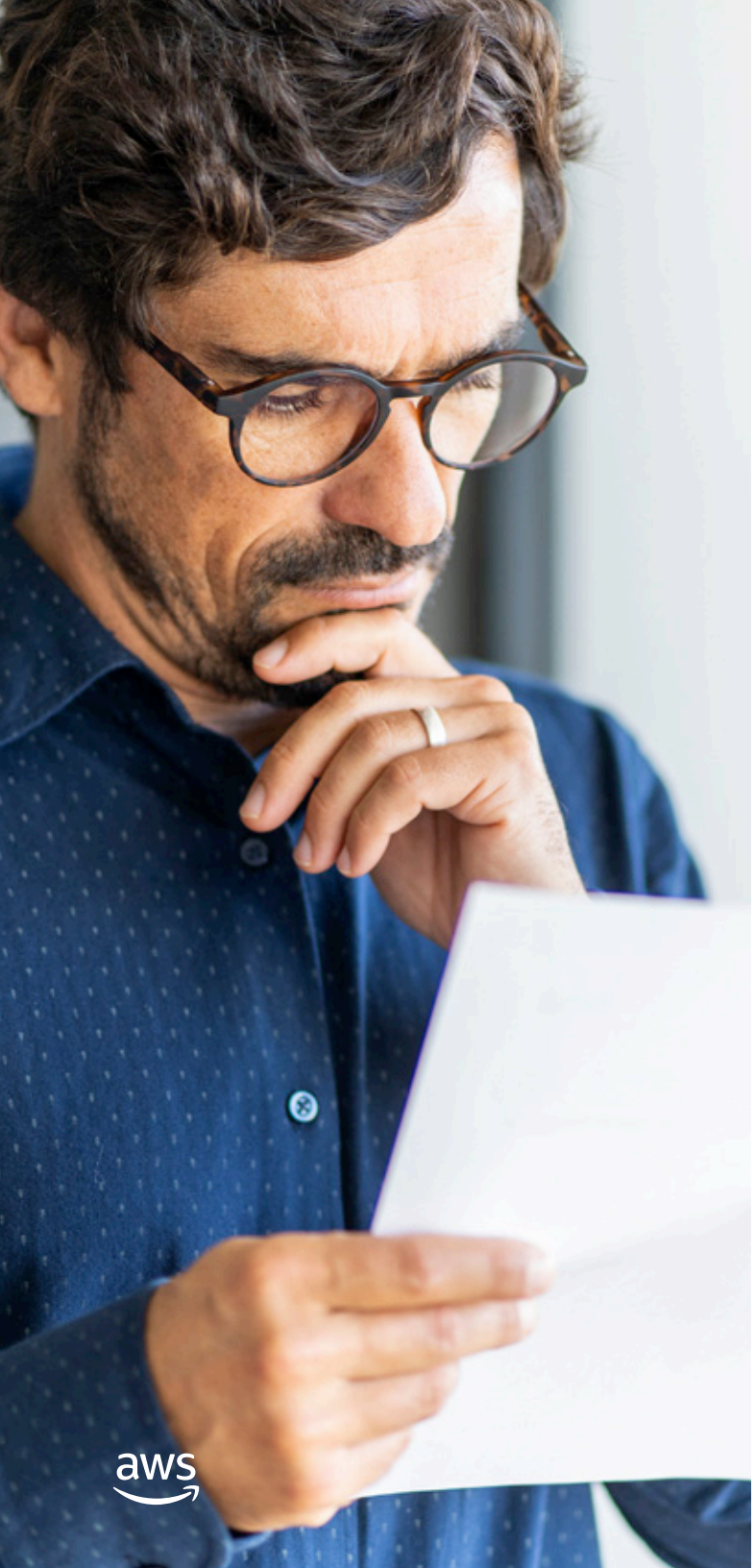
Automate document data extraction and analysis

Extract data and derive critical insights from documents—quickly, accurately, and at scale with AWS AI services now enhanced by generative AI



Many of today's organizations have a large number of documents that are critical to business processes. Legacy optical character recognition (OCR) systems for extracting and analyzing data are manual, cost-intensive, prone to errors, and difficult to scale. This problem is prevalent across many industries:

Financial services organizations must examine documents, such as loan or credit applications, insurance forms, and insurance claims, with the highest degree of accuracy and care, often turning to manual review to extract sensitive or critical information, such as mortgage rates or credit scores.

Healthcare and life sciences organizations are fighting an uphill battle against an ever-growing mountain of documents and forms, searching for and analyzing data essential to clinical trial research and patient diagnoses in order to treat their patients more accurately.

Public sector entities are forced to tie up their already-strained resources to process data from documents, such as applications, for a wide variety of public services.

Intelligent document processing (IDP) is changing the equation. Powered by generative AI and machine learning (ML), it can help organizations extract text from millions of documents, understand the sentiment within data or the relationships between the entities, and even include a human step to validate, correct, or augment the results for higher accuracy and compliance. IDP solutions can help empower employees, clients, and customers of these businesses with fast, easy access to the information they need—while helping them discover new insights from data, expedite decision cycles, and augment their generative AI use case further.

In this eBook, we will take a deeper look at how IDP can provide practical business benefits—and show you how services from Amazon Web Services (AWS) can help achieve impactful outcomes.

Table of contents

The benefits of intelligent document processing

Drive business process efficiency

Manual document processing is expensive and time-consuming. You need to allocate resources to process large volumes of documents, slowing down business process efficiency. Moreover, your employees are doing manual stare-and-compare, going page by page, reducing worker morale. **AWS IDP** helps you overcome these challenges by automating document processing using artificial intelligence (AI) and ML, allowing you to optimize resource allocation to high-value tasks, enabling faster business decision making and reducing expenses.

Scale quickly

Legacy OCR solutions are often difficult to scale, especially with variability in document types and volumes. AWS IDP helps you scale quickly with large volumes and different document types. You can scale up or down in a few clicks, depending on business needs.

Accelerate your data strategy

Organizations need to sift through volumes of documents in order to tease out important data points such as customer name, date of birth, and ID number. Manual document processing is not only time-consuming and resource-intensive, but also prone to errors. Bad data can result in a poor customer experience. Using AWS IDP, you can curtail and even eliminate errors associated with manual data entry, improving customer satisfaction by providing data quickly and accurately.

Get the most from intelligent document processing with AWS

At AWS, we're constantly innovating on behalf of our customers to deliver the broadest and deepest set of ML capabilities for builders of all levels of expertise. Intelligent document processing solutions at AWS allow you to process documents automatically with high accuracy, focus resources on higher-value tasks, scale quickly, and serve more customers.

There are several flexible approaches for implementing an AI-based IDP solution depending on the stage of your document processing maturity, business landscape, and desired business outcomes. For organizations that are looking to get started with an IDP solution today, with no ML expertise required, AWS offers services such as **Amazon Textract**, **Amazon Comprehend**, **Amazon Augmented AI** (Amazon A2I), and **Amazon Kendra**. Combined together or used separately, these AWS services offer a compelling solution to quickly and accurately classify, extract, and analyze data to drive business agility, scale on demand, and serve a larger set of customers.

Alternatively, organizations that want to develop their own ML models for text extraction and analysis can use **Amazon SageMaker**, a fully managed service that helps business analysts, data scientists, and ML developers prepare data and build, train, and deploy ML models quickly. With **Amazon Bedrock** organizations can now easily add generative AI capabilities like question and answering and summarization to their document processing workflows. The data science team can be engaged for foundation model (FM) selection, evaluation, and tuning based on the generative AI use case and link generative AI and FM modules with AWS IDP (through chaining, for example) to create an end-to-end document processing pipeline.

In the next few sections, we will review the benefits and general use cases for these services. Later, we will show how financial services, healthcare organizations, and contact centers are successfully leveraging one or more AWS services for IDP.

"Amazon Textract helped support 80% of Kabbage's PPP applicants to receive a fully automated lending experience and reduced approval times from multiple days to a median speed of four hours. Kabbage became the second largest PPP lender in the nation by application volume…serving over 297,000 small businesses and preserving an estimated 945,000 jobs across America."[1]

Anthony Sabelli, Head of Data Science, Kabbage

Easily extract the data you need from virtually any document

For organizations looking to extract and analyze data automatically from volumes of documents, **Amazon Textract** is here to help. Amazon Textract uses ML to extract handwriting, printed text, and data from scanned documents. It goes beyond traditional OCR technology—which requires manual configuration that must be updated each time a form is changed—by accurately extracting text, forms, tables, and other data without the need for any manual effort of custom code.

With Amazon Textract, you can quickly automate activities across complex documents such as loan applications, insurance forms, tax documents, enrollment forms, or medical claims. Once the information from these documents is captured, you can move into the next stage of your business process faster—leading to more productive employees and happier customers.

Saving nearly a million jobs through faster business loans

Kabbage,* a data and technology company in the financial services sector, currently provides more than 500,000 small businesses with access to more than $16 billion of working capital. The Paycheck Protection Program (PPP), a component of the US government's response to economic struggles related to the COVID-19 pandemic, presented Kabbage with a unique opportunity. But it also offered a challenge: Kabbage had never issued a Small Business Administration (SBA) loan, and it needed a fast, automated way to process PPP applications.[1]

[1] Sabelli, A., "How Kabbage Improved the PPP Lending Experience with Amazon Textract," AWS Machine Learning Blog
* Company name has changed to American Express Business Blueprint

"Black Knight is focused on using the power of artificial intelligence (AI) to create innovative solutions that address the needs of the mortgage industry. Our AI product, called AIVA, drives efficiency across the Black Knight ecosystem by eliminating stare-and-compare work and by applying ML and advanced business rules to lower loan processing costs. The Black Knight ML team worked with the Amazon Machine Learning Solutions Lab to provide faster and more reliable processing behind Black Knight's mortgage origination solutions, allowing our lender clients to be more competitive and efficient, providing operating leverage to further accelerate their growth."[2]

Rich Gagliano, President, Origination Technologies for Black Knight, Inc.

Transform documents into usable data

Amazon Comprehend is a natural language processing (NLP) service that uses ML to find insights and relationships in text. The service identifies the language of the text, extracts key phrases, places, people, or brands, understands the sentiment in the text, and automatically organizes a collection of text files by topic. You can train Amazon Comprehend to identify entities relevant to your organization, categorize documents, and assign relevant labels to text.

BLACK KNIGHT®

Improve accuracy and reduce costs in document extraction with machine learning

Black Knight, Inc. is a premier provider of technology solutions in the mortgage and home equity lending industry. The company provides solutions that reduce manual processes, support compliance and quality, mitigate risk, and deliver significant cost savings to their clients, including many of the largest mortgage lenders in the United States. AIVA, Black Knight's AI product developed in partnership with Amazon and leveraging **Amazon Textract**, allows for the classification and extraction of large volumes of content and a variety of documents. With this AI-powered solution, Black Knight expects a reduction in the time it takes to process bank statements, as well as a reduction in the overall number of bank statements passing through humans for processing. This process is repeatable for other documents and use cases within the mortgage domain.[2]

[2] Zhu, J., et al., "Black Knight Continued Innovation with Document Extraction using AWS Machine Learning Capabilities," AWS for Industries Blog

"For cancer patients and the researchers dedicated to curing them, time is the limiting resource. [Deploying Amazon Comprehend Medical] is a vital step toward getting researchers rapid access to the information they need when they need it so they can find actionable insights to advance lifesaving therapies for patients."[3]

Matthew Trunnell, CIO, Fred Hutchinson Cancer Research Center

Extract potentially lifesaving insights from medical text

Amazon Comprehend Medical extends the power of NLP to healthcare and life sciences organizations. The service makes it easy to use ML to extract relevant medical information from unstructured text. Using Amazon Comprehend Medical, you can quickly and accurately gather critical data—such as medical condition and medication dosage, strength, and frequency—from a variety of sources, like doctors' notes, clinical trial reports, and patient health records.

Amazon Comprehend Medical can also link the detected information to medical ontologies, such as ICD-10-CM or RxNorm, so it can easily be used by downstream healthcare applications.

Fred Hutch Cancer Center

Using IDP to accelerate the development of cancer treatments

Fred Hutchinson Cancer Research Center* is an institute with the lofty goal of curing cancer by 2025—and it's using Amazon Comprehend Medical to help make its vision a reality. The organization needed a better method of using document data to develop clinical trials and connect researchers with the right patients faster. Amazon Comprehend Medical helped provide the solution, reducing the time needed to sift through and label unstructured clinical data from hours to seconds.[3]

[3] Amazon Comprehend Customers and Partners

* Company name has changed to The Fred Hutchinson Cancer Center

> **"We are excited about Amazon Augmented AI because it allows us to take advantage of machine learning while still applying human judgement. That's a game changer for us."[4]**
>
> Chris Suter, Head of Cloud Platforms & Innovation, National Health Service Business Services Authority

Facilitate faster, more efficient human review

Due to regulatory requirements and other unique business needs, many organizations must still rely on manual processes to analyze documents in certain use cases, such as sensitive documents in healthcare, handwritten documents, or insurance claims. This generally does not mean that IDP cannot be deployed at all—only that manual review of the work done by IDP systems is required at some level.

Amazon A2I makes it easy to build and manage human reviews for IDP and other AI applications. The service provides built-in human review workflows for common use cases, such as content moderation and text extraction from documents. Using Amazon A2I, you can send any document to a human for review to ensure the text, phrase, or information is processed correctly. Additionally, you can use this human review information to retrain your ML model to provide better accuracy downstream.

NHS

Integrating machine learning and human judgment in healthcare

The National Health Service Business Services Authority (NHSBSA) provides a range of support services to healthcare organizations, contractors, and patients in the United Kingdom. As part of those services, NHS processes 54 million paper prescriptions per month. The organization requires an IDP solution that can do that work quickly—but can also allow fast, easy human intervention where necessary.[4]

[4] Amazon Augmented AI Customers

"Prior to our deployment of Kendra, we attempted to create logical navigation paths to guide our customers to the information they needed; however, none produced the desired level of customer experience (CX) and could lead to frustration or worse, customers making less than optimal decisions about their health benefits. Amazon Kendra plugged in effortlessly to our existing portal platform and freed up staff members to focus on higher value activities to drive CX improvements. While we're still early in our use of Amazon Kendra, we've already seen it support thousands of positive customer interactions and deliver higher quality answers in just the first few days."[5]

Devin Parsons, Former VP & Head of Digital Transformation, bswift, a CVS Company

Find accurate and relevant information faster

Nearly half the time, users fail to find the information they need to excel because the information is scattered across their organization in the form of documents and other unstructured data. This negatively impacts customer experiences as well as workforce productivity.

Amazon Kendra delivers intelligent search and natural language understanding capabilities, enabling employees and customers to ask natural language questions to find relevant information contained within documents. Amazon Kendra will quickly surface relevant content as a suggested answer instead of a list of links, so employees get the information they need quickly and easily.

bswift®

Making enterprise knowledge searchable through machine learning

Since its inception, bswift has focused on using technology to simplify the administration of healthcare, reduce costs, and connect customers to the resources they need to achieve their health ambitions.[5]

[5] Amazon Kendra Customers

Amazon SageMaker

Organizations that want to develop their own ML models for text extraction and analysis can use **Amazon SageMaker**, a fully managed service that helps business analysts, data scientists, and ML developers prepare data and build, train, and deploy ML models quickly. For text extraction, SageMaker offers built-in algorithms, such as BlazingText, pretrained models available through **Amazon SageMaker JumpStart**, and the ability to develop your own text-processing algorithms. Regardless of which option you choose, SageMaker provides all the end-to-end tools you need for ML—so you can easily develop high-quality text processing models.

Amazon Bedrock

Organizations can use **Amazon Bedrock** to accelerate development of generative AI applications with FMs from AI21 Labs, Anthropic, Cohere, Meta, Stability AI, and Amazon through an API without managing their infrastructure. **Amazon Titan** FMs are also built to detect and remove harmful content in the data, reject inappropriate content in the user input, and filter model outputs that contain inappropriate content, such as hate speech, profanity, and violence.

SIEMENS

Building custom IDP solutions with Amazon SageMaker

Siemens Financial Services provides business-to-business financial solutions, offering equipment, transport, and automobile leasing services. It uses an NLP model it developed with SageMaker to extract information from documents and accelerate investment due diligence—reducing the time to summarize documents from 12 hours to 30 seconds.[6]

THOMSON REUTERS

Helping customers find the information they need

Thomson Reuters, the world's leading source of news and information for professional markets, designed an NLP capability in the context of a question-answering application, which delivers accurate information to customers and allows them to simplify and derive more value from their work.[7]

[6] "The Total Cost of Ownership (TCO) of Amazon SageMaker," AWS, 2020
[7] Amazon SageMaker Customers

AWS IDP solutions for specific industries and use cases

AWS IDP solutions use ML to help organizations of all sizes and industries extract and analyze document data faster. But in this section, we will identify and explore a few industries and use cases where the technology is particularly beneficial.

Financial services organizations

Documents at financial services organizations, such as mortgage applications and insurance forms, contain sensitive information that must be analyzed with absolute precision. The smallest of errors could result in the wrongful denial of a loan or a costly overpayment of interest.

With Amazon Textract, financial organizations can extract text, forms, and tables from documents. They can then use Amazon Comprehend to analyze entities, phrases, or key-value pairs within the extracted text and Amazon A2I to facilitate manual review of sensitive or nuanced results. For example, an organization deploying all of these services could pull key phrases—like company names, social security numbers, and interest rates—from documents using Amazon Textract, analyze the information to discover patterns and opportunities with Amazon Comprehend, and perform manual accuracy checks through Amazon A2I. They can then make this extracted and analyzed content searchable throughout the organization using Amazon Kendra. A financial services customer can also use the Textract Tables feature to extract all the debit and credit transactions from a scanned bank statement document while using generative AI to create a tally (sum) of specific transaction types (such as debits only).

Healthcare organizations

By deploying both Amazon Textract and Amazon Comprehend Medical, healthcare organizations can quickly and easily extract data from tables within documents—and then group key-value pairs or entities to understand their relationships within the document. Healthcare organizations use these solutions to gain a more holistic view of insights contained within various documents. For example, a provider or researcher can analyze symptoms from multiple documents or patient histories to better understand how particular symptoms relate to a specific diagnosis. With the combination of IDP and generative AI, organizations can quickly and accurately identify sensitive document types to create a summary of patient history.

Contact centers and customer support teams

Organizations across nearly every industry can use AWS IDP services to augment their contact centers and customer support teams. They can use Amazon Comprehend to extract and analyze important text from emails and support tickets. Amazon Comprehend also enables sentiment analysis, such as finding positive or negative reviews and identifying specific features associated with happy or unhappy customers. Having faster and more comprehensive access to this information will help increase understanding and incorporate feedback to improve customer experiences, products, and services.

Unlocking the value of your documents

Ready to unlock more value from your documents? AWS provides a range of options that can help you start realizing the benefits of AI-powered IDP today.

You can leverage the **AWS Professional Services organization**, the **AWS ISV Accelerate Program**, and **AWS Systems Integrators**, a global team of experts that can help you deploy IDP and other services to realize your desired business outcomes when using the AWS Cloud.

Or, you can **train your developers** and data scientists to build custom IDP models and gain a stronger understanding of ML in general. Our training initiatives use the same curriculum we use at Amazon, and many courses are available on demand and at no cost. We can help everyone in your organization—executives, developers, and data scientists alike—become more proficient in ML.

Finally, you can **contact us** directly for more information on IDP or visit the **AI Use Case Explorer** to discover other AI use cases for your organization.

No matter what path you choose, one thing is certain: To remain competitive in today's business world, you can't afford to continue to use manual, time-consuming methods of document processing. By setting your data free and transforming it into actionable insights, you can boost productivity, cut down on costs, and improve customer satisfaction.

Explore more ways IDP can benefit your business ›