AWS TextTract is an OCR tool to accurately extract text from a variety of documents to turn it into ordered/structured data to use. By automatically identifying form labels and values, it is able to extract information from these documents without compromising the structure of the data.

This allows for documents to be processed in hours instead of weeks through manual labor and data entry. Additional features of this tool include smart search indexes, and the ability to work on a variety of types of documents like: health forms, tax forms, invoices, etc.

This is a useful link for setting up and using amazon textract: <https://www.youtube.com/playlist?list=PL5KTLzN85O4LpL7cWsFHFDsKaXtrbBgPi>

Specifically this video (vid 3) and below: <https://www.youtube.com/watch?v=YP2Lk6fqKQM&list=PL5KTLzN85O4LpL7cWsFHFDsKaXtrbBgPi&index=3>

The standard flow is that when a user uploads a file into the amazon s3 buckets (similar to a google drive) , it will auto trigger an AWS lambda function to execute amazon textract.

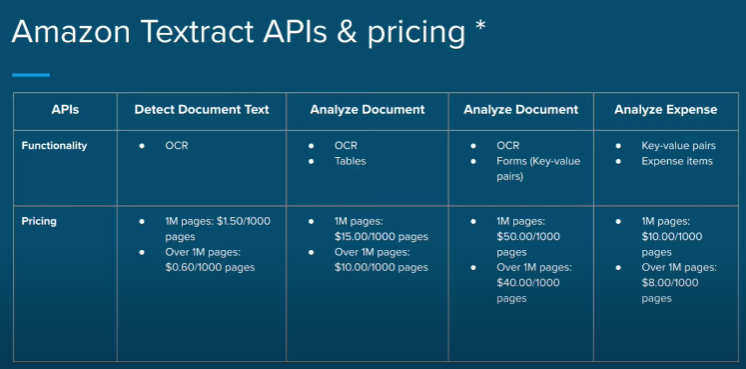
This summarizes the workflow trigger that we will need to setup outside of coding.

The videos show the code in python, the code is under this repository:

<https://github.com/srcecde/aws-tutorial-code/tree/master/textract>

The following types are supported by amazon textract: PNG, JPEG, PDF

They can be submitted as an S3 object or as a byte array. For asynchronous APIs, it has to be through S3 objects.



The reason Amazon Textract is used compared to other options is that our platform is already heavily integrated with AWS services and so including another one of AWS’s services into the workflow will make it easy to integrated and scalable. We will also not be required to implement this from scratch and no machine learning expertise is required.

Textract is also able to extract tabular data, forms, and expense-related documents.

Limitations:

