AWS re:Invent

SVS304-R

Build a serverless engine to process large-scale documents

Leo Drakopoulos
Solutions Architect
Amazon Web Services

AWS re:Invent

aws

Documents are important

Primary tool of recordkeeping, communicating, collaborating, and transactions


Finance


Insurance


Real Estate


Accounting


Tax Management


Medical


Legal

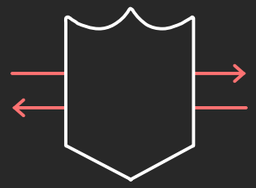

Business Management


Education

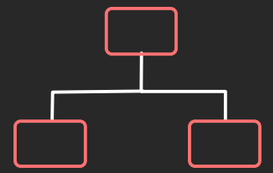

And many more…

The need for processing documents

Search
and discovery



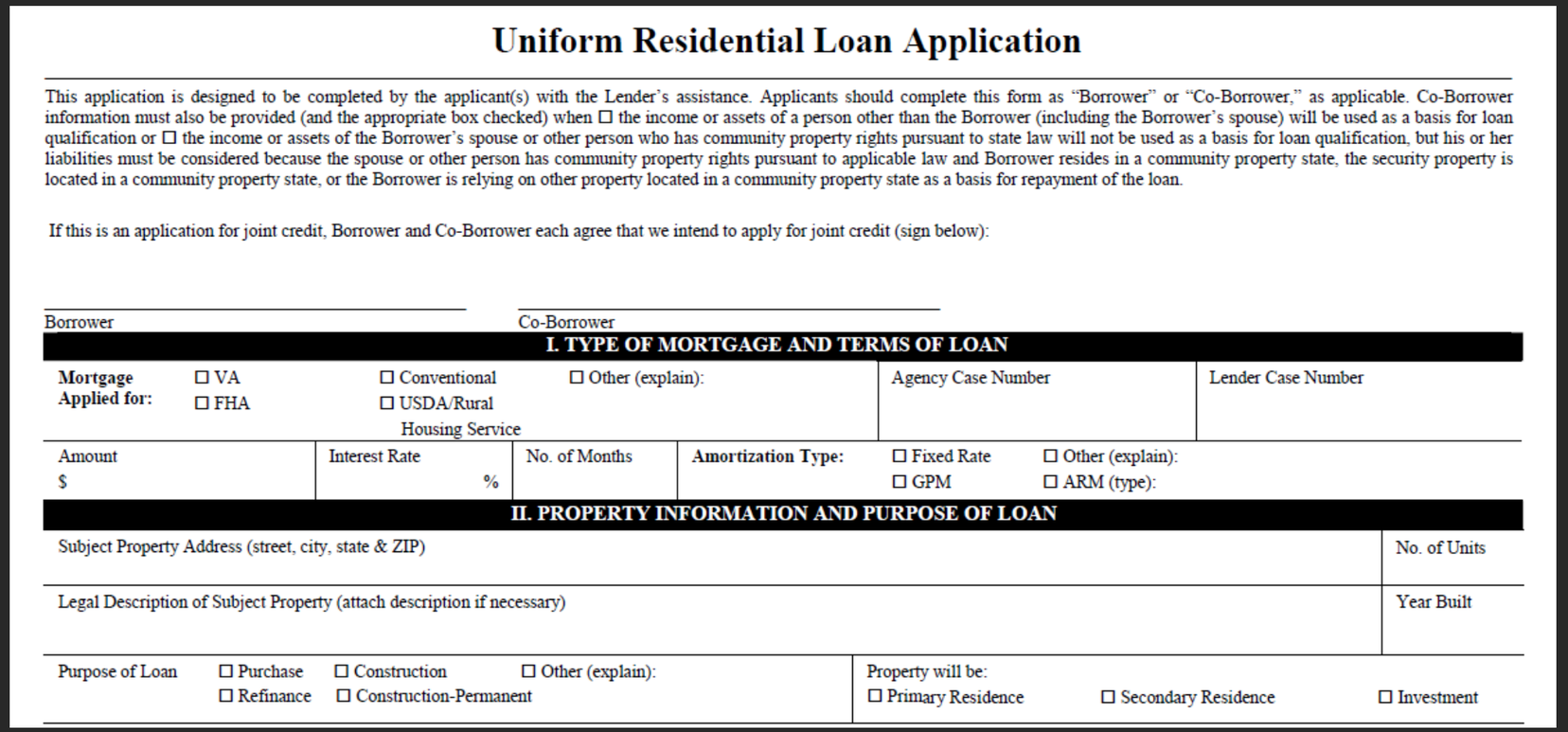
Compliance
and control



Business
process automation

Challenges with processing documents

16.3 million US mortgage applications ($2.1 trillion) in 2016



Uniform Residential Loan Application

This application is designed to be completed by the applicant(s) with the Lender's assistance. Applicants should complete this form as "Borrower" or "Co-Borrower," as applicable. Co-Borrower information must also be provided (and the appropriate box checked) when ☐ the income or assets of a person other than the Borrower (including the Borrower's spouse) will be used as a basis for loan qualification or ☐ the income or assets of the Borrower's spouse or other person who has community property rights pursuant to state law will not be used as a basis for loan qualification, but his or her liabilities must be considered because the spouse or other person has community property rights pursuant to applicable law and Borrower resides in a community property state, the security property is located in a community property state, or the Borrower is relying on other property located in a community property state as a basis for repayment of the loan.

If this is an application for joint credit, Borrower and Co-Borrower each agree that we intend to apply for joint credit (sign below):

Borrower _____ Co-Borrower _____

I. TYPE OF MORTGAGE AND TERMS OF LOAN

Mortgage Applied for:	☐ VA ☐ FHA	☐ Conventional ☐ USDA/Rural Housing Service	☐ Other (explain):	Agency Case Number	Lender Case Number

Amount $	Interest Rate %	No. of Months	**Amortization Type:**	☐ Fixed Rate ☐ GPM	☐ Other (explain): ☐ ARM (type):

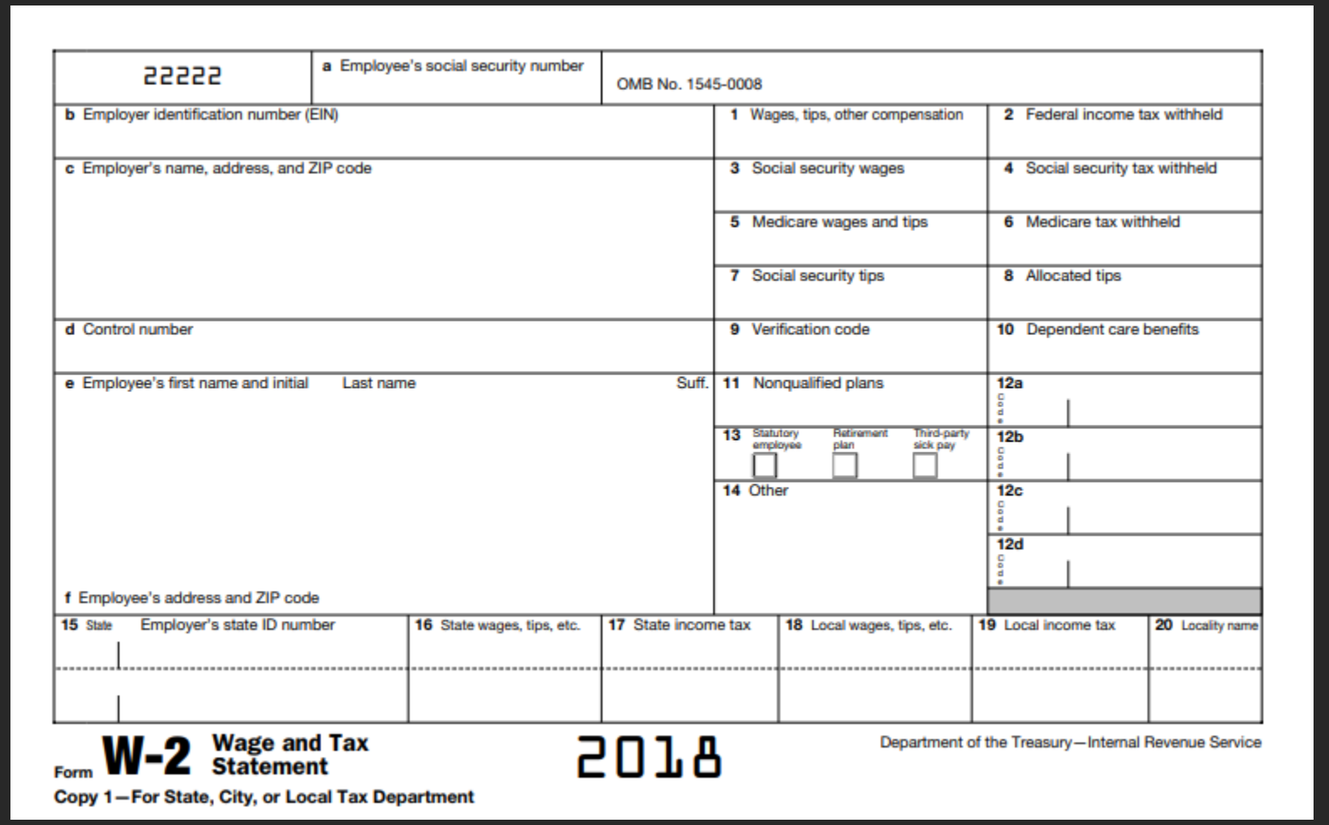
II. PROPERTY INFORMATION AND PURPOSE OF LOAN

Subject Property Address (street, city, state & ZIP)	No. of Units
Legal Description of Subject Property (attach description if necessary)	Year Built

Purpose of Loan	☐ Purchase ☐ Refinance	☐ Construction ☐ Construction-Permanent	☐ Other (explain):	Property will be: ☐ Primary Residence	☐ Secondary Residence	☐ Investment

*Mortgage Bankers Association 2016 HMDA

About 240 million W-2 tax forms will be processed for FY 2018 in the US



22222	a Employee's social security number	OMB No. 1545-0008		
b Employer identification number (EIN)			1 Wages, tips, other compensation	2 Federal income tax withheld
c Employer's name, address, and ZIP code			3 Social security wages	4 Social security tax withheld
			5 Medicare wages and tips	6 Medicare tax withheld
			7 Social security tips	8 Allocated tips
d Control number			9 Verification code	10 Dependent care benefits
e Employee's first name and initial Last name Suff.			11 Nonqualified plans	12a
			13 Statutory employee / Retirement plan / Third-party sick pay	12b
			14 Other	12c
				12d
f Employee's address and ZIP code				

15 State Employer's state ID number	16 State wages, tips, etc.	17 State income tax	18 Local wages, tips, etc.	19 Local income tax	20 Locality name

Form **W-2** Wage and Tax Statement 2018 Department of the Treasury—Internal Revenue Service

Copy 1—For State, City, or Local Tax Department

*IRS—https://www.irs.gov/individuals/w-2-verification-code

How documents are processed today

Manual processing

Optical Character Recognition (OCR)

Rules and template-based extraction

Challenges for processing documents

Manual processing

Expensive

Prone to errors

Time-consuming

Challenges for processing documents
Optical Character Recognition

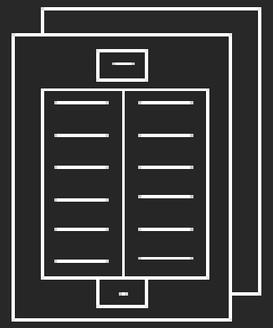
Simple documents only

Prone to errors

Flat bag of words

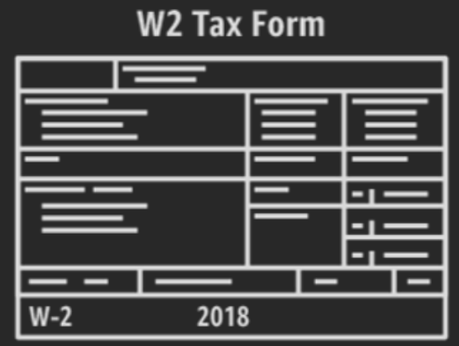
What problem does Amazon Textract solve?

Amazon Textract features



Text extraction

Table extraction

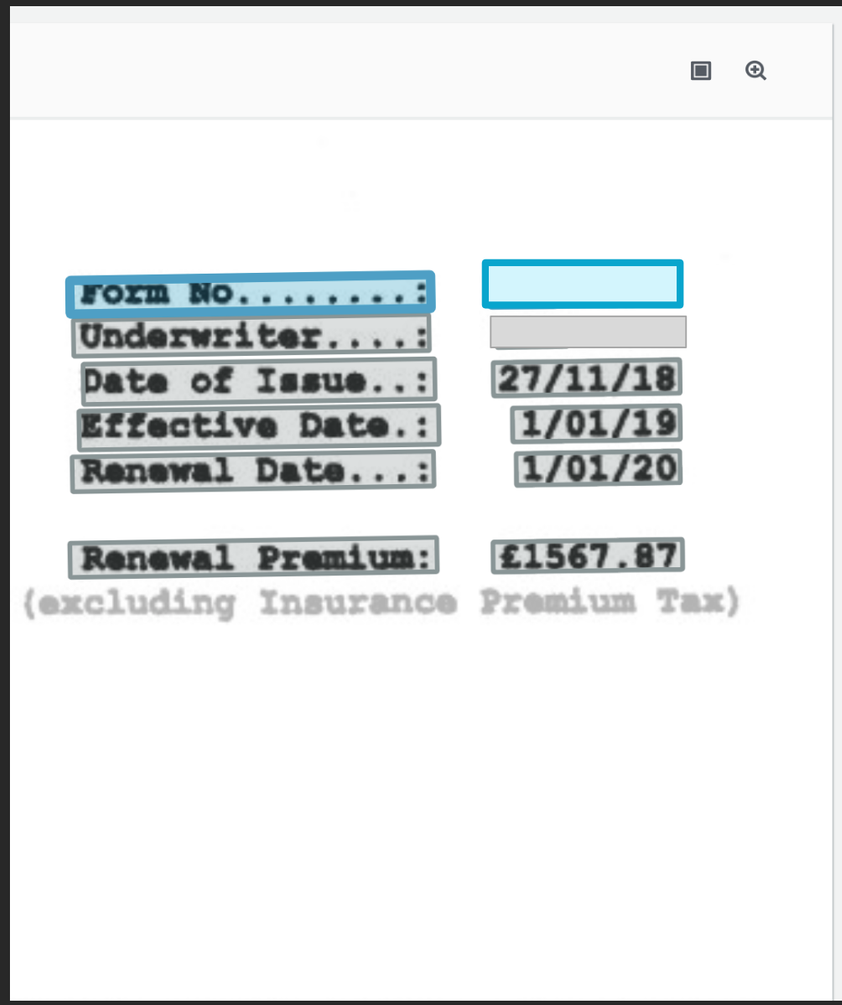


Form extraction

Amazon Textract in the AWS Management Console

Amazon Textract

Amazon Textract



Raw text | Forms | Tables

Search

Agency:

SCHEDULE: Commercial Combined

Insurance Premium Tax:

Underwriter

Effective Date.:
1/01/19

Form No

Renewal Date.
1/01/20

Date of Issue. :
27/11/18

Renewal Premium:
11567.87

Amazon Textract demo

What are the different Amazon Textract APIs

Amazon Textract—text extraction API

DetectDocumentText

Request

Name	Description
Document	Blob or Amazon S3 object

Response

Name	Description
Blocks	List of blocks identified from the document
ID	Unique ID of the unit
Relationships	CHILD
Block type	PAGE, LINE, WORD
Pages	Contains number of pages in the document

Amazon Textract—forms extraction API

AnalyzeDocument with “forms” as FeatureTypes parameter

Request

Name	Description
Document	Blob or Amazon S3 object
FeatureTypes	FORMS

Response

Name	Description
Blocks	List of blocks identified from the document
ID	Unique ID of the unit
Relationships	KEY, VALUE, CHILD
Block type	PAGE, KEY_VALUE_SET
Pages	Contains number of pages in the document

Amazon Textract—table extraction API

AnalyzeDocument with “table” as FeatureTypes parameter

Request

Name	Description
Document	Blob or Amazon S3 object
FeatureTypes	TABLES

Response

Name	Description
Blocks	List of blocks identified from the document
ID	Unique ID of the unit
Relationships	CHILD
Block type	PAGE, TABLE, CELL
Pages	Contains number of pages in the document

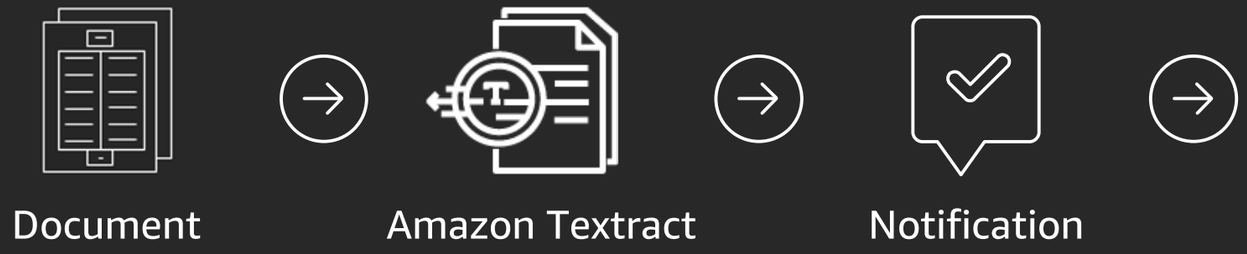
Amazon Textract

Sync and async

Synchronous

Supports single-page documents such as images (e.g., mobile capture)

Asynchronous



For multi-page documents up to 3,000 pages

Parsing JSON response

```
{
   "Blocks": [
      {
         "BlockType": "string",
         "ColumnIndex": number,
         "ColumnSpan": number,
         "Confidence": number,
         "EntityTypes": [ "string" ],
         "Geometry": {
            "BoundingBox": {
               "Height": number,
               "Left": number,
               "Top": number,
               "Width": number
            },
            "Polygon": [
               {
                  "X": number,
                  "Y": number
               }
            ]
         },
         "Id": "string",
         "Page": number,
         "Relationships": [
            {
               "Ids": [ "string" ],
               "Type": "string"
            }
         ],
         "RowIndex": number,
         "RowSpan": number,
         "Text": "string"
      }
   ],
   "DocumentMetadata": {
      "Pages": number
   },
   "JobStatus": "string",
   "NextToken": "string",
   "StatusMessage": "string",
   "Warnings": [
      {
         "ErrorCode": "string",
         "Pages": [ number ]
      }
   ]
}
```

JSON response parser library

```
# Print detected text
for item in response["Blocks"]:
    if item["BlockType"] == "LINE":
        print (item["Text"])
```

=>

```
# Print detected text
for page in doc.pages:
    for line in page.lines:
        print(line.text)
```

```
# Print fields
for field in page.form.fields:
    print("Field: Key: {}, Value: {}".format(field.key.text, field.value.text))

# Get field by key
key = "Phone Number:"
field = page.form.getFieldByKey(key)
if(field):
    print("Field: Key: {}, Value: {}".format(field.key, field.value))

# Search fields by key
key = "address"
fields = page.form.searchFieldsByKey(key)
for field in fields:
    print("Field: Key: {}, Value: {}".format(field.key, field.value))
```

https://github.com/aws-samples/amazon-textract-response-parser

How can I process documents at large scale with Amazon Textract?

Throttling & TPS limits

Amazon Textract Limits

Amazon Textract has the following limits that you can change.

Resource	Default Limit
Transactions per second per account for synchronous operations: • AnalyzeDocument • DetectDocumentText	In each Region that Amazon Textract supports – 0.25
Transactions per second per account for all *Start* (asynchronous) operations: • StartDocumentAnalysis • StartDocumentTextDetection	In each Region that Amazon Textract supports – 0.25
Transactions per second per account for all *Get* (asynchronous) operations: • GetDocumentAnalysis • GetDocumentTextDetection	In each Region that Amazon Textract supports – 2
Maximum number of asynchronous jobs per account that can simultaneously exist	In each Region that Amazon Textract supports – 2

Throttling & TPS

0.25 TPS => 21,600 documents/day

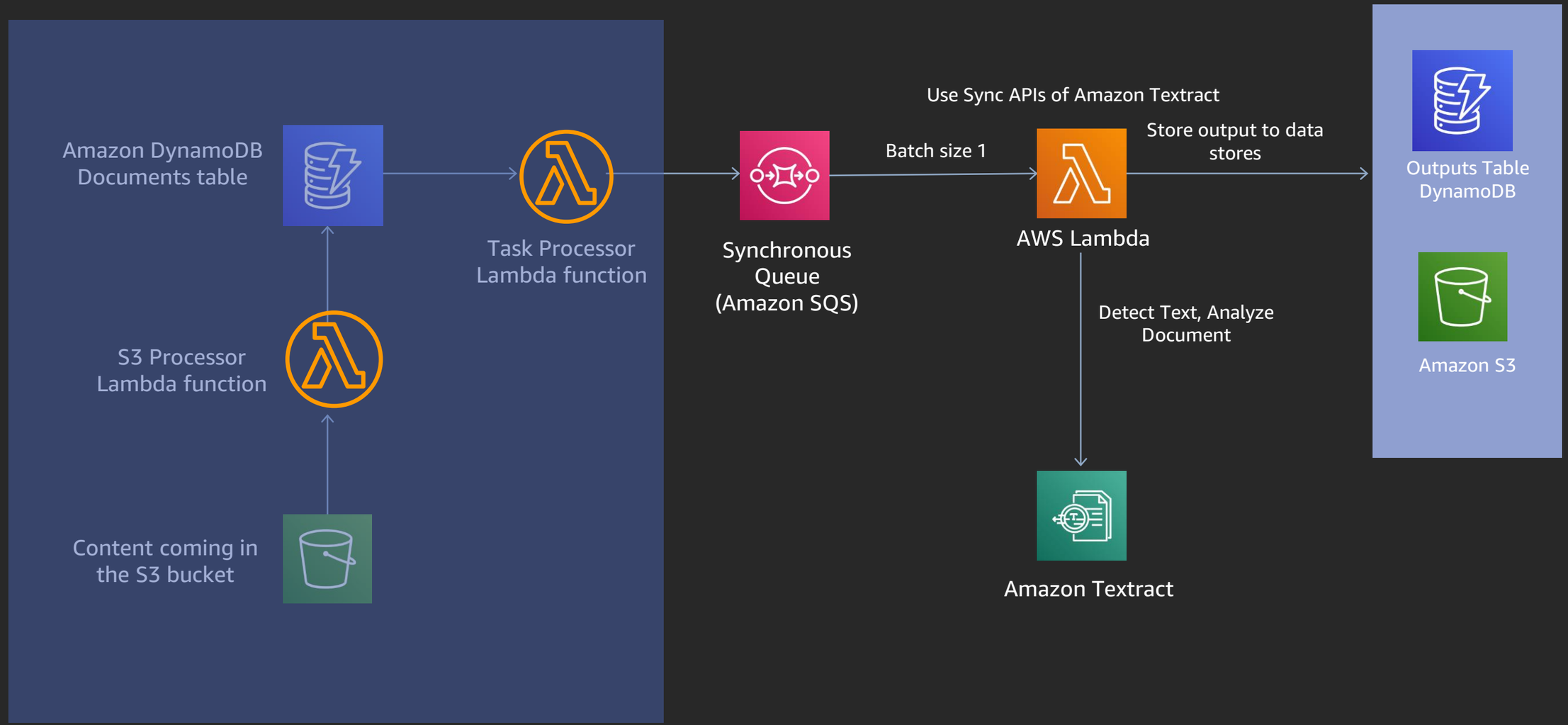
- How many documents do you need to process, and how quickly do you need to get them done

- Are you using sync APIs or async APIs

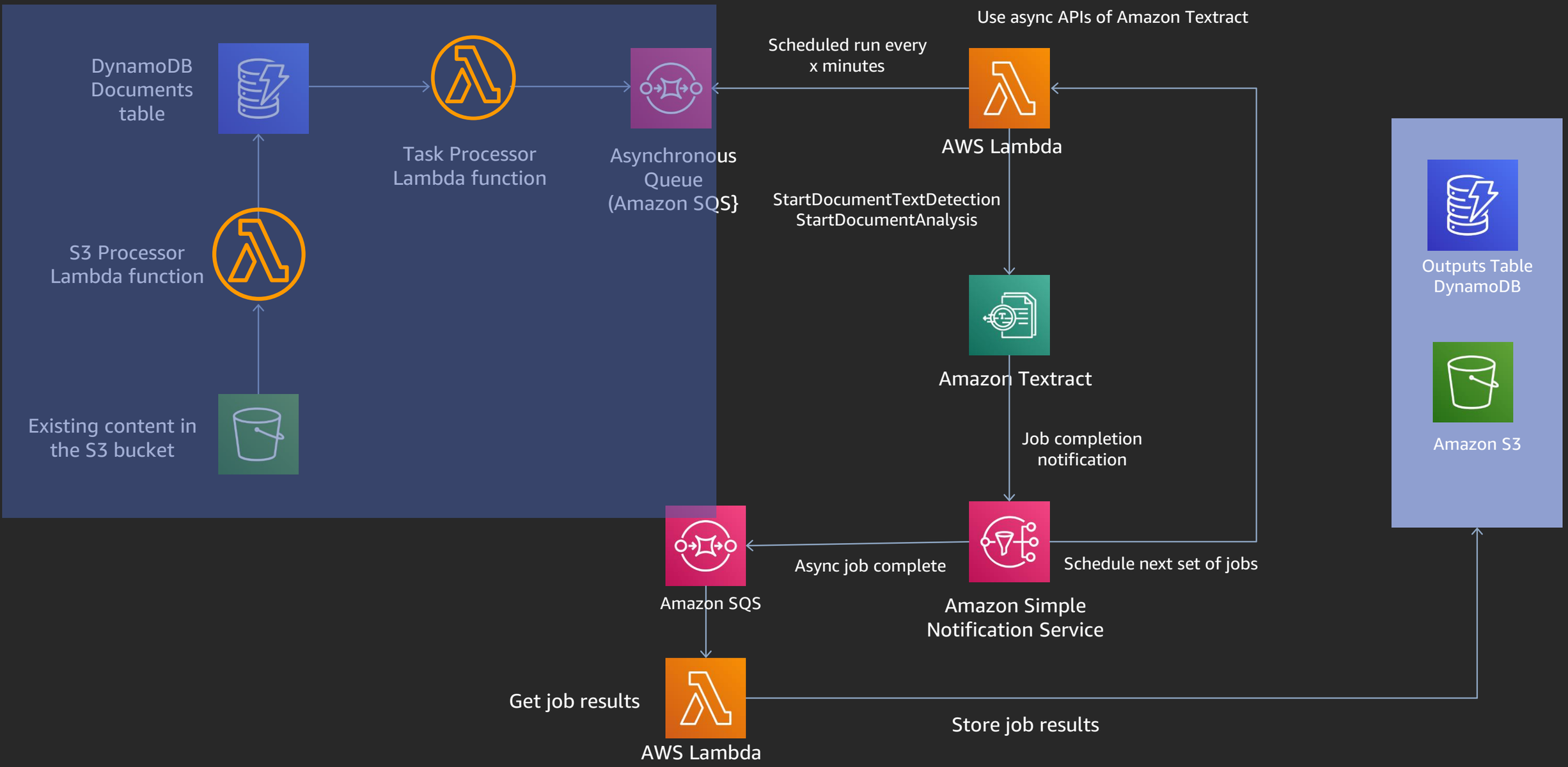
- Request limit increase for TPS and concurrent job limits based on your desired throughput

Workshop architecture

Module 1: Synchronous-processing architecture



Module 2: Asynchronous-processing architecture



Workshop demo

AWS re:Invent

aws

Workshop link

http://bit.ly/serverless-document-processing

Resources

Blogpost
https://aws.amazon.com/blogs/machine-learning/automatically-extract-text-and-structured-data-from-documents-with-amazon-textract/

Webinar
https://www.youtube.com/watch?v=aBaoS4c4-Yo

Large-scale document processing, reference architecture and sample implementation
https://github.com/aws-samples/amazon-textract-serverless-large-scale-document-processing

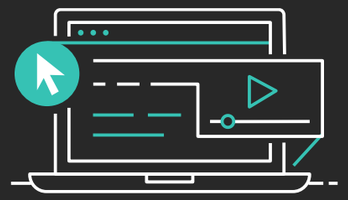
Code samples
https://github.com/aws-samples/amazon-textract-code-samples

JSON response parser
https://github.com/aws-samples/amazon-textract-response-parser

Batch processing tool
https://github.com/aws-samples/amazon-textract-textractor

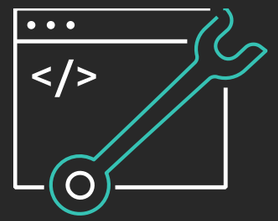
Learn serverless with AWS Training and Certification

Resources created by the experts at AWS to help you learn modern application development



Free, on-demand courses on serverless, including

- Introduction to Serverless Development
- Getting into the Serverless Mindset
- AWS Lambda Foundations
- Amazon API Gateway for Serverless Applications
- Amazon DynamoDB for Serverless Architectures



Additional digital and classroom trainings cover modern application development and computing

Visit the Learning Library at https://aws.training

aws training and certification

Thank you!

AWS re:Invent

aws

Please complete the session survey in the mobile app.