## Lab 5-06: Automating Document Insights with Amazon Textract

|  |
| --- |
| **Introduction**  Amazon Textract is an AWS service that extracts text, key-value pairs (forms), and tables from scanned documents (PDFs/images) using machine learning, with no manual data entry required. It can produce structured JSON output that your applications can read to automate workflows such as extracting invoice fields, reading forms, or pulling data from receipts.  **Challenge**  Your company receives many scanned invoices and supplier forms as PDFs. Manually copying invoice numbers, dates, totals, and table line items into the accounting system takes too long and is error-prone. You will use Amazon Textract to automatically extract text, form fields, and tables from uploaded documents and save the structured results (JSON) to an S3 bucket so downstream systems (or people) can consume them.  **Lab Diagram**  **Solution**  **Step 1: Create an S3 bucket to hold documents**   1. Open the AWS Console.      1. Search S3.      1. Click on Create bucket.      1. Name it like textract-docs-yourname (must be unique).      1. Leave default settings for the lab and click Create bucket.     **Step 2: Upload a sample document**   1. Open your new bucket.      1. Click Upload.      1. Pick a PDF or JPEG/PNG image (for example, invoice1.pdf).      1. Then click Upload.     **Step 3: Open Amazon Textract Console**   1. In the AWS Console search, type Textract and open Amazon Textract.     **Step 4: Use “Analyze document (forms & tables)” demo**   1. In the Textract console, choose Analyze document (or Document analysis).      1. Click on Choose document.      1. Upload your PDF, invoice1.pdf.      1. Choose analysis features: check Forms (key-value pairs) and Tables.      1. Click Apply configuration (or Start analysis).      1. Wait a few seconds, Textract will process the file.     **Step 5: View and download results**   1. When complete, the console displays recognized text, detected key/value fields, and tables.      1. Review the extracted fields (invoice number, date, totals) and table items.      1. Click Download result (JSON) to save structured output.     **Step 6: Inspect JSON**   1. Open the downloaded JSON file in a text editor.      1. Look for Blocks → entries with BlockType equal to KEY\_VALUE\_SET (forms) and TABLE / CELL (tables).      1. Use these keys in your app or manual process to map fields into your accounting system. |