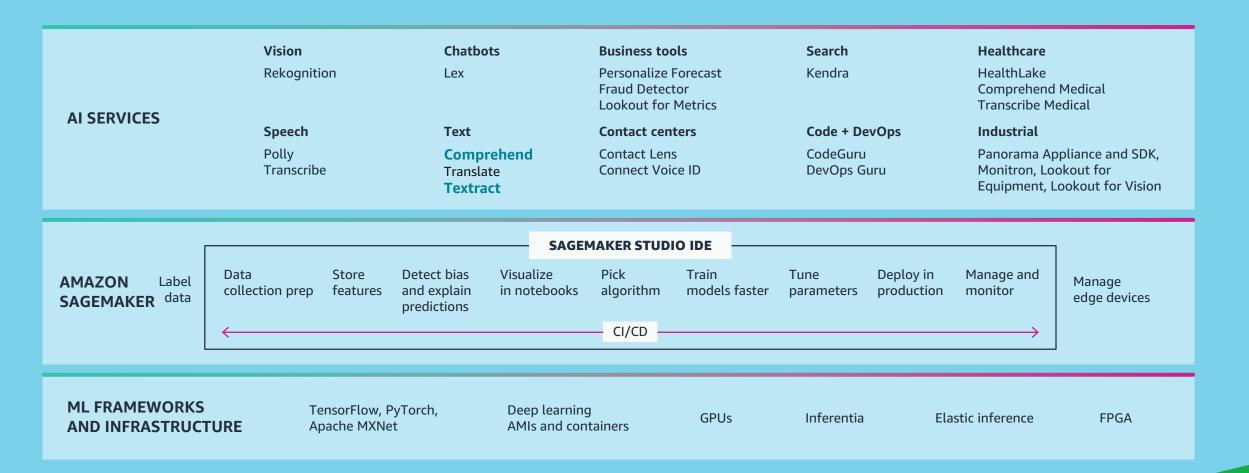


AI Workflow Automation for Document Processing

Raj Pathak

Solutions Architect, Amazon Web Services

The AWS ML stack Broadest and most complete set of machine learning capabilities







Business operations

Insurance



Real estate



Accounting



Tax management



Documents are a primary business tool across industries



Medical



Legal



Education



And many more...

How documents are processed today



Manual processing



💢 Error prone

X Time consuming



Traditional Optical Character Recognition (OCR)



💢 Error prone

Dump of text



Rules and templatebased extraction

Example 2 Limited by OCR accuracy

Development and management overhead

(XX) Templates are brittle

Why do these challenges exist

CERTIFICATE OF LIABILITY INSURANCE IS ISSUED AS A MATTER OF INFORMATION ONLY AND CONFERS NO RIGHTS UPON THE CER I NOT APPRIMATIVELY OR NEGATIVELY AREND, EXTERD OR ALTER THE COVERAGE APPO looking. We use words such as anticipates, believes, expects, future looking statements. Forward-looking statements reflect manageme Actual results could differ materially for a variety of reasons, inclu Flagstaff FLG and we also manufacture and sell electronic devices and produce media content. Generally, we recognitions us sell from our inventory as product sales and recognize our net flates of revenue of items sold service sales. We seek to increase unit alse across our stores, through increased product selection, acr 320 So Cambridge Ln Flagstaff, AZ 86001 categories. We also offer other services such as compute, storage, and database offerings, fulfillment. 928-774-5747 Our financial focus is on long-term, sustainable growth in free cash flows. Free cash flows are d increasing operating income and efficiently managing accounts receivable, inventory, accounts payabl Subtotal: expenditures, including our decision to purchase or lease property and equipment. Increases in operatu-result from increases in sales of products and services and efficiently managing our operating costs, po-investments we make in longer-term strategic initiatives, including capital expenditures focused on im-experience. To increase sales of products and services, we focus on improving all aspects of the custor Net Sales: lowering prices, improving availability, offering faster delivery and performance times, increasing sele-content, increasing product categories and service offerings, expanding product information, improvin-reliability, and earning customer trust. See "Results of Operations — Non-GA-8F risancial Measures" Total: \$3.89 Sold Items: Paid: We seek to reduce our variable costs per unit and work to leverage our fixed costs. Our variable and content costs, payment processing and related transaction costs, picking, packaging, and preparing transportation, customer service support, costs necessary to run AWS, and a portion of our marketing of the cost of the Debit include the costs necessary to build and run our technology infrastructure; to build, enhance, and add f stores, web services, electronic devices, and digital offerings; and to build and optimize our fulfillment facilities. Variable costs generally change directly with a less volume, whale fixed costs generally are departing needs, geographic expansion, category expansion, and other factors. To decrease our variable and enable us to lower prices for customers, we seek to increase our direct sourcing, increase discount MASSACHUSETTS DRIVER LICENSE 0.05 LAGSTAFF, 4.05 0.14 Tax/Fee Total: 01/20/2020 Jorge Souza

Documents are varied, and requirements for data are unique

Simply extracting data has become table stakes, adding insights and structure is what will provide organizations value from their documents

Item 7. Management's Discussion and Analysis of Financ

This Annual Report on Form 10-K includes forward-looking Litigation Reform Act of 1995. All statements other than statement: industry prospects, or future results of operations or financial posi-

changes in global economic conditions and customer spending, wo commerce and cloud services the amount that Imazon com inves investments, the mix of products and services sold to customers, the services, the extent to which we owe income or other taxes, compet operating results, international growth and expansion, the outcom proceedings, fulfillment, sortation, delivery, and data center optim demand, the degree to which we enter into, maintain, and develop

acquisitions and strategic transactions, payments risks, and risks of global economic climate and additional or unforeseen effects from risks and uncertainties, as well as other risks and uncertainties tha management's expectations, are described in greater detail in Item

information on our non-GAAP free cash flows financial measures.

Forward-Looking Statements

and digital content subscriptions.

reduce defects in our processes. To m

our high inventory velocity means a

The operating cycle is the mun accounts payable days.

Because of our model we are a

Solving these challenges with Intelligent Document Processing

Let's start by modelling the process into there unique phases of the document processing lifecycle

Ingestion

Extraction & Classification

Post-Processing and Transformation

Storage and Workflow Automation











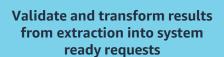






Ingest documents into centralized document repository from different sources (email, upload, fax, scan etc.)

Extract data and classify documents



Ingest data into backend systems for storage or workflow automation

Ingestion

The first step in the IDP lifecycle, ingestion allows us to store the document, and tag it with the appropriate metadata preparing it for processing



Digital Upload



Email or Fax



Document Scan







Using Amazon S3 as a landing zone for documents allows us to

- Tag documents with important metadata (timestamp, sender, document format) with Amazon S3 Object Tagging
- Allows for versioning and encryption of documents
- Provides low cost of storage, and storage tiering for archival data
- 11 9's of data durability
- WORM configurations can be applied to documents
- Built in integrations to AWS Services

Extraction & Classification

The next step in our IDP workflow, Extraction and Classification, we will extract data off different documents and classify our results



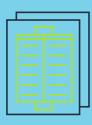


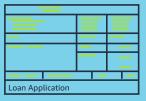


Amazon Comprehend

Using Amazon Textract will allow for extraction of data from complex documents and Amazon Comprehend will allow for granular classification and insight generation from extracted data

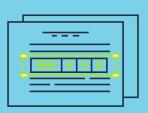
Amazon Textract capabilities





Text

Forms



Invoices

and receipts

Identity

documents

Specialized

documents

Tables

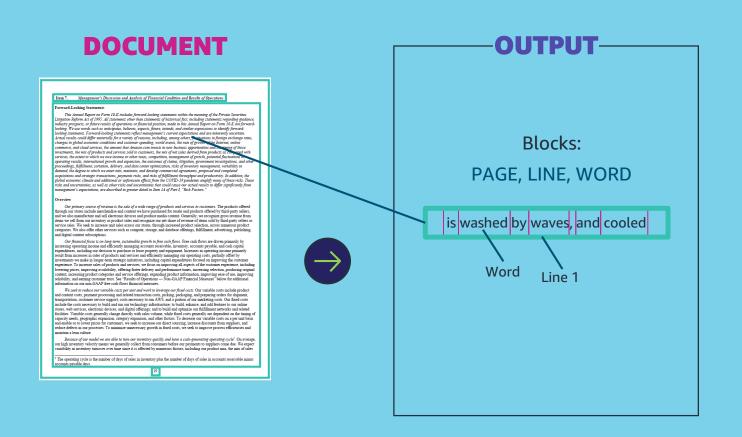




Handwriting

Text extraction

Optimal for dense text extraction with industry leading OCR accuracy





Outputs detected text in 3 hierarchy blocks: Page, Lines and Word



Bounding box for each Line and Word provide visual cues for post-processing



Included Confidence scores enable informed decision making for your workflows

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

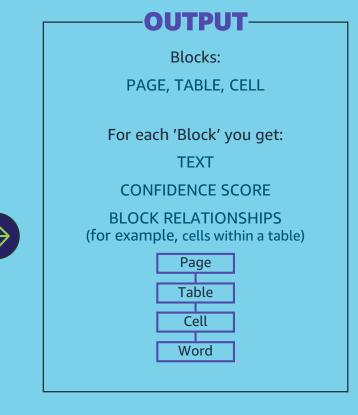
Table extraction

Extract tables from documents while preserving data structure and relationships

DOCUMENT

Previous employment history

Start date	End date	Employer name	Position held	Reason for leaving
1/15/2009	6/30/2013	Any company	Head Baker	Family relocated
8/15/2013	Present	Example corp.	Baker	N/A, current employer





Outputs recognized tables with relationships data intact



Intelligently groups cells within tables and words within each cell

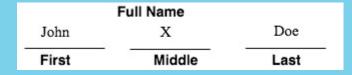


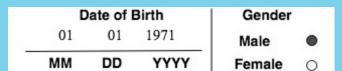
Output also includes confidence scores, geometry info, and row/column indexes

Form extraction

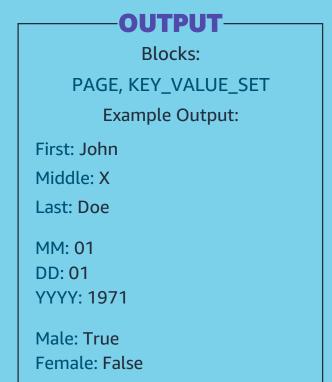
Extract form data from documents as key-value pairs to preserve document structure

DOCUMENT











Outputs form field name (Key) and field value name (Value) with relationship data intact



Captures logical groupings, relationships, and glyphs



Output also includes confidence scores, and geometry info

Invoices and receipts

Specialized support to process invoices and receipts at scale

DOCUMENT





Summary Fields:

Vendor Name: WHOLE FOODS MARKET

Subtotal (SUBTOTAL): \$3.50 Net Sales (OTHER): \$3.50

Tax/Fee (TAX): \$0.39 Sold Items (OTHER): 1

Paid (OTHER):

Debit (OTHER): \$3.89

Tax/Fee Total (TAX): \$0.39

Total (TOTAL): \$3.89

Line Items:

ITEM: Pizza Slice

PRICE: \$3.50



Outputs headline amounts, line item details and inferred fields (like Vendor Name)



Supports any style of invoice or receipt



No templates or configuration required

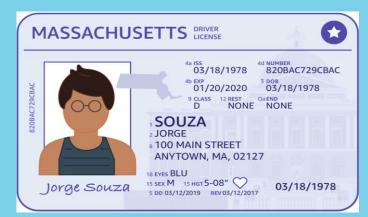


Identity documents



Specialized support for identity documents

DOCUMENT





OUTPUT

First Name: JORGE Last Name: SOUZA

Middle Name:

Address Line1: 100 MAIN STREET

Address Line 2: City: ANYTOWN

State: MA

Document Number: 820BAC729CBAC

Expiration Date: 01/20/2020 Date of Birth: 03/18/1978 ID Type: Driver License Date of Issue: 03/18/1978 Issued By: MASSACHUSETTS

Class: D

Restrictions: NONE Endorsements: NONE



95%+ accuracy for US driver licenses and passports



No templates or configuration required



Outputs normalized field names and supports implied elements



Amazon Comprehend IDP capabilities



Named Entity Recognition (NER)



Document Classification



PII Detection and Redaction

Entity Detection

Amazon.com, Inc. is located in Seattle, WA and was founded July 5th, 1994 by Jeff
Bezos. Known to the most customer obsessed organization, it welcomes thousands of customers and partners to one of its flagship events AWS re:Invent every year.

Amazon.com, Inc. Entity: ORGANIZATION

Seattle, WA Entity: LOCATION

July 5th, 1994 Entity: DATE

Jeff Bezos Entity: PERSON

re:Invent Entity: EVENT

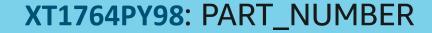
Entities detected: Person, Organization, Location, Date, Quantity, Title, Commercial Item, Event, Other

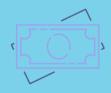
Custom Entity Detection

Customize entity detection to your specific requirements by training an AutoNLP model



"I never received the shipment for part number XT1764PY98"





"Trading of ETFs was halted today"

ETF: INVESTMENT_VEHICLE



"I liked the shwarma but loved the hummus"

shwarma, humus: MENU_ITEMS

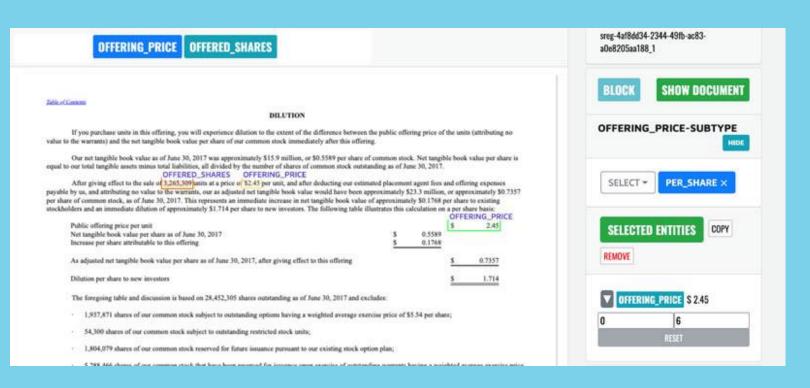
Custom Entity Detection with spatial information



Custom NER enhancements + document format support

Comprehend entity detection support for semi-structured documents

- Use positional and structural context in addition to natural language context to build NER Models
- Bullets, tables, forms, and more
- Don't have to develop 'flattening' strategies
- Pre-integrates Amazon Textract and Amazon Comprehend



Custom Classification

Classify documents to your specific requirements by training an AutoNLP classification model





Multi-class

- Classes are mutually exclusive
- Classification into one class
- Each model allows 1000 classes

Multi-label

- Classes are not mutually exclusive
- Classification into one or more classes
- Each model allows 100 unique classes

PII Detection

Hi, my name is <u>John Doe</u>. For verification, the last 4 digits of my social are <u>6789</u> and my dob is 01/01.

I paid for my credit card 1111-0000-1111-0000 last week from my bank account XXXXXXX1111 with the routing number XXXXXX0000. The check was mailed from 100 Main Street, Anytown, WA 98121.

Please confirm receipt by calling me at 206-555-0199 or emailing at john.doe@anycompany.com.

John Doe Entity: NAME

6789 Entity: SSN

01/01 Entity: DATE_TIME

1111-0000-1111-0000 Entity: CREDIT_DEBIT_NUMBER

XXXXXX1111 Entity: BANK_ACCOUNT_NUMBER

XXXXX0000 Entity: BANK_ROUTING

100 Main Street, Anytown, WA 98121 Entity: ADDRESS

206-555-0199 Entity: PHONE

PII handing options: Detect, Redact, Mask

Detect

Hi, my name is <u>John Doe</u>. For verification, the last 4 digits of my social are 6789 and my dob is 01/01.

I paid for my credit card 1111-0000-1111-0000 last week from my bank account XXXXXXX1111 with the routing number XXXXXX0000. The check was mailed from 100 Main Street, Anytown, WA 98121.

Please confirm receipt by calling me at 206-555-0199 or emailing at john.doe@anycompany.com.

Redact

Hi, my name is [NAME]. For verification, the last 4 digits of my social are [SSN] and my dob is [DATE_TIME].

I paid for my credit card
[CREDIT_DEBIT_NUMBER] last week
from my bank account
[BANK_ACCOUNT_NUMBER] with the
routing number [BANK_ROUTING].
The check was mailed from
[ADDRESS].

Please confirm receipt by calling me at [PHONE] or emailing at [EMAIL].

Mask

Hi, my name is *******. For verification, the last 4 digits of my social are **** and my dob is *****.

I paid for my credit card

************** last week from my
bank account ******** with the
routing number *******. The check
was mailed from

Putting it together

Let's understand where each API can be used as apart of our extraction and classification workflow









Identify based on document type which Textract extraction API is most appropriate (Text Detection, Forms, Tables, Invoices, ID etc.)

If the extracted data from Amazon Textract is sufficient for the required use case send results to system of record storage and/or post-processing







Amazon Comprehend

If further granularity is required from the documents such as identifying document class, identifying specific entities or redacting PII use the appropriate Amazon Comprehend API on extracted data from Amazon Textract

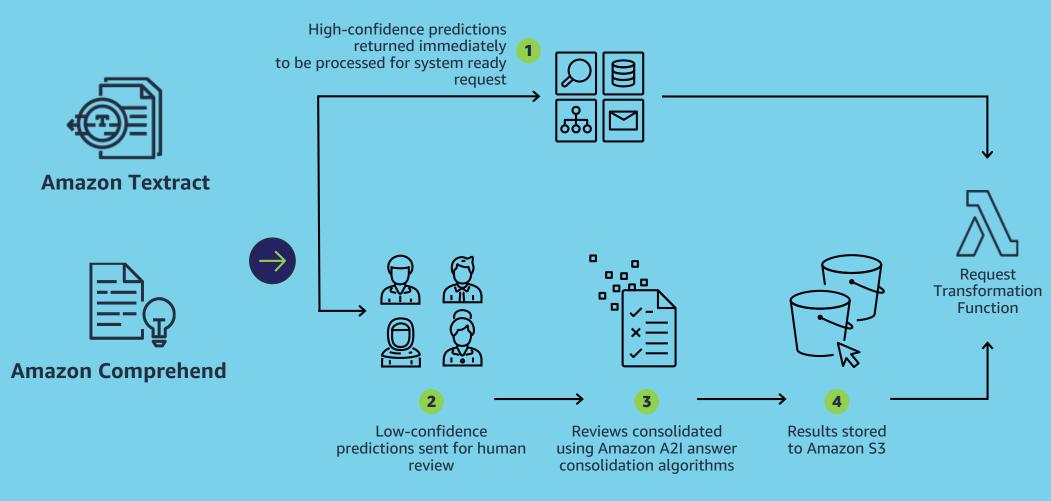




System of Record Storage

Post-Processing and Transformation

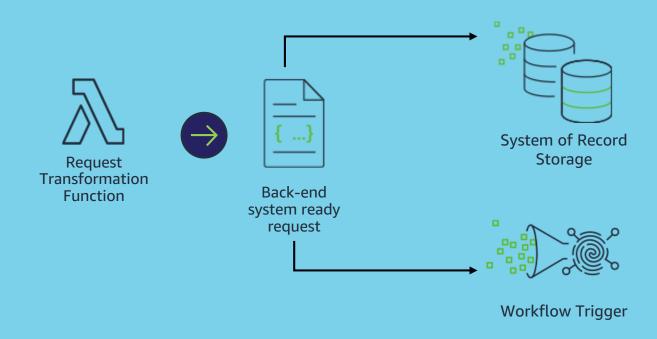
The next step in our IDP workflow, Post-Processing and Transformation, ensures our output data is of high quality and is ready to ingest into our back-end storage and workflows



23

Storage and Workflow Automation

The final step in our IDP workflow, Storage and Workflow Automation, acts as the final stop for our document data, it is used to store the document in a system of record, or used to trigger back-end automation



An end to end IDP example visualized - Insurance Policy Servicing



Broker fills out an semistructured insurance policy servicing document and emails document to Insurance carrier







Document is sent to **Amazon S3** for storage and is tagged with sender, timestamp metadata





Text is extracted from the document, along with any tables or key-value pairs using **Amazon Textract**











Document is classified as a Policy servicing document and tagged, and required servicing action is identified using **Amazon** Comprehend





The extracted data from **Amazon Textract** and **Amazon Comprehend** is transformed into a policy servicing API request





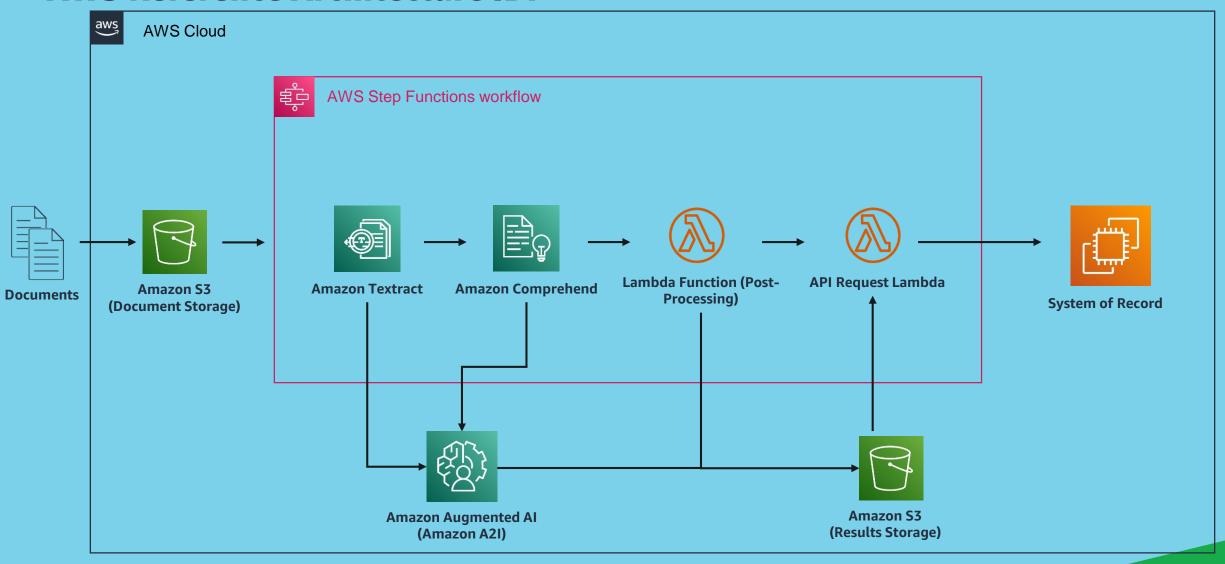
The changed policy is stored in the policy admin system





The servicing request is sent to policy management system API to trigger a policy servicing action

AWS Reference Architecture IDP



26



Thank you!