

GROUP A PROJECT - INTELLIGENT DOCUMENT PROCESSING WITH AWS AI SERVICES.

MEMBERS

JIDE OLOKO

CHINAZAEKPERE OBIMDIKE

ADEMIDE OLANREWAJU

CHIOMA EGWUIBE

Contents

- ❖ Introduction
- ❖ AWS services used
- ❖ System Architecture - Architecture Diagram
- ❖ Data Governance- Security and Policies Challenges
- ❖ Implementation
- ❖ Challenges

INTRODUCTION

Imagine a world where your documents could understand themselves, extracting key information and automating tasks that once required hours of manual labor. This is the power of Intelligent Document Processing (IDP).

IDP harnesses the capabilities of artificial intelligence and machine learning to analyze and understand unstructured or semi-structured documents. By automating the process of extracting data and information, IDP can significantly improve efficiency, reduce errors, and provide valuable insights.

Key benefits of IDP include:

- **Enhanced Productivity:** Streamline operations and reduce manual labor.
- **Improved Accuracy:** Minimize errors and ensure data quality.
- **Faster Decision-Making:** Gain valuable insights from your documents to make informed decisions.
- **Cost Savings:** Reduce operational costs associated with manual data processing.

IDP applications span various industries, including but not limited to:

- **Financial Services:** Processing invoices, contracts, and insurance claims.
- **Healthcare: Analyzing:** medical records, prescriptions, and insurance claims.
- **Legal:** Extracting information from legal documents like contracts and court filings.
- **Human Resources:** Processing resumes, job applications, and employee records.

- **Customer Service:** Automating customer support tasks.

By embracing IDP, businesses can unlock the potential of their documents and drive innovation.

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

AWS SERVICES USED:

AMAZON Textract: Amazon Textract is a machine learning service that uses optical character recognition (OCR) to automatically extract text, handwriting, and data from scanned documents, forms, and tables. It's a powerful tool for automating document-intensive processes and extracting valuable information from unstructured data.

AMAZON COMPREHEND: Amazon Comprehend is an NLP service that extracts information from unstructured text. It identifies entities, sentiment, and topics, and can redact sensitive data. It's fully managed and can process millions of documents quickly.

AMAZON AUGMENTED AI (A2I): Amazon Augmented AI (A2I) streamlines the process of adding human review to IDP workflows. It integrates with Textract and Comprehend, automating tasks like document preparation and review assignment. This frees up developers to focus on building the core logic of their applications.

AMAZON SAGEMAKER: A fully managed machine learning platform that provides developers with all the tools they need to build, train, deploy, and manage machine learning models. It simplifies the process of machine learning, making it accessible to a wider range of developers.

AMAZON S3 BUCKET: An S3 bucket is a storage unit in Amazon Simple Storage Service (S3) that can store and retrieve any amount of data, from any place, at any time. It's designed for storing and retrieving any amount of data, from any place, at any time.

AMAZON IAM: is a fundamental service in AWS that controls who can access and what they can do within your AWS account. It's like the gatekeeper of your AWS environment, ensuring that only authorized users and resources have the necessary permissions.

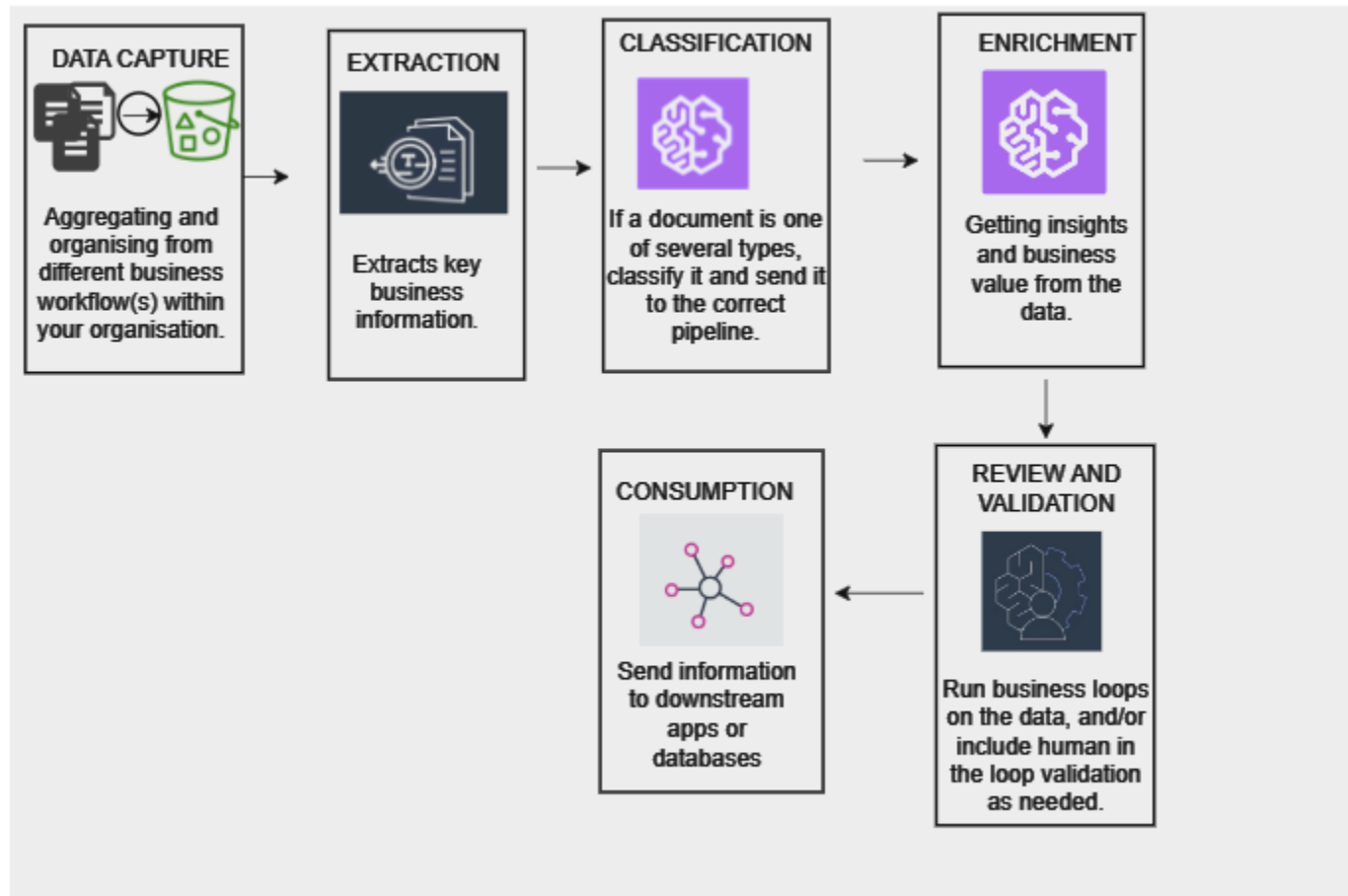
CLOUDFORMATION: A powerful tool that lets you define and manage your AWS infrastructure using a simple declarative language. It automates the provisioning, updating, and deletion of resources, making it easier to deploy and manage your applications on AWS.

AMAZON QUICK SIGHT: A cloud-based business intelligence (BI) service from AWS that empowers businesses to make data-driven decisions. It offers a user-friendly interface for creating interactive dashboards, analyzing data, and uncovering valuable insights.

SYSTEM ARCHITECTURE

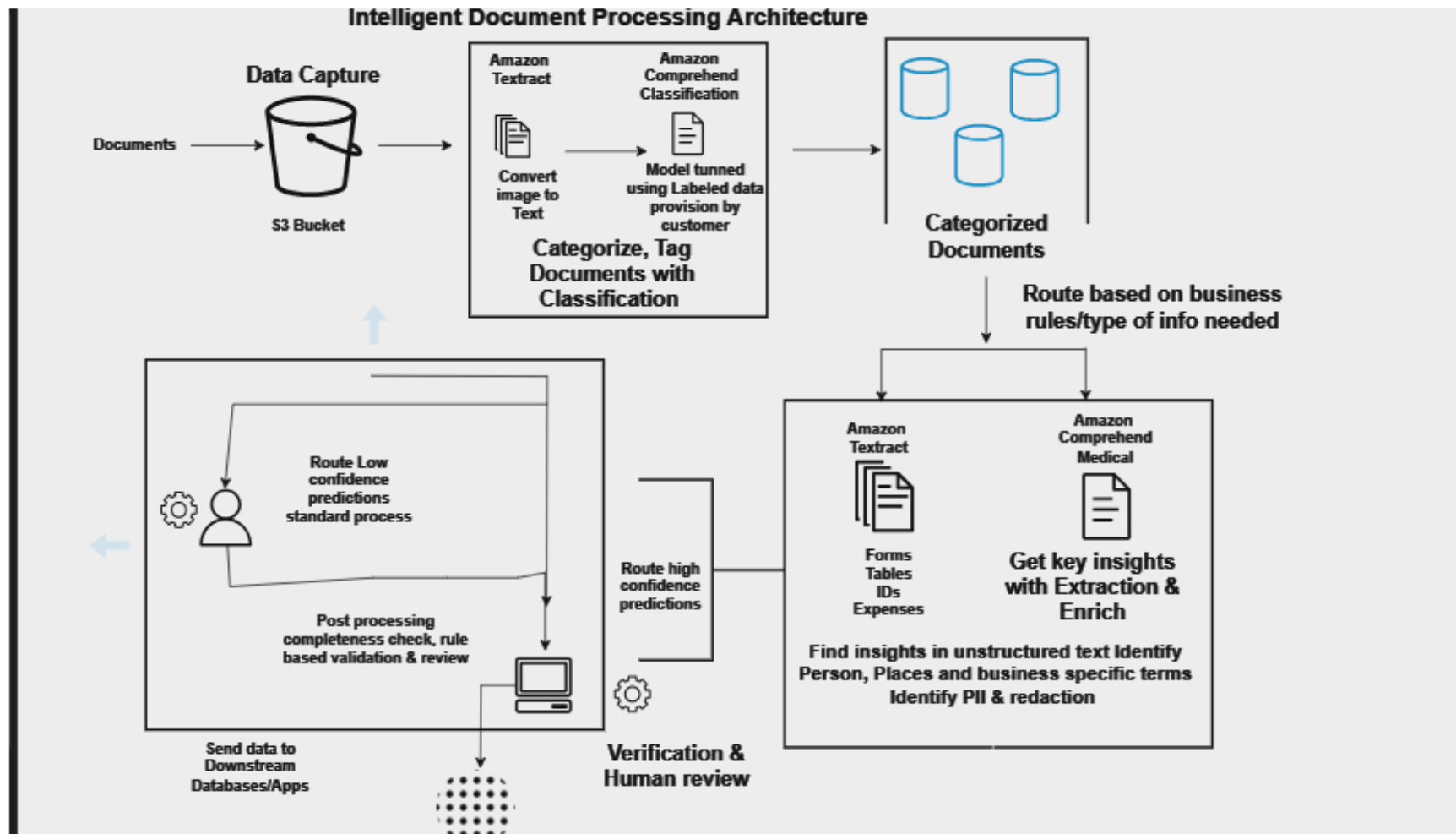
The architecture diagram below shows the stages of an Intelligent Document Processing workflow. It starts with a data capture stage to securely store, aggregate different types (pdf, jpeg, png, tiff), formats, and layouts of documents. The next stage is classification, this is where you categorize your documents (for example categories such as contracts, claim forms, invoices, receipts and so on) followed by document extraction. In the extraction stage, you can extract meaningful business information from your documents. This extracted data is often used to gather insights via data analysis or sent to downstream systems such as databases or transactional systems. The following stage is enrichment, at this stage documents can be enriched by redacting PII data, custom business term extraction, and so on. Finally, in the review/verification stage you can include a human workforce for document reviews to ensure the outcome is accurate

INTELLIGENT DOCUMENT PROCESSING WORKFLOW



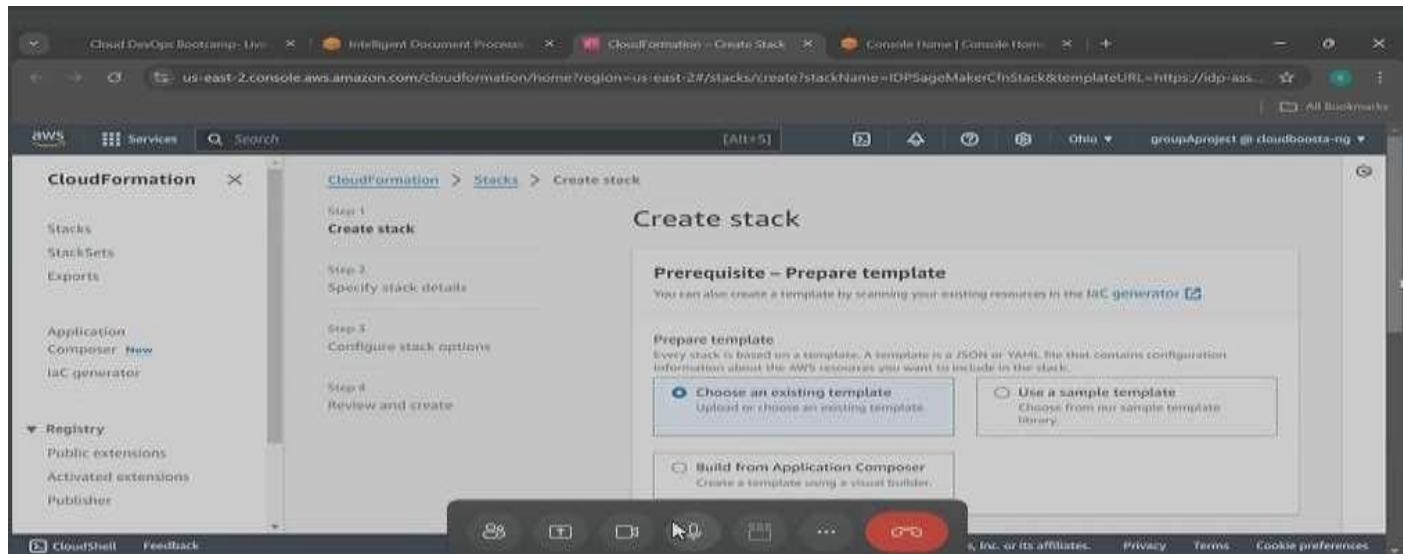
Data Governance- Security and Policies

- **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 of third-party services.

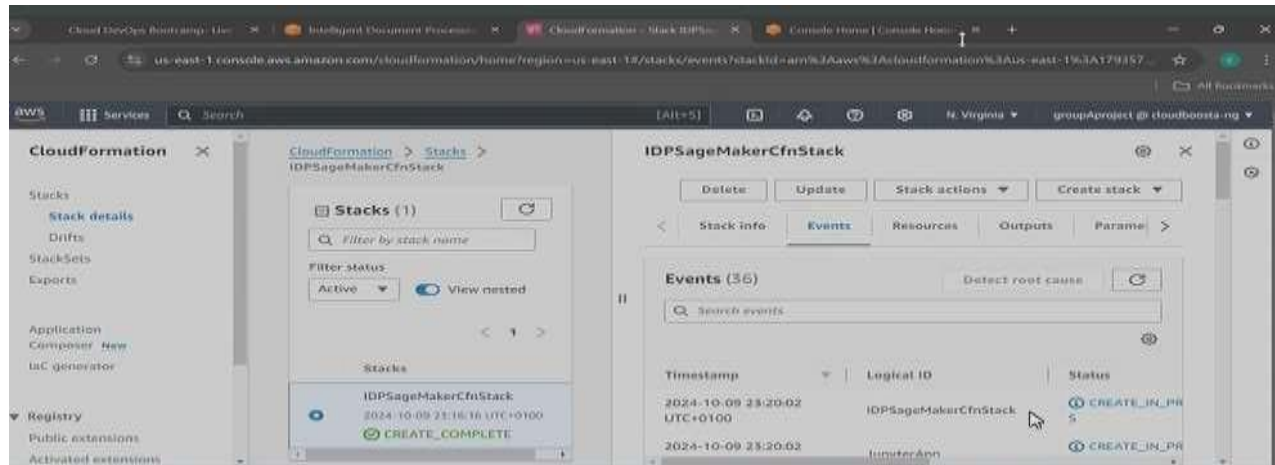


GETTING STARTED

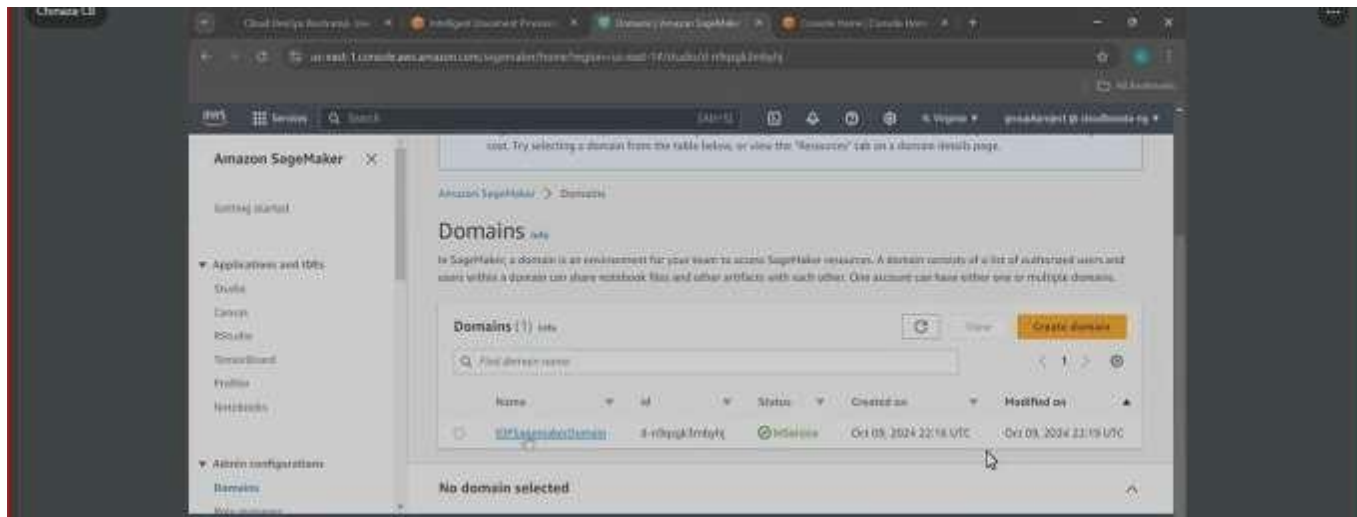
1. Creation of stack via CloudFormation.



2. Ensure that the stack is created successfully.



3. On Amazon SageMaker, Domain is created automatically



4. The studio is launched via the domain, then we can access our notebooks to run the script.



Document Extraction

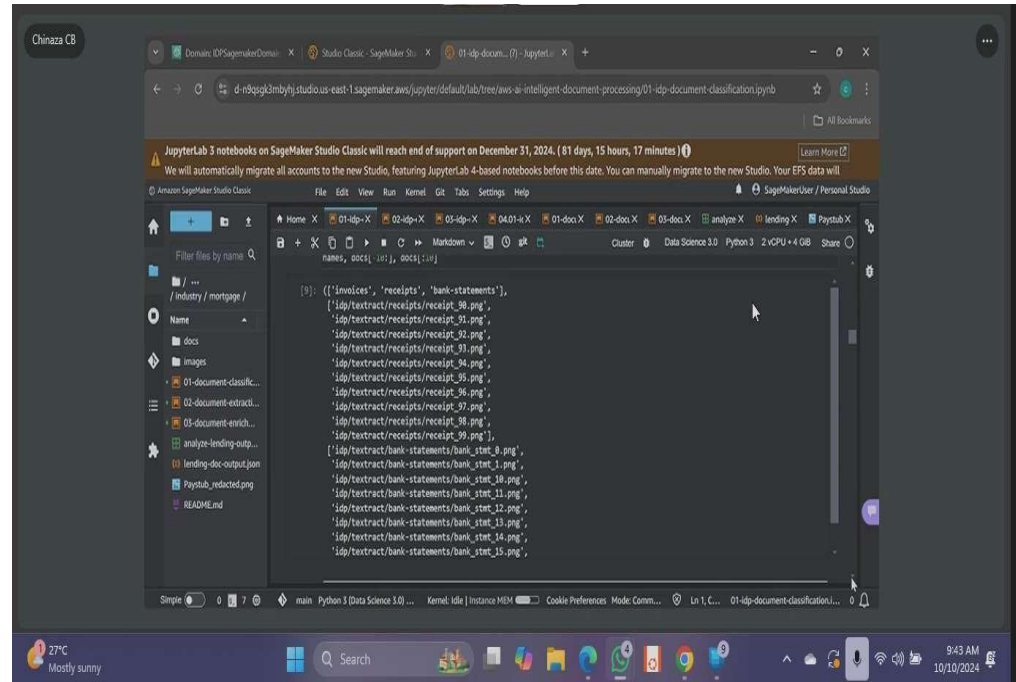
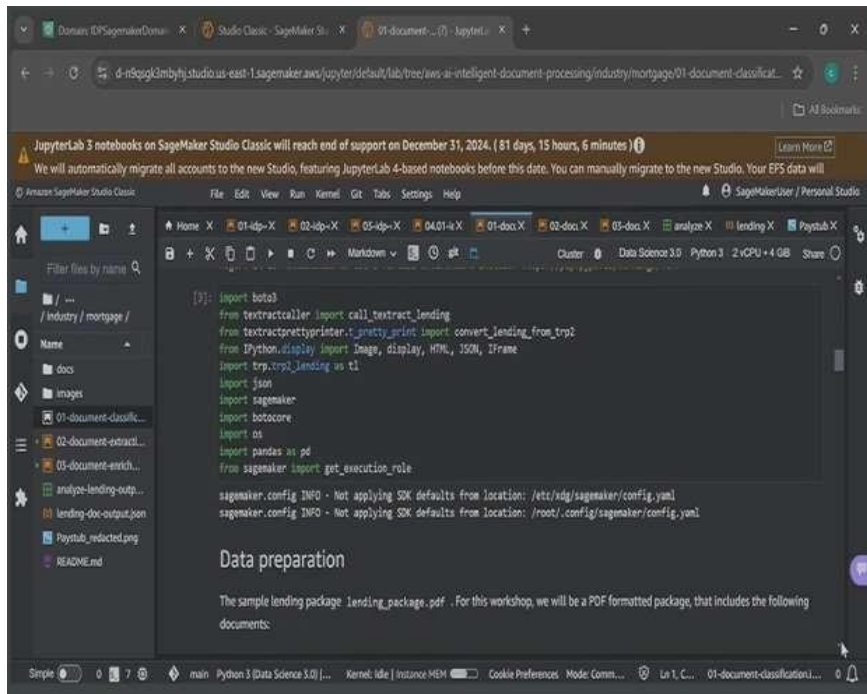


DOCUMENT EXTRACTION ARCHITECTURE.

DOCUMENT EXTRACTION

- Setup Notebook By Deploying Cloud formation 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 bucket.

CLOOBOOSTA ACADEMY



DOCUMENTS EXTRACTED VIA AMAZON TEXTTRACT'S "DETECT_DOCUMENT_TEXT" API

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

CLOOBOOSTA ACADEMY

EXTRACTING TABLE FROM A STRUCTURED DATA

4190 MARYLYNN CAUSEWAY, HEATHCOTE, NJ 07931

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/ID.

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 account 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	

IDENTITY DOCUMENT EXTRACTION



Home X 01-idp-document-classification X 02-idp-document-extraction.ip Cluster Data Science

parse out the required information from the response JSON from Textract.

```
[18]: for id_docs in response_passport['IdentityDocuments']:
      id_doc_kvs={}
      for field in id_docs['IdentityDocumentFields']:
          id_doc_kvs[field['Type']]['Text'] = field['ValueDetection']['Text']

      display(JSON(id_doc_kvs, root='ID Document Key-values', expanded=True))
```

▼ ID Document Key-values:

```
FIRST_NAME: "ROSALEZ"
LAST_NAME: "ALEJANDRO"
MIDDLE_NAME: ""
SUFFIX: ""
CITY_IN_ADDRESS: ""
ZIP_CODE_IN_ADDRESS: ""
STATE_IN_ADDRESS: ""
STATE_NAME: ""
DOCUMENT_NUMBER: "918268822"
EXPIRATION_DATE: "31 JAN 2029"
DATE_OF_BIRTH: "15 JUN 1988"
```

main Python 3 (Data Science 3.0) | ... Kernel: Idle | Instance MEM Cookie Preferences Mode: Comm... Lr

EXTRACTION WITH TEXTTRACT 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/SN_John_Doe.jpg

Job status: IN_PROGRESS

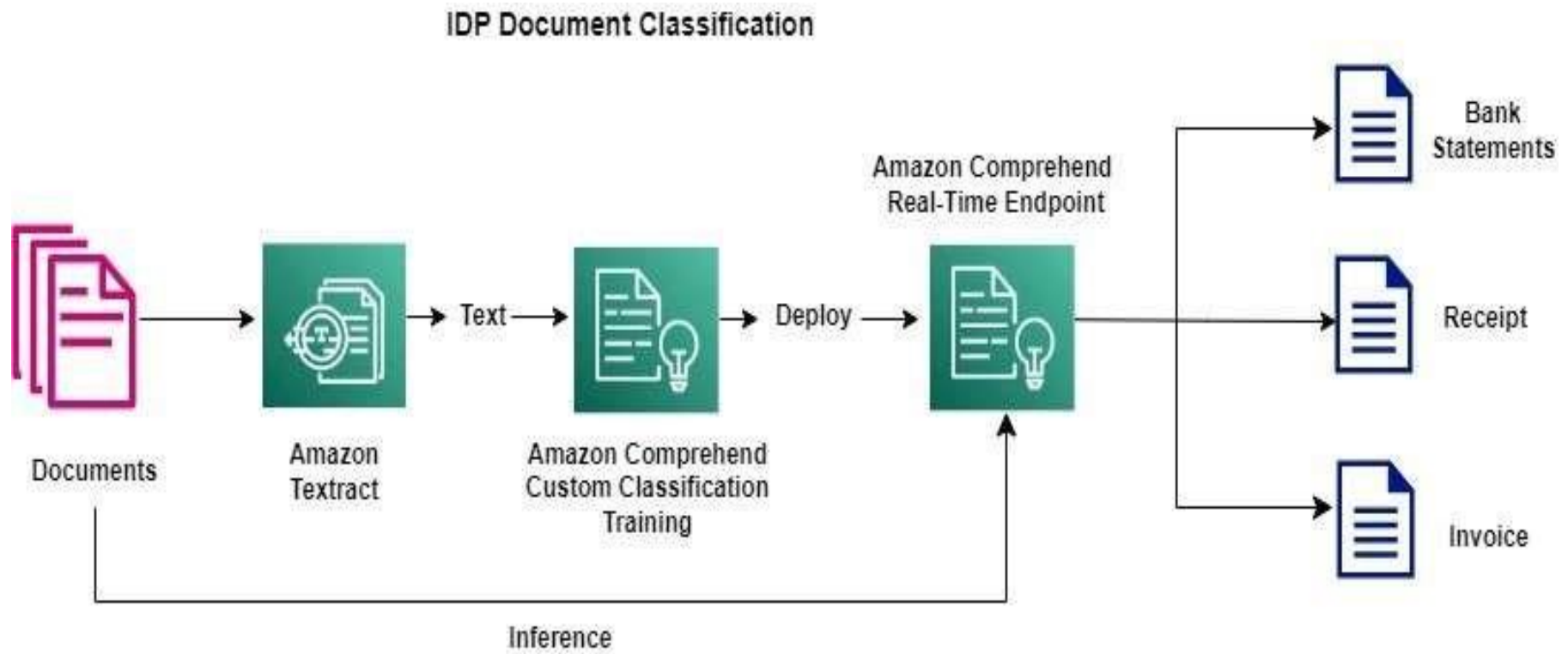
Job status: IN_PROGRESS

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

DOCUMENT CLASSIFICATION ARCHITECTURE



DOCUMENT CLASSIFICATION

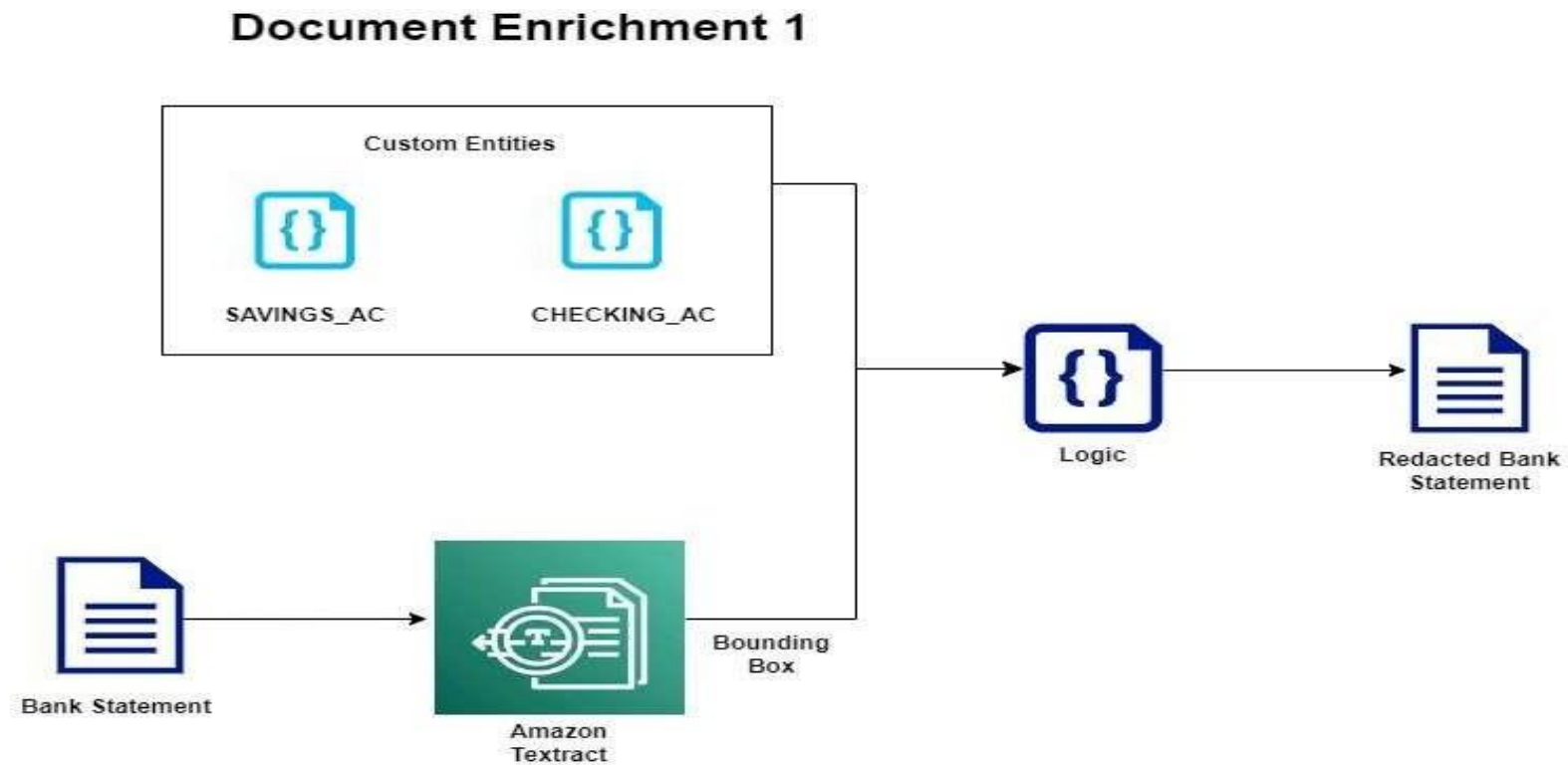
- Prepare a CSV training dataset for AmazonComprehend custom classifier training.
- This data is written into a csv file and uploaded into an s3 bucket and be used as training data.
- We trained a custom classifier using Amazon Comprehend's Custom Classification feature and the labeled data CSV file 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.

	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...

DOCUMENT CLASSIFIED USING AMAZON COMPREHEND CUSTOM CLASSIFICATION MODEL

IDP DOCUMENT ENRICHMENT ARCHITECTURE



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.

CLOOBOOSTACADEMY

- 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.
- Used the bounding box geometry to annotate the document and redact the customer's name and address.

The screenshot displays a web application interface for document processing. At the top, there are three tabs: '01-idp-document-classification', '02-idp-document-extraction', and '03-idp-document-enrichment'. Below the tabs, the application shows two side-by-side document previews. The left preview, titled 'Unredacted Document', shows a bank statement for Paulo Santos with his name and address visible. The right preview, titled 'Redacted Document', shows the same statement but with the customer's name and address redacted with black boxes. Both documents include a 'Summary of your accounts' table.

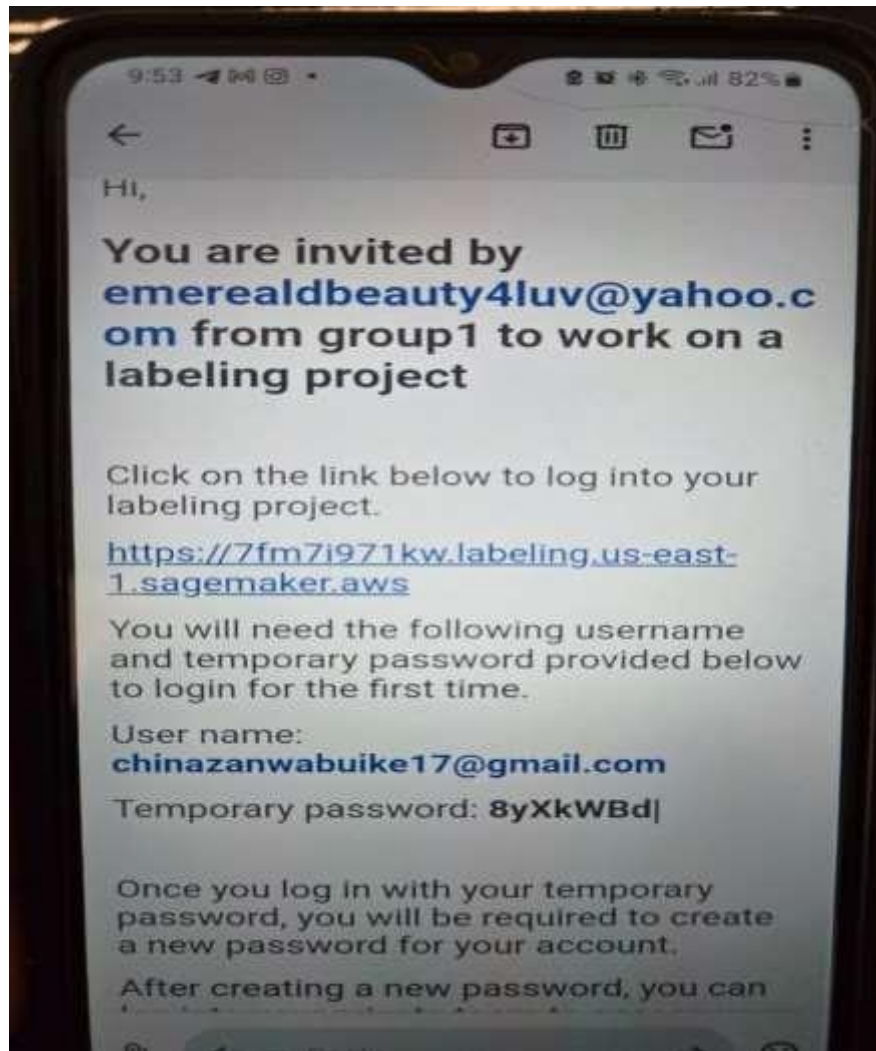
ACCOUNT NAME	ACCOUNT NUMBER	BALANCE (\$)	DETAILS ON
CHECKING	005278099679	10,137.64	page 1
Total checking and money market savings accounts		\$10,137.64	
SAVINGS	527809967936	14,500.11	page 3
Total savings accounts		\$14,500.11	

The Customer's Name and Address Has Been Redacted.

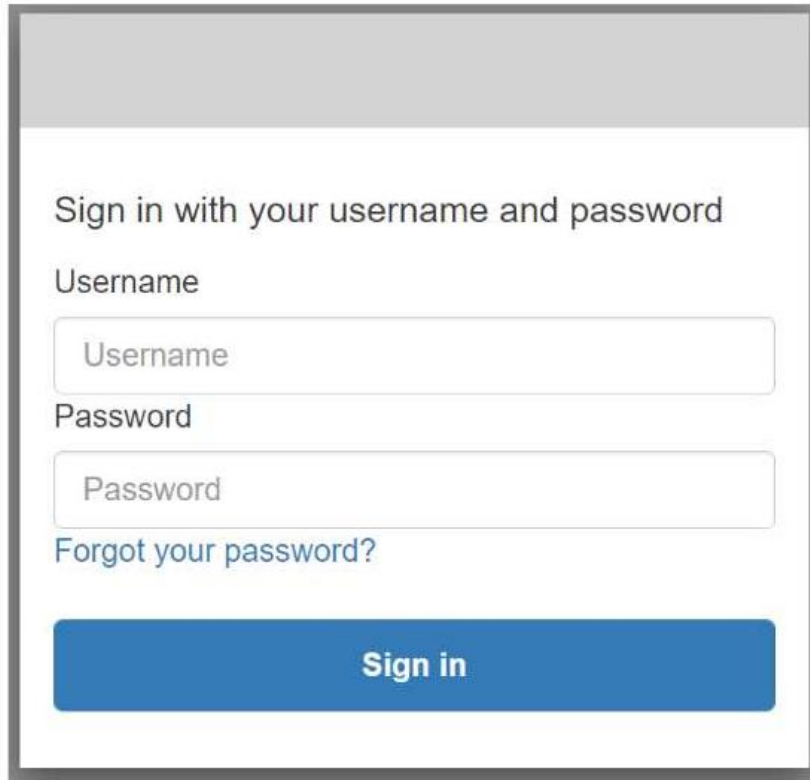
DOCUMENT REVIEW AND VERIFICATION (AMAZON AUGUMENTED AI)

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

- Worker task template
 - Human review workflow
 - Human loop
-
- 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 set up 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



An Email invite to login into the human review portal.



Sign in with your username and password

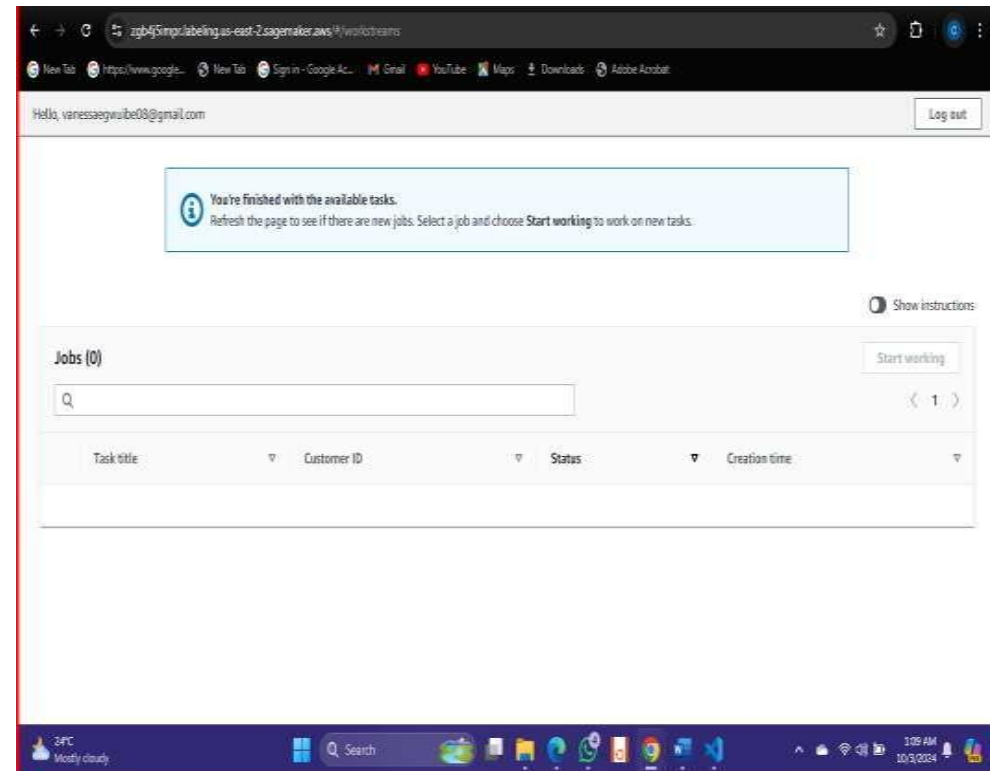
Username

Password

[Forgot your password?](#)

Sign in

LOG IN INTO HUMAN REVIEW PORTAL

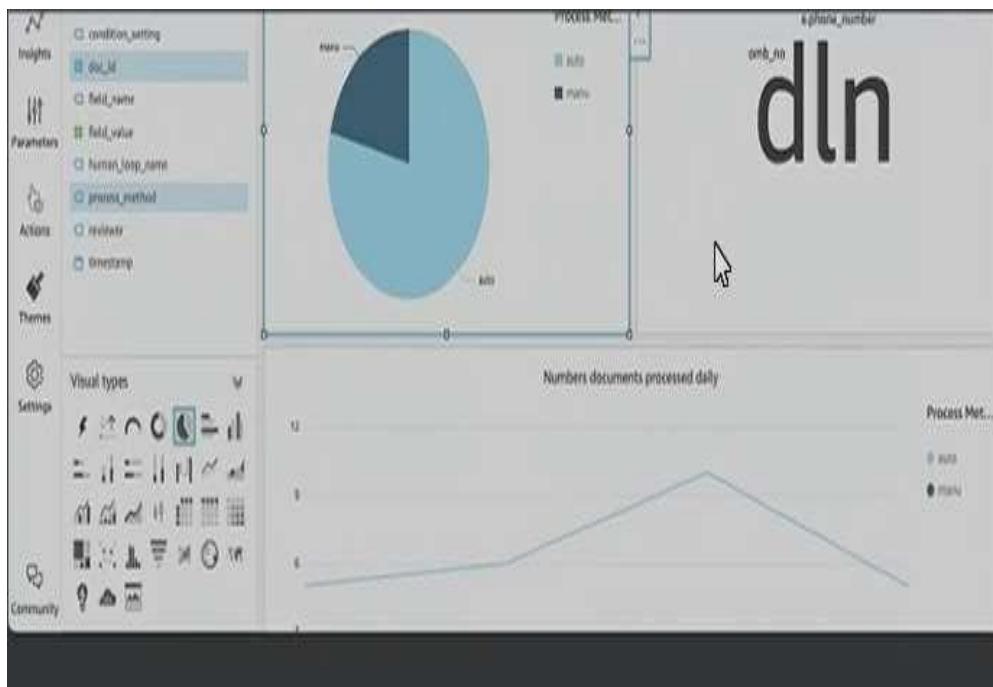


HUMAN REVIEW AFTER TASK IS DONE

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.



CHALLENGES

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

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