you can optimize for your specific use cases while maintaining the benefits of the integrated workflow.

The following diagram illustrates the solution architecture, showing the default Bedrock Data Automation workflow (Pattern-1).


AWS Cloud
Region
Authenticated user
Client application
1
Amazon S3 Input bucket (PDFs)
Amazon EventBridge
2
AWS Lambda Queue Sender
Amazon SQS Queue
AWS Lambda Queue Processor
3
Amazon DynamoDB Concurrent counter
AWS Lambda Tracker
AWS AppSync API
6
Amazon DynamoDB Tracking table
4
GenAI IDP Web UI
Amazon Cognito
Amazon CloudFront
Amazon S3 Web app bucket
7
AWS Step Functions: Amazon Bedrock Data Automation workflow
AWS Lambda Invoke Amazon Bedrock Data Automation
Amazon Bedrock Data Automation Project
AWS Lambda Process Results
AWS Lambda Summarization
Amazon S3 Output bucket (JSON files)
5
Amazon S3 Evaluation baseline bucket
AWS Lambda Evaluation

Refer to the GitHub repo for additional details and processing patterns.

Some of the key features of the solution include:

- **Serverless architecture** – Built on AWS Lambda, AWS Step Functions, and other serverless technologies for queueing, concurrency management, and retries to provide automatic scaling and pay-per-use pricing for production workloads of many sizes
- **Generative AI-powered document packet splitting and classification** – Intelligent document classification using Amazon Bedrock Data Automation or Amazon Bedrock multimodal FMs, including support for multi-document packets and packet splitting
- **Advanced AI key information extraction** – Key information extraction using Amazon Bedrock Data Automation or Amazon Bedrock multimodal FMs
- **Multiple processing patterns** – Choose from pre-built patterns optimized for different workloads with different configurability, cost, and accuracy requirements, or extend the solution with additional patterns:
    - Pattern 1 – Uses Amazon Bedrock Data Automation, a fully managed service that offers rich out-of-the-box features, ease of use, and straightforward per-page pricing. This pattern is recommended for most use cases.
    - Pattern 2 – Uses Amazon Textract and Amazon Bedrock with Amazon Nova, Anthropic’s Claude, or custom fine-tuned Amazon Nova models. This pattern is ideal for complex documents requiring custom logic.
    - Pattern 3 – Uses Amazon Textract, Amazon SageMaker with a fine-tuned model for classification, and Amazon Bedrock for extraction. This pattern is ideal for documents requiring specialized classification.

We expect to add more pattern options to handle additional real-world document processing needs, and to take advantage of ever-improving state-of-the-art capabilities:

- **Few-shot learning** – Improve accuracy for classification and extraction by providing few-shot examples to guide the AI models
- **Confidence assessment** – AI-powered quality assurance that evaluates extraction field confidence, used to indicate documents for human review
- **Human-in-the-loop (HITL) review** – Integrated workflow for human review of low-confidence extractions using Amazon SageMaker Augmented AI (Amazon A2I), currently available for Pattern 1, with support for Patterns 2 and 3 coming soon
- **Web user interface** – Responsive web UI for monitoring document processing, viewing results, and managing configurations
- **Knowledge base integration** – Query processed documents using natural language through Amazon Bedrock Knowledge Bases
- **Built-in evaluation** – Framework to evaluate and improve accuracy against baseline data
- **Analytics and reporting database** – Centralized analytics database for tracking processing metrics, accuracy trends, and cost optimization across document workflows, and for analyzing extracted document content using Amazon Athena
- **No-code configuration** – Customize document types, extraction fields, and processing logic through configuration, editable in the web UI

- **Developer-friendly python package** – For data science and engineering teams who want to experiment, optimize, or integrate the IDP capabilities directly into their workflows, the solution's core logic is available through the idp_common Python package

Prerequisites

Before you deploy the solution, make sure you have an AWS account with administrator permissions and access to Amazon and Anthropic models on Amazon Bedrock. For more details, see Access Amazon Bedrock foundation models.

Deploy the GenAI IDP Accelerator

To deploy the GenAI IDP Accelerator, you can use the provided AWS CloudFormation template. For more details, see the quick start option on the GitHub repo. The high-level steps are as follows:

1. Log in to your AWS account.
2. Choose **Launch Stack** for your preferred AWS Region:

Region	Launch Stack
US East (N. Virginia)	Launch Stack
US West (Oregon)	Launch Stack

1. Enter your email address and choose your processing pattern (default is Pattern 1, using Amazon Bedrock Data Automation).
2. Use defaults for all other configuration parameters.
3. Deploy the stack.

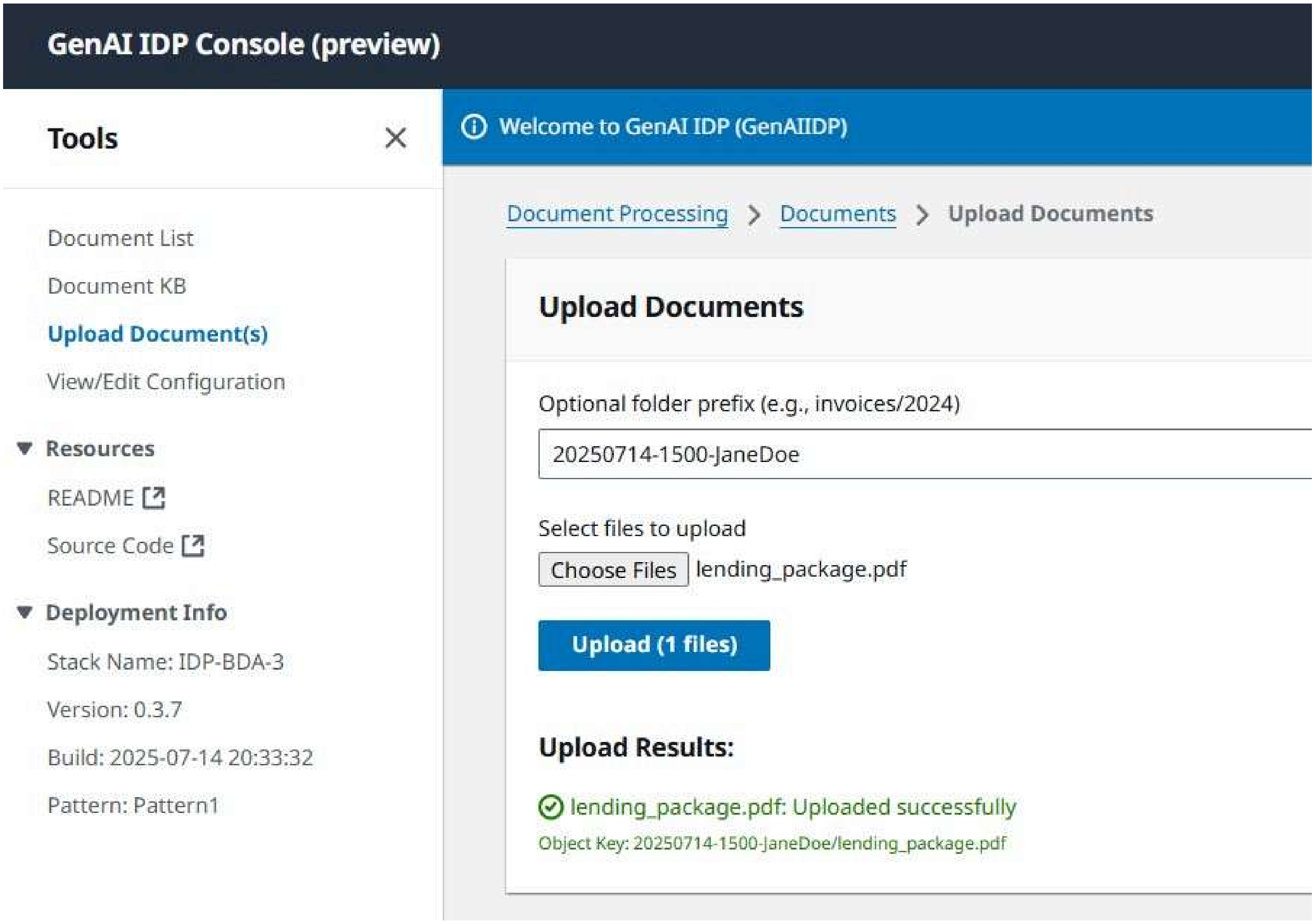
The stack takes approximately 15–20 minutes to deploy the resources. After deployment, you will receive an email with login credentials for the web interface.

Process documents

After you deploy the solution, you can start processing documents:

1. Use the web interface to upload a sample document (you can use the provided sample: lending_package.pdf).

In production, you typically automate loading your documents directly to the Amazon Simple Storage Service (Amazon S3) input bucket, automatically triggering processing. To learn more, see Testing without the UI.


GenAI IDP Console (preview)

Tools

Document List

Document KB

Upload Document(s)

View/Edit Configuration

Resources

README

Source Code

Deployment Info

Stack Name: IDP-BDA-3

Version: 0.3.7

Build: 2025-07-14 20:33:32

Pattern: Pattern1

Welcome to GenAI IDP (GenAIIDP)

Document Processing > Documents > Upload Documents

Upload Documents

Optional folder prefix (e.g., invoices/2024)

20250714-1500-JaneDoe

Select files to upload

Choose Files lending_package.pdf

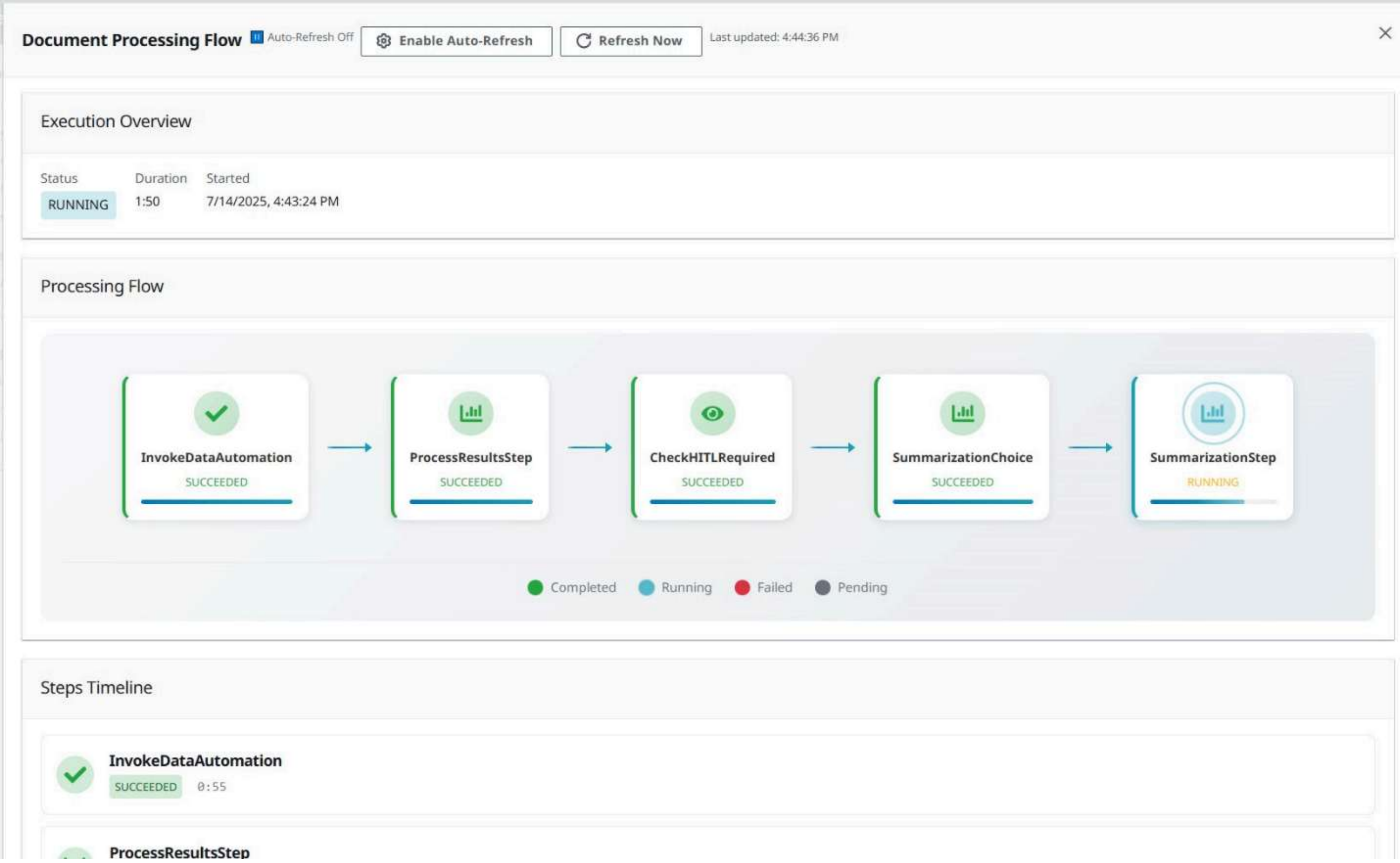
Upload (1 files)

Upload Results:

lending_package.pdf: Uploaded successfully

Object Key: 20250714-1500-JaneDoe/lending_package.pdf

1. Select your document from the document list and choose **View Processing Flow** to watch as your document flows through the pipeline.



Execution Overview

Status	Duration	Started
RUNNING	1:50	7/14/2025, 4:43:24 PM

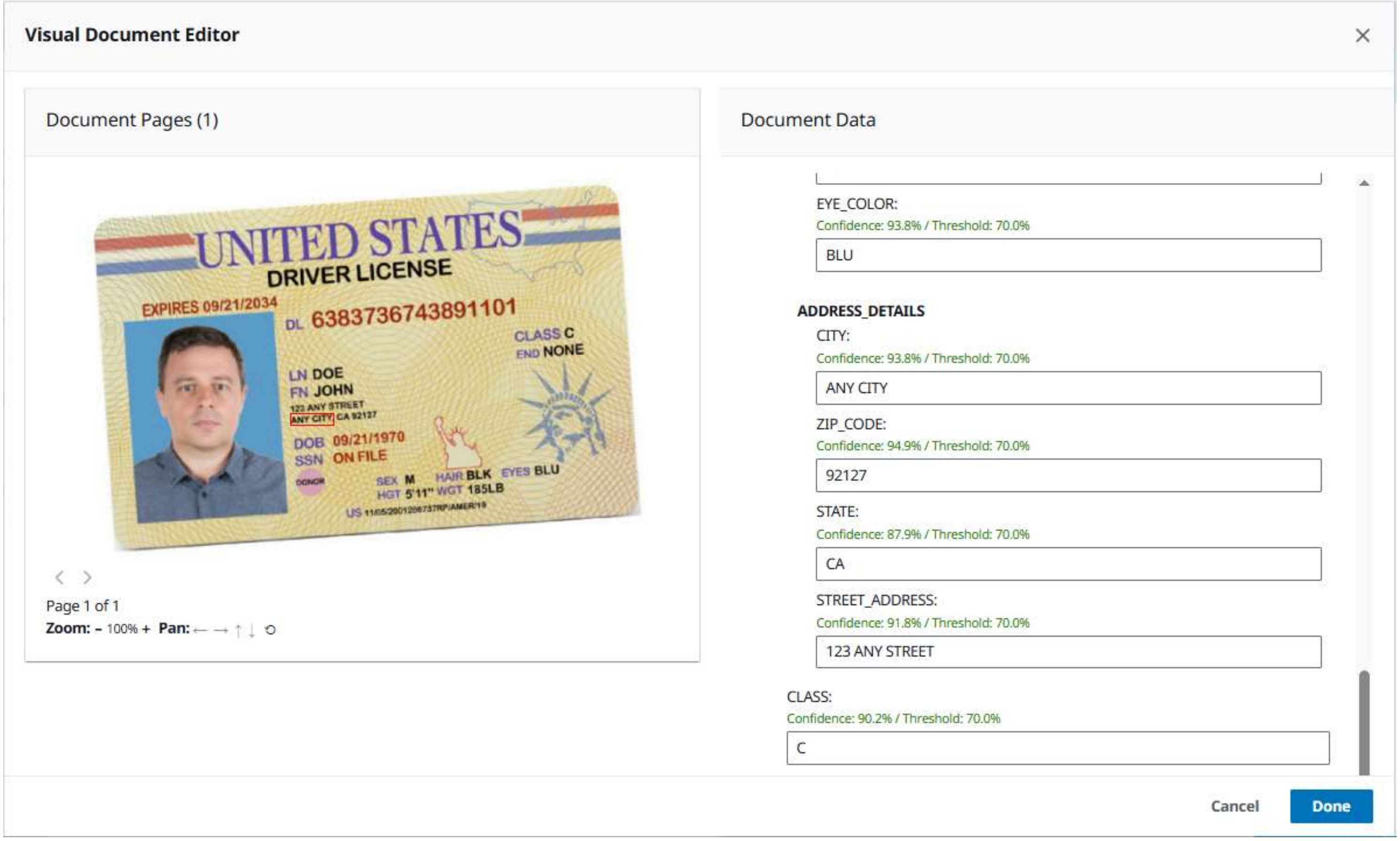
Processing Flow

Steps Timeline

✓ InvokeDataAutomation
SUCCEEDED 0:55

ProcessResultsStep

2. Examine the extracted data with confidence scores.



Document Data

EYE_COLOR:
Confidence: 93.8% / Threshold: 70.0%
BLU

ADDRESS_DETAILS

CITY:
Confidence: 93.8% / Threshold: 70.0%
ANY CITY

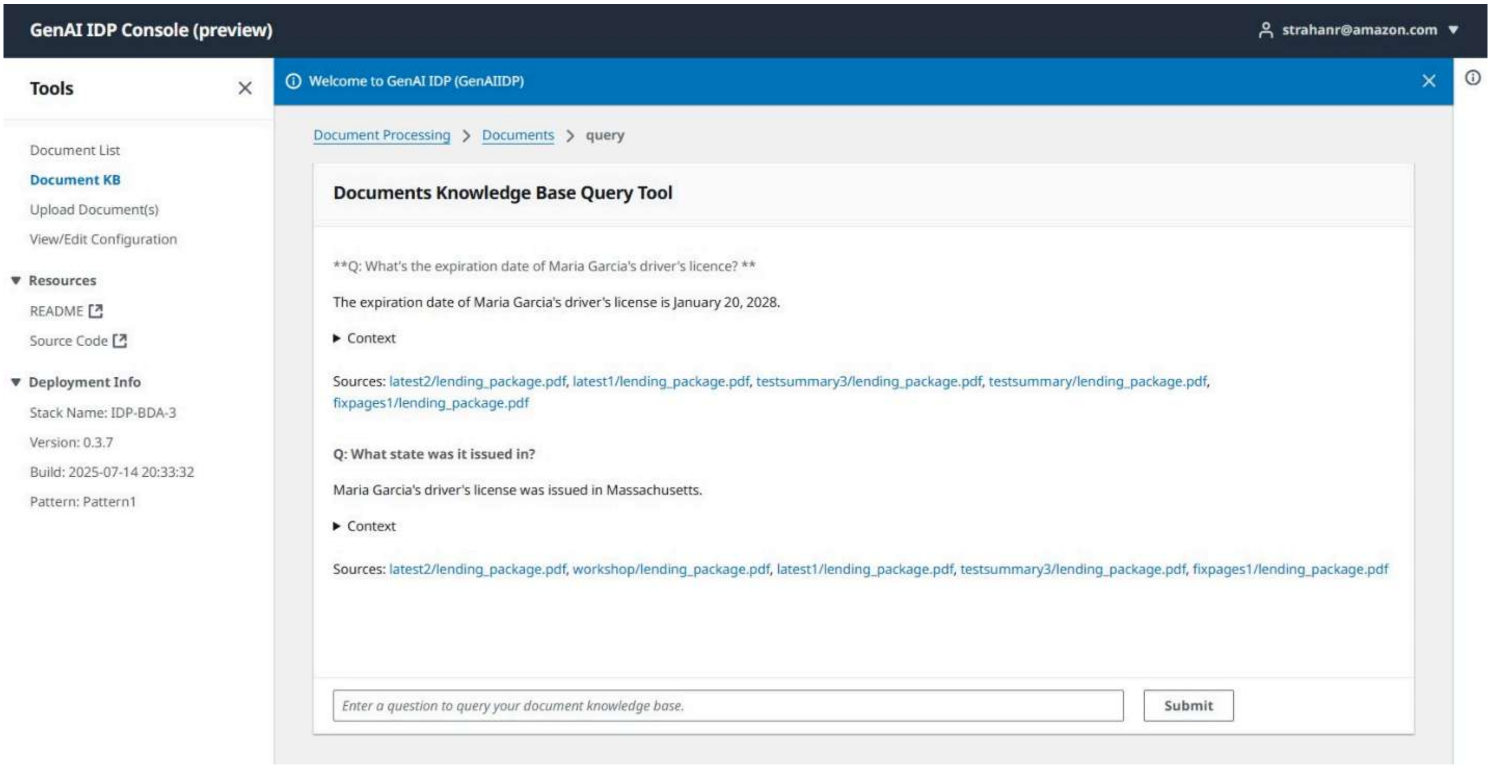
ZIP_CODE:
Confidence: 94.9% / Threshold: 70.0%
92127

STATE:
Confidence: 87.9% / Threshold: 70.0%
CA

STREET_ADDRESS:
Confidence: 91.8% / Threshold: 70.0%
123 ANY STREET

CLASS:
Confidence: 90.2% / Threshold: 70.0%
C

3. Use the knowledge base feature to ask questions about processed content.


GenAI IDP Console (preview)
strahanr@amazon.com
Tools
Document List
Document KB
Upload Document(s)
View/Edit Configuration
Resources
README
Source Code
Deployment Info
Welcome to GenAI IDP (GenAIIDP)
Document Processing > Documents > query
Enter a question to query your document knowledge base.
Submit

Stack Name: IDP-BDA-3

Version: 0.3.7

Build: 2025-07-14 20:33:32

Pattern: Pattern1

Documents Knowledge Base Query Tool

**Q: What's the expiration date of Maria Garcia's driver's licence? **

The expiration date of Maria Garcia's driver's license is January 20, 2028.

▸ Context

Sources: latest2/lending_package.pdf, latest1/lending_package.pdf, testsummary3/lending_package.pdf, testsummary/lending_package.pdf, fixpages1/lending_package.pdf

Q: What state was it issued in?

Maria Garcia's driver's license was issued in Massachusetts.

▸ Context

Sources: latest2/lending_package.pdf, workshop/lending_package.pdf, latest1/lending_package.pdf, testsummary3/lending_package.pdf, fixpages1/lending_package.pdf

Alternative deployment methods

You can build the solution from source code if you need to deploy the solution to additional Regions or build and deploy code changes.

We hope to add support for AWS Cloud Development Kit (AWS CDK) and Terraform deployments. Follow the GitHub repository for updates, or contact AWS Professional Services for implementation assistance.

Update an existing GenAI IDP Accelerator stack

You can update your existing GenAI IDP Accelerator stack to the latest release. For more details, see Updating an Existing Stack.

Clean up

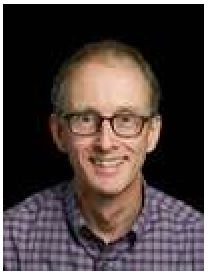
When you're finished experimenting, clean up your resources by using the AWS CloudFormation console to delete the IDP stack that you deployed.

Conclusion

In this post, we discussed the GenAI IDP Accelerator, a new approach to document processing that combines the power of generative AI with the reliability and scale of AWS. You can process hundreds or even millions of documents to achieve better results faster and more cost-effectively than traditional approaches.

Visit the GitHub repository for detailed guides and examples and choose **watch** to stay informed on new releases and features. AWS Professional Services and AWS Partners are available to help with implementation. You can also join the GitHub community to contribute improvements and share your experiences.

About the Authors



Bob Strahan is a Principal Solutions Architect in the AWS Generative AI Innovation Center.

Joe King is a Senior Data Scientist in the AWS Generative AI Innovation Center.

Mofijul Islam is an Applied Scientist in the AWS Generative AI Innovation Center.

Vincil Bishop is a Senior Deep Learning Architect in the AWS Generative AI Innovation Center.

David Kaleko is a Senior Applied Scientist in the AWS Generative AI Innovation Center.

Rafal Pawlaszek is a Senior Cloud Application Architect in the AWS Generative AI Innovation Center.

Spencer Romo is a Senior Data Scientist in the AWS Generative AI Innovation Center.

Vamsi Thilak Gudi is a Solutions Architect in the AWS World Wide Public Sector team.

Acknowledgments

We would like to thank Abhi Sharma, Akhil Nooney, Aleksei Iancheruk, Ava Kong, Boyi Xie, Diego Socolinsky, Guillermo Tantachuco, Ilya Marmur, Jared Kramer, Jason Zhang, Jordan Ratner, Mariano Bellagamba, Mark Aiyer, Niharika Jain, Nimish Radia, Shean Sager, Sirajus Salekin, Yingwei Yu, and many others in our expanding community, for their unwavering vision, passion, contributions, and guidance throughout.

Like Share

Comments

Log in to comment