Module 3 - Data Extraction

Important
Make sure you have performed the steps described in the **Module 1 - GenAI Introduction for IDP** section before beginning this module.

Extracting Data From Documents

Many IDP use cases involve extracting valuable content from scanned documents and writing this data to a database or sending it to a downstream process via API. A few example use cases include processing patient intake forms, processing purchase orders and shipping records, and processing loan applications. For these use cases, it can be time-consuming and error-prone to convert text to SQL or JSON formats for sending to these downstream systems. GenAI has the ability to understand the meaning in text documents and format the output correctly. This reduces the complexity in building applications to process documents, making it easier to create automated workflows that streamline manual processes.

Document Extraction Lab

To get started with the lab, navigate to the gen-ai folder in the file menu, then into the Bedrock folder. Open the notebook titled `03-idp-genai-data-extraction.ipynb`.

Once the kernel starts, follow along with the instructions in the notebook through the end of **Document Extraction**.

Background

Now that you have experienced how to leverage Amazon Textract to do document extraction, we will now demonstrate how to leverage foundation models in order to do data extraction on both structured and unstructured data. To simplify the requests to Amazon Bedrock, we are using a library called Rhubarb which will do all the boilerplate code and data processing. In the lab we will be using alot of system prompts that have been provided by Rhubarb. To view them you can check here.

Conclusion

Before you finish, be sure to visit the cleanup section to remove any resources you no longer need.

Previous Next