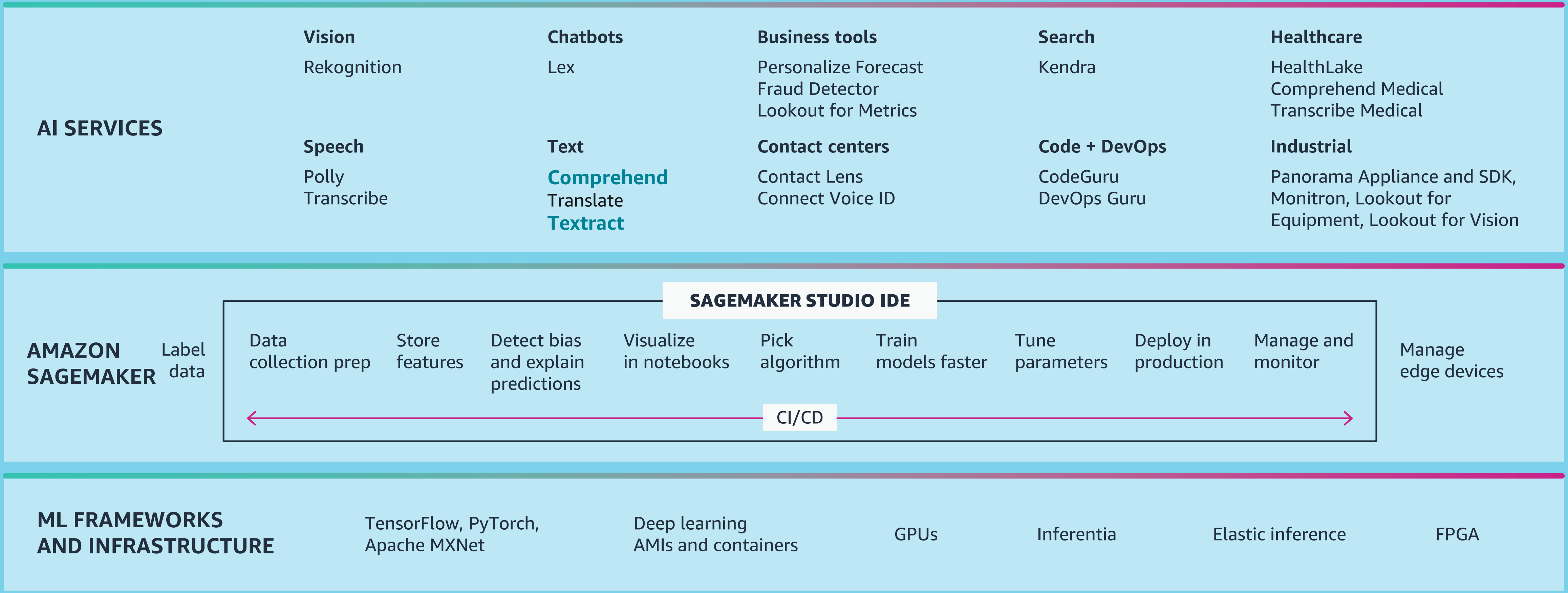
AI Workflow Automation for Document Processing

Automating data extraction from documents with AWS AI Services

Raj Pathak
Solutions Architect, Amazon Web Services

The AWS ML stack

Broadest and most complete set of machine learning capabilities



AI SERVICES

Vision
Rekognition

Chatbots
Lex

Business tools
Personalize Forecast
Fraud Detector
Lookout for Metrics

Search
Kendra

Healthcare
HealthLake
Comprehend Medical
Transcribe Medical

Speech
Polly
Transcribe

Text
Comprehend
Translate
Textract

Contact centers
Contact Lens
Connect Voice ID

Code + DevOps
CodeGuru
DevOps Guru

Industrial
Panorama Appliance and SDK, Monitron, Lookout for Equipment, Lookout for Vision

AMAZON SAGEMAKER

Label data

SAGEMAKER STUDIO IDE

Data collection prep
Store features
Detect bias and explain predictions
Visualize in notebooks
Pick algorithm
Train models faster
Tune parameters
Deploy in production
Manage and monitor

CI/CD

Manage edge devices

ML FRAMEWORKS AND INFRASTRUCTURE

TensorFlow, PyTorch, Apache MXNet
Deep learning AMIs and containers
GPUs
Inferentia
Elastic inference
FPGA

Documents are a primary business tool across industries

Finance

Insurance

Real estate

Accounting

Tax management

Business operations

Medical

Legal

Education

And many more…

How documents are processed today

Manual processing

- Expensive
- Error prone
- Time consuming

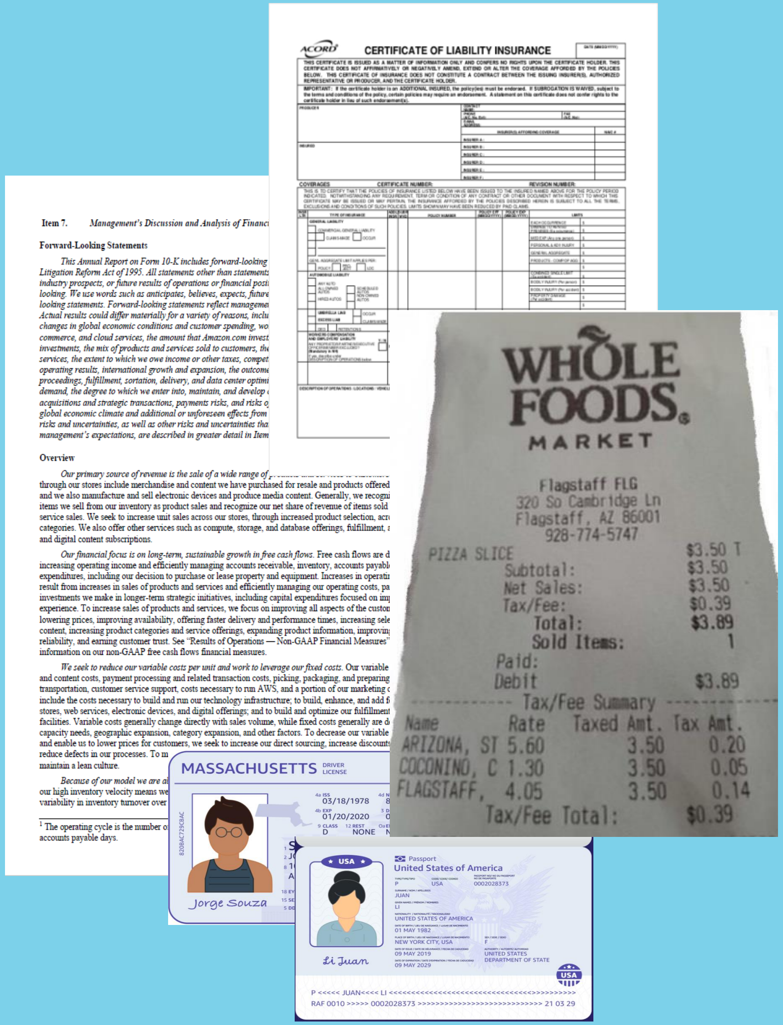
Traditional Optical Character Recognition (OCR)

- Simple documents only
- Error prone
- Dump of text

Rules and template-based extraction

- Limited by OCR accuracy
- Development and management overhead
- Templates are brittle

Why do these challenges exist



ACORD

CERTIFICATE OF LIABILITY INSURANCE

DATE (MM/DD/YYYY)

THIS CERTIFICATE IS ISSUED AS A MATTER OF INFORMATION ONLY AND CONFERS NO RIGHTS UPON THE CERTIFICATE HOLDER. THIS CERTIFICATE DOES NOT AFFIRMATIVELY OR NEGATIVELY AMEND, EXTEND OR ALTER THE COVERAGE AFFORDED BY THE POLICIES BELOW. THIS CERTIFICATE OF INSURANCE DOES NOT CONSTITUTE A CONTRACT BETWEEN THE ISSUING INSURER(S), AUTHORIZED REPRESENTATIVE OR PRODUCER, AND THE CERTIFICATE HOLDER.

IMPORTANT: If the certificate holder is an ADDITIONAL INSURED, the policy(ies) must be endorsed. If SUBROGATION IS WAIVED, subject to the terms and conditions of the policy, certain policies may require an endorsement. A statement on this certificate does not confer rights to the certificate holder in lieu of such endorsement(s).

PRODUCER

INSURED

INSURER(S) AFFORDING COVERAGE

INSURER A:
INSURER B:
INSURER C:
INSURER D:
INSURER E:
INSURER F:

COVERAGES CERTIFICATE NUMBER: REVISION NUMBER:

THIS IS TO CERTIFY THAT THE POLICIES OF INSURANCE LISTED BELOW HAVE BEEN ISSUED TO THE INSURED NAMED ABOVE FOR THE POLICY PERIOD INDICATED. NOTWITHSTANDING ANY REQUIREMENT, TERM OR CONDITION OF ANY CONTRACT OR OTHER DOCUMENT WITH RESPECT TO WHICH THIS CERTIFICATE MAY BE ISSUED OR MAY PERTAIN, THE INSURANCE AFFORDED BY THE POLICIES DESCRIBED HEREIN IS SUBJECT TO ALL THE TERMS, EXCLUSIONS AND CONDITIONS OF SUCH POLICIES. LIMITS SHOWN MAY HAVE BEEN REDUCED BY PAID CLAIMS.

TYPE OF INSURANCE | POLICY NUMBER | POLICY EFF (MM/DD/YYYY) | POLICY EXP (MM/DD/YYYY) | LIMITS

GENERAL LIABILITY
□ COMMERCIAL GENERAL LIABILITY
□ CLAIMS-MADE □ OCCUR
GEN'L AGGREGATE LIMIT APPLIES PER:
□ POLICY □ PRO-JECT □ LOC

EACH OCCURRENCE $
DAMAGE TO RENTED PREMISES (Ea occurrence) $
MED EXP (Any one person) $
PERSONAL & ADV INJURY $
GENERAL AGGREGATE $
PRODUCTS - COMP/OP AGG $

AUTOMOBILE LIABILITY
□ ANY AUTO
□ ALL OWNED AUTOS □ SCHEDULED AUTOS
□ HIRED AUTOS □ NON-OWNED AUTOS

COMBINED SINGLE LIMIT (Ea accident) $
BODILY INJURY (Per person) $

□ UMBRELLA LIAB □ OCCUR
□ EXCESS LIAB □ CLAIMS-MADE
DED □ RETENTION $

WORKERS COMPENSATION AND EMPLOYERS' LIABILITY
ANY PROPRIETOR/PARTNER/EXECUTIVE OFFICER/MEMBER EXCLUDED? □
(Mandatory in NH)
If yes, describe under DESCRIPTION OF OPERATIONS below

DESCRIPTION OF OPERATIONS / LOCATIONS / VEHICLES

Item 7. *Management's Discussion and Analysis of Financi*

Forward-Looking Statements

This Annual Report on Form 10-K includes forward-looking
Litigation Reform Act of 1995. All statements other than statements
industry prospects, or future results of operations or financial posi
looking. We use words such as anticipates, believes, expects, future
looking statements. Forward-looking statements reflect managemen
Actual results could differ materially for a variety of reasons, inclu
changes in global economic conditions and customer spending, wo
commerce, and cloud services, the amount that Amazon.com invest
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 outcome
proceedings, fulfillment, sortation, delivery, and data center optimi
demand, the degree to which we enter into, maintain, and develop
acquisitions and strategic transactions, payments risks, and risks o
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

Overview

Our primary source of revenue is the sale of a wide range of p
through our stores include merchandise and content we have purchased for resale and products offered
and we also manufacture and sell electronic devices and produce media content. Generally, we recogni
items we sell from our inventory as product sales and recognize our net share of revenue of items sold
service sales. We seek to increase unit sales across our stores, through increased product selection, acr
categories. We also offer other services such as compute, storage, and database offerings, fulfillment, a
and digital content subscriptions.

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
expenditures, including our decision to purchase or lease property and equipment. Increases in operati
result from increases in sales of products and services and efficiently managing our operating costs, pa
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 custon
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-GAAP Financial Measures"
information on our non-GAAP free cash flows financial measures.

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 c
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 sales volume, while fixed costs generally are d
capacity 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 discounts
reduce defects in our processes. To m
maintain a lean culture.

Because of our model we are al
our high inventory velocity means we
variability in inventory turnover over

[1] The operating cycle is the number o
accounts payable days.

WHOLE FOODS
MARKET

Flagstaff FLG
320 So Cambridge Ln
Flagstaff, AZ 86001
928-774-5747

PIZZA SLICE	$3.50 T
Subtotal:	$3.50
Net Sales:	$3.50
Tax/Fee:	$0.39
Total:	$3.89
Sold Items:	1
Paid:	
Debit	$3.89

Tax/Fee Summary

Name	Rate	Taxed Amt.	Tax Amt.
ARIZONA, ST	5.60	3.50	0.20
COCONINO, C	1.30	3.50	0.05
FLAGSTAFF,	4.05	3.50	0.14
Tax/Fee Total:			$0.39

MASSACHUSETTS DRIVER LICENSE

820BAC729CBAC

4a ISS 03/18/1978
4b EXP 01/20/2020
9 CLASS D
12 REST NONE

Jorge Souza

USA

Passport
United States of America

Type/Type/Tipo: P
Code/Code/Codigo: USA
Passport No/No du passeport/No de pasaporte: 0002028373
Surname/Nom/Apellidos: JUAN
Given names/Prénom/Nombres: LI
Nationality/Nationalité/Nacionalidad: UNITED STATES OF AMERICA
Date of birth/Lieu de naissance/Lugar de nacimiento: 01 MAY 1982
Place of birth/Lieu de naissance/Lugar de nacimiento: NEW YORK CITY, USA
Sex/Sexe/Sexo: F
Date of issue/Date de délivrance/Fecha de caducidad: 09 MAY 2019
Authority/Autorité/Autoridad: UNITED STATES DEPARTMENT OF STATE
Date of expiration/Date d'expiration/Fecha de caducidad: 09 MAY 2029

Li Juan

USA

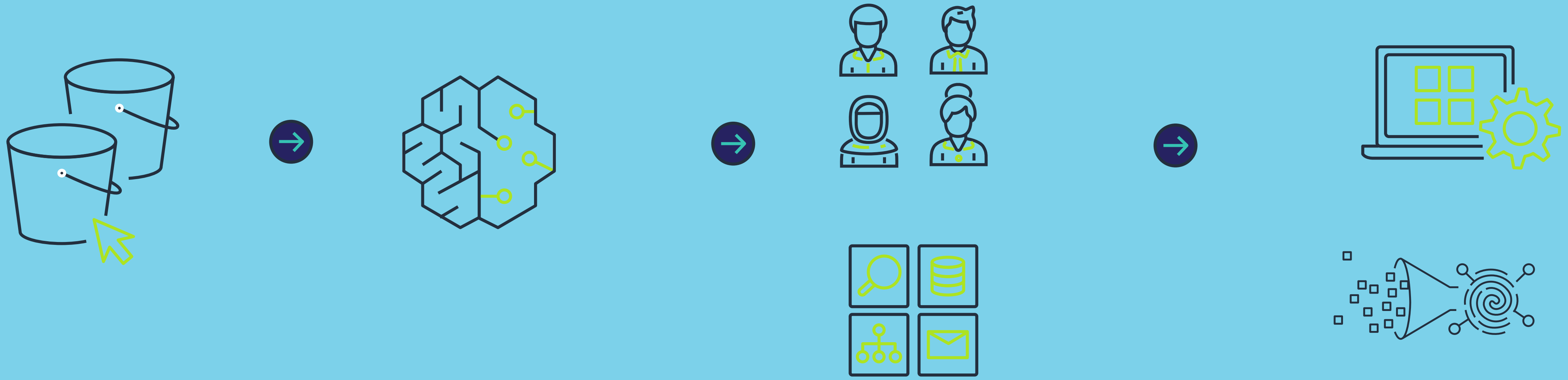
P <<<<< JUAN<<<< LI <<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<>>>>>>>>>>
RAF 0010 >>>>> 0002028373 >>>>>>>>>>>>>>>>>>>>>>>>>>>>> 21 03 29

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

Solving these challenges with Intelligent Document Processing

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



Ingestion

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

Extraction & Classification

Extract data and classify documents

Post-Processing and Transformation

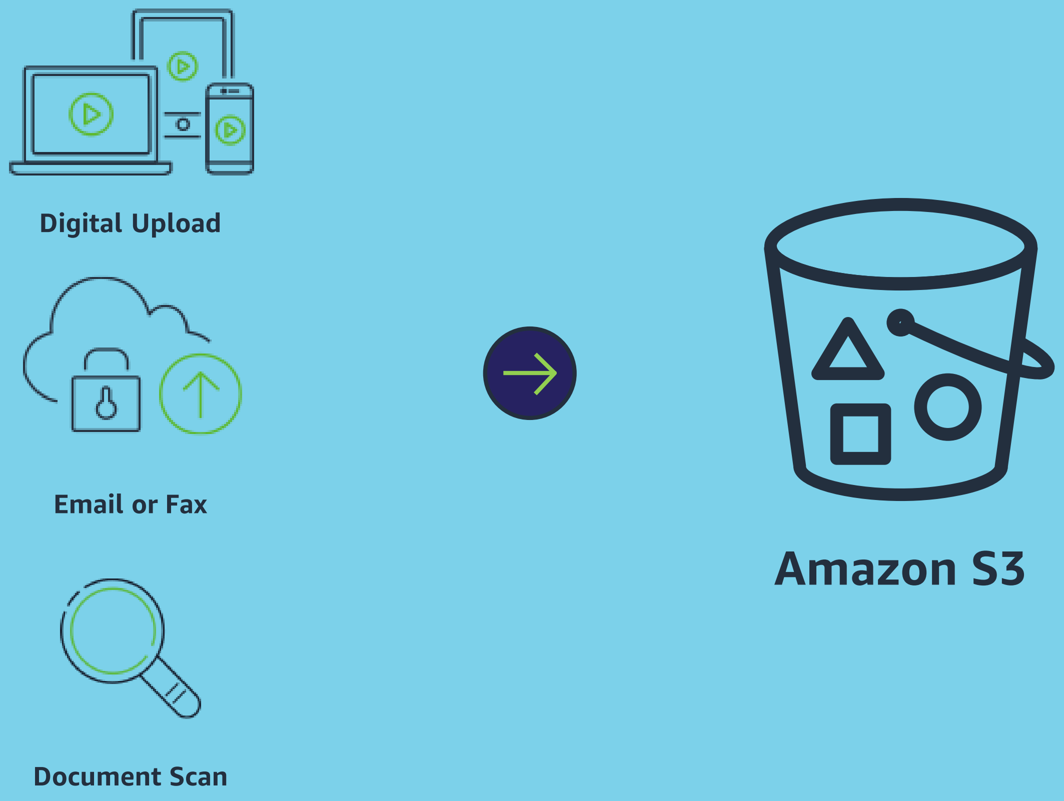
Validate and transform results from extraction into system ready requests

Storage and Workflow Automation

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

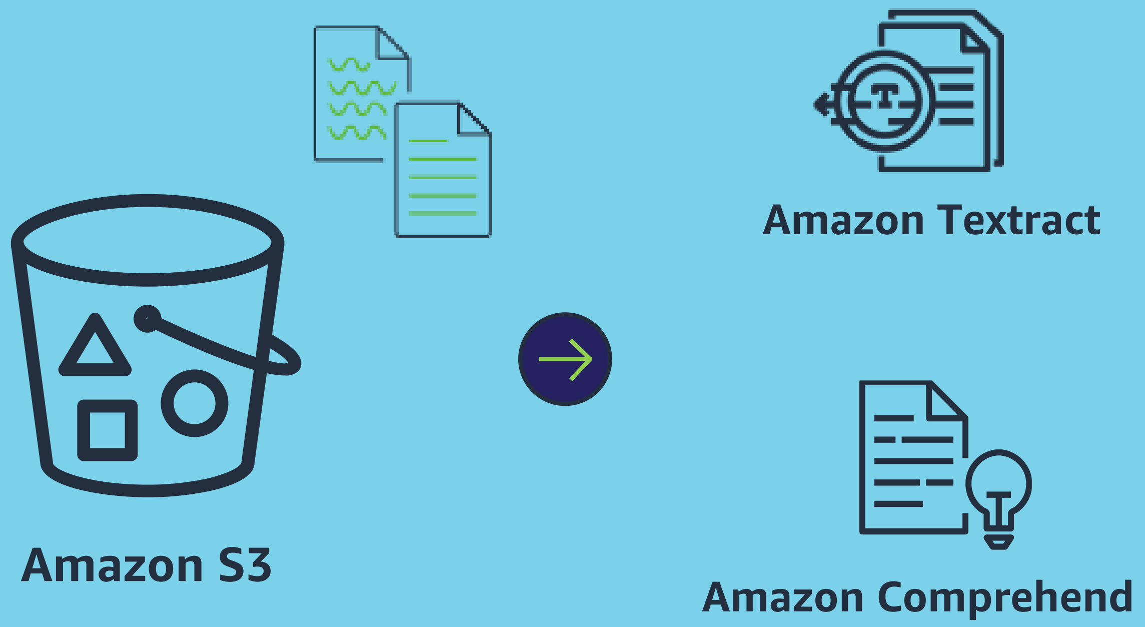
Amazon S3

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



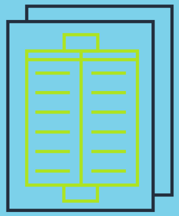
Amazon S3

Amazon Textract

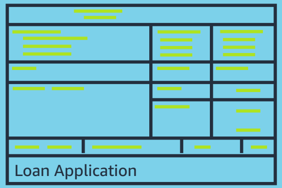
Amazon Comprehend

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

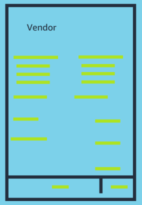


Text

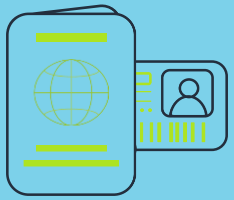


Forms

Invoices
and receipts

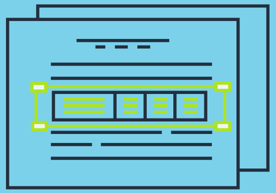


Identity
documents



Specialized documents

Handwriting

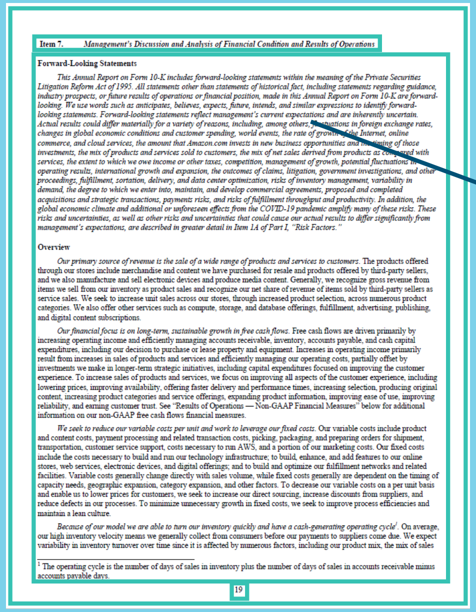


Tables

Text extraction

Optimal for dense text extraction with industry leading OCR accuracy

DOCUMENT



Item 7. *Management's Discussion and Analysis of Financial Condition and Results of Operations*

Forward-Looking Statements

This Annual Report on Form 10-K includes forward-looking statements within the meaning of the Private Securities Litigation Reform Act of 1995. All statements other than statements of historical fact, including statements regarding guidance, industry prospects, or future results of operations or financial position, made in this Annual Report on Form 10-K are forward-looking. We use words such as anticipates, believes, expects, future, intends, and similar expressions to identify forward-looking statements. Forward-looking statements reflect management's current expectations and are inherently uncertain. Actual results could differ materially for a variety of reasons, including, among others, fluctuations in foreign exchange rates, changes in global economic conditions and customer spending, world events, the rate of growth of the Internet, online commerce, and cloud services, the amount that Amazon.com invests in new business opportunities and the timing of those investments, the mix of products and services sold to customers, the mix of net sales derived from products as compared with services, the extent to which we owe income or other taxes, competition, management of growth, potential fluctuations in operating results, international growth and expansion, the outcomes of claims, litigation, government investigations, and other proceedings, fulfillment, sortation, delivery, and data center optimization, risks of inventory management, variability in demand, the degree to which we enter into, maintain, and develop commercial agreements, proposed and completed acquisitions and strategic transactions, payments risks, and risks of fulfillment throughput and productivity. In addition, the global economic climate and additional or unforeseen effects from the COVID-19 pandemic amplify many of these risks. These risks and uncertainties, as well as other risks and uncertainties that could cause our actual results to differ significantly from management's expectations, are described in greater detail in Item 1A of Part I, "Risk Factors."

Overview

Our primary source of revenue is the sale of a wide range of products and services to customers. The products offered through our stores include merchandise and content we have purchased for resale and products offered by third-party sellers, and we also manufacture and sell electronic devices and produce media content. Generally, we recognize gross revenue from items we sell from our inventory as product sales and recognize our net share of revenue of items sold by third-party sellers as service sales. We seek to increase unit sales across our stores, through increased product selection, across numerous product categories. We also offer other services such as compute, storage, and database offerings, fulfillment, advertising, publishing, and digital content subscriptions.

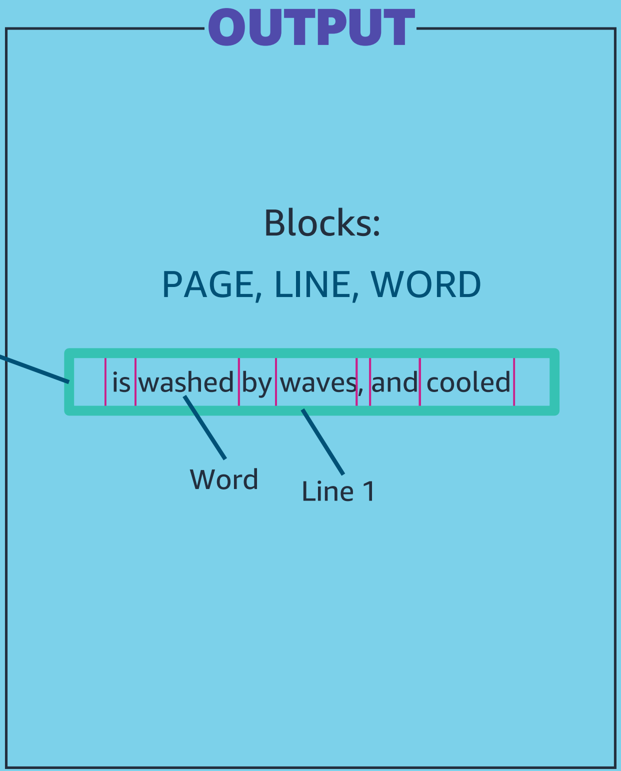
Our financial focus is on long-term, sustainable growth in free cash flows. Free cash flows are driven primarily by increasing operating income and efficiently managing accounts receivable, inventory, accounts payable, and cash capital expenditures, including our decision to purchase or lease property and equipment. Increases in operating income primarily result from increases in sales of products and services and efficiently managing our operating costs, partially offset by investments we make in longer-term strategic initiatives, including capital expenditures focused on improving the customer experience. To increase sales of products and services, we focus on improving all aspects of the customer experience, including lowering prices, improving availability, offering faster delivery and performance times, increasing selection, producing original content, increasing product categories and service offerings, expanding product information, improving ease of use, improving reliability, and earning customer trust. See "Results of Operations — Non-GAAP Financial Measures" below for additional information on our non-GAAP free cash flows financial measures.

We seek to reduce our variable costs per unit and work to leverage our fixed costs. Our variable costs include product and content costs, payment processing and related transaction costs, picking, packaging, and preparing orders for shipment, transportation, customer service support, costs necessary to run AWS, and a portion of our marketing costs. Our fixed costs include the costs necessary to build and run our technology infrastructure; to build, enhance, and add features to our online stores, web services, electronic devices, and digital offerings; and to build and optimize our fulfillment networks and related facilities. Variable costs generally change directly with sales volume, while fixed costs generally are dependent on the timing of capacity needs, geographic expansion, category expansion, and other factors. To decrease our variable costs on a per unit basis and enable us to lower prices for customers, we seek to increase our direct sourcing, increase discounts from suppliers, and reduce defects in our processes. To minimize unnecessary growth in fixed costs, we seek to improve process efficiencies and maintain a lean culture.

Because of our model we are able to turn our inventory quickly and have a cash-generating operating cycle[1]. On average, our high inventory velocity means we generally collect from consumers before our payments to suppliers come due. We expect variability in inventory turnover over time since it is affected by numerous factors, including our product mix, the mix of sales

[1] The operating cycle is the number of days of sales in inventory plus the number of days of sales in accounts receivable minus accounts payable days.

19



OUTPUT

Blocks:

PAGE, LINE, WORD

is washed by waves, and cooled

Word

Line 1

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

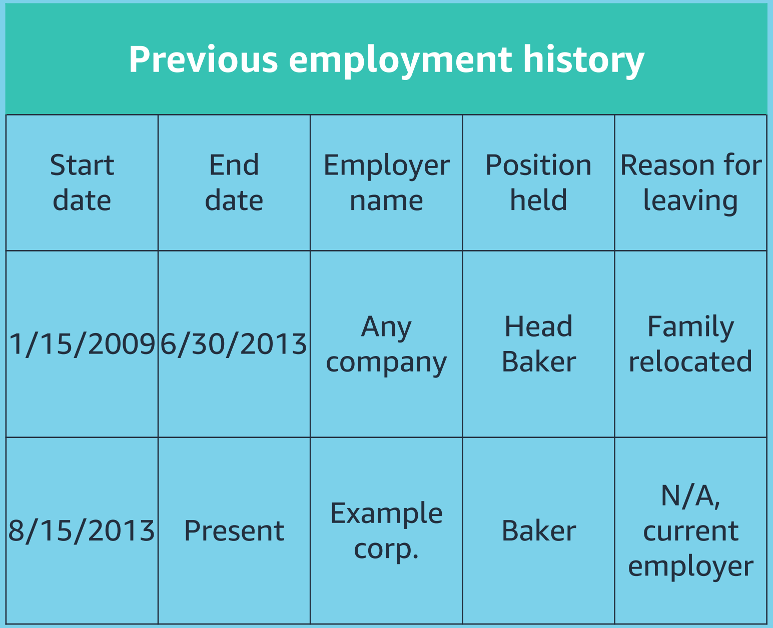
Bounding box for each Line and Word provide visual cues for post-processing

Included Confidence scores enable informed decision making for your workflows

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

OUTPUT

Blocks:

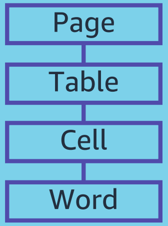
PAGE, TABLE, CELL

For each 'Block' you get:

TEXT

CONFIDENCE SCORE

BLOCK RELATIONSHIPS (for example, cells within a table)



✓ Outputs recognized tables with relationships data intact

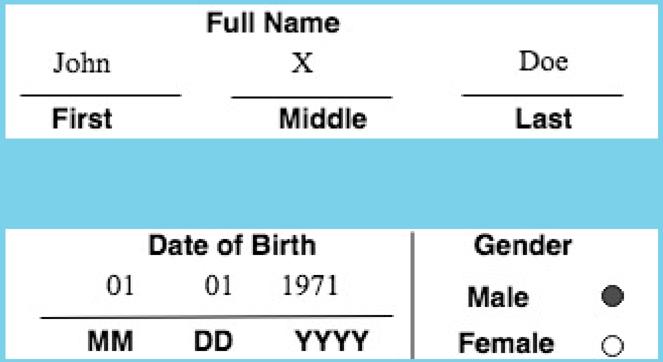
✓ Intelligently groups cells within tables and words within each cell

✓ 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



Full Name

John	X	Doe
First	Middle	Last

Date of Birth			Gender	
01	01	1971	Male	●
MM	DD	YYYY	Female	○

→

OUTPUT

Blocks:

PAGE, KEY_VALUE_SET

Example Output:

First: John

Middle: X

Last: Doe

MM: 01
DD: 01
YYYY: 1971

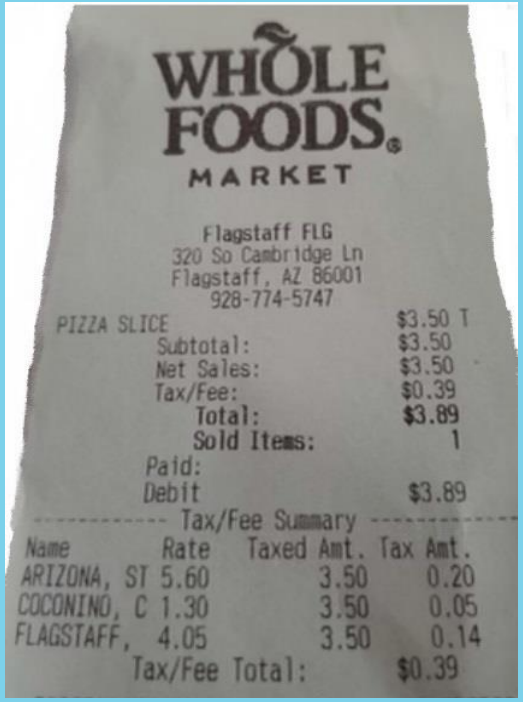
Male: True
Female: False

- 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



WHOLE FOODS MARKET

Flagstaff FLG
320 So Cambridge Ln
Flagstaff, AZ 86001
928-774-5747

PIZZA SLICE	$3.50 T
Subtotal:	$3.50
Net Sales:	$3.50
Tax/Fee:	$0.39
Total:	$3.89
Sold Items:	1
Paid:	
Debit	$3.89

Tax/Fee Summary

Name	Rate	Taxed Amt.	Tax Amt.
ARIZONA, ST	5.60	3.50	0.20
COCONINO, C	1.30	3.50	0.05
FLAGSTAFF,	4.05	3.50	0.14

Tax/Fee Total: $0.39

OUTPUT

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

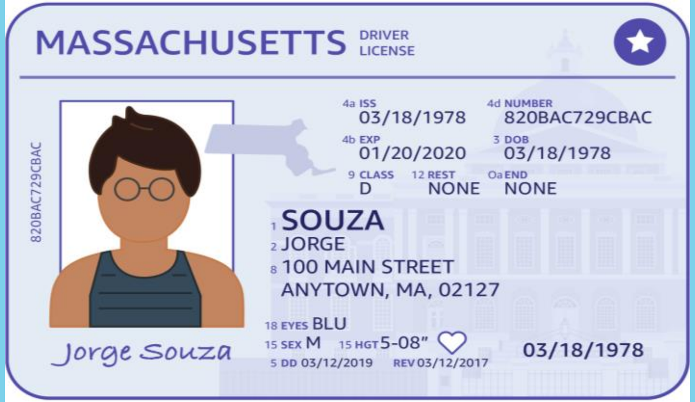
✓ No templates or configuration required

Identity documents

NEW!

Specialized support for identity documents

DOCUMENT



MASSACHUSETTS DRIVER LICENSE

820BAC729CBAC

4a ISS 03/18/1978

4d NUMBER 820BAC729CBAC

4b EXP 01/20/2020

3 DOB 03/18/1978

9 CLASS D

12 REST NONE

Oa END NONE

1 SOUZA

2 JORGE

8 100 MAIN STREET
ANYTOWN, MA, 02127

18 EYES BLU

15 SEX M

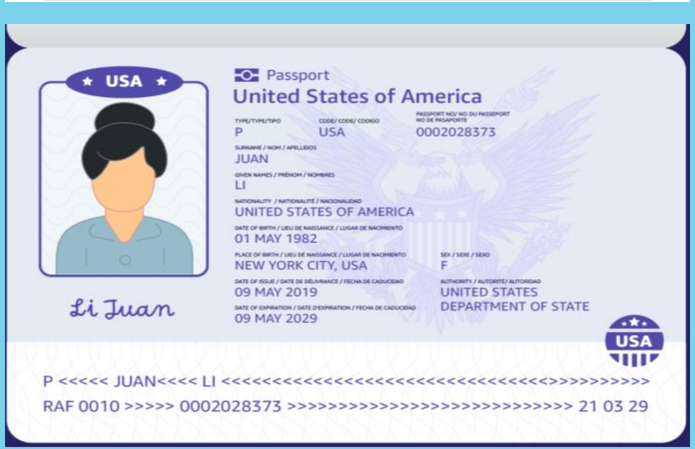
15 HGT 5-08"

5 DD 03/12/2019

REV 03/12/2017

03/18/1978

Jorge Souza



USA

Passport

United States of America

TYPE/TYPE/TIPO P

CODE/ CODE/ CODIGO USA

PASSPORT NO/ NO DU PASSEPORT NO DE PASAPORTE 0002028373

SURNAME / NOM / APELLIDOS JUAN

GIVEN NAMES / PRÉNOM / NOMBRES LI

NATIONALITY / NATIONALITÉ / NACIONALIDAD UNITED STATES OF AMERICA

DATE OF BIRTH / LIEU DE NAISSANCE / LUGAR DE NACIMIENTO 01 MAY 1982

PLACE OF BIRTH / LIEU DÉ NAISSANCE / LUGAR DE NACIMIENTO NEW YORK CITY, USA

SEX / SEXE / SEXO F

DATE OF ISSUE / DATE DE DÉLIVRANCE / FECHA DE CADUCIDAD 09 MAY 2019

AUTHORITY / AUTORITÉ/ AUTORIDAD UNITED STATES DEPARTMENT OF STATE

DATE OF EXPIRATION / DATE D'EXPIRATION / FECHA DE CADUCIDAD 09 MAY 2029

Li Juan

USA

P <<<<< JUAN<<<< LI <<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<>>>>>>>>>
RAF 0010 >>>>> 0002028373 >>>>>>>>>>>>>>>>>>>>>>>>>>>>>> 21 03 29

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
thousands of customers	Entity: QUANTITY
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**”

XT1764PY98: PART_NUMBER

“Trading of **ETFs** was halted today”

ETF: INVESTMENT_VEHICLE

“I liked the **shwarma** but loved the **hummus**”

shwarma, humus: MENU_ITEMS

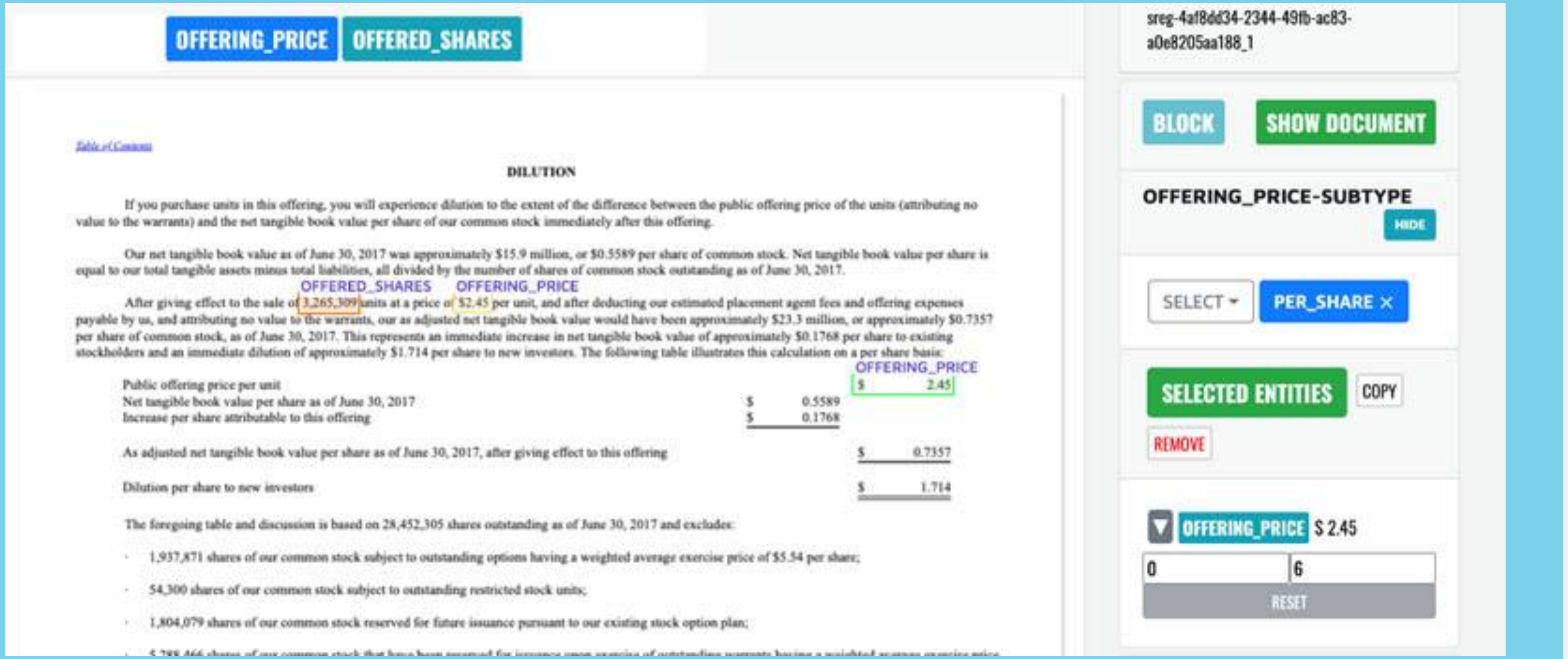
Custom Entity Detection with spatial information

NEW!

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



DILUTION

If you purchase units in this offering, you will experience dilution to the extent of the difference between the public offering price of the units (attributing no value to the warrants) and the net tangible book value per share of our common stock immediately after this offering.

Our net tangible book value as of June 30, 2017 was approximately $15.9 million, or $0.5589 per share of common stock. Net tangible book value per share is equal to our total tangible assets minus total liabilities, all divided by the number of shares of common stock outstanding as of June 30, 2017.

After giving effect to the sale of 3,265,309 units at a price of $2.45 per unit, and after deducting our estimated placement agent fees and offering expenses payable by us, and attributing no value to the warrants, our as adjusted net tangible book value would have been approximately $23.3 million, or approximately $0.7357 per share of common stock, as of June 30, 2017. This represents an immediate increase in net tangible book value of approximately $0.1768 per share to existing stockholders and an immediate dilution of approximately $1.714 per share to new investors. The following table illustrates this calculation on a per share basis:

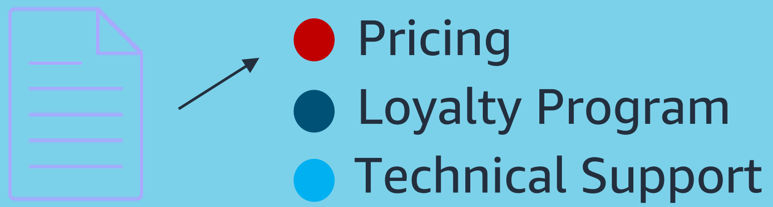
Public offering price per unit		$ 2.45
Net tangible book value per share as of June 30, 2017	$ 0.5589	
Increase per share attributable to this offering	$ 0.1768	
As adjusted net tangible book value per share as of June 30, 2017, after giving effect to this offering		$ 0.7357
Dilution per share to new investors		$ 1.714

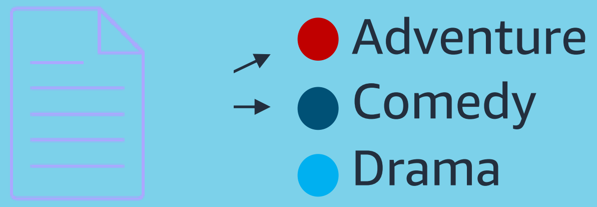
The foregoing table and discussion is based on 28,452,305 shares outstanding as of June 30, 2017 and excludes:

- 1,937,871 shares of our common stock subject to outstanding options having a weighted average exercise price of $5.54 per share;
- 54,300 shares of our common stock subject to outstanding restricted stock units;
- 1,804,079 shares of our common stock reserved for future issuance pursuant to our existing stock option plan;
- 5,288,466 shares of our common stock that have been reserved for issuance upon exercise of outstanding warrants having a weighted average exercise price

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 John Doe. 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 XXXXXX1111 with the routing number XXXXX0000. 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
john.doe@anycompany.com	Entity: EMAIL

PII handing options: Detect, Redact, Mask

Detect

Hi, my name is John Doe. 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 XXXXXX1111 with the routing number XXXXX0000. 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

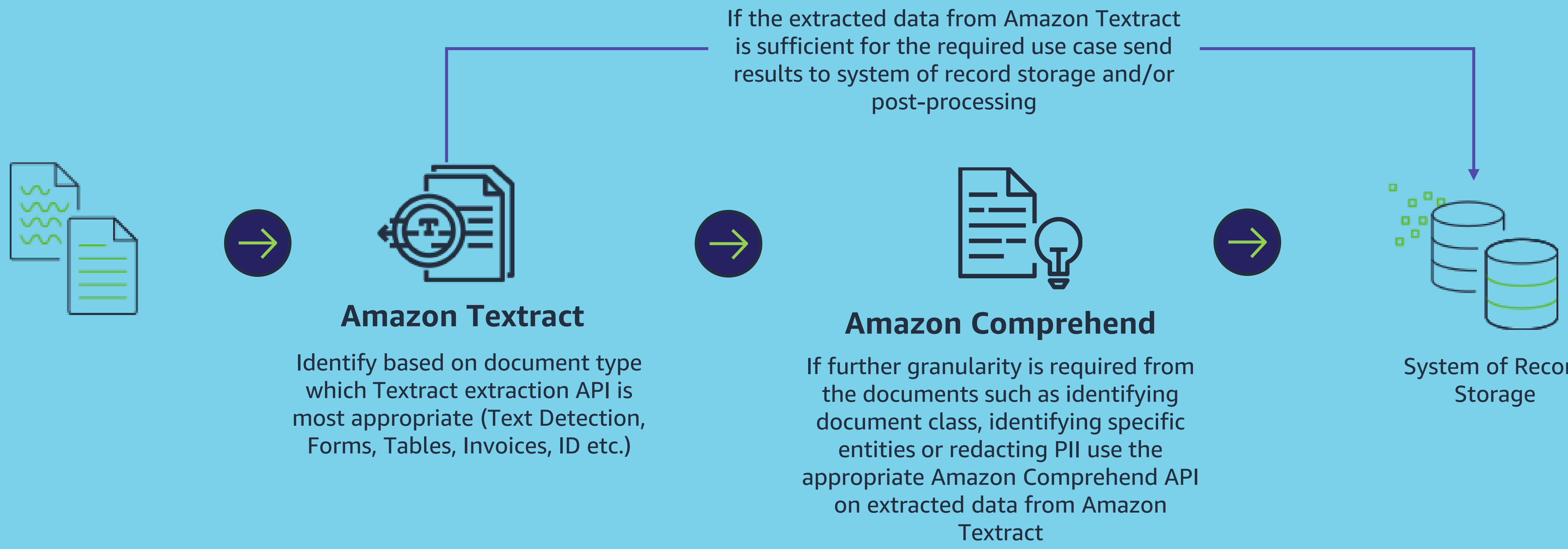
Hi, my name is ********. For verification, the last 4 digits of my social are **** and my dob is *****.

I paid for my credit card ******************* last week from my bank account ********** with the routing number **********. The check was mailed from *************************.

Please confirm receipt by calling me at ************ or emailing at ***********************.

Putting it together

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



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 Textract

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

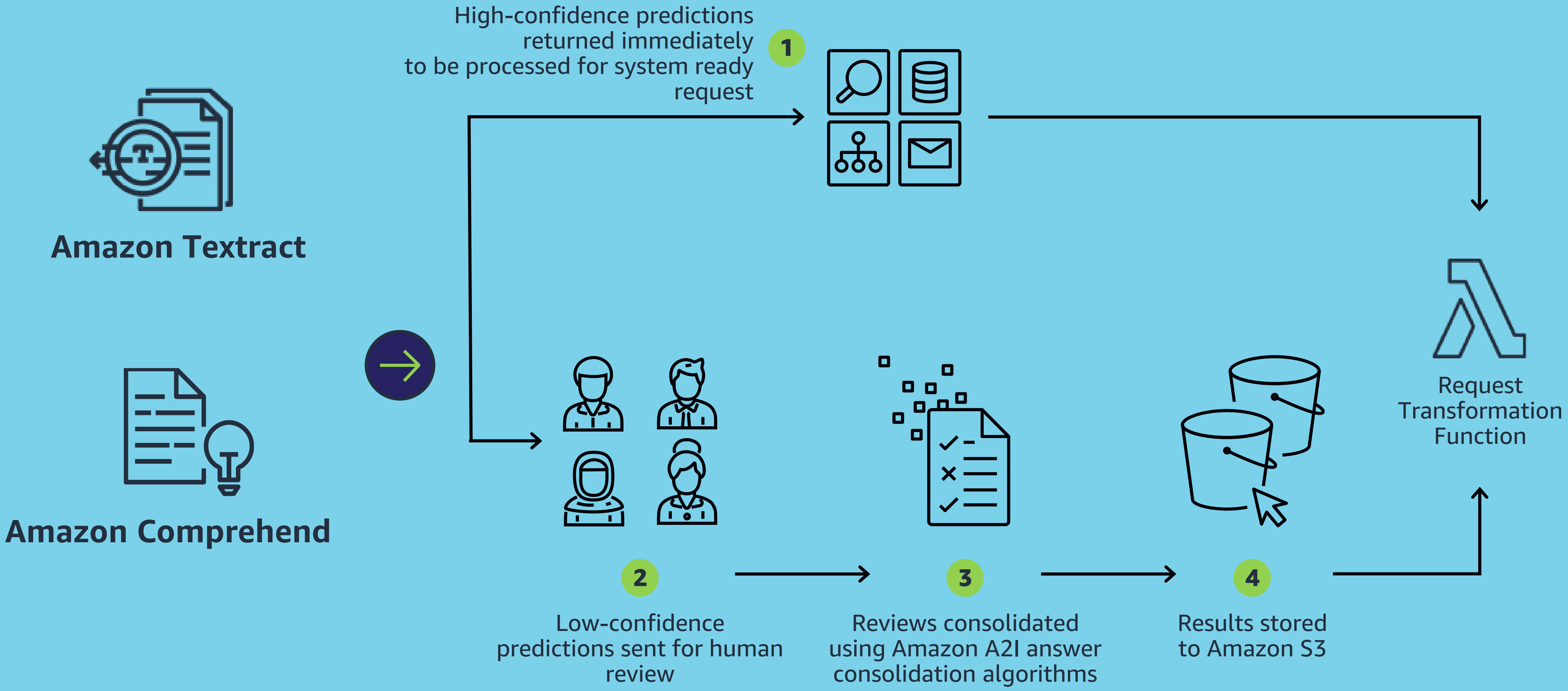
Amazon Comprehend

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



Amazon Textract

Amazon Comprehend

1 High-confidence predictions returned immediately to be processed for system ready request

2 Low-confidence predictions sent for human review

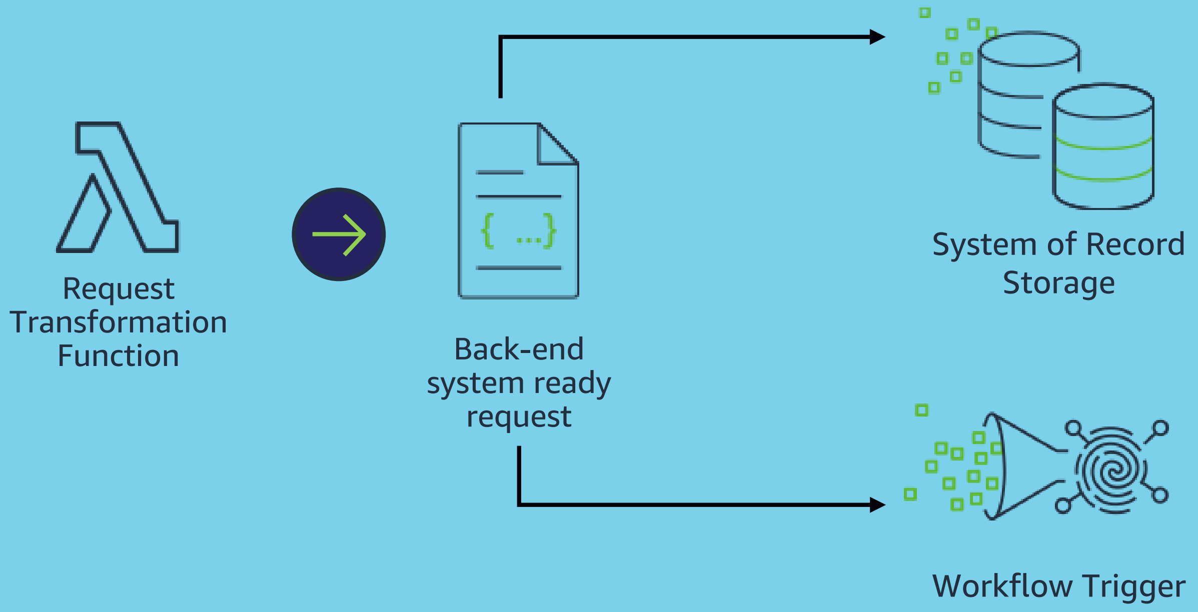
3 Reviews consolidated using Amazon A2I answer consolidation algorithms

4 Results stored to Amazon S3

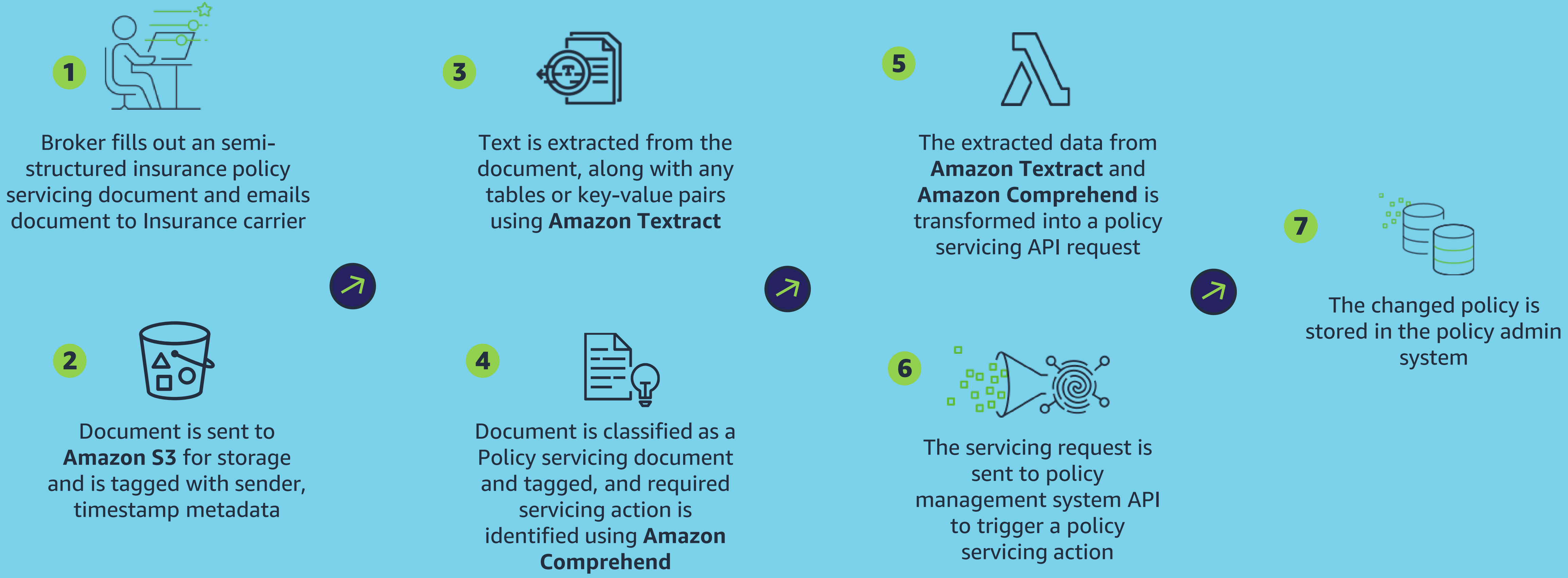
Request Transformation Function

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


Request Transformation Function
{ ...}
Back-end system ready request
System of Record Storage
Workflow Trigger

An end to end IDP example visualized – Insurance Policy Servicing



1 Broker fills out an semi-structured insurance policy servicing document and emails document to Insurance carrier

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

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

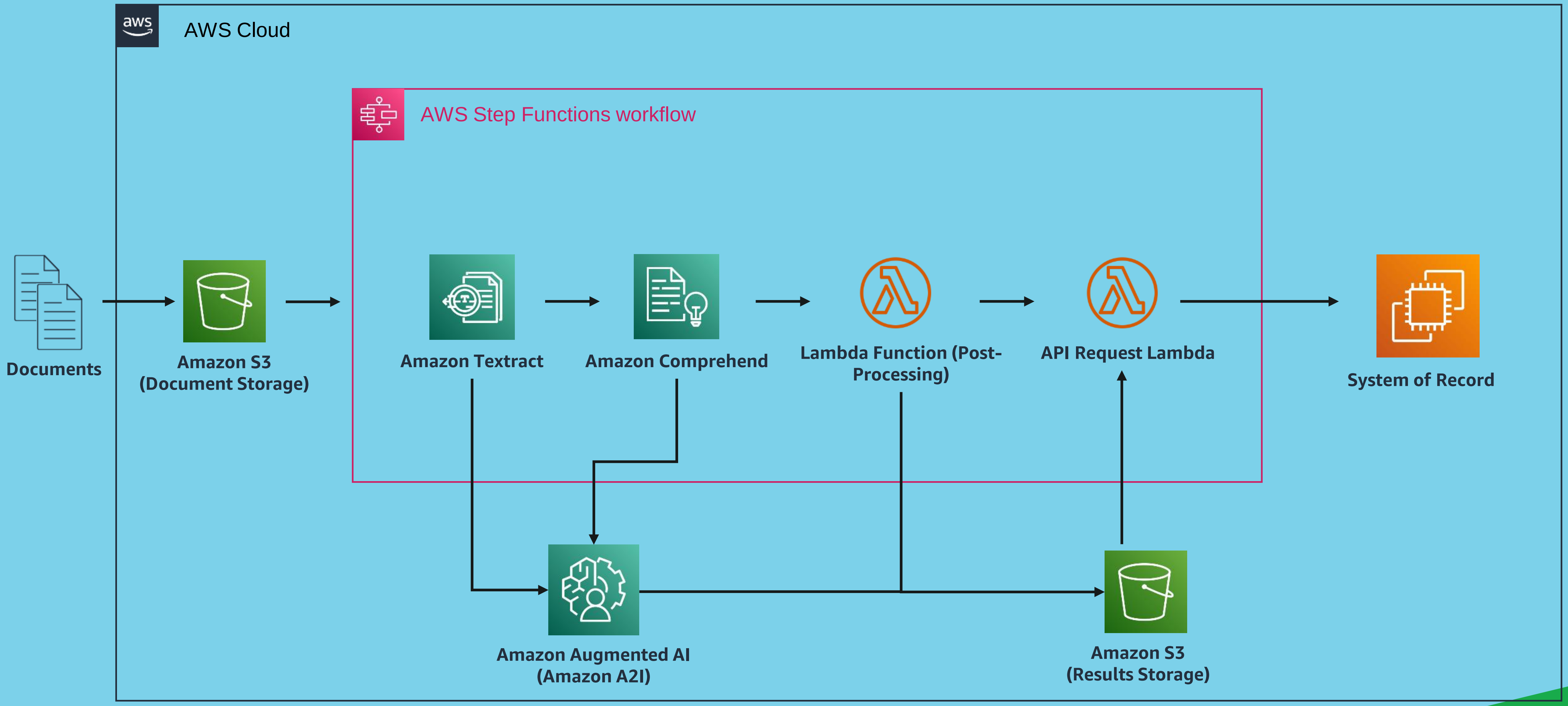
4 Document is classified as a Policy servicing document and tagged, and required servicing action is identified using **Amazon Comprehend**

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

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

7 The changed policy is stored in the policy admin system

AWS Reference Architecture IDP



Thank you!