INTELLIGENT DOCUMENT PROCESSING WITH AWS AI SERVICES

GROUP 1

- JIDE OLOKO
- CHINAZAEKPERE OBIMDIKE
- ADEMIDE OLANREWAJU
- CHIOMA EGWUIBE

Cloudboosta

01

Solution Introduction

02

Implementation

03

Challenges & Recommendations

THE CONCEPT OF INTELLIGENT DOCUMENT PROCESSING (IDP)

Intelligent Document Processing (IDP) uses AI to automate document processing.
This boosts efficiency, accuracy, and decision-making.
It's revolutionizing how businesses handle document-intensive tasks.

Cloudboosta

THE BENEFITS OF INTELLIGENT DOCUMENT PROCESSING

Increased efficiency

Automate manual tasks and reduce processing time.

Improved accuracy

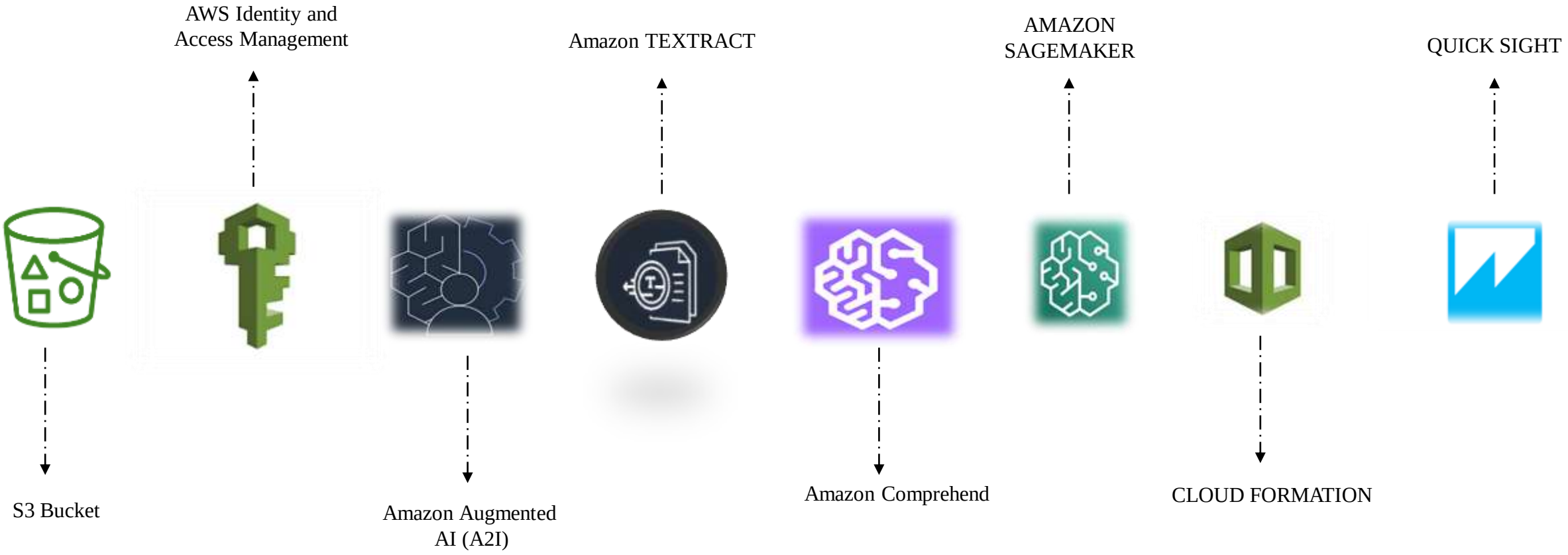
Achieve higher accuracy in data extraction and analysis.

Enhanced decision-making

Gain valuable insights from document data to make informed decisions.

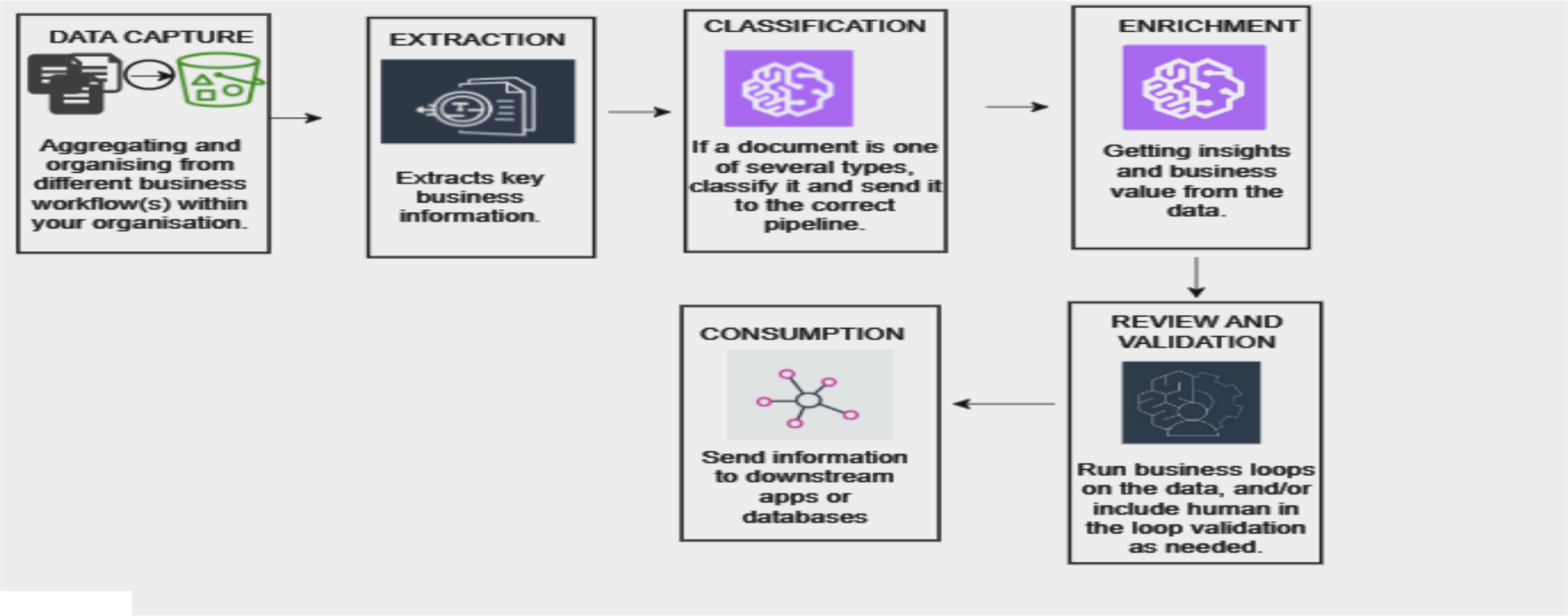
Cloudboosta

AWS AI SERVICES DEPLOYED FOR THIS PROJECT



Cloudboosta

INTELLIGENT DOCUMENT PROCESSING (IDP) WORKFLOW



Cloudboosta

Data Governance- Security and Policies

- **IAM Roles and Policies:** Use granular permissions and MFA.
- **Data Encryption:** Encrypt data at rest and in transit.
- **Access Control:** Use ACLs and security groups.
- **Data Loss Prevention:** Implement DLP measures.
- **Logging and Monitoring:** Track activity and detect anomalies.
- **Vulnerability Management:** Scan for vulnerabilities and apply patches.
- **Incident Response:** Have a plan and conduct regular drills.
- **Compliance:** Ensure compliance with relevant regulations.
- **Third-Party Integrations:** Evaluate security practices of third-party services.

Cloudboosta

01

Solution Introduction

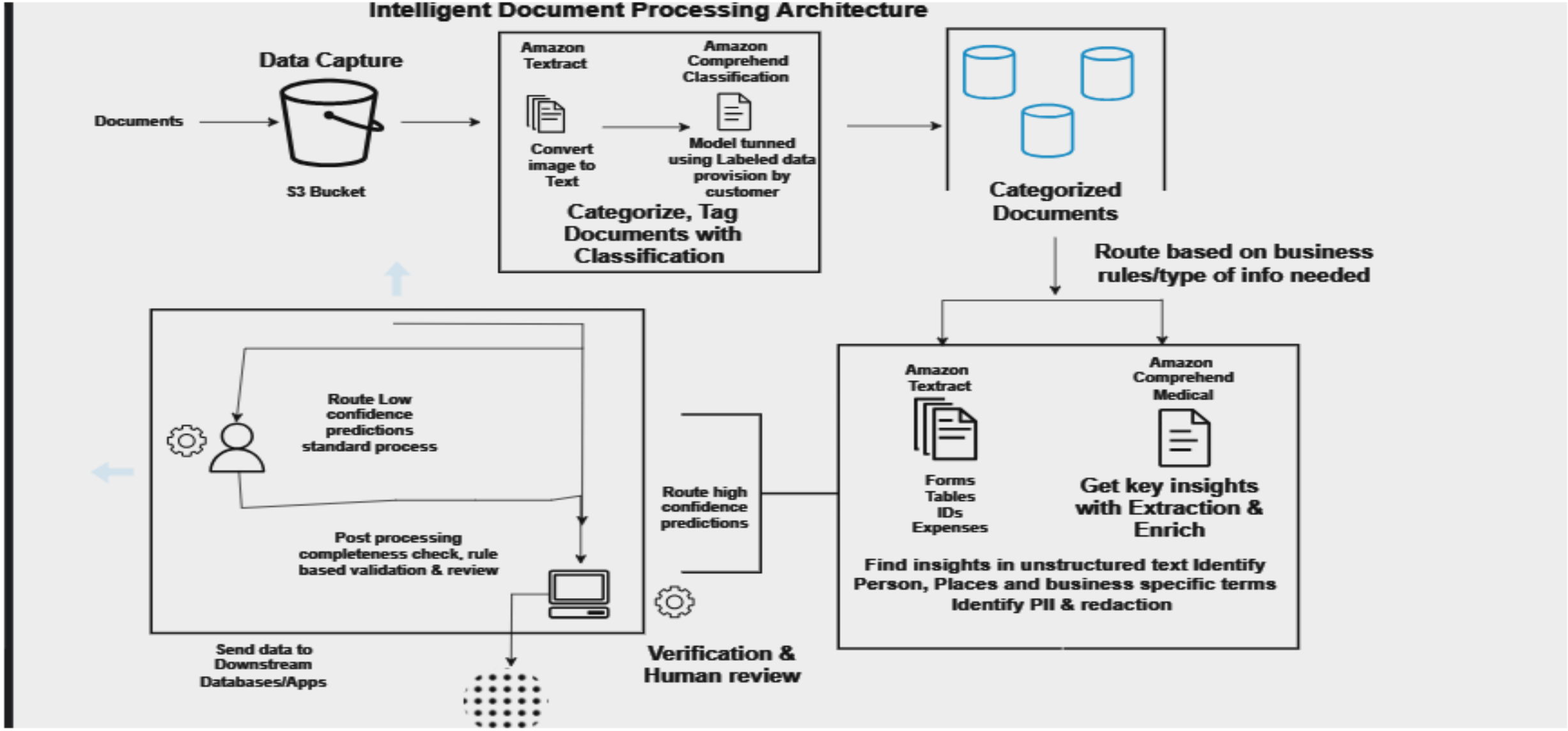
02

Implementation

03

Challenges & Recommendations

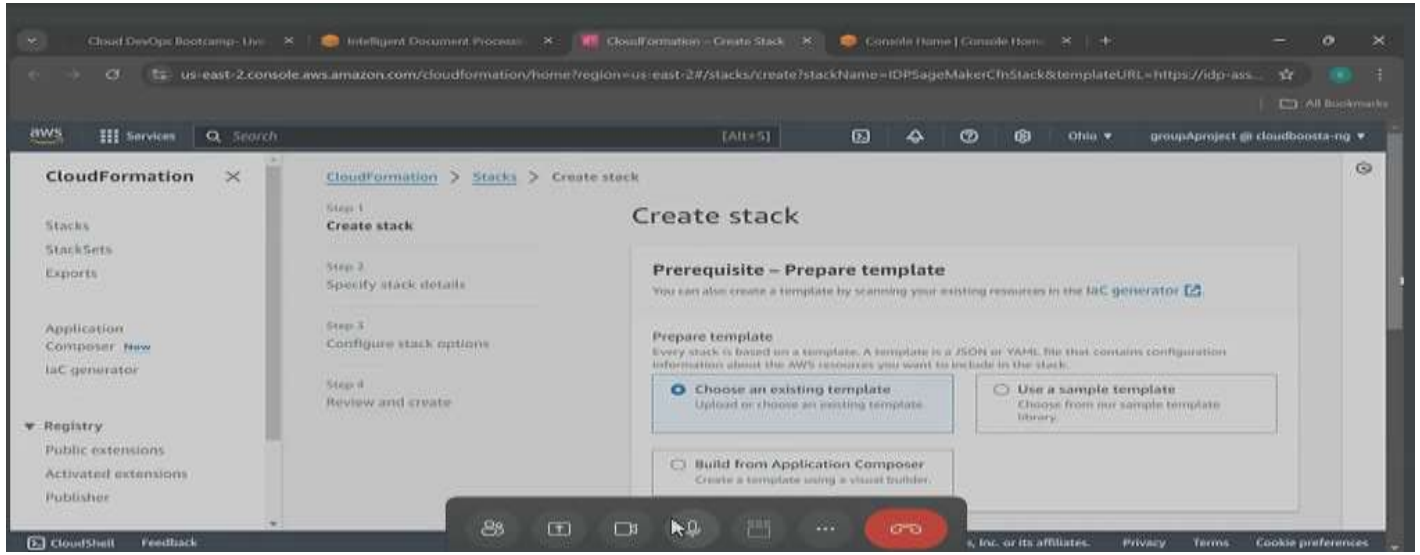
INTELLIGENT DOCUMENT PROCESSING ARCHITECTURE



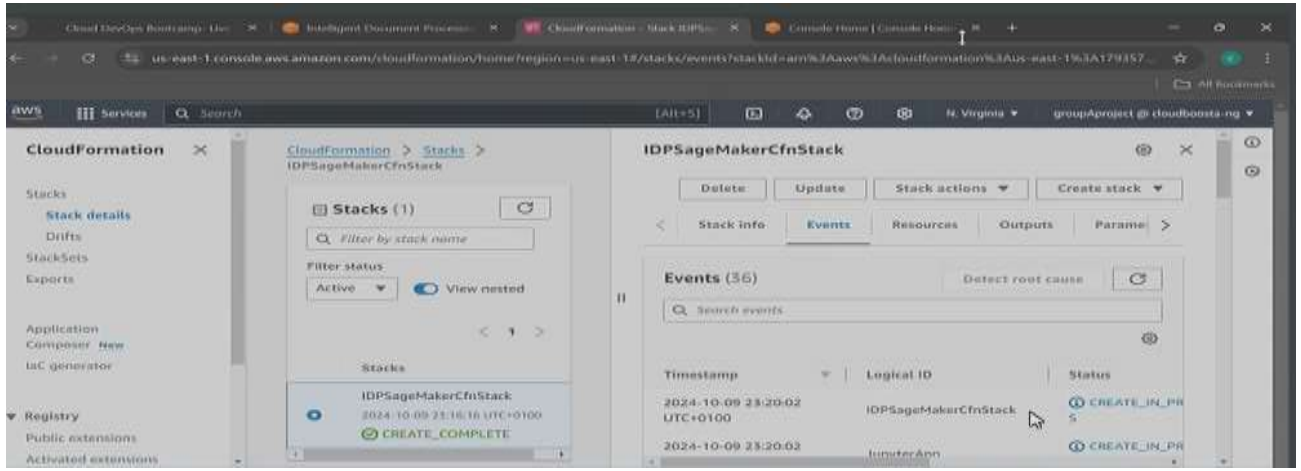
Cloudboosta

GETTING STARTED

1. Creation of stack via Cloudformation



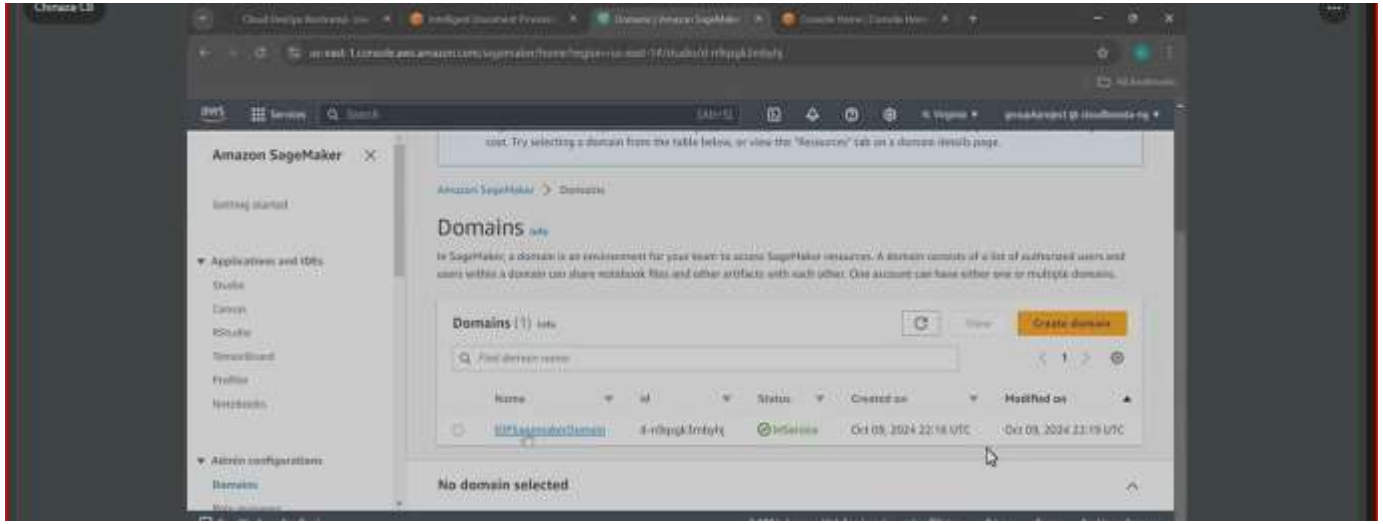
Ensure Stack is Created Successfully.



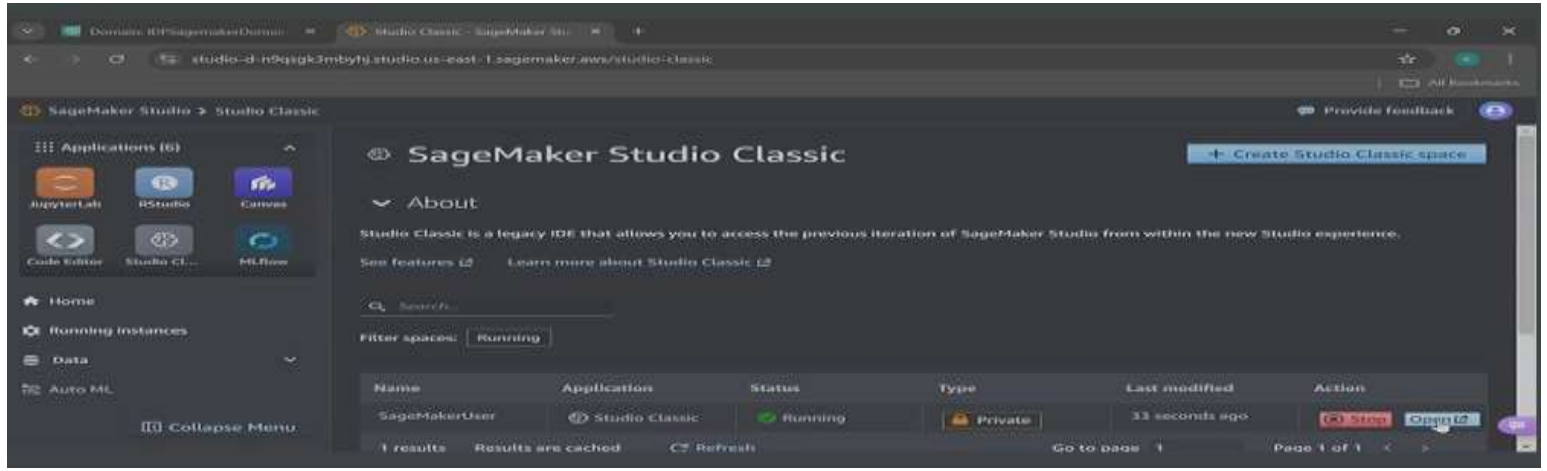
Cloudboosta

GETTING STARTED

1. On Amazon SageMaker, Domain is created automatically



The Studio is Launched via the Domain. Then we can access our Notebook to run our scripts.



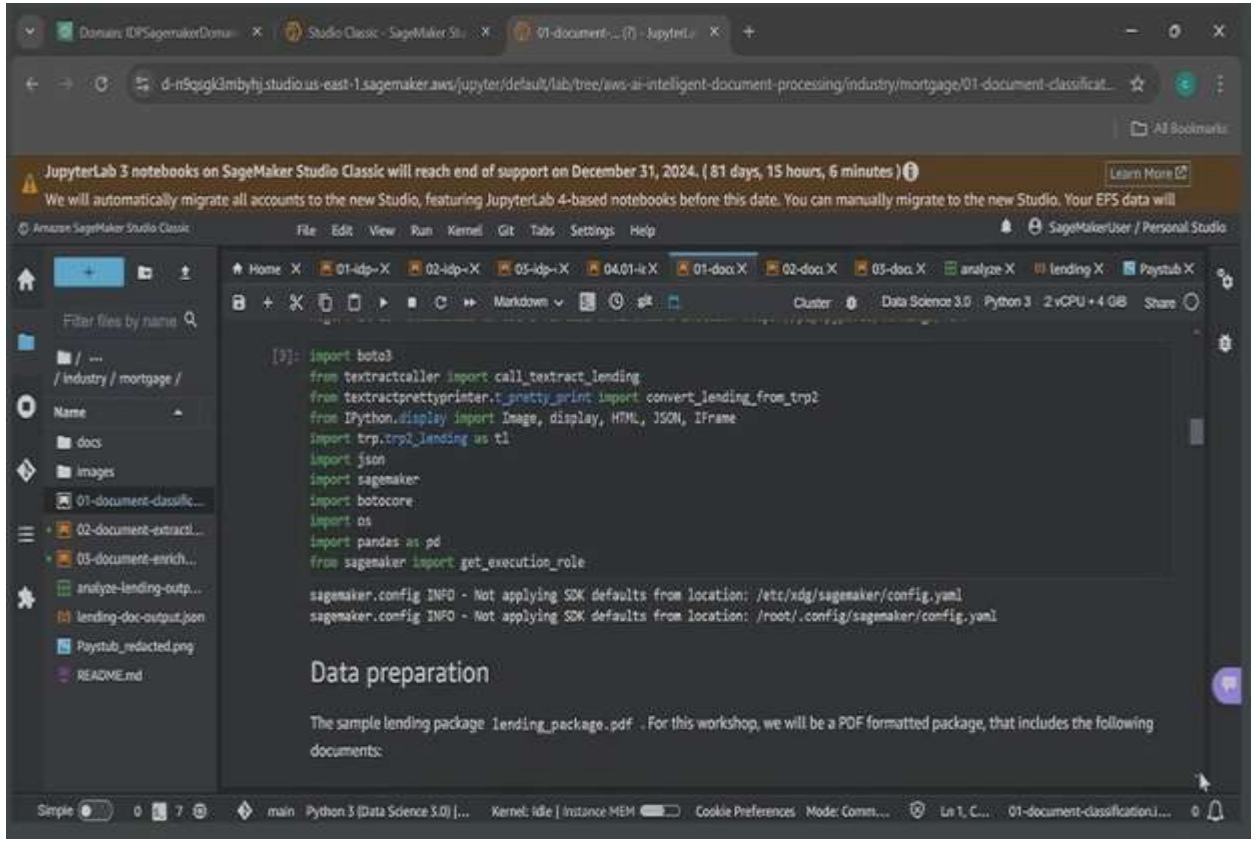
Cloudboosta

1. IDP Architecture
2. Getting Started
3. Level 1:Document Extraction
4. Level 2: Document Classification
5. Level 3: Document Enrichment
6. Level 4: Document Review And Verification (A2I)

DOCUMENT EXTRACTION ARCHITECTURE

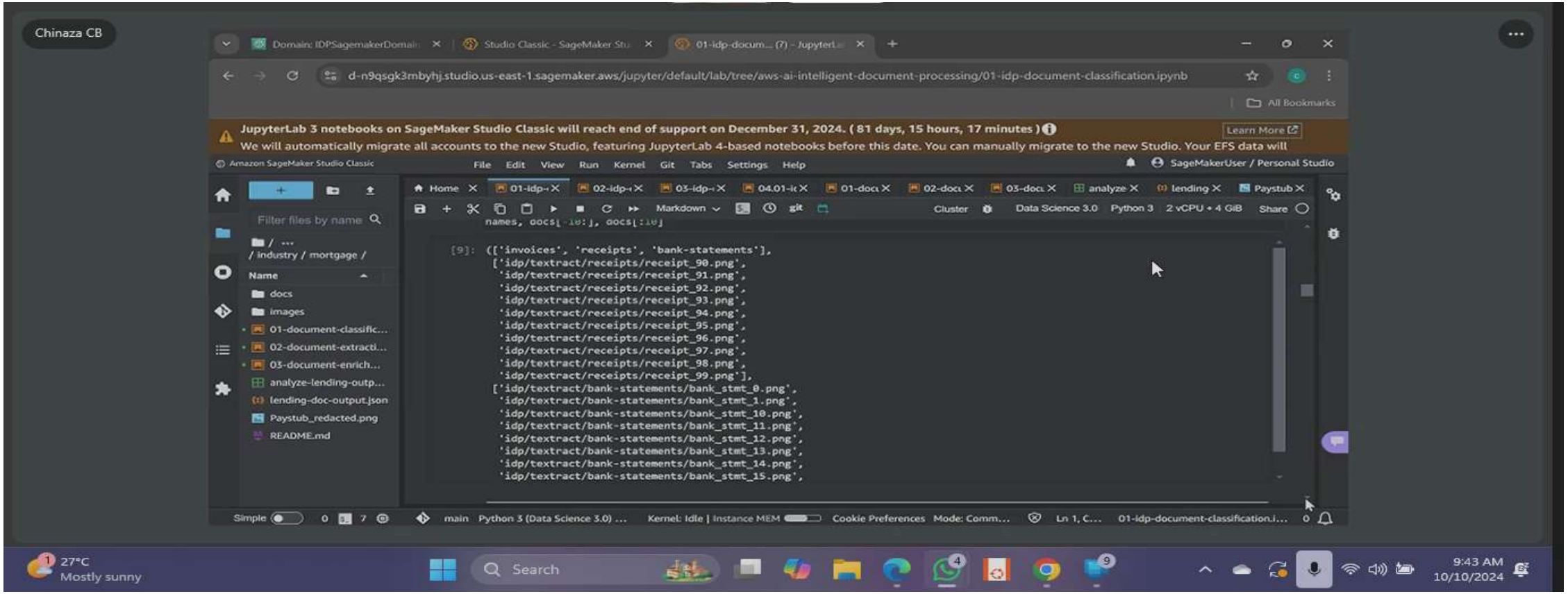
DOCUMENT EXTRACTION

- Setup Notebook By Deploying Cloudformation Stacks On Amazon Sagemaker.
- The Data Was Prepared And Uploaded Into An S3 Bucket.
- Amazon Textract's "detect_document_text" API to extract the raw text information for all the documents in S3



Cloudboosta

DOCUMENT EXTRACTED VIA AMAZON TEXTRACT'S "DETECT_DOCUMENT_TEXT" API



Cloudboosta

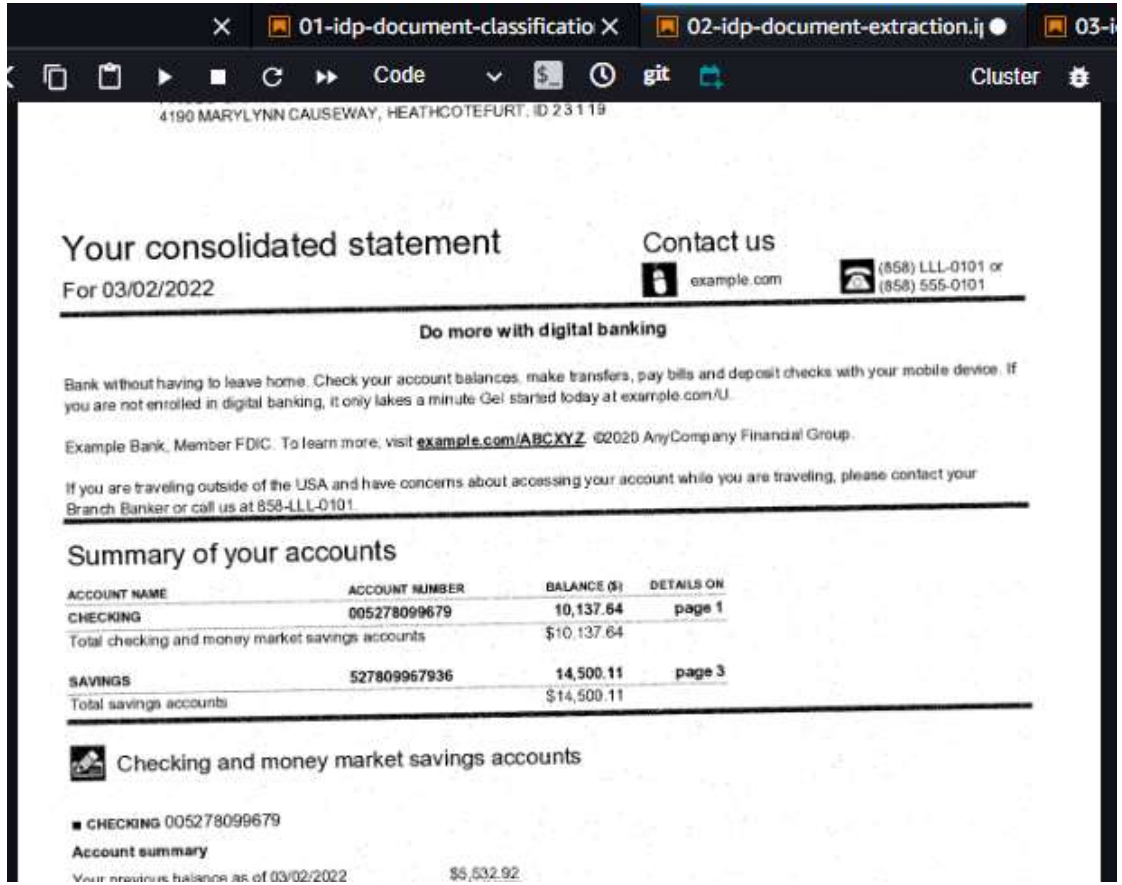
DOCUMENT EXTRACTION: Data Extracted

Below are some of the Kinds of data that was extracted:

- Unstructured Data Extraction
- Semi-Structured Data Extraction
- Structured Data Extraction
- Extraction with Textract Queries
- Signature Detection
- Invoices and Receipts Extraction
- Identity Documents Extraction

Cloudboosta

EXTRACTING TABLE FROM A STRUCTURED DATA



01-idp-document-classificatio | 02-idp-document-extraction.ij | 03-i

Code | git | Cluster

4190 MARYLYNN CAUSEWAY, HEATHCOTEFURT, ID 23119

Your consolidated statement

For 03/02/2022

Contact us

example.com

(858) LLL-0101 or (858) 555-0101

Do more with digital banking

Bank without having to leave home. Check your account balances, make transfers, pay bills and deposit checks with your mobile device. If you are not enrolled in digital banking, it only takes a minute. Get started today at example.com/U.

Example Bank, Member FDIC. To learn more, visit example.com/ABCXYZ. ©2020 AnyCompany Financial Group.

If you are traveling outside of the USA and have concerns about accessing your account while you are traveling, please contact your Branch Banker or call us at 858-LLL-0101.

Summary of your accounts

ACCOUNT NAME	ACCOUNT NUMBER	BALANCE ($)	DETAILS ON
CHECKING	005278099679	10,137.64	page 1
Total checking and money market savings accounts		$10,137.64	
SAVINGS	527809967936	14,500.11	page 3
Total savings accounts		$14,500.11	

Checking and money market savings accounts

■ CHECKING 005278099679

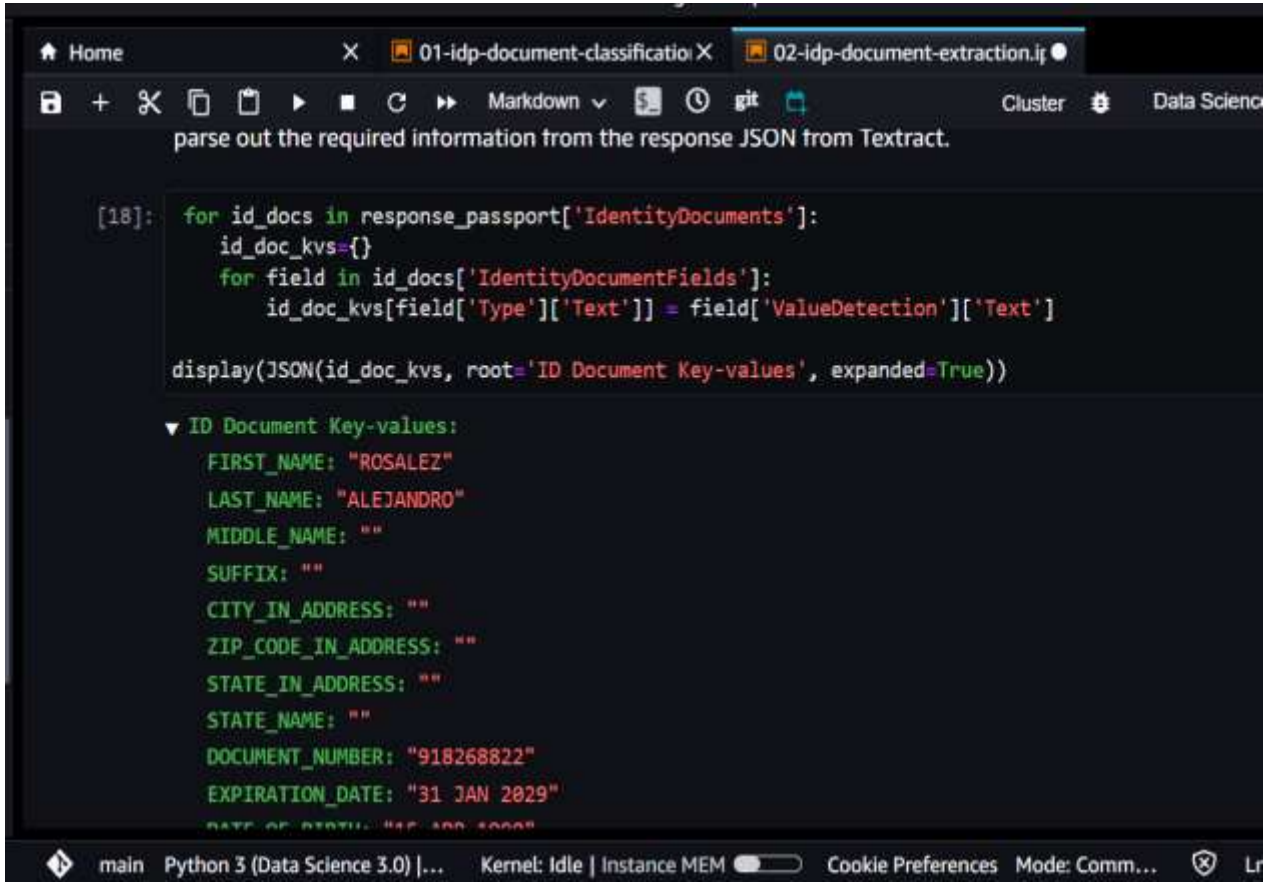
Account summary

Your previous balance as of 03/02/2022 $5,532.92

	0	1	2	3
0	ACCOUNT NAME	ACCOUNT NUMBER	BALANCE ($)	DETAILS ON
1	CHECKING	005278099679	10,137.64	page 1
2	Total checking and money market	savings accounts	$10,137.64	
3	SAVINGS	527809967936	14,500.11	page 3
4	Total savings accounts		$14,500.11	

Cloudboosta

IDENTITY DOCUMENT EXTRACTION



Cloudboosta

EXTRACTION WITH TEXTRACT QUERIES.

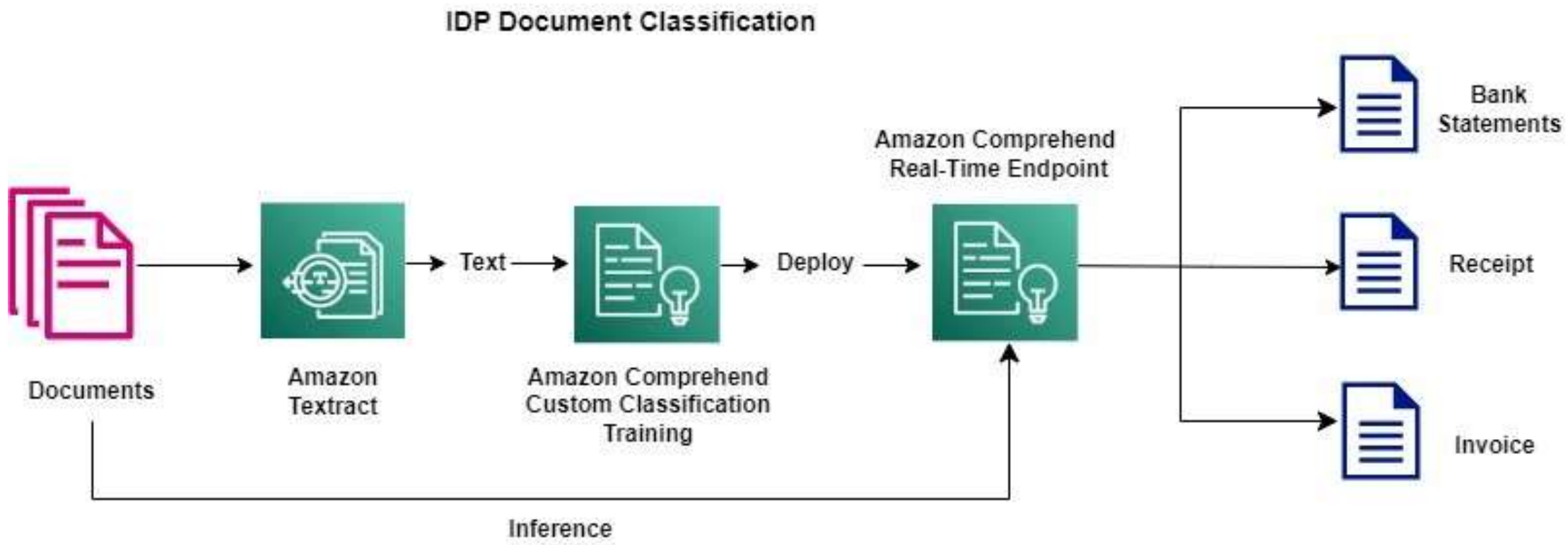
```
[23]: # Main code for execution
      # -----
      job_id = start_analyze_job(s3_bucket, object_key)

      # Monitor the job status
      print("Started analyze job with id: {}, document is: {}".format(job_id, object_key))
      if(is_job_complete(job_id)):
          ssn_response = get_job_results(job_id)

      # Print the result
      print_result_in_document(ssn_response)
```

```
Started analyze job with id: 72549deebf1601365bfadaa63f6e942953e668b348a692fe6f2d95a0a36f8c8e, document is: sample-files/S
SN_John_Doe.jpg
Job status: IN_PROGRESS
Job status: IN_PROGRESS
Job status: SUCCEEDED
Result page recieved: 1
|-------------------------|----------------|-------------|
| What is the name on SSN? | SSN_OWNER_NAME | JOHN DOE    |
| What is SSN number?      | SSN_NUMBER     | 123-45-6789 |
```

Cloudboosta

1. IDP Architecture
2. Getting Started
3. Level 1:Document Extraction
4. Level 2: Document Classification
5. Level 3: Document Enrichment
6. Level 4: Document Review And Verification (A2I)

DOCUMENT CLASSIFICATION ARCHITECTURE



Cloudboosta

DOCUMENT CLASSIFICATION

- Prepare a CSV training dataset for Amazon Comprehend custom classifier training.
- This data is written into a csv file and uploaded into an s3 bucket and be used as a training data.
- We trained a custom classifier using Amazon Comprehend's Custom Classification feature and the labeled data CSV file we created.
- We used Amazon Comprehend's custom classification model to classify sample documents asynchronously using the start_document_classification_job API.
- We specified DocumentReadAction and used Amazon Textract's DETECT_DOCUMENT_TEXT option. This enabled Amazon Comprehend to automatically extract text and classify it.
- The next step is to use the Amazon Comprehend real-time endpoint to classify these documents.

Cloudboosta

DOCUMENT CLASSIFIED USING AMAZON COMPREHEND CUSTOM CLASSIFICATION MODEL

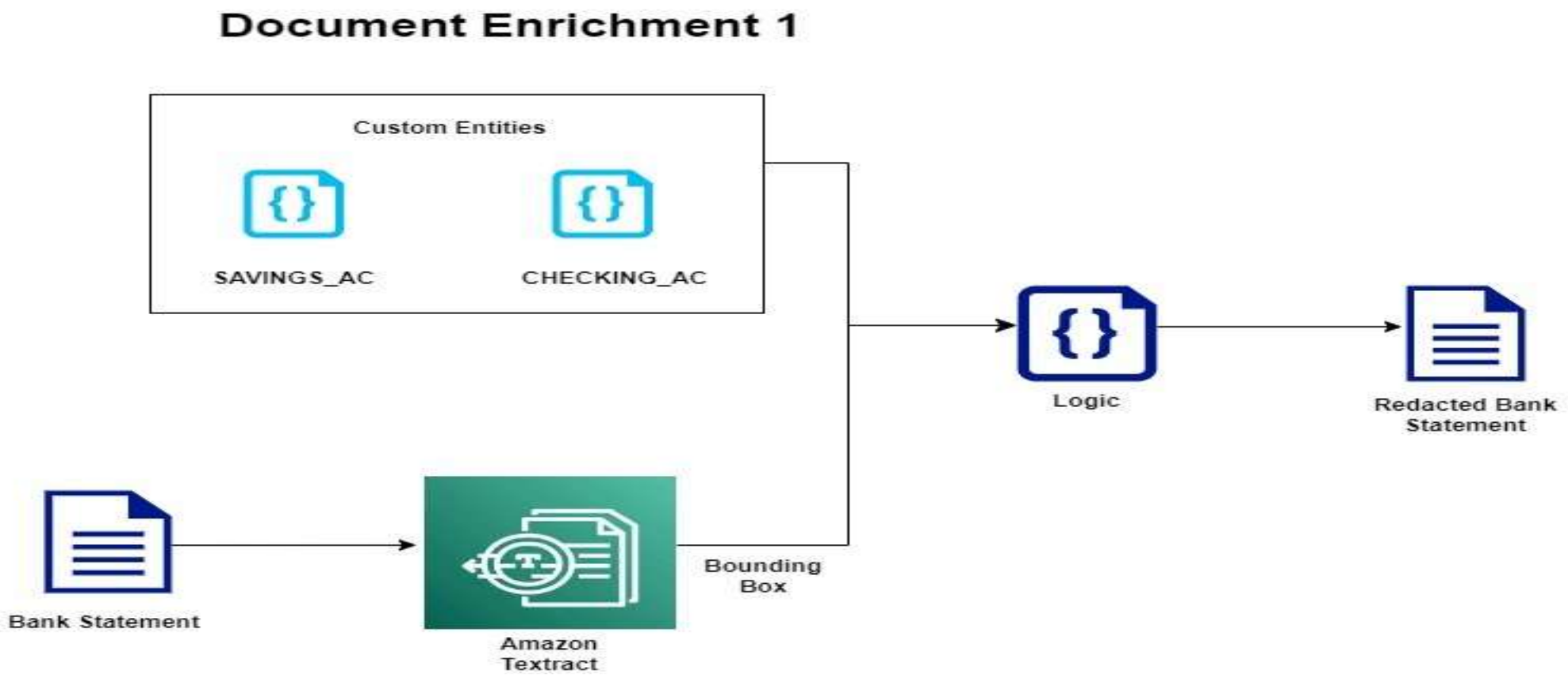
[27]:

	Document	DocType	Confidence	s3path	DocText
0	document_9.png	receipts	0.9999	idp/comprehend/classified-docs/receipts/docume...	THE AIML StORE\n1234 SOMEWHERE RD\nPOWAY, CALI...
1	document_1.png	bank-statements	1.0000	idp/comprehend/classified-docs/bank-statements...	Page 1 of 5 03/02/2022\nDC 1090001004290\nAnyC...
2	document_10.png	receipts	1.0000	idp/comprehend/classified-docs/receipts/docume...	THE AIML StORE\n1234 SOMEWHERE RD\nPOWAY, CALI...
3	document_5.png	invoices	0.9999	idp/comprehend/classified-docs/invoices/docume...	INVOICE\nAnyCompany Hardwares LLC\nDATE\nMay 2...
4	document_3.png	bank-statements	1.0000	idp/comprehend/classified-docs/bank-statements...	Page 1 of 5 03/02/2022\nDC 1090001004290\nAnyC...
5	document_7.png	invoices	0.9999	idp/comprehend/classified-docs/invoices/docume...	INVOICE\nAnyCompany Hardware\nDATE\nDec 09, 20...
6	document_2.png	bank-statements	1.0000	idp/comprehend/classified-docs/bank-statements...	Page 1 of 5 03/02/2022\nDC 1090001004290\nAnyC...
7	document_6.png	invoices	1.0000	idp/comprehend/classified-docs/invoices/docume...	INVOICE\nAnyCompany Manufacturing\nDATE\nDec 2...
8	document_4.png	bank-statements	1.0000	idp/comprehend/classified-docs/bank-statements...	Page 1 of 5 03/02/2022\nDC 1090001004290\nAnyC...
9	document_0.png	bank-statements	1.0000	idp/comprehend/classified-docs/bank-statements...	Page 1 of 5 03/02/2022\nDC 1090001004290\nAnyC...

Cloudboosta

1. IDP Architecture
2. Getting Started
3. Level 1:Document Extraction
4. Level 2: Document Classification
5. Level 3: Document Enrichment
6. Level 4: Document Review And Verification (A2I)

IDP DOCUMENT ENRICHMENT ARCHITECTURE



Cloudboosta

DOCUMENT ENRICHMENT

- We picked a bank statement from our list of documents, then got the S3 location of the document and then perform the actions below:
- Used Amazon Textract to get the geometry information i.e. the bounding boxes, of all the lines in the document
- Used the extracted text above to identify the entities CHECKING_AC and SAVINGS_AC, using Comprehend custom entity recognizer
- Found the bounding box for the CHECKING_AC and SAVINGS_AC words from the Textract response
- Use the bounding box geometry to annotate the document and redact the customer name and address.

Cloudboosta

DOCUMENT ENRICHMENT

Home | 01-idp-document-classificatio | 02-idp-document-extraction.ip | 03-idp-document-enrichment.

Markdown | git | Cluster | Data Science 3.0 | Python 3 | 2 vCPU + 4 GiB

[23]:

Unredacted Document

Page 1 of 5 03/02/2022
DC 1090001004290

AnyCompany Financial Group

999-99-99-99 16789 3 C 001 11 S 66 002
PAULO SANTOS
4190 MARYLYNN CAUSEWAY, HEATHCOTEFURT, ID 2 3 1 1 9

Your consolidated statement

For 03/02/2022

Contact us
example.com
(858) LLL-0101 or (858) 555-0101

Do more with digital banking

Bank without having to leave home. Check your account balances, make transfers, pay bills and deposit checks with your mobile device. If you are not enrolled in digital banking, it only takes a minute Get started today at example.com/U.

Example Bank, Member FDIC. To learn more, visit example.com/ABCXYZ. ©2020 AnyCompany Financial Group.

If you are traveling outside of the USA and have concerns about accessing your account while you are traveling, please contact your Branch Banker or call us at 858-LLL-0101.

Summary of your accounts

ACCOUNT NAME	ACCOUNT NUMBER	BALANCE ($)	DETAILS ON
CHECKING	005278099679	10,137.64	page 1
Total checking and money market savings accounts		$10,137.64	
SAVINGS	527809967936	14,500.11	page 3
Total savings accounts		$14,500.11	

Checking and money market savings accounts

■ CHECKING 005278099679

Redacted Document

Page 1 of 5 03/02/2022
DC 1090001004290

AnyCompany Financial Group

999-99-99-99 16789 3 C 001 11 S 66 002
PAULO SANTOS
4190 MARYLYNN CAUSEWAY, HEATHCOTEFURT, ID 2 3 1 1 9

Your consolidated statement

For 03/02/2022

Contact us
example.com
(858) LLL-0101 or (858) 555-0101

Do more with digital banking

Bank without having to leave home. Check your account balances, make transfers, pay bills and deposit checks with your mobile device. If you are not enrolled in digital banking, it only takes a minute Get started today at example.com/U.

Example Bank, Member FDIC. To learn more, visit example.com/ABCXYZ. ©2020 AnyCompany Financial Group.

If you are traveling outside of the USA and have concerns about accessing your account while you are traveling, please contact your Branch Banker or call us at 858-LLL-0101.

Summary of your accounts

ACCOUNT NAME	ACCOUNT NUMBER	BALANCE ($)	DETAILS ON
CHECKING	[redacted]	10,137.64	page 1
Total checking and money market savings accounts		$10,137.64	
SAVINGS	[redacted]	14,500.11	page 3
Total savings accounts		$14,500.11	

Checking and money market savings accounts

■ CHECKING [redacted]

Cloudboosta

The Customer's Name And Address Has Been Redacted

DOCUMENT REVIEW AND VERIFICATION (AMAZON AUGUMENTED AI)

To incorporate Amazon A2I into human review workflows, we needed the following resources:

- Worker task template
- Human review workflow
- Human loop

Cloudboosta

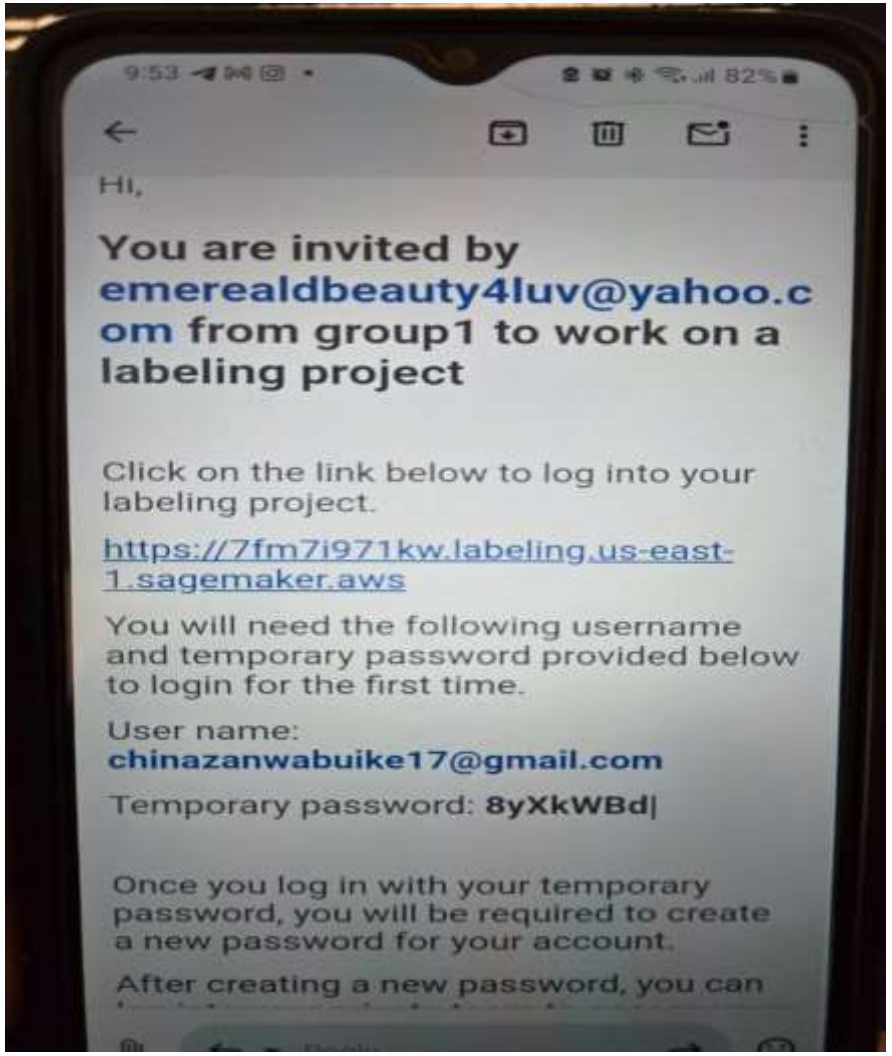
DOCUMENT REVIEW AND VERIFICATION (AMAZON AUGUMENTED AI)

- We used an amazon s3 bucket to store data for A2I workers.

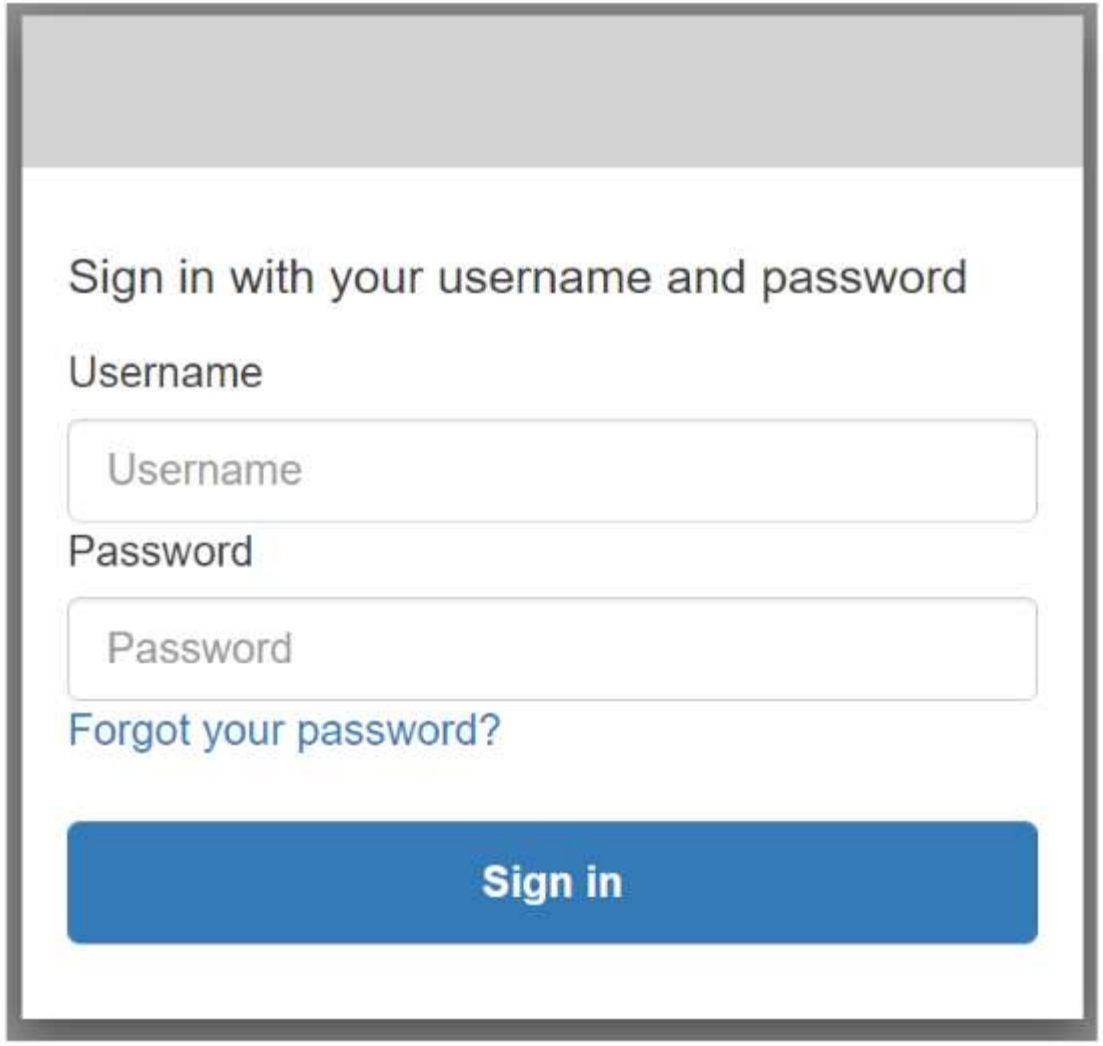
- We created a human review workflow via Augmented AIG which is found in the left panel of the amazon Sagemaker console.

- We then setup the A2I WorkFlow definition, while calling Amazon Textract's Analyze Document API including the A2I paramters in the HumanLoopConfig, and Provided the A2I workflow ARN to be used by Amazon Textract.

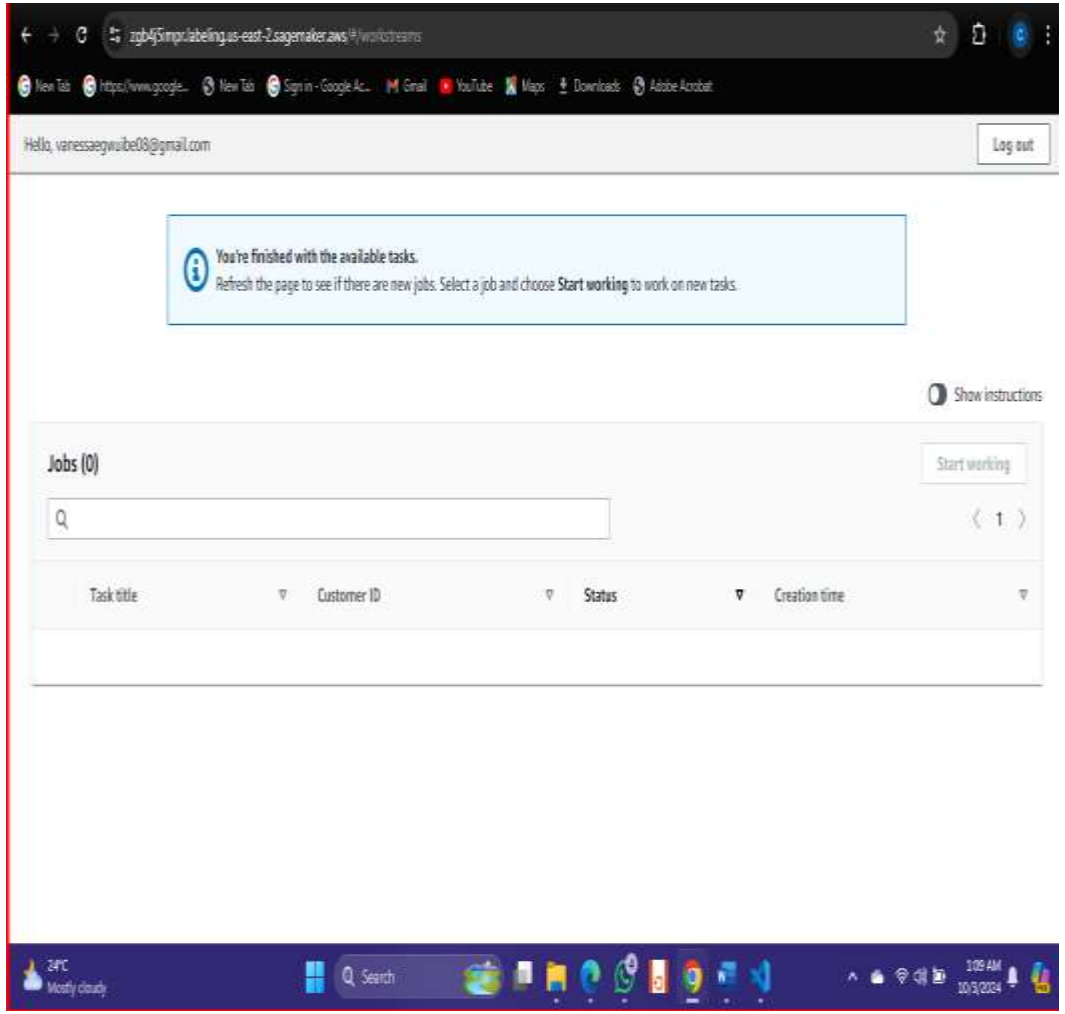
- We logged into the labelling/human review portal after we had received an email with a link to the Labeling/human review portal with details on how to login and a portal URL



DOCUMENT REVIEW AND VERIFICATION (AMAZON AUGUMENTED AI)



LOG IN INTO HUMAN REVIEW PORTAL



HUMAN REVIEW AFTER TASK IS DONE.

DATA VISUALIZATION

- The data was then visualized using QUICKSIGHT, which is one of the services offered by AWS as a visualization tool.

- In this project, we used Pie Chart, and Line Chart for visualization.

01

Solution Introduction

02

Implementation

03

Challenges & Recommendations

CHALLENGES

Below are some of the Challenges we faced while deploying the project:

- Outdated Codes
- Ran out of Finance
- Using same region by multiple users
- Time Constraints
- Unfamiliar Territory

Cloudboosta

THANK YOU