# Document extraction and summarization

**Estimated Duration: 45 minutes**

## Introduction

Document processing is a critical function for businesses that need to convert unstructured files into usable data. This process traditionally requires manual extraction and review, which can be slow and error-prone. Automation allows organizations to transform documents and presentations into structured formats and generate concise summaries, streamlining workflows and improving data accessibility.

## Use Case

Imagine an organization with a growing repository of PDF documents containing critical business data, such as reports, contracts, and presentations. Currently, employees need to manually review each document to find relevant information, a process that is both time-consuming and prone to human error.

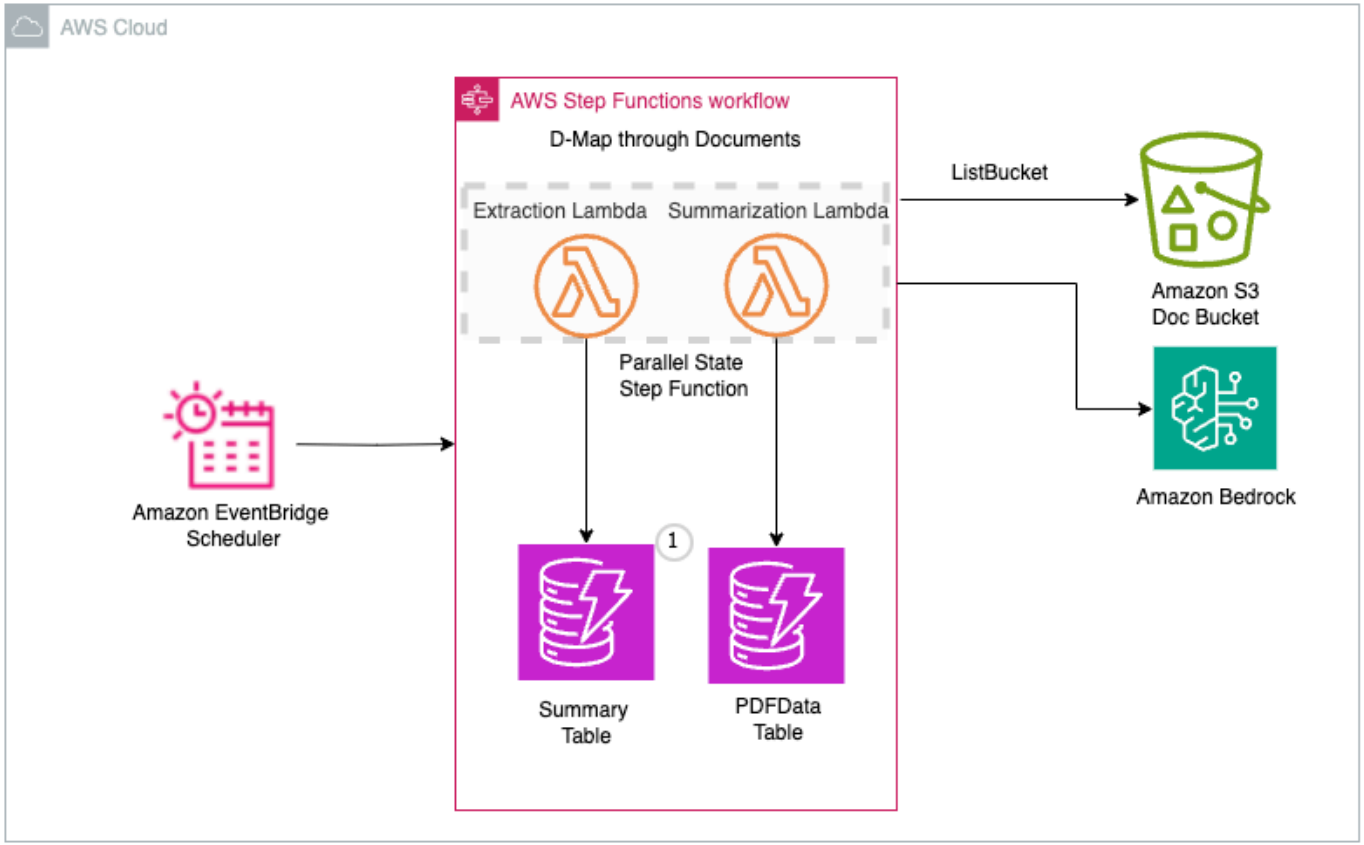
The goal of this workshop is to develop a solution that automates the extraction of key data from these documents and generates concise summaries, allowing users to quickly access the most pertinent information. By automating this process, the organization can significantly reduce the time and effort required for document review, enhance data accessibility, and enable more informed decision-making. This solution can be applied across various industries, such as legal,

financial, or educational sectors, where vast amounts of document-based data need to be processed efficiently.

## Architecture

## PDF Extraction and Summarization Workflow:

The following is an architecture of how the summarization is done in the module. As this is a batch summarization process, an EventBridge scheduler is scheduled to kick off the workflow at regular intervals or one time. The scheduler invokes a Step Functions workflow which iterates on the documents you might have in the configured S3 bucket. During the process, each document is summarized as well as some key data is extracted and stored in Amazon DynamoDB for further




## What you will accomplish

- Learn how to use generative AI and Serverless to extract and summarize PDF documents

## Services in this module

- Amazon S3 - Object Storage built to retrieve any amount of data from anywhere
- AWS Step Functions - Visual workflows for distributed applications
- AWS Lambda - Serverless compute service; Run code without thinking about servers or clusters

- Amazon Bedrock - The easiest way to build and scale generative AI applications with foundation models

## LLMs in this module

- Anthropic Claude 3 Haiku - The fastest and most affordable model in its intelligence class

## What's included in this module

The extraction and summarization code are in the following AWS Lambda functions:

- **documentextractionstack-DocumentSummaryLambda-**: This Lambda function processes PDFs stored in S3 by extracting text from each page and generating summaries. It then stores these summaries along with metadata in a DynamoDB table, creating a searchable and accessible content repository.
- **documentextractionstack-DocumentExtractionLambda-**: This Lambda function processes PDFs stored in an S3 bucket, extracts contents from the document based on the schema provided. It then stores the extracted information along with the metadata in a DynamoDB table.

---

Previous Next