PDF Document Translation – Web Based Project
(Project Overview and Progress)

I have been making this project as a mandatory coursework of cloud computing for session 2023-2024 under prof. Simone Merlini and Prof. Nicolo Marchesi.

The overview of the Project is described Below.

As an initial though, what needed to be done to make this project is drawn as an architecture in the below figure.

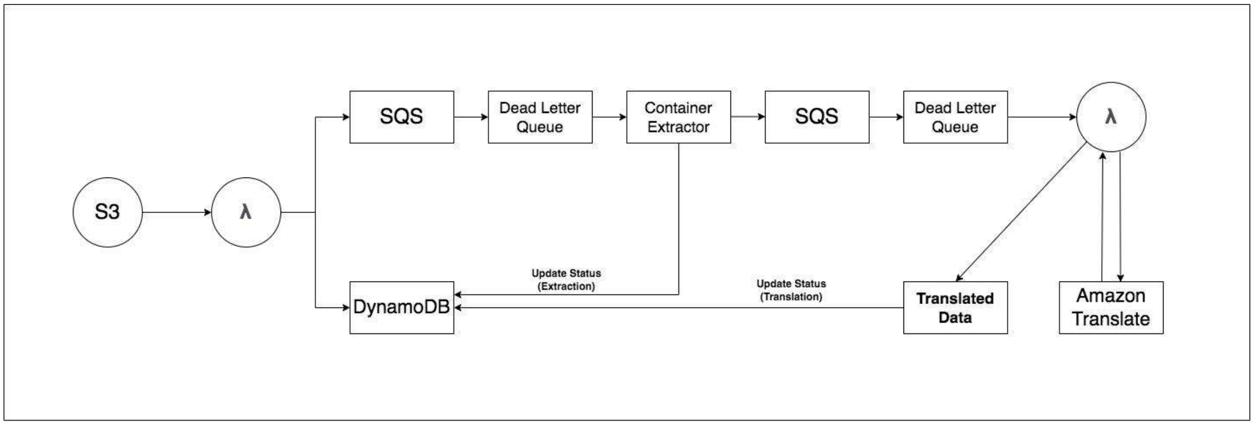


fig.no.1

Description:

This project is designed to provide an automated solution for translating PDF documents from Italian to English using AWS services. The architecture effectively integrates various AWS components to ensure a streamlined and efficient translation process.

The workflow begins with the user uploading an Italian PDF document to an S3 bucket. This action triggers a Lambda function, which extracts the document's text using Amazon Textract. The extracted text is then sent to an Amazon SQS queue to ensure reliable and asynchronous processing.

The text in the SQS queue is picked up by a containerized extraction service that performs further text processing if necessary. Any issues during this stage are handled by a Dead Letter Queue (DLQ) for troubleshooting and retries. The processed text is then placed in another SQS queue, awaiting translation.

The next Lambda function retrieves the text from the SQS queue and invokes Amazon Translate to convert the text from Italian to English. The translated text is then stored in a DynamoDB table to track the status and manage the workflow.

Once the translation is complete, the text is compiled into a new PDF document. The final translated PDF is stored back in an S3 bucket, making it available for the user to download. The system ensures robust monitoring and error handling through the use of DynamoDB for status updates and DLQs for error management.

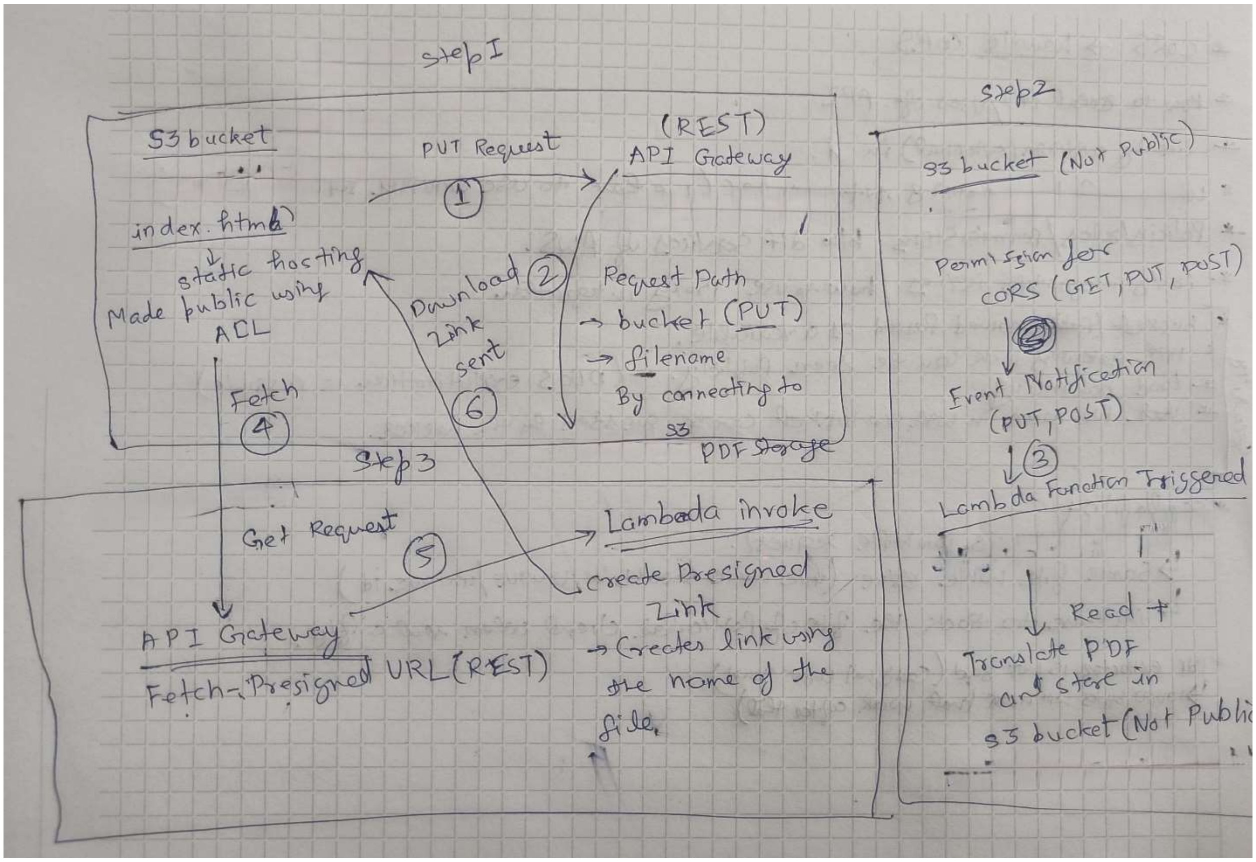
This architecture demonstrates the effective use of AWS cloud services to automate and streamline the document translation process, providing users with a reliable and scalable solution for converting PDF documents between languages.

## Now, while making this project some changes come up in the architecture.

As this is a web-based project, we made a web page where the user can upload the document and can also download it. Hence for this we have to introduce one more S3 bucket to host the website.

Use of API Gateway

Below is the Architecture of current work done:



Description:

1. User Upload (Step 1):

A user uploads an Italian PDF document through a web page hosted on a public S3 bucket. This bucket hosts the static website and is configured to be publicly accessible via an Access Control List (ACL).

The web page makes a PUT request to an API Gateway endpoint.

2. PDF Storage (Step 2):

The API Gateway processes the request and stores the uploaded PDF in a non-public S3 bucket designated for PDF storage.

The S3 bucket is configured with appropriate CORS permissions to handle GET, PUT, and POST requests.

An event notification is triggered by the PUT operation, which invokes a Lambda function.

3. Text Extraction and Translation:

The Lambda function reads the uploaded PDF, extracts the text, and translates it from Italian to English using Amazon Translate.

The translated text is then compiled into a new PDF document.

The new PDF is stored in another non-public S3 bucket dedicated to storing translated PDFs.

4. Download Link Generation:

The Lambda function creates a presigned URL for the translated PDF, ensuring secure access to the document.

The presigned URL is sent back through the API Gateway.

5. User Notification and Download (Step 3):

The user receives the presigned URL and can send a GET request to this URL via the web interface.

The translated PDF is fetched from the S3 bucket using the presigned URL.

The web page displays a download link, allowing the user to download the translated PDF document.

This architecture effectively uses AWS services such as S3, Lambda, API Gateway, and Amazon Translate to automate and streamline the document translation process. The integration of a user-friendly web interface ensures a seamless experience, allowing users to easily upload documents and download the translated versions.

Now as the project is working /doing it original function i.e. translating the PDF document and can download it.

Now, currently working on the internal processes which is described in the initial architecture (fig. no. 1) to make the project Salable, Robust and secure because the idea is to make this Project available for public use.