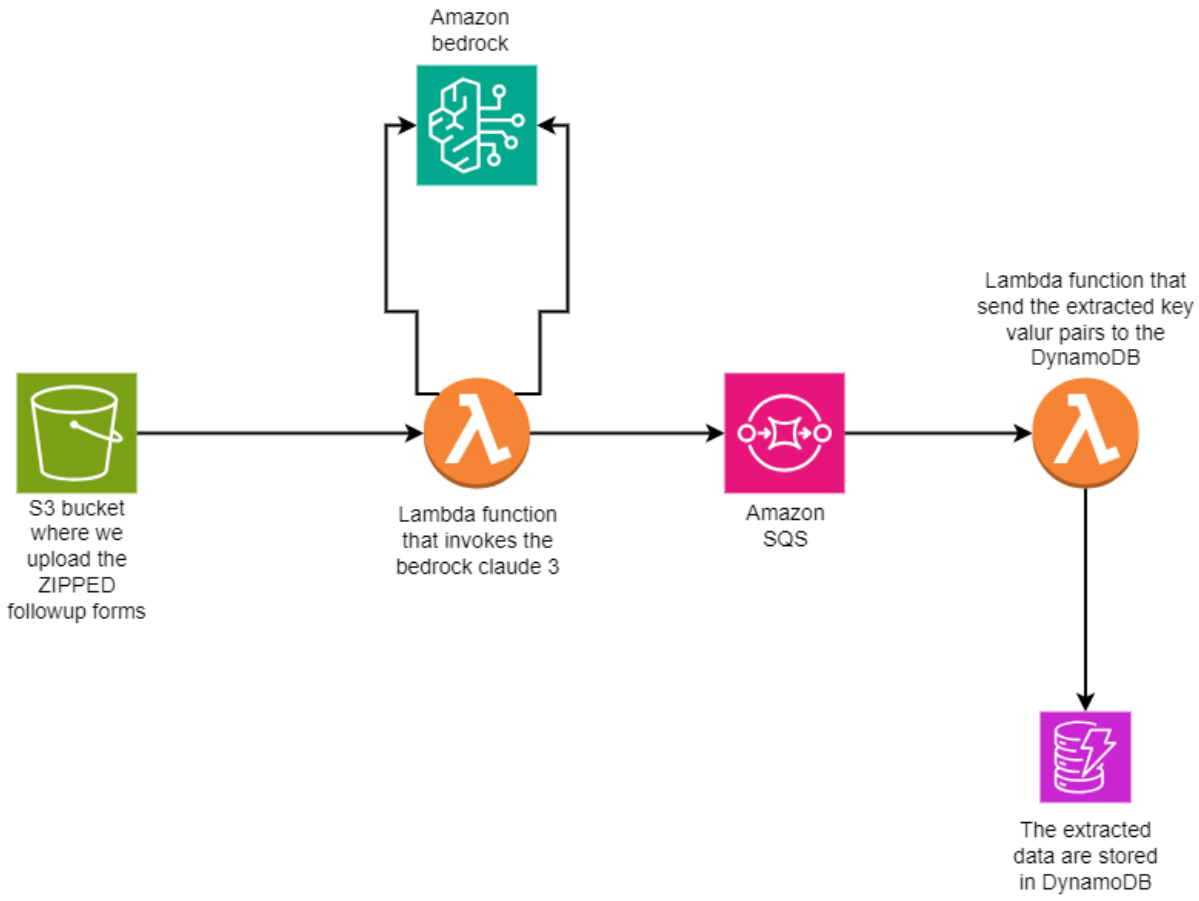
OCR Project Documentation:

Automating the Manual Entry System for Stroke Patient Follow-Up Forms

The flow diagram for the process with Bedrock



AWS resource names

S3 Bucket input - bedrock-claude3-idp-12345

Lambda invoking the bedrock - invoke_bedrock_claude3

Lambda invoking the DynamoDB saving - insert_into_dynamodb

SQS (Simple Queue Service) - bedrock-idp-extracted-data

S3 Bucket output - bedrock-claude3-idp-output

DynamoDB - patient_forms

Introduction

This project automates the manual entry system for stroke patient follow-up forms using Optical Character Recognition (OCR). The goal is to simplify the data extraction process from scanned documents, primarily multi-page PDFs, which will be converted into images before feeding into Amazon Bedrock for text extraction. The extracted data will be stored in DynamoDB for further analysis and retrieval. The system leverages AWS services, such as Lambda, S3, SQS, and DynamoDB, to handle the entire workflow seamlessly.

Architecture Overview

The architecture is based the following components:

- Amazon S3: Stores scanned documents uploaded by users.
- Amazon Bedrock: Processes images and extracts text from scanned documents.
- Amazon SQS (Simple Queue Service): Acts as a message broker to pass extracted data between services.
- AWS Lambda Functions: Handles the invocation of Amazon Bedrock, data processing, and uploading extracted data to DynamoDB.
- DynamoDB: Stores the extracted data in key-value pairs for easy access and retrieval.

Step-by-Step Process

Document Upload:

- The stroke patient follow-up forms are initially ins multi-page PDF documents. However, Amazon Bedrock does not support PDFs directly, which necessitates converting these documents into images for processing.
- Since each PDF contains multiple pages, converting them results in several individual image files. To manage these multiple images efficiently, they are compressed into a single ZIP file. This ZIP file, containing all the images corresponding to the PDF, is then uploaded to the designated S3 bucket.

Document Conversion:

Users can use one of the following online tools to convert pdfs to ZIP files.

https://www.ilovepdf.com/pdf_to_jpg

https://acrobat.adobe.com/link/acrobat/pdf-to-image?x_api_client_id=adobe_com&x_api_client_location=pdf_to_image

Lambda Function (Bedrock Invocation):

- Once the ZIP file is uploaded, an S3 event notification triggers a Lambda function.
- The Lambda function invokes Amazon Bedrock and passes the image data in base64 format as input.
- Bedrock processes the image data using the Claude 3 model, which extracts the text and sends the response in JSON format.

Amazon SQS (Message Queuing):

- The Lambda function sends the extracted data from Bedrock to an SQS queue for further processing.
- This decouples the text extraction step from the data storage step, ensuring smooth handling of multiple tasks in parallel.

Lambda Function (Data Parsing):

- Another Lambda function is triggered by the SQS queue, which parses the JSON data containing the extracted text.
- The parsed data is then inserted into DynamoDB in the form of key-value pairs, ensuring that the extracted data is structured for easy retrieval.

Data Storage in DynamoDB:

- The extracted and parsed data is stored in DynamoDB, allowing for further analysis and access by authorized personnel.
- Each record in DynamoDB corresponds to a single stroke patient's form, with the extracted fields organized for efficient querying.

User Instructions

To use this OCR system, follow the steps outlined below:

1. Convert PDF to Images and create a Zip file
   Use an online converter to split the multi-page PDF into individual JPEG images. And download the compressed ZIP file.

2. Upload the ZIP File to S3:
   Log in to the AWS S3 console.
   Navigate to the designated S3 bucket for stroke patient follow-up forms.
   Upload the ZIP file containing the images.
   Once the file is uploaded, it will trigger the OCR process automatically.

3. Retrieving Extracted Data and Access DynamoDB:
   After the data extraction process is completed, the parsed data is stored in DynamoDB.
   Log in to the AWS DynamoDB console and navigate to the table storing the OCR-extracted data.

4. Querying Data:
   You can search for a specific patient's data by querying the DynamoDB table using the appropriate key (such as the patient ID or form ID).

Conclusion

This system automates the tedious task of manually entering stroke patient follow-up forms, improving both the speed and accuracy of data entry. By leveraging AWS services like Lambda, S3, SQS, Bedrock, and DynamoDB, we have created a scalable and reliable solution that can handle multi-page documents with ease.