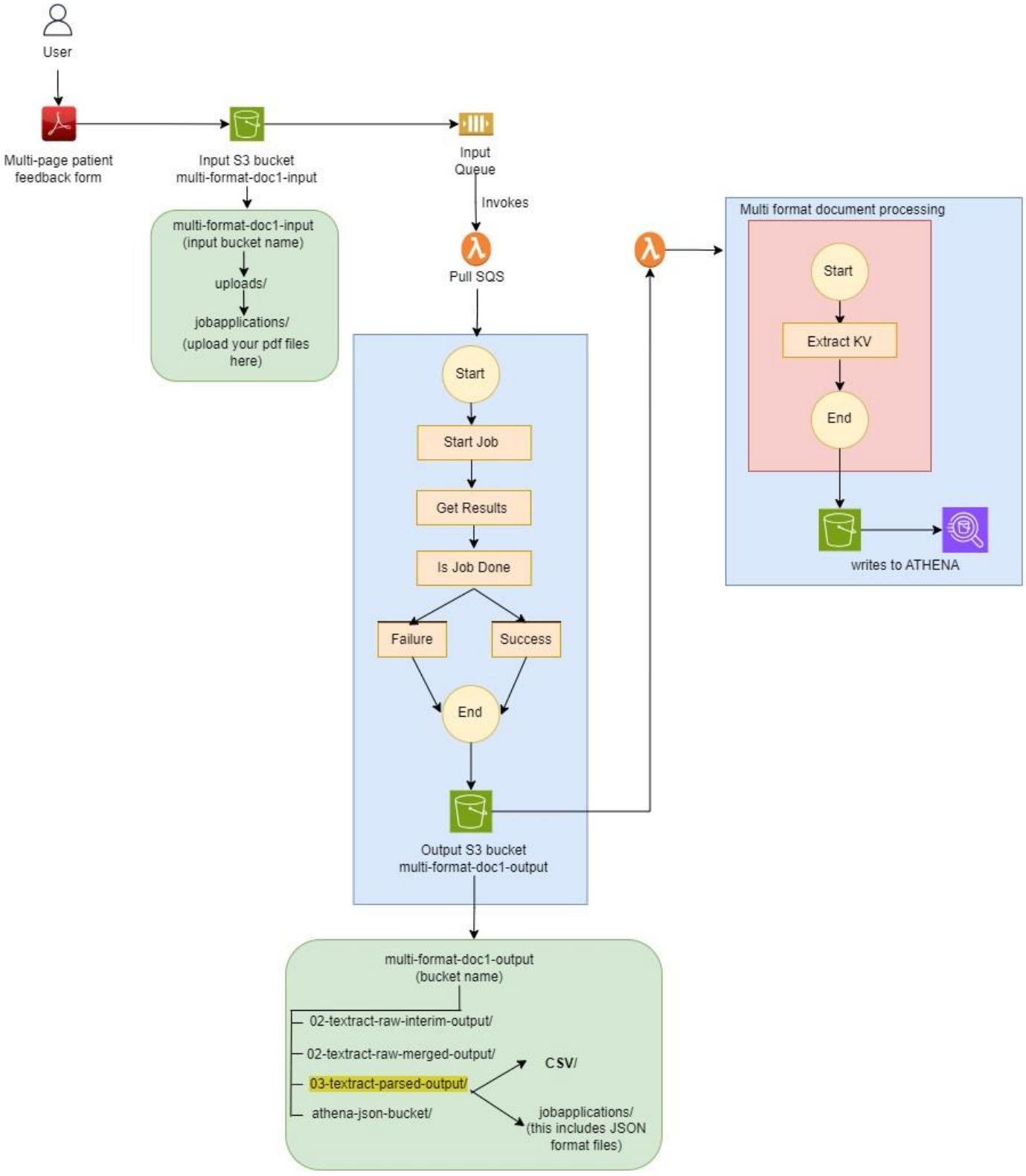
## OCR Workflow Diagram



## AWS resource names

- S3 Bucket input - multi-format-doc1-input
- Lambda function that stores the main code – multi-format-doc1-stack-Text-ExtractKeyValueLambda-YBBYnUVq7JVW
- Other lambda functions related to the stack of this pipeline(not changed from the original)
  multi-format-doc1-stack-TextExtract-StartJobLambda-dbNRsRdJCCQB

  multi-format-doc1-stack-TextExtra-GetResultsLambda-KvBilanklarV

  multi-format-doc1-stack-E-StartTextProcessingStack-Co1Fg4KtQ6lz

  multi-format-doc1-stack-IngestS-PullInputSQSLambda-rQRhu1HhK7p4
- S3 Bucket output – multi-format-doc1-output

## OCR Workflow Documentation

### Overview:

This document provides a detailed explanation of the OCR (Optical Character Recognition) workflow depicted in the diagram. The workflow is designed to process multi-page patient feedback forms and extract structured data, which can then be accessed in various formats such as CSV and JSON. Below is the step of the process, from uploading documents to retrieving the final output.

### Step-by-Step Workflow Description:

- Multi-page Patient Feedback Form Upload:
  Input: The user uploads a multi-page patient feedback form to the Input S3 Bucket named multi-format-doc1-input.

  Instructions:
  - Navigate to the S3 bucket multi-format-doc1-input.
  - Place your PDF files in the uploads/jobapplications/ directory within this bucket.
- S3 Input Bucket (multi-format-doc1-input):

The S3 bucket acts as the storage location where the uploaded forms are initially placed.

Directory Structure:

- uploads/: The root directory where user files are stored.
- jobapplications/: A subdirectory specifically for uploading PDF files to be processed.

- Processing Initiation:

  Once the document is uploaded, it triggers the Input Queue.

  AWS Lambda Function pulls messages from this queue and initiates the document processing job.

- Document Processing:

  Start Job: The process begins by starting a job that extracts key-value pairs (KVs) and other necessary data from the document.

  Get Results: The extracted data is then processed and checked for completion.

  Is Job Done?: The system verifies if the job is completed successfully.

  Failure: If the job fails, the process ends, and an error message is returned.

  Success: If the job is successful, the extracted data is sent to the Output S3 Bucket named multi-format-doc1-output.

- Output S3 Bucket (multi-format-doc1-output):

  The output bucket stores the processed files in different formats.

  Directory Structure:

  02-textract-raw-interim-output/: Contains raw outputs from the Textract process.

  02-textract-raw-merged-output/: Contains merged raw outputs.

  03-textract-parsed-output/: Contains the final parsed output, available in CSV format and JSON formats.

  athena-json-bucket/: Contains JSON format files.

- Final Output Access:

  CSV Files:

  You can find the CSV files in the 03-textract-parsed-output/csv/ directory within the multi-format-doc1-output bucket.

  JSON Files:

  JSON files are available in the 03-textract-parsed-output/jobapplications directory within the same output bucket.

- Multi-format Document Processing:

  The workflow includes a step where the processed data is written to Athena for further querying and analysis.

  Instructions:

  Use AWS Athena to run SQL queries against the structured data stored in the output S3 bucket.

**Summary:**

This workflow provides an automated pipeline for processing multi-page patient feedback forms, extracting data, and storing it in various formats for further analysis. The user needs to upload documents to the specified input bucket, and they can retrieve processed outputs from the designated directories in the output bucket.