



WOR102

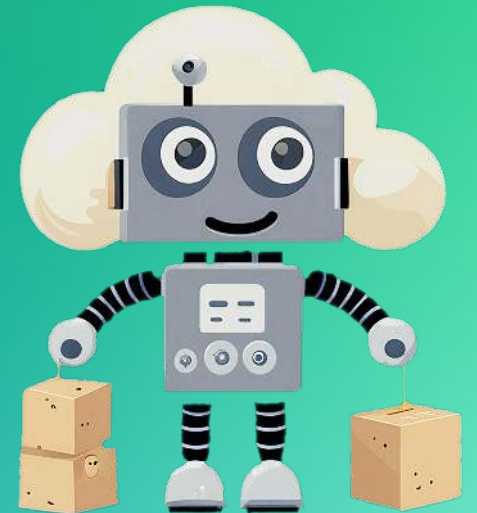
Extracting Information From Complex Documents with Amazon Bedrock

Nikita Kozodoi, PhD

(he/him)
Sr Applied Scientist
Amazon Web Services

Aiham Taleb, PhD

(he/him)
Applied Scientist
Amazon Web Services



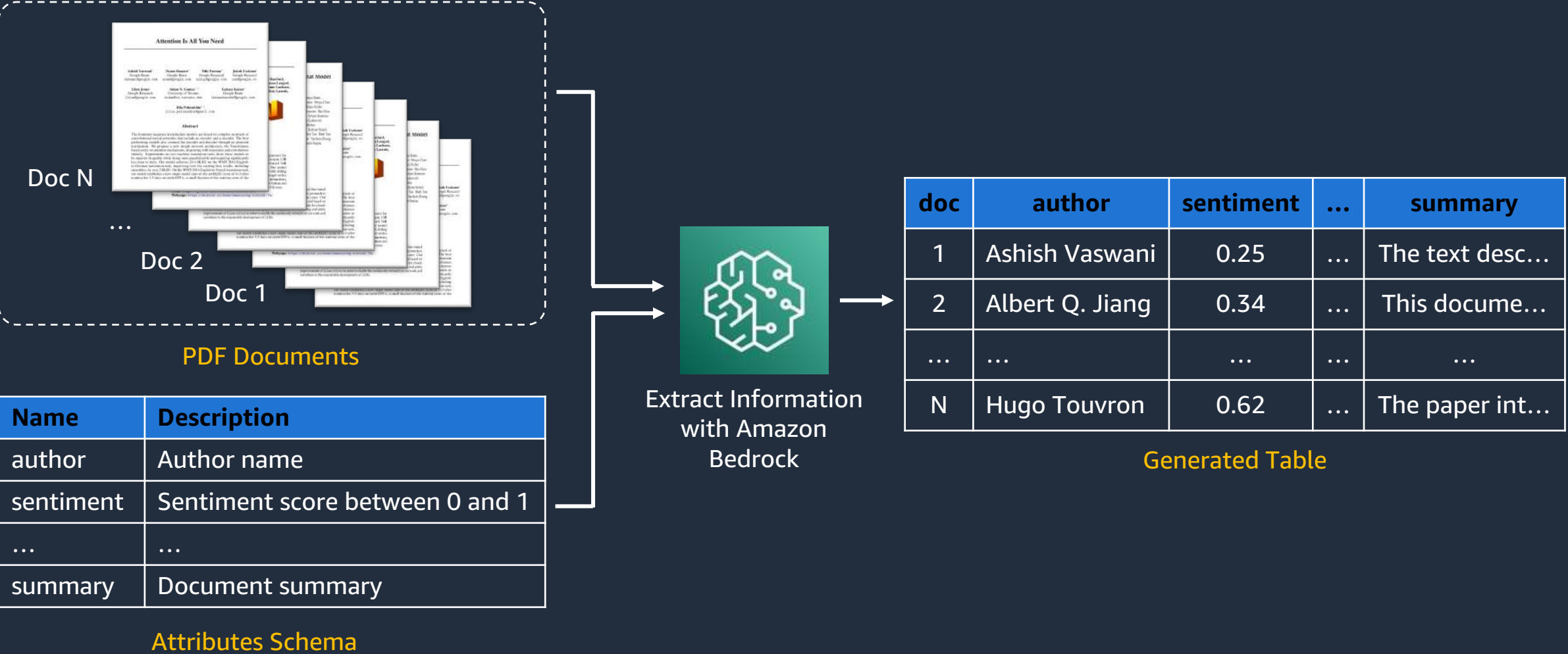
Overview



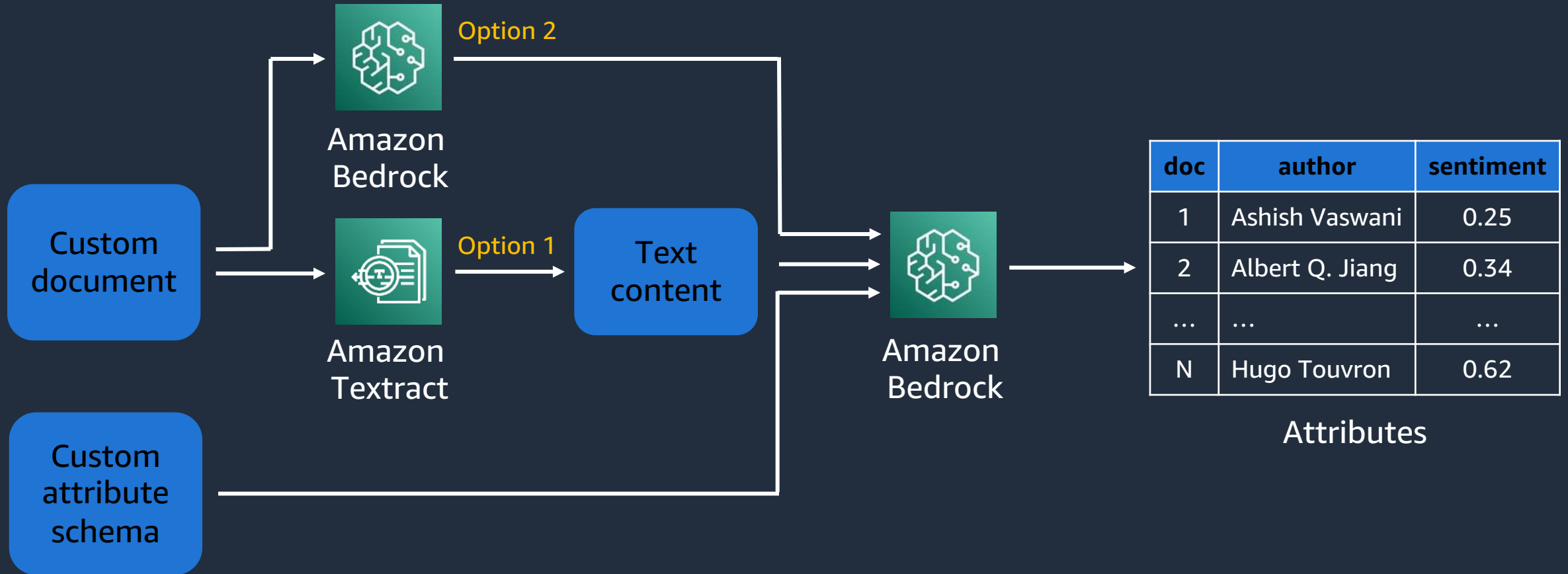
Recurring Business Task

-  Transform custom documents into a structured database
 - **Extracting entities:** e.g., names, dates, IDs, etc.
 - **Assigning scores:** e.g., sentiment, readability, urgency, etc.
 - **Adding free-form content:** e.g., summary, suggested response, etc.
-  Example use cases:
 - Extract information from **financial news articles** (e.g., sentiment, summary, date)
 - Get entities from **customer emails** (e.g., issue type, urgency score, suggested reply)
 - Extract medical & nutrition entities from **research papers** (e.g., chemical elements, drug names)

Suggested Approach



High-Level Architecture



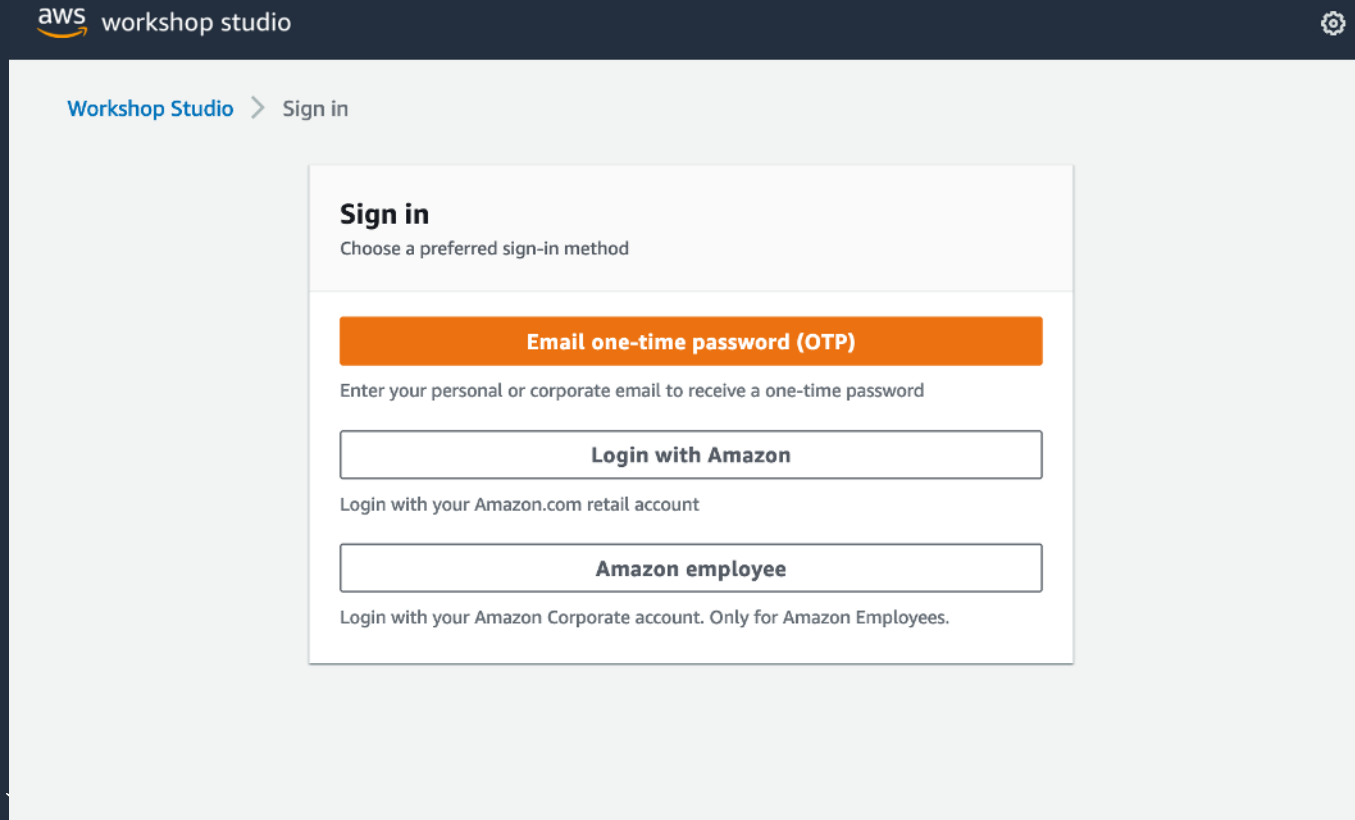
Getting Started



1. Join the Workshop Session

Open this link on your laptop and log in with the one-time-password

<https://bit.ly/3Nyd3a4>



The screenshot shows the AWS Workshop Studio sign-in interface. At the top left is the AWS logo followed by the text "workshop studio". At the top right is a gear icon. Below the header, there is a breadcrumb trail: "Workshop Studio > Sign in". The main content area is a white box with a light gray border. Inside, the heading "Sign in" is followed by the instruction "Choose a preferred sign-in method". There are three sign-in options, each in a white box with a thin gray border. The first option, "Email one-time password (OTP)", has an orange header bar. Below it, the text says "Enter your personal or corporate email to receive a one-time password". The second option, "Login with Amazon", has a white header bar. Below it, the text says "Login with your Amazon.com retail account". The third option, "Amazon employee", has a white header bar. Below it, the text says "Login with your Amazon Corporate account. Only for Amazon Employees."

aws workshop studio

Workshop Studio > Sign in

Sign in
Choose a preferred sign-in method

Email one-time password (OTP)
Enter your personal or corporate email to receive a one-time password

Login with Amazon
Login with your Amazon.com retail account

Amazon employee
Login with your Amazon Corporate account. Only for Amazon Employees.

2. Open the Workshop Environment

Open the AWS console from the workshop page (left sidebar)

The screenshot shows the AWS Workshop Studio interface. On the left sidebar, under 'AWS account access', the link 'Open AWS console' is highlighted with a purple box and a purple arrow. The main content area displays the 'Event dashboard' for the 'AWS General Immersion Day'. It includes event information such as start time (9/23/2022 01:13 AM) and duration (12). Below this, there is a 'Workshop' section with a table listing AWS services and topics. The footer contains copyright information: '© 2008 - 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy policy Terms of use'.

Title	Complexity level	AWS services	Topics
AWS General Immersion Day	100	Amazon EC2 Auto Scaling, Amazon Relational Database Service (Amazon RDS), Amazon Simple Storage Service (Amazon S3)	Compute, Databases, Networking & Content Delivery, Storage



The screenshot shows the AWS Console Home interface. The top navigation bar includes 'Console Home' and 'Add widgets'. The main content area displays several widgets: 'Welcome to AWS' with links for 'Getting started with AWS', 'Training and certification', and 'What's new with AWS?'; 'AWS Health' showing 'No health data'; and 'Cost and usage'. The footer contains links for 'Feedback', 'Looking for language selection? Find it in the new Unified Settings', and copyright information: '© 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences'.

2. Open the Workshop Environment

Open the AWS console and switch to the *us-west-2* region if you are not already there

The screenshot shows the AWS Management Console interface. At the top, the 'Services' bar includes icons for Amazon SageMaker, S3, EC2, CodeCommit, Elastic Container Registry, Amazon OpenSearch Service, CloudFormation, Amazon Bedrock, and Lambda. The 'Console Home' page displays a 'Recently visited' section with a grid of service tiles: Amazon Bedrock, IAM, CloudWatch, Lambda, S3, Amazon SageMaker, EC2, CloudFormation, Amazon Kendra, Amazon OpenSearch Service, DynamoDB, and Amazon Transcribe. A 'View all services' link is at the bottom. The right sidebar, titled 'Welcome to AWS', contains links for 'Getting started with AWS', 'Training and certification', and 'What's new with AWS?'. The top right corner shows the current region as 'Oregon' with a dropdown menu open, listing various AWS regions. 'US West (Oregon)' is highlighted in orange, corresponding to the 'us-west-2' region mentioned in the text.

Region	Region Code
US East (N. Virginia)	us-east-1
US East (Ohio)	us-east-2
US West (N. California)	us-west-1
US West (Oregon)	us-west-2
Asia Pacific (Mumbai)	ap-south-1
Asia Pacific (Osaka)	ap-northeast-3
Asia Pacific (Seoul)	ap-northeast-2
Asia Pacific (Singapore)	ap-southeast-1
Asia Pacific (Sydney)	ap-southeast-2
Asia Pacific (Tokyo)	ap-northeast-1
Canada (Central)	ca-central-1
Europe (Frankfurt)	eu-central-1
Europe (Ireland)	eu-west-1

2. Open the Workshop Environment

Under *services*, search for *Amazon SageMaker*

Search results for 'sagemaker'

Services (2)

Features (1)

Documentation (39,311)

Marketplace (106)

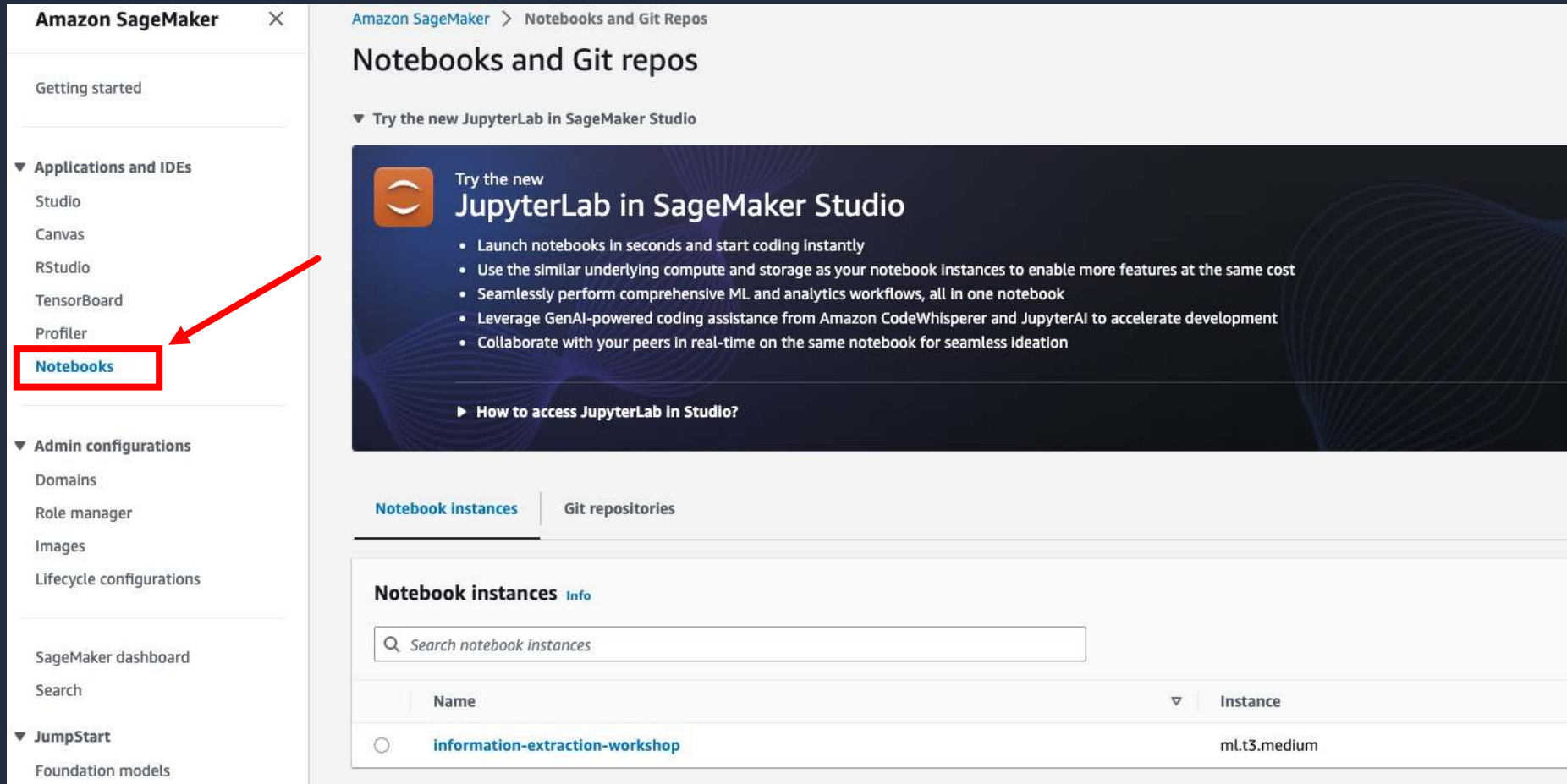
Services



Amazon SageMaker
Build, Train, and Deploy Machine Learning Models

2. Open the Workshop Environment

In the side bar, find the *Applications and IDEs* section, and click on *Notebooks*



The screenshot shows the Amazon SageMaker console interface. On the left, the 'Applications and IDEs' section is expanded, and the 'Notebooks' link is highlighted with a red box and a red arrow. The main content area displays the 'Notebooks and Git repos' page, which includes a banner for 'Try the new JupyterLab in SageMaker Studio' and a list of features. Below the banner, there are tabs for 'Notebook instances' and 'Git repositories'. The 'Notebook instances' tab is active, showing a search bar and a table with one instance named 'information-extraction-workshop'.

Amazon SageMaker X

Getting started

▼ Applications and IDEs

- Studio
- Canvas
- RStudio
- TensorBoard
- Profiler
- Notebooks**

▼ Admin configurations

- Domains
- Role manager
- Images
- Lifecycle configurations

SageMaker dashboard

Search

▼ JumpStart

- Foundation models

Amazon SageMaker > Notebooks and Git Repos

Notebooks and Git repos

▼ Try the new JupyterLab in SageMaker Studio

 Try the new
JupyterLab in SageMaker Studio

- Launch notebooks in seconds and start coding instantly
- Use the similar underlying compute and storage as your notebook instances to enable more features at the same cost
- Seamlessly perform comprehensive ML and analytics workflows, all in one notebook
- Leverage GenAI-powered coding assistance from Amazon CodeWhisperer and JupyterAI to accelerate development
- Collaborate with your peers in real-time on the same notebook for seamless ideation

► How to access JupyterLab in Studio?

Notebook instances | Git repositories

Notebook instances Info

🔍 Search notebook instances

	Name	Instance
☐	information-extraction-workshop	m1.t3.medium

2. Open the Workshop Environment

A Notebook instance should already be provisioned

If the notebook status is "Stopped", click **Actions** => **Start**

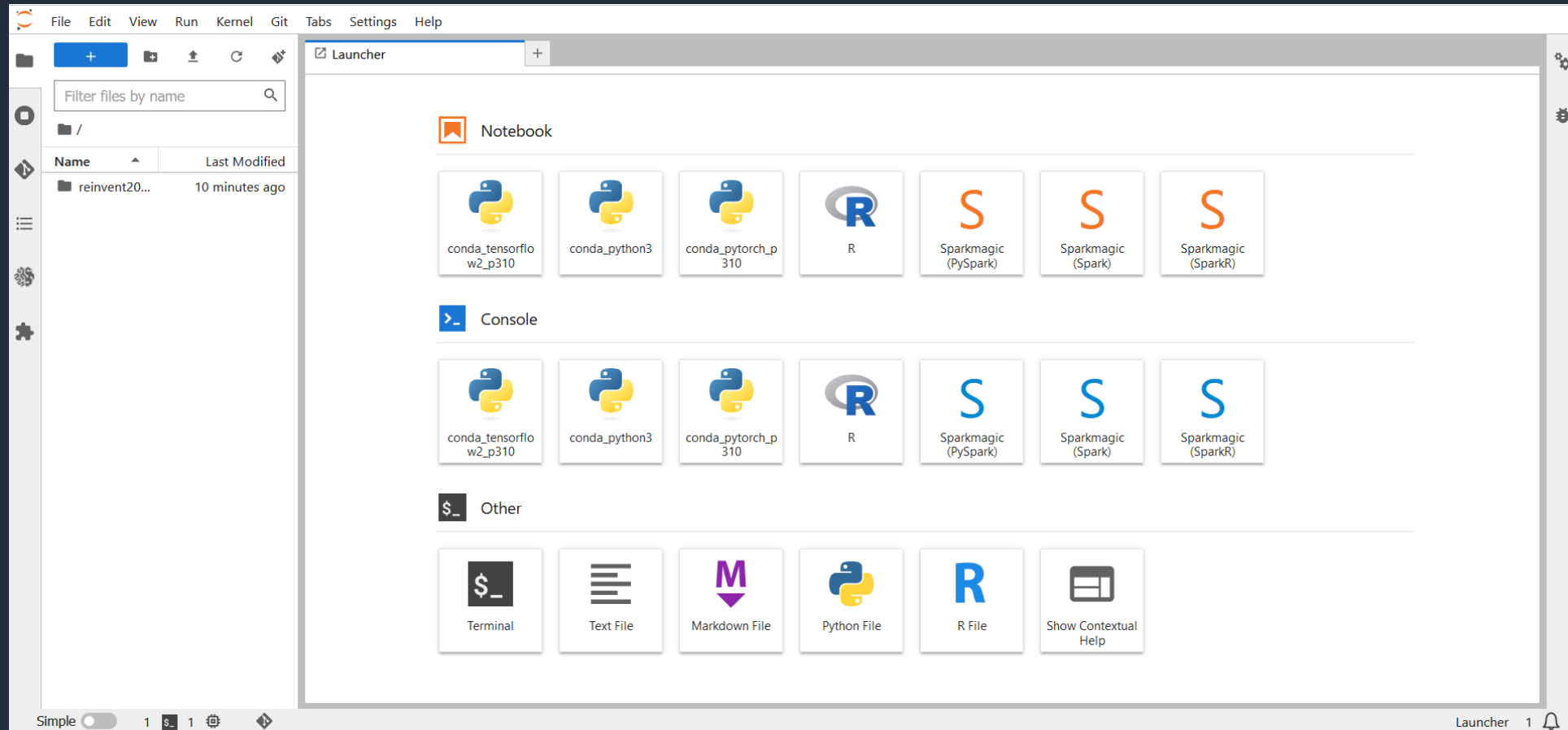
Click on **Open JupyterLab** (on the right side of the provisioned Notebook instance under **Actions**)

The screenshot displays the Amazon SageMaker console's 'Notebook instances' page. On the left is a sidebar with navigation links: 'Dashboard', 'Search', and 'SageMaker Domain' (with sub-links 'Studio' and 'RStudio'). The main content area shows a table of notebook instances. The table has columns for 'Name', 'Instance', 'Creation time', 'Status', and 'Actions'. A single instance named 'ThreeDEndToEnd-NB' is listed with instance type 'ml.g4dn.2xlarge' and creation time 'Nov 23, 2021 20:55 UTC'. Its status is 'InService'. In the 'Actions' column, there are two buttons: 'Open Jupyter' and 'Open JupyterLab'. A red arrow points from the text 'Click here' to the 'Open JupyterLab' button, which is also highlighted with a red box.

Name	Instance	Creation time	Status	Actions
ThreeDEndToEnd-NB	ml.g4dn.2xlarge	Nov 23, 2021 20:55 UTC	InService	Open Jupyter Open JupyterLab

2. Open the Workshop Environment

You will be redirected to a new browser tab that looks like this:



3. Work Through the Lab

Start working through the notebooks one by one (together with us or at your own pace)

You will find three notebooks:

- **00_prerequisites.ipynb**: contains workshop prerequisites [**~5-10 min**]
- **01_process_pdf.ipynb**: processes PDF documents and saves processed files [**~30 min**]
- **02_extract_info.ipynb**: extracts custom information from the processed files [**~30 min**]

Hands-On Session

Go through *00_prerequisites.ipynb* [~10 min]



Hands-On Session

Work on *01_process_pdf.ipynb* [~30 min]

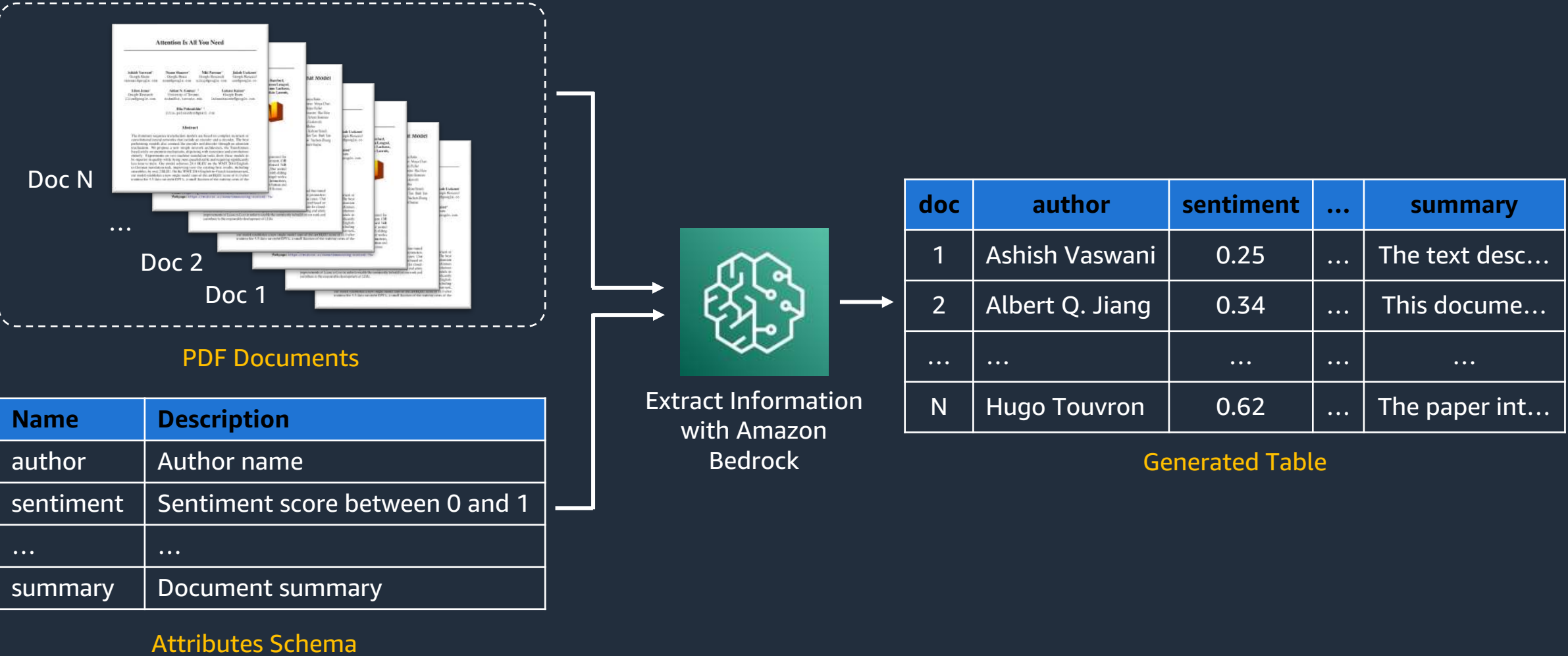


Hands-On Session

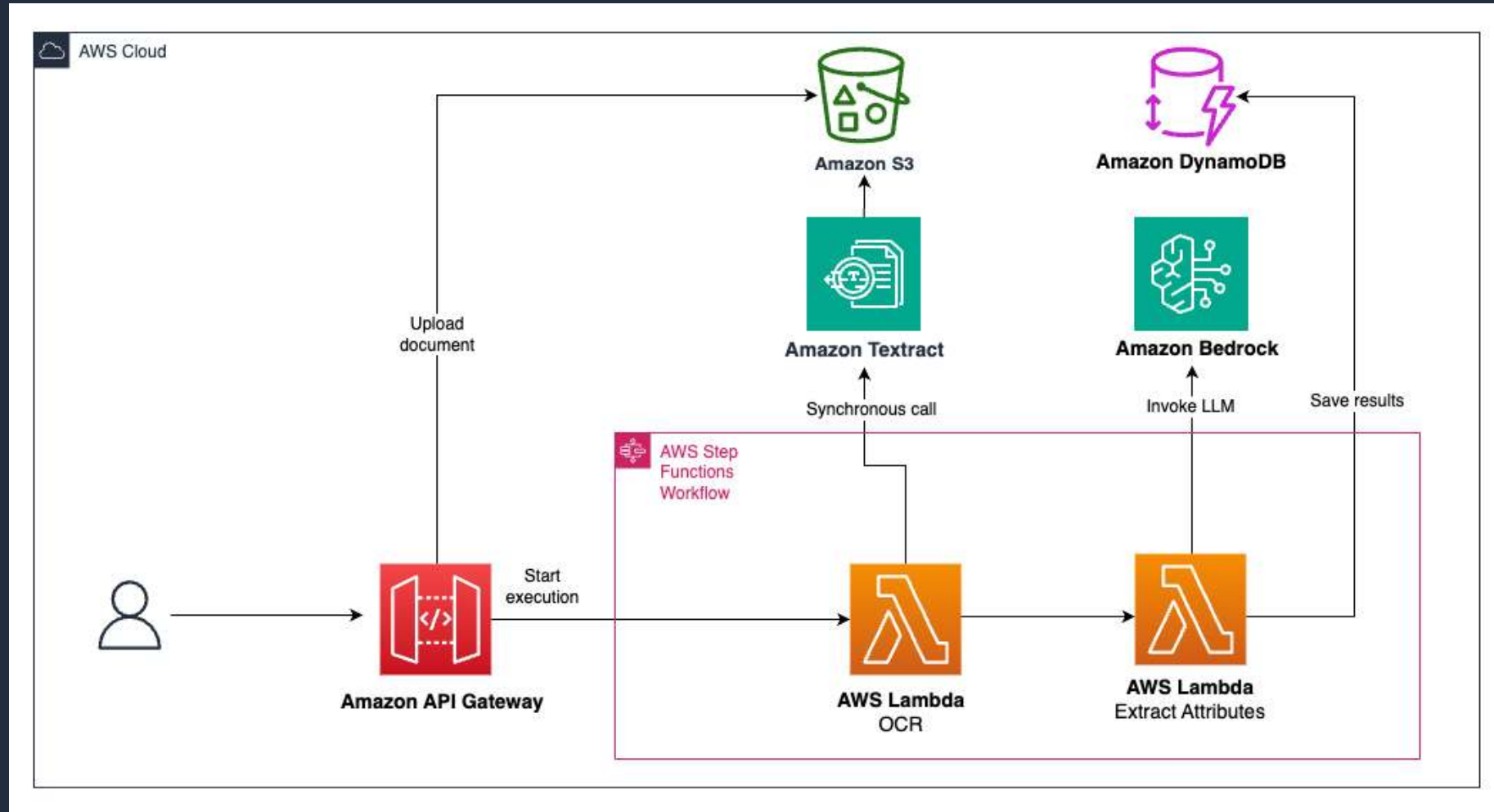
Work on *02_extract_attributes.ipynb* [~30 min]



Solution Recap



Scaling Up





Thank you!

Nikita Kozodoi, PhD

(he/him)
Sr Applied Scientist
Amazon Web Services

Aiham Taleb, PhD

(he/him)
Applied Scientist
Amazon Web Services



Please complete the
session survey.



Appendix



LLM Prompt Template

```
Human: You are an expert AI assistant who can extract information from texts.  
Read the document below and extract the entities or assign scores listed below in <entities></entities> tags.  
The answer must contain the extracted entities and scores in JSON format. Do not include any other  
information in the answer.  
If the entity has multiple values, provide them as a list in this format: ['value1', 'value2', 'value3'].
```

```
Here is an example:  
<example>  
...  
</example>
```

← Few-shot examples

```
Document:  
<document>  
{document}  
</document>
```

← Input document(s)

```
Entities and scores to be extracted:  
<entities>  
{entities}  
</entities>
```

← User-provided entities in this format:

- {name_1}: {description_1} (must be {type_1})
- {name_2}: {description_2} (must be {type_2})
- ...

```
Assistant:
```

Extracting Different Information Types

-  Explicit well-defined entities:
 - Document title
 - Product price
-  Implicit entities and scores:
 - Sentiment score
 - Priority
-  Complex generated content:
 - Document summary
 - Translation

Problem Setting

-  Existing tools like **Amazon Comprehend** use entity recognition models with limitations:
 - Pre-trained models only support a **fixed limited set of entities**
 - Using custom entities requires **collecting annotated data**
 - Adding or modifying an entity requires **retraining the model**
-  Solution: LLM-based entity recognition and scoring
 - **Does not require labeled data** (optional few-shot examples)
 - Supports **custom set of entities** without retraining
 - Goes beyond NER: also allows assigning **numeric scores** (e.g. sentiment)
 - Research shows LLAMA-based NER **matches performance** of fine-tuned BERT (Goel et al. 2023)