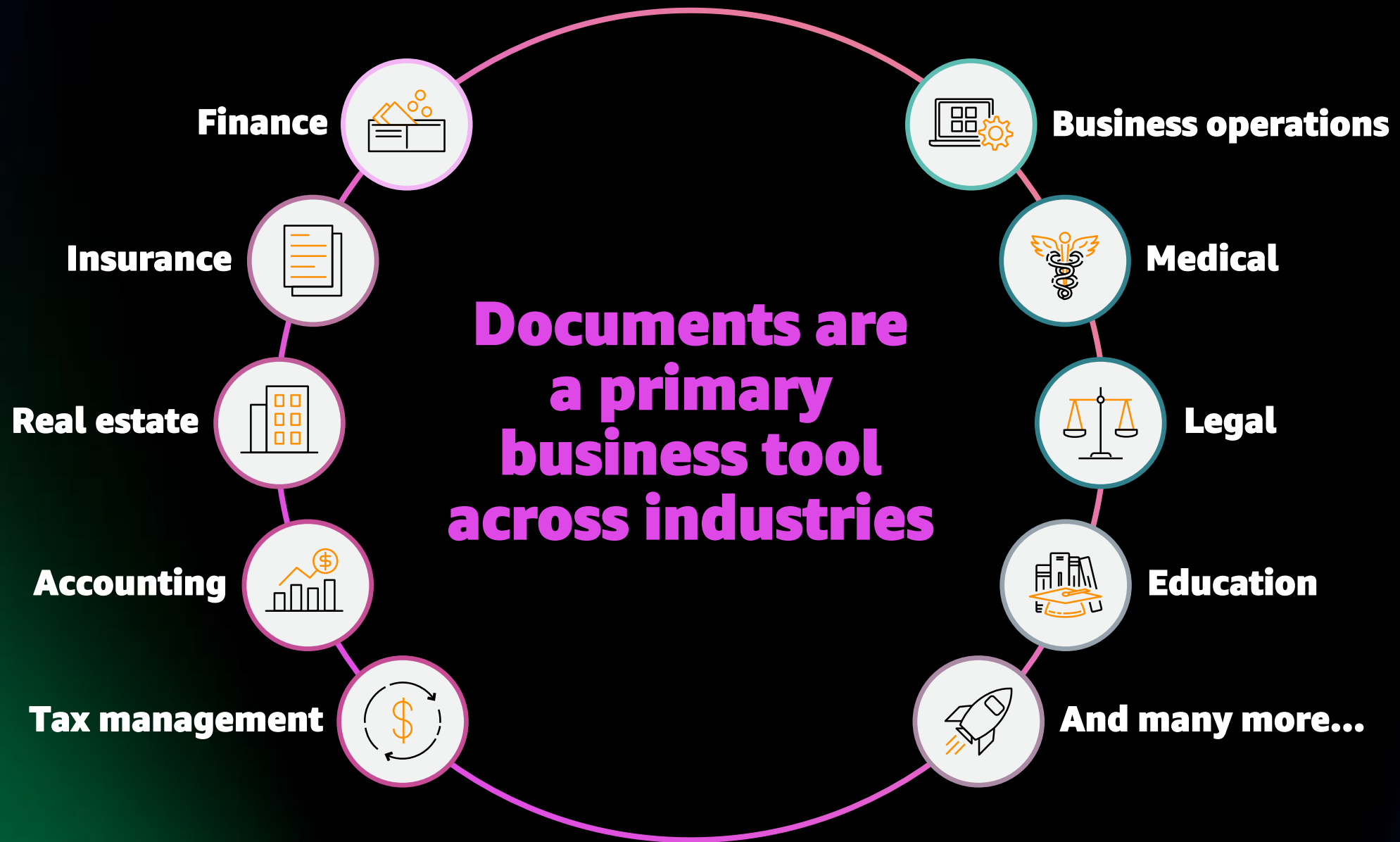# aws INNOVATE

## MODERN APPLICATIONS EDITION

20 October, 2022

aws

# Intelligent unstructured document processing

Santhosh Urukonda

Prototyping Architect, AISPL

aws

Finance

Insurance

Real estate

Accounting

Tax management

**Documents are a primary business tool across industries**

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
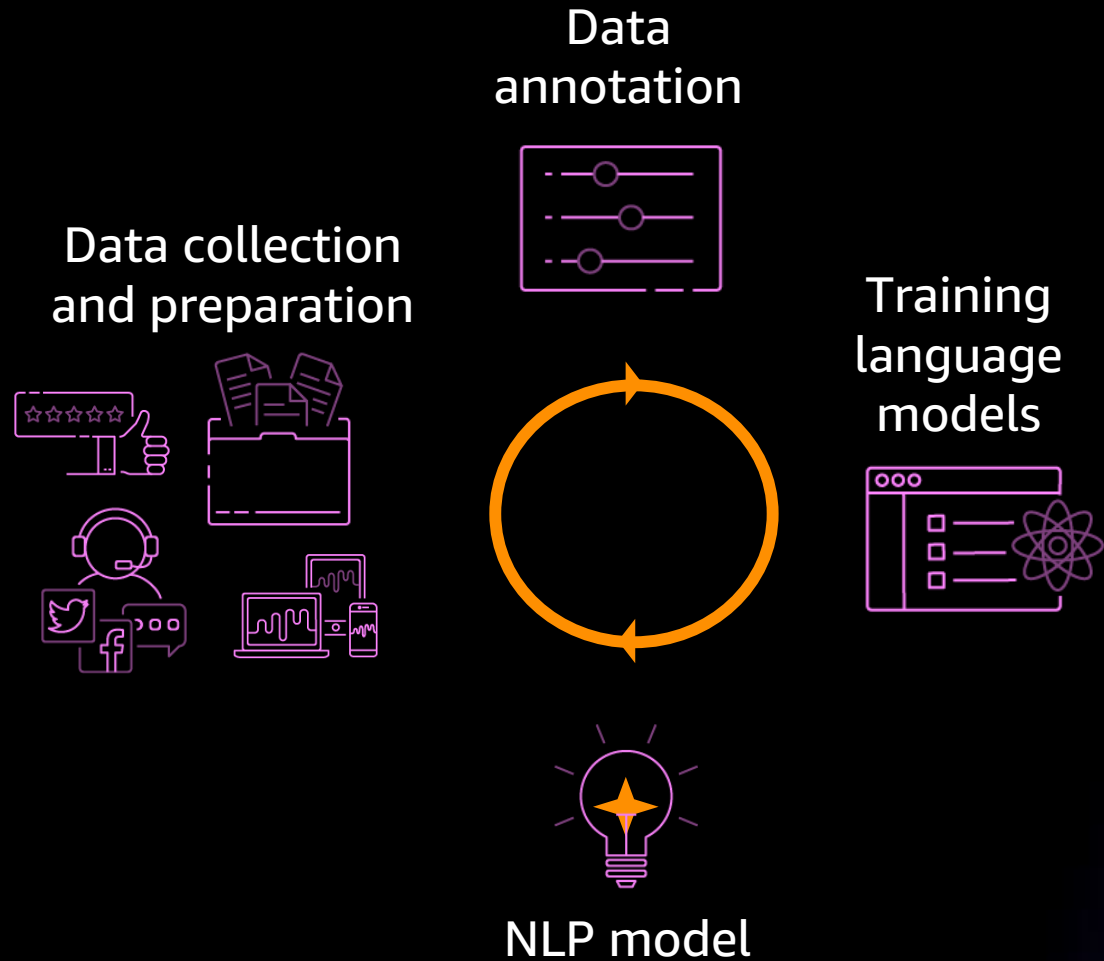- ❌ Error prone
- ❌ Dump of text

## Rules and template-based extraction

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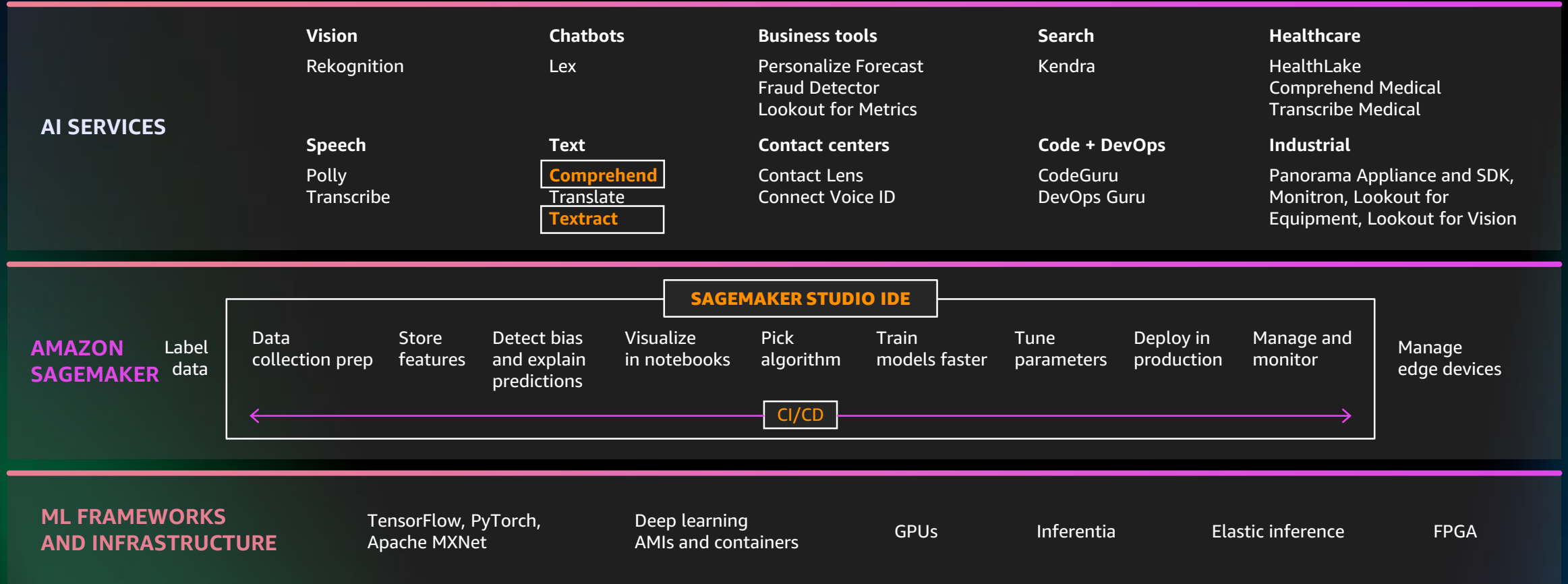- ❌ Templates are brittle

# Customer challenges with building NLP models

- Machine learning skills

- Labeled training data

- Model training effort
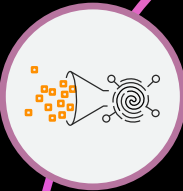
- Model management in production

Data annotation

Data collection and preparation

Training language models

NLP model

# 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

**SAGEMAKER STUDIO IDE**

Label data

| 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

aws

# Amazon Textract

**Go beyond OCR with accurate, versatile information extraction**

**No ML experience required**
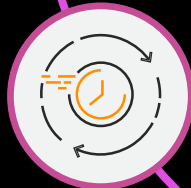
## Amazon Textract

Automatically extract printed text, handwriting, and data from any document

**Reduce manual effort**

**Lower costs**

# Amazon Comprehend



**Documents**

Email, chat, social, phone calls and more

→

**Amazon Comprehend**

Automatically extract insights from text

→

**Entities**
+ Custom Entities

**Key Phrases**

**PII**
(Personally Identifiable Information)

**Sentiment**

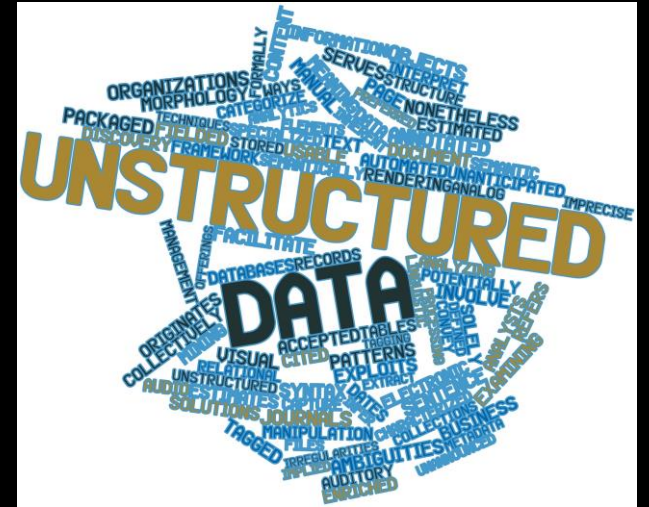**Document Classification**

**Topics**

**Language**

**Syntax**

**Events**

# Use case

- Process unstructured documents like product reviews
- Data is not in fixed/organized format
- Multiple file types (doc, pdf, image etc.)
- Technical jargon specific to product
- Product feature comparison
- Overall sentiment of the product

# AWS services used

**1** Extract text from documents

Amazon Textract

**2** Entity detection & sentiment

Amazon Comprehend
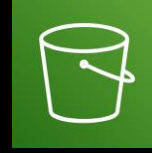
**3** Invoke other AWS services & run custom rules

AWS Lambda
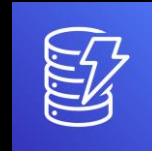
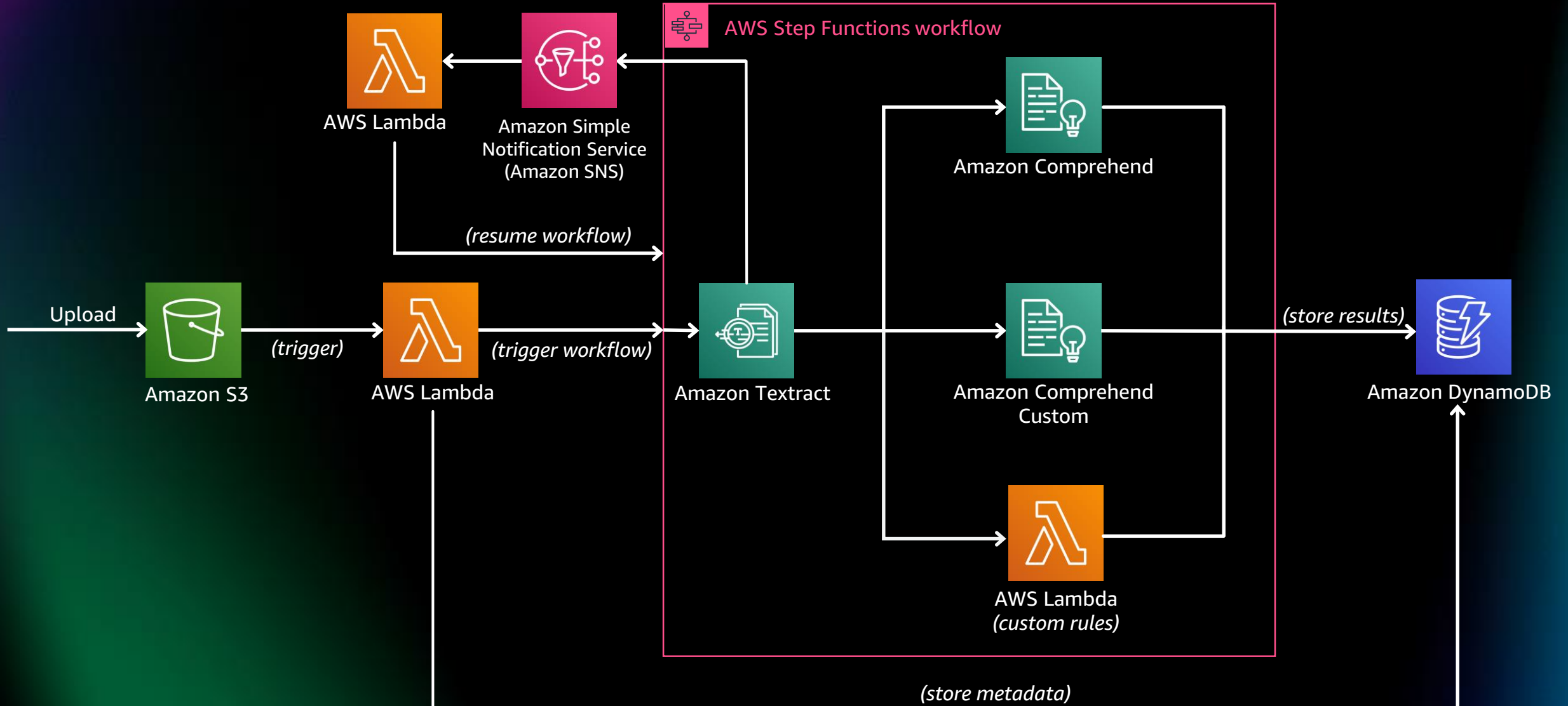**4** Orchestrate workflow

AWS Step Functions

**5** Document storage

Amazon S3

**6** Store results & job tracking

Amazon DynamoDB

# Architecture



AWS Step Functions workflow

Upload → Amazon S3 —(trigger)→ AWS Lambda —(trigger workflow)→ Amazon Textract → Amazon Comprehend / Amazon Comprehend Custom / AWS Lambda (custom rules) —(store results)→ Amazon DynamoDB

(resume workflow)

AWS Lambda ← Amazon Simple Notification Service (Amazon SNS) ← 

(store metadata)

# Demo

# Recap

- Unstructured document processing challenges

- AWS AI services (Amazon Textract & Amazon Comprehend)

- Use case

- System architecture

- Solution walkthrough

# Additional resources

Amazon Textract Blogs



https://go.aws/3259Sm8

Amazon Comprehend Blogs



https://go.aws/3qxDVwm

# Visit the Modern Applications resource hub

Dive deeper with these resources to help you develop an effective plan for your modernization journey.

- Build modern applications on AWS

- Business value of cloud modernization

- An introduction to event-driven architectures

- Accelerate full-stack web and mobile app development

- Determining the total cost of ownership: Comparing serverless and server-based technologies

- Building event-driven architectures with AWS

- Continuous learning, continuous modernization



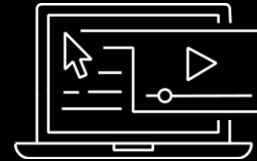https://tinyurl.com/modern-apps-aws

Visit resource hub

# AWS Training and Certification

**Get started with Free Digital Training for you and your team today**



Achieve key milestones and plan your next steps with the AWS Modern Application skills training

Access 500+ free digital courses with AWS Skill Builder

Earn an industry-recognized credential:
AWS Certified Developer – Associate
AWS Certified DevOps – Professional

Create a self-paced learning roadmap
AWS ramp-up guide - Developer
AWS ramp-up guide - DevOps

aws

# Thank you for attending AWS Innovate
# Modern Applications Edition

We hope you found it interesting! A kind reminder to **complete the survey.**
Let us know what you thought of today's event and how we can improve the event experience for you in the future.

aws-apj-marketing@amazon.com

twitter.com/AWSCloud

facebook.com/AmazonWebServices

youtube.com/user/AmazonWebServices

slideshare.net/AmazonWebServices

twitch.tv/aws

# Thank you!