
AWS
re:Invent
NOV. 27 – DEC. 1, 2023 | LAS VEGAS, NV

AIM322-R

Intelligent document processing with generative AI for public sector

Sonali Sahu
Sr. Manager Specialist SA
AWS

Cameron Williams
Release Train Engineer
Booz | Allen | Hamilton

Agenda

Introduction
- Brief overview on intelligent document processing (IDP)

How to build large-scale document processing
- How Booz Allen built their document processing pipeline
- Large-scale document processing architecture

How generative AI can assist IDP
- Build document processing pipeline with generative AI

Wrap-up and next steps

Intelligent document processing (IDP)

STATE OF THE UNION

Accurate

Industry-leading accuracy among major on-premises and cloud vendors

Trustworthy

Elevate your security, compliance, and data privacy posture

Production ready

SLA-backed availability with high elasticity and scalability

Pay as you go

Pay for only what you use

Automate with confidence

Paytm extracts user data from documents with 97% accuracy

Improve business process efficiency

Elevance Health automated **90%** of claims processing workflow

Lower costs

Customers achieved **73%** ROI with AWS IDP (Forrester Study)

Core IDP capabilities

IDP CAPABILITIES WITH AMAZON TEXTRACT

OCR

Layout

Forms

Entities

Handwriting and signatures

Tables

Queries/ custom queries

Domain-specific APIs (e.g., lending)

Intelligent document processing demonstration

Need for an AI powered-IDP solution at VA

Use case

- 2022 PACT Act expanded benefits for over 5 million veterans
- The average benefits application packet contains 100+ documents averaging 13–15 pages each that can take ~30 minutes to review per document
- Expedite disability benefits determination by reducing the manual review of documents
- Extract raw text from the benefits application documents and enable a search engine that helps to quickly find and review the documents

Why Amazon IDP

- Available on GovCloud and is FedRAMP High
- Ability to scale as required to process 50–80 million pages per day
- High accuracy with both print and handwritten texts
- Reduce time spent reviewing data by highlighting results using Amazon Textract geometry data

Results/impact

- 400 million documents
- 10 billion+ pages processed
- Largest implementation of Amazon Textract in under one year

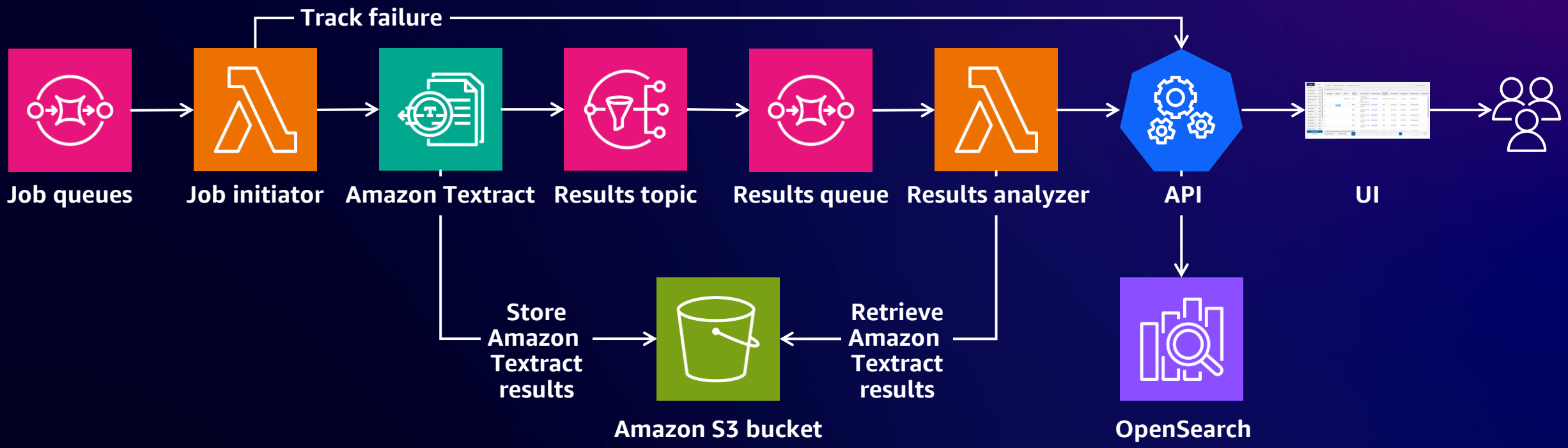
- Estimated 9x efficiency boost to document processing and manual review time

- Smart search engine to search and filter the content, minimizing number of documents to review

- Enabling future innovation through ML applications including generative AI by leveraging the extracted text

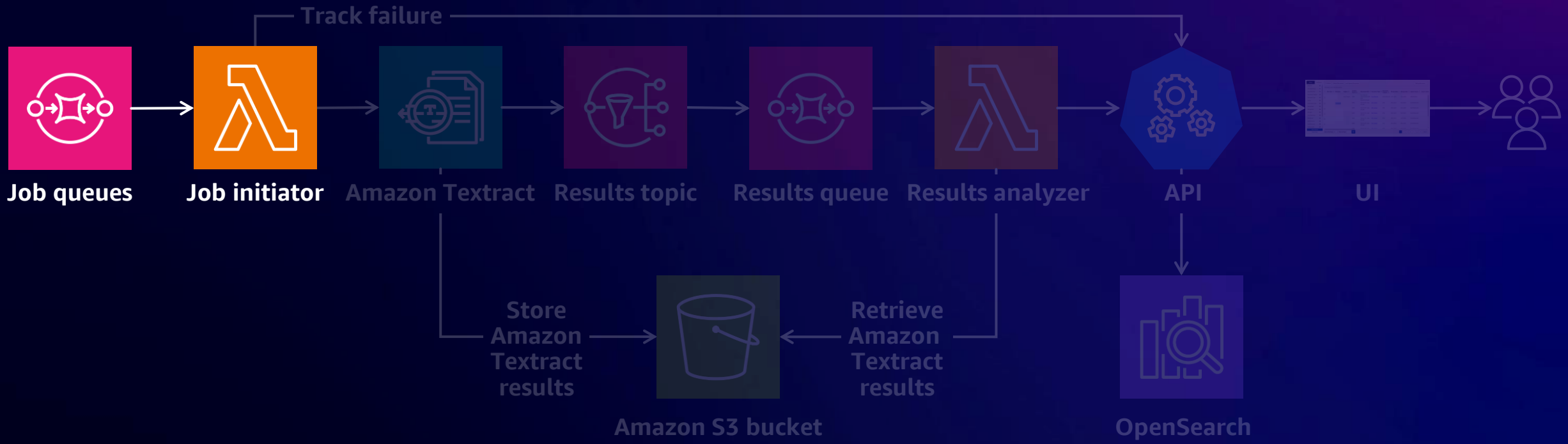
Questions

Our IDP workflow



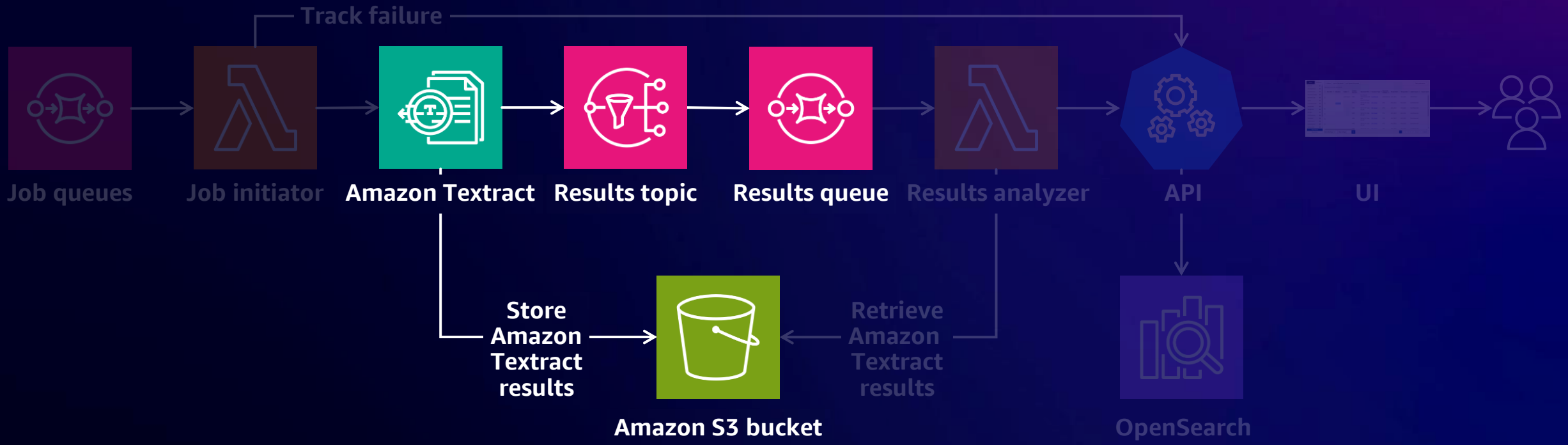
- System is FISMA High
- In GovCloud, all FedRAMP High
- Translation: services are secure and reliable

Our IDP workflow: Job queues and initiator



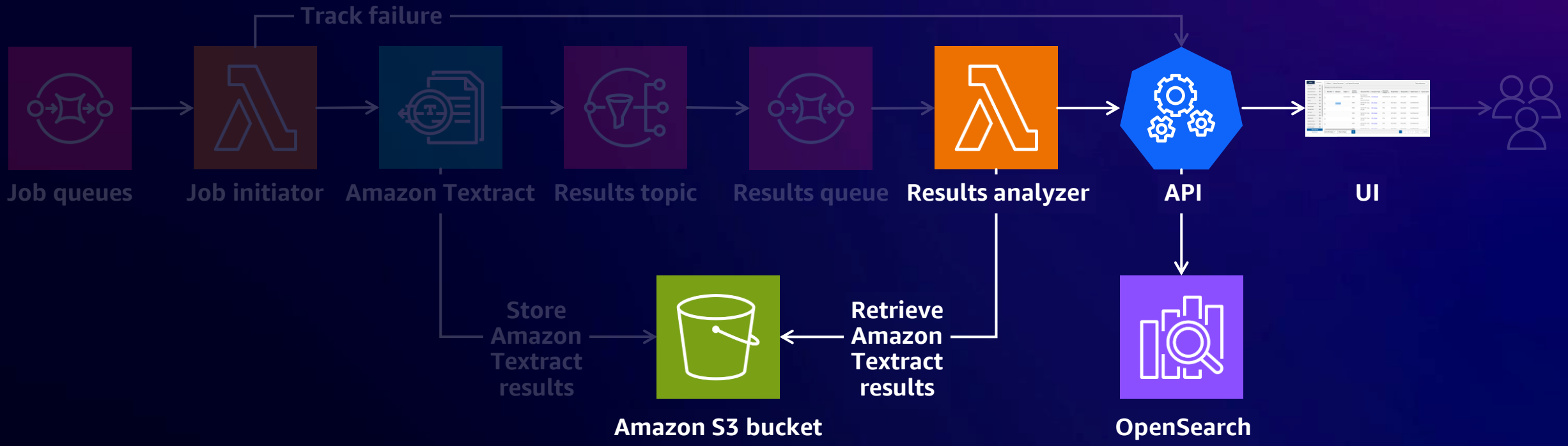
- Job queues invoke job initiator via event source mapping
- Amazon SQS and AWS Lambda are scalable, durable, and easy to integrate
- Consideration: balance message creation with rate limits

Our IDP workflow: Amazon Textract



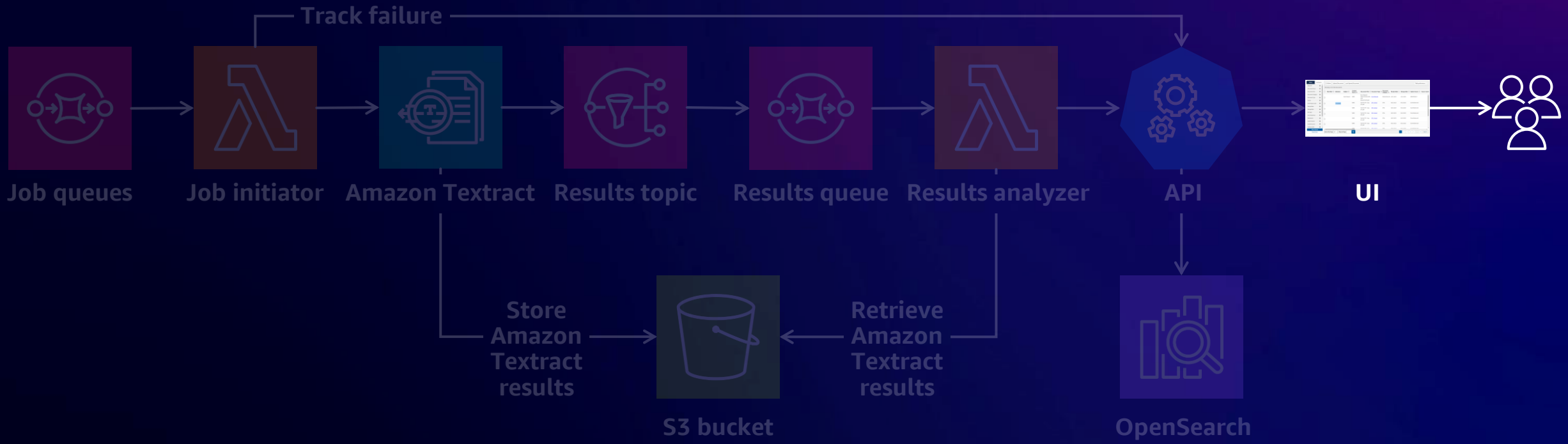
- Amazon Textract reads from and writes to S3 bucket using Lambda IAM role
- Synchronous (single-page) vs. asynchronous (multipage) operations
  - Synchronous operations return the result on the response, simplifying architecture
  - Asynchronous sends status to SNS (results topic)

Our IDP workflow: Results analyzer



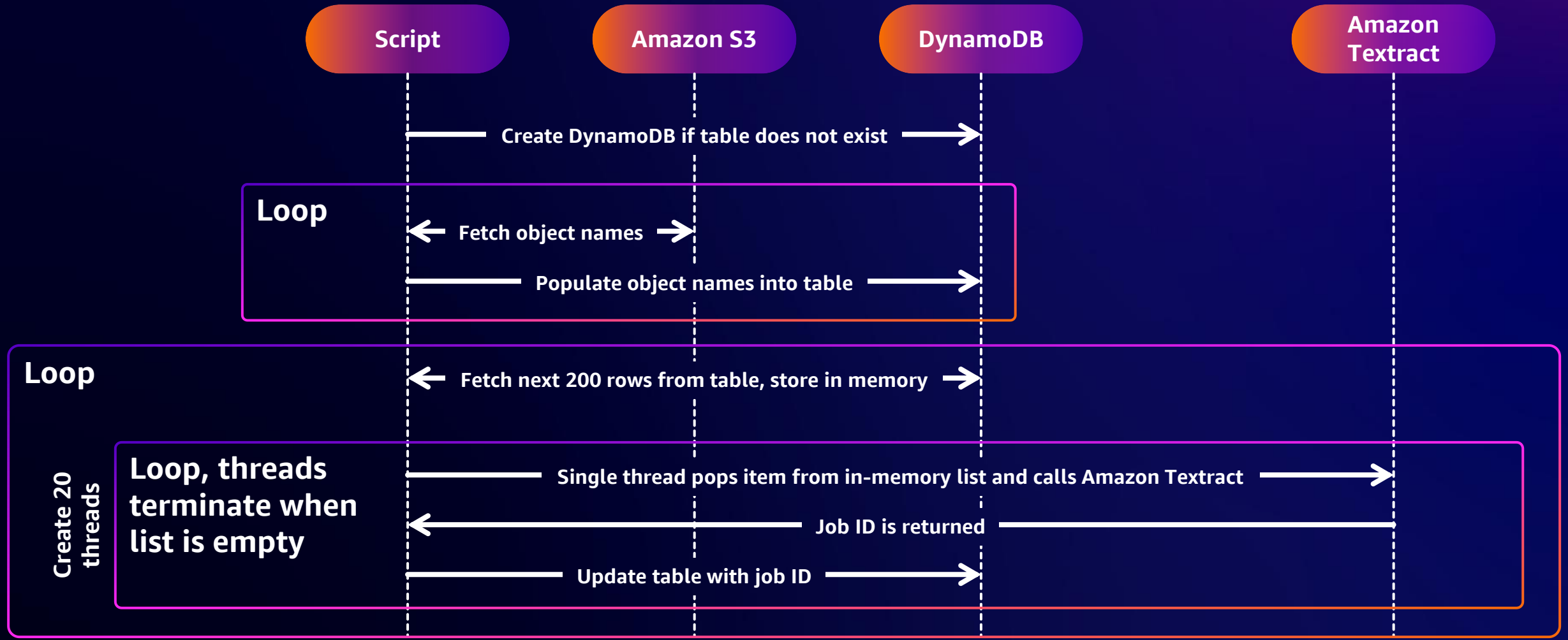
- Results analyzer retrieves all pages of blocks from S3
- Maps results to custom internal schema and persists to RESTful API
- Store data in OpenSearch and S3

Our IDP workflow: User interaction

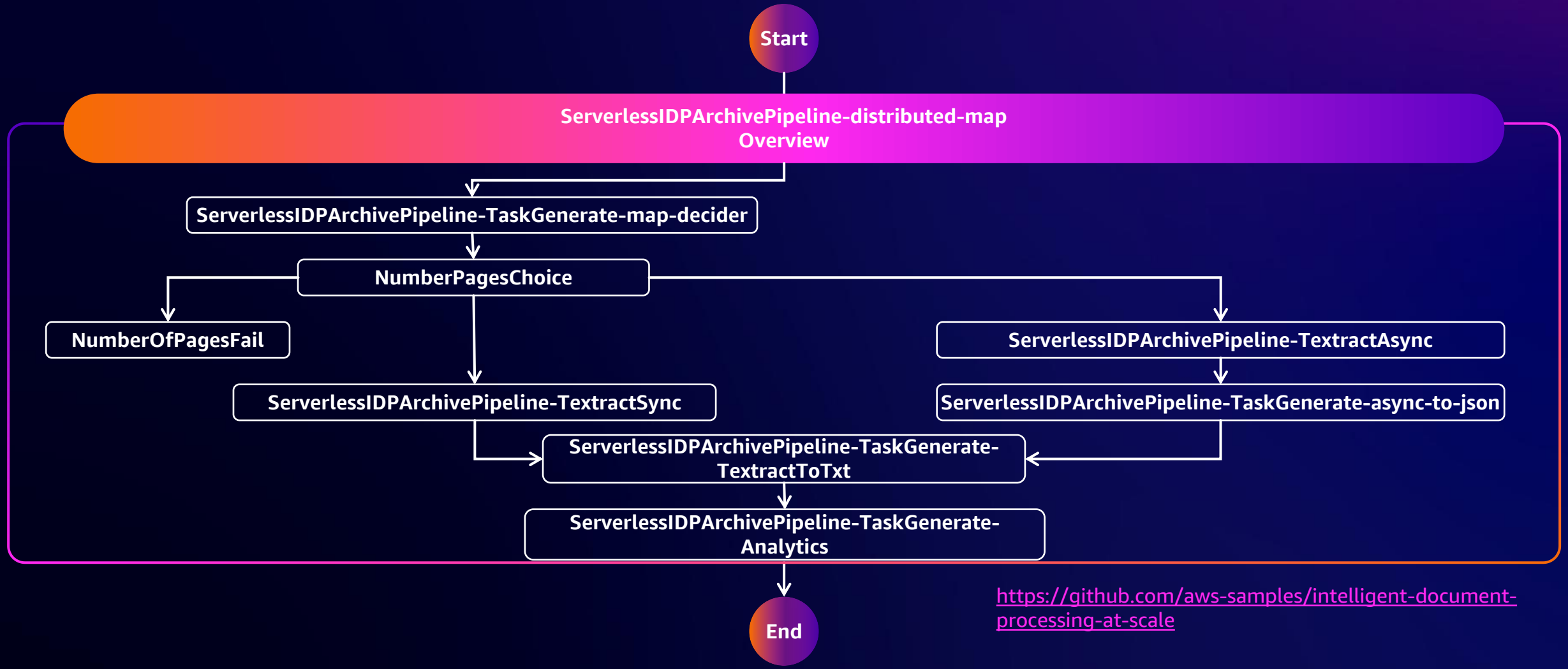


Demonstration

Large-scale design flow



Large-scale serverless deployment – IDP



https://github.com/aws-samples/intelligent-document-processing-at-scale

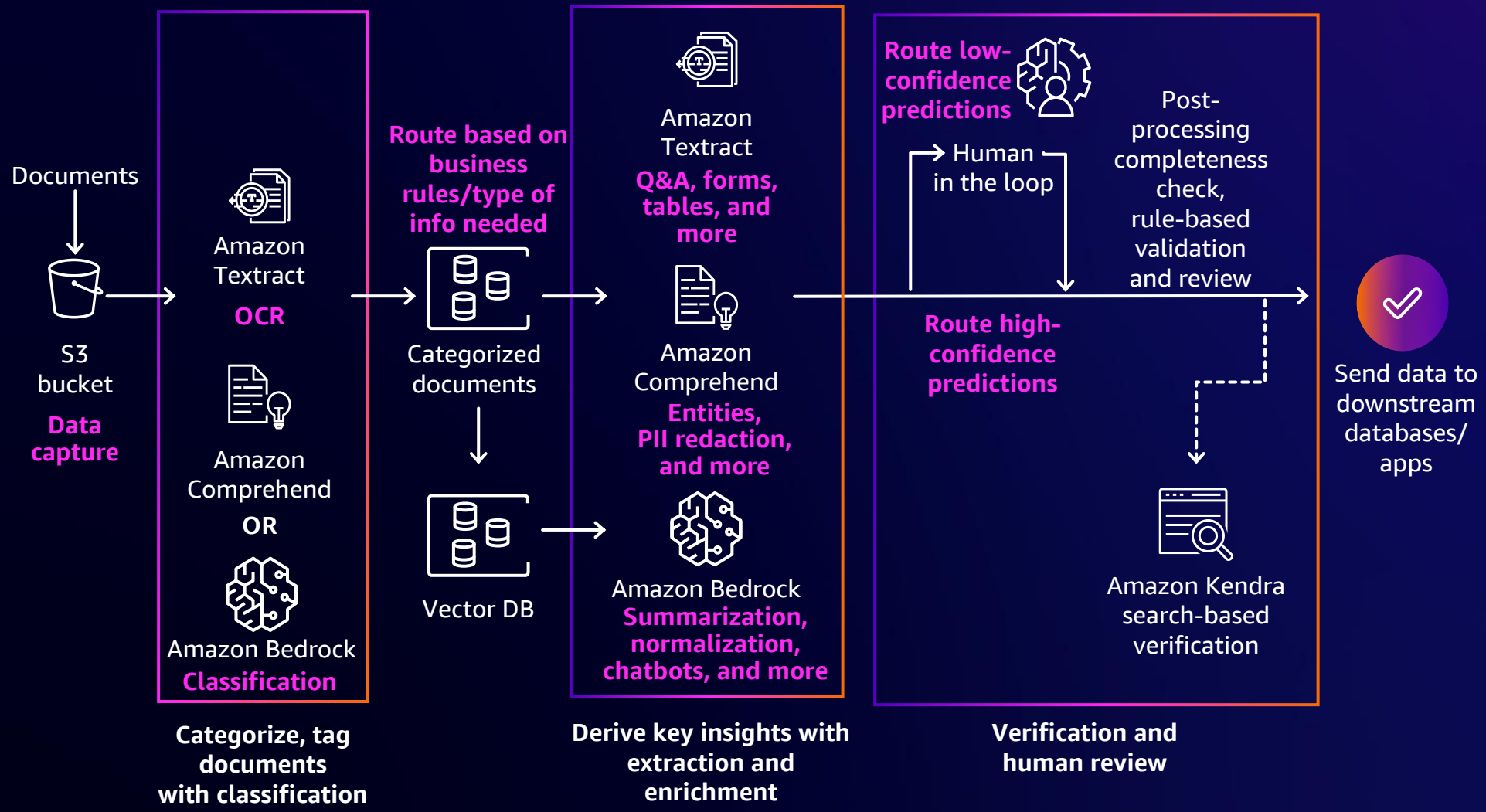
Questions

Moving forward with generative AI

- Built prototype to answer predefined questions
- Collaborating with AWS on personally identifiable information recognition models as a step toward keeping generative AI solutions compliant
- Future possibilities
  - Prepopulate long, complex forms based on model findings
  - Generate recommended adjudication of claims
  - Racial and gender bias reports

AI is here; let's generate a better world!

Document pipeline with AWS IDP and generative AI



How to get started

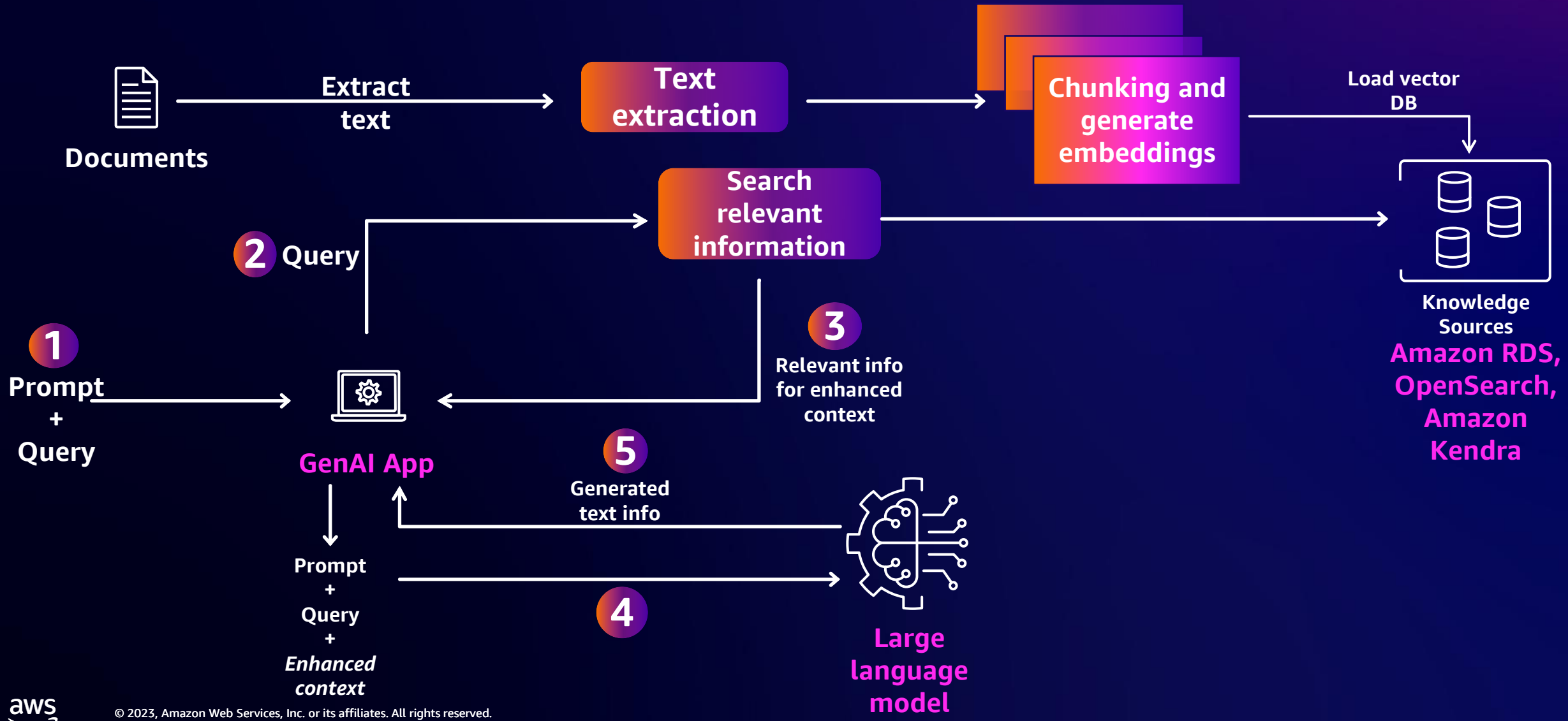
Developers can easily embed AI-powered functionality from Textract and Comprehend into your business workflows and apps

Engage your data science team for FM selection, evaluation and tuning based on your generative AI use case

Link your gen AI/FM modules with AWS IDP (e.g., through chaining) to create an end-to-end document processing pipeline

Using retrieval-augmented generation (RAG)

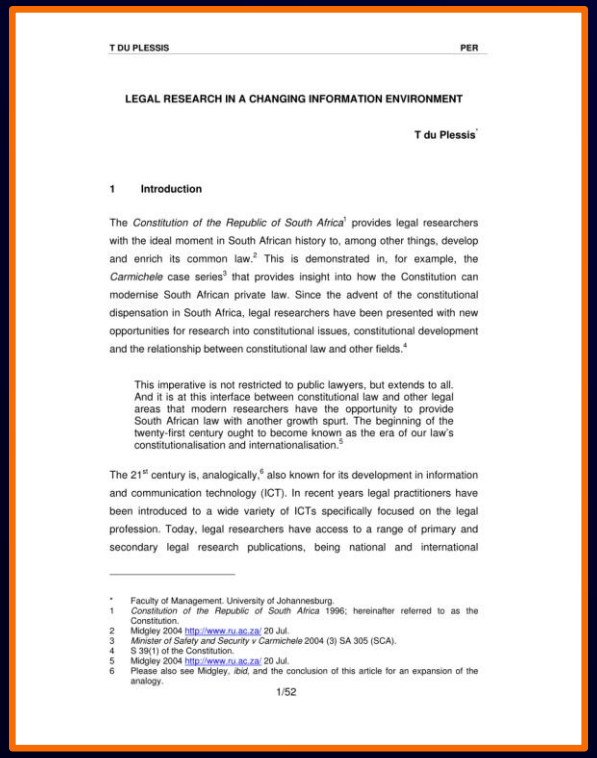
AUGMENT PROMPTS WITH RELEVANT DATA IN CONTEXT



Layout detection for improved LLM accuracy

LAUNCHED!

DOCUMENT



T DU PLESSIS PER

LEGAL RESEARCH IN A CHANGING INFORMATION ENVIRONMENT

T du Plessis*

1 Introduction

The *Constitution of the Republic of South Africa*[1] provides legal researchers with the ideal moment in South African history to, among other things, develop and enrich its common law.[2] This is demonstrated in, for example, the *Carmichele* case series[3] that provides insight into how the Constitution can modernise South African private law. Since the advent of the constitutional dispensation in South Africa, legal researchers have been presented with new opportunities for research into constitutional issues, constitutional development and the relationship between constitutional law and other fields.[4]

> This imperative is not restricted to public lawyers, but extends to all. And it is at this interface between constitutional law and other legal areas that modern researchers have the opportunity to provide South African law with another growth spurt. The beginning of the twenty-first century ought to become known as the era of our law's constitutionalisation and internationalisation.[5]

The 21st century is, analogically,[6] also known for its development in information and communication technology (ICT). In recent years legal practitioners have been introduced to a wide variety of ICTs specifically focused on the legal profession. Today, legal researchers have access to a range of primary and secondary legal research publications, being national and international

* Faculty of Management. University of Johannesburg.
1 *Constitution of the Republic of South Africa* 1996; hereinafter referred to as the Constitution.
2 Midgley 2004 http://www.ru.ac.za/ 20 Jul.
3 *Minister of Safety and Security v Carmichele* 2004 (3) SA 305 (SCA).
4 S 39(1) of the Constitution.
5 Midgley 2004 http://www.ru.ac.za/ 20 Jul.
6 Please also see Midgley, *ibid*, and the conclusion of this article for an expansion of the analogy.

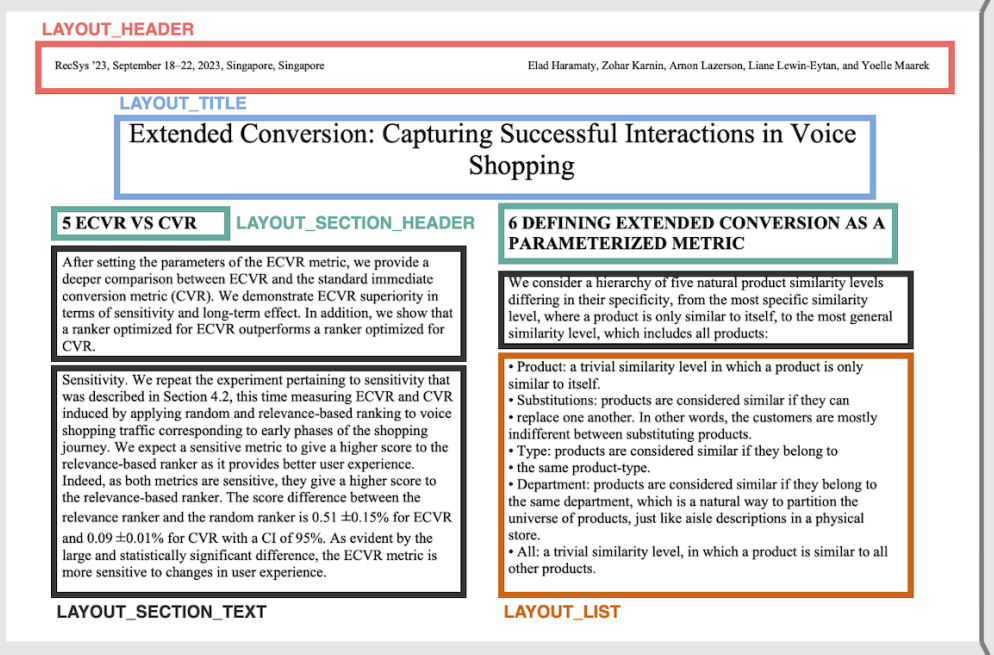
1/52

OUTPUT

Header: T DU PLESSIS
Header: PER
Title: LEGAL RESEARCH IN A CHANGING INFORMATION ENVIRONMENT
Paragraph: 1 Introduction
Paragraph: The Constitution of the Republic of South Africa provides legal . . .
Paragraph: This imperative is not redacted to public . . .
Paragraph: The 21st century is analogically also known for . . .
Footer: Faculty of
Footer: Contributors

- Extract titles, paragraphs, footer, headers, etc.
- Orders output in a way humans would read (reading order)
- Reduced post-processing when extracting information from dense documents

Example




```
RecSys '23, September 18–22, 2023, Singapore, Singapore

Elad Haramaty, Zohar Karnin, Arnon Lazerson, Liane Lewin-Eytan, and Yoelle Maarek

Extended Conversion: Capturing Successful Interactions in Voice
Shopping

5 ECVR VS CVR

After setting the parameters of the ECVR metric, we provide a
deeper comparison between ECVR and the standard immediate
conversion metric (CVR). We demonstrate ECVR superiority in
terms of sensitivity and long-term effect. In addition, we show that
a ranker optimized for ECVR outperforms a ranker optimized for
CVR.

Sensitivity. We repeat the experiment pertaining to sensitivity that
was described in Section 4.2, this time measuring ECVR and CVR
induced by applying random and relevance-based ranking to voice
shopping traffic corresponding to early phases of the shopping
journey. We expect a sensitive metric to give a higher score to the
relevance-based ranker as it provides better user experience.
Indeed, as both metrics are sensitive, they give a higher score to
the relevance-based ranker. The score difference between the
relevance ranker and the random ranker is 0.51 +0.15% for ECVR
and 0.09 0.01% for CVR with a CI of 95%. As evident by the
large and statistically significant difference, the ECVR metric is
more sensitive to changes in user experience.

6 DEFINING EXTENDED CONVERSION AS A
PARAMETERIZED METRIC

We consider a hierarchy of five natural product similarity levels
differing in their specificity, from the most specific similarity
level, where a product is only similar to itself, to the most general
similarity level, which includes all products:

......
```

***Extracted text trimmed for brevity*

https://aws.amazon.com/blogs/machine-learning/amazon-textracts-new-layout-feature-introduces-efficiencies-in-general-purpose-and-generative-ai-document-processing-tasks/

Questions

Getting started

Get started with Amazon Bedrock

IDP hands-on Workshop with Layout

Get started with hands-on Amazon Bedrock workshop

Get started with hands-on IDP OpenSearch workshop

Thank you!

Please complete the session survey in the mobile app

Sonali Sahu

sonalsah@amazon.com

Cameron Williams