

Overview:

The project involved parsing a PDF to extract required data and store it in the database for further use. The client is a clinical research organization and supports the medical industry with data research and regulatory submissions. The solution enabled the client to extract the required information: key-value pairs and tabular data and store it in the repository.

Client details:

Name: Confidential | Industry: Transportation & Logistics | Location: USA

Technologies:

Pytesseract, opencv, invoice2data, azure-storage, celery, RabbitMQ

Modules Used:

- **Pytesseract:** This is used in Non-Searchable PDF parsing. A python wrapper around Google's Tesseract performs Optical Character Recognition; i.e. gives a string output for text in images.
- **OpenCV**: This is used in Non-Searchable PDF parsing. It performs Computer Vision operations on an image.
- **invoice2data:** This is used in Searchable PDF parsing. It uses YAML templates to parse the data from searchable PDFs and returns the result as CSV, JSON or XML.
- **Azure-storage**: It is used for storing the PDFs as we needed a structured storage. If the requirement is an unstructured storage then Azure Blob Storage was used.
- **Celery & RabbitMQ:** It is used to handle processing of tens of thousands of PDFs using workers and message queue. Multiple queues with different priorities help to process high priority PDFs as soon as possible.



Project Description:

Earlier, the client used to manually search through previous PDF Documents to filter out its requirements. It lead to cost overruns and was inefficient. After analyzing the client's need, team@Mindfire offered to develop a solution that would help the client to parse a PDF and extract the relevant info. There are two modules available to parse PDFs. Some salient features are:

Searchable PDF:

- Searchable PDFs result from the application of OCR (Optical Character Recognition) to scanned PDFs or other image-based documents.
- The solution parses invoices of different companies using the YAML Template. Python's invoice2data library uses YAML templates (containing keys and the respective regex to map the values to these keys) to extract key-value pairs and uses the JSON data for further operations.

Non- Searchable PDF:

- Non- Searchable PDFs are documents that cannot be searched by text and are created by saving the PDF file as an image.
- HTML 5 Canvas marks the dimensions of the data to extract, and the JSON File stores the individual data dimension.
- For parsing, the PDF is converted to an image. OpenCV reads the image and contour detection to identify the text.
- JSON Template locates the data where images appear, and the OpenCV uses filter operations to operate on the cropped image.
- Tesseract processes the cropped image for string output which, in turn, is stored in a JSON file for further operations.

After the initial launch of the application, the team faced two major challenges:

Bulk Document Parsing



OCR Output Efficiency

- Earlier, the accuracy of the output was 80% as the OCR faced difficulties in interpreting certain letters like B, 3, S, \$, O, O etc.
- The accuracy was improved by the use of image processing before sending the image to OCR Engine. This resulted in excellent results with 99% accuracy.

Scalability

- After the initial launch of the application, the time taken to parse a single PDF page was approximately 3 sec. The client's requirement was about 10,000 pages a day, and the current speed did not suffice.
- The processing speed was improved with the help of multiprocessing. The team implemented this with the help of RabbiMQ, a queue server.
- The RabbiMQ parses the invoices in parallel. The application picks up the attachments from the mail, stores them in the azure server, and adds the task in the queue. The module picks up the pending tasks from the queue and processes them in parallel.



Architecture:





Workflow:



Screenshots:



Screenshot 1: YAML is created for a particular vendor based on the format

Bulk Document Parsing



Template Content: eco_staff_1.yml		Create / Select Templa	ate
sauer too Suf	-	Search	
heids:			
static_amount: 10"		Generic Template Y	
subtotal			
sum_amount_bright_amt:		and staff 1 and	
· W		eco_son_r.ym	1
sun_amount_tax_amt			
· W		eco_staff_2.yml	1
total_invoice_ant:			
dete:			<
invoice_number:			
ship_to_name:			
bil_to_name:			
job_marber:			
keywords			
1			
· ·			
a construction of the second se			
lines:			
start:			
end	*		

Screenshot 2: Searchable Parser: Template. It has the required fields created by users in YAML according to invoice format.

Å	RADESTA Annu Processor Annu Processo		INVO	NCE II 222465 23	in Chrok Provide 3 to the off god by Twe to an arrivation Theory water water Twe State water Twe Water Twe State Office water heat of the State of Theory and State of the State State of the State State of the State State State of the State State State of the State State State of the State State State State State State of the State St	Line (60 Am Los), San Girl And Tanz Hone of 21 Britation research and 20 Am Service San American San American San American San American San American San American San American San American San
BILL TO: Thank you for yo	Electris Link, 3nd. 21.755 2-45 M, Bidg #13 Spring, TX 77388					
CATE	INVOICE NO.	PACE	ACCOUNT NO.	TERMS:		
PERIOD	CONTRACTOR & C	NO. OF EL	1 10000	1424283	BATE	440437
P0/30b Numbe Week Worked	91 07/38/2021					
Bell, Clube Lee Januar Jan				48.00	25.60	\$1.064
Carbin, Nathan A	Gam			43.00	26.60	\$5,004
Carbin, Nathan A	lan.			.50	39.90	818
Evilantei, Jose Cor Acquiar Pay	sepcian			38.00	19.60	\$744
Extende, Jose Cor Auriking/Tel Ext	Negocian Raime			3.00	10-00	\$33
Lanco Hintz, Willi Angular Pay	an Waldomar			32.00	19.60	9627
Hadina, Christian Regular Pay	1			43.00	25.50	\$1,039
Medina, Christian Overtime	1			29.50	38.25	\$1,128
Nuncio, Jonathan Regular Pay	Alejandro			43.00	19.60	6784
Nuncio, Jonathan Overtine	Algandro			17.00	29.40	\$403
Oyanola, Emmar Regular Pay	tuel Bartscheke			43.00	19.60	\$254
Overtime	ruel berndete			23.00	29.40	5568
Regular Pay				39.00	21.90	Sec. 1

Screenshot 3: Marked invoice

Bulk Document Parsing



IMAGE DIMENSIO	NS						
Width Multiplier:	4.5	Height Multiplier:	4.2				
MAGE BORDER -							
Border Top:	1	Border Bottom:	1	Border Left:	4	Border Right:	4
MAGE BLUR							
Kernel Size:	5	X Deviation:	15				
NOISE REMOVAL							
Parameter 1:	15	Parameter 2:	15	Parameter 3:	21		

Screenshot 4: Filter Settings for JSON



Screenshot 5: Marked Dimensions: JSON Template

interaction and an and a second			
Desour_value: INVOICE			
1			
S Desired as a desired of a			
"Sense_number": 3,			
317. 47.			
γr: 397.			
X21: "130",		i.	
		Go to OCR Thainer RUN	
Results:			0
Results:			0
Results: DATE: 07/23/2021 Sales TER DAVISE NUMBER:	STLL TO AND: Flatte Link, Tor.	ECHIT TO NAME: Remit To: TradeSTAD. Inc.	0
Results: DATE: 07/23/2003 SUPPLIER_DWOICE_NUMBER: TOTAL_INVOICE_AN: \$15,265-43	BILL_TO_NAME: Electra Link, Inc.	REMIT_TO_MAMME: Remit To: TradeSTAR, Inc.	0
Results: 0ATE: 87/23/2001 SUPPLIE_DWOICE_NUMBER: TOTAL_DWOICE_AM: \$15,255.43	BILL_TO_NAME: Electra Link, Inc.	REMIT_TO_NAME: Remit To: TradeSTAR, Inc.	0
Results: 04TE: 67/23/2003 SAPPLIER_DW0ICE_NAMER: T0TAL_DW0ICE_AM: \$15,255.43	Riu_10_XVVC: Electra Link, Inc.	REMIT_TO_MAME: Remit To: TradeSTAR, Inc.	0
Results: DATE: 07/23/2003 SARPLICE_INVOICE_WHEER: TOTAL_DEVOICE_AM: \$15,255.43 TABULAR DATA	BELL_TO_ANNE: Electra Link, Enc.	REFET_TO_MANNE: Remit To: TradeSTAR, Inc.	0
Results: DATE: 47/23/2003 SIMPLIE: RWOOLE WHERE: TOTAL_INVELCE_ME: \$15,355.43	REL_TO_ANNE: Electra Line, Inc.	RENIT_TO_NAME: Resis To: TradeSTAR, Inc.	0
Results: ont::::0723/3021 safeLTR:_DWOLCE_weekR:: ToTAL_INVOLCE_WEER: toTAL_INVOLCE_WE:::525,326.43 TABULAR GATA ROW 1.	BILLTO_NNNE: Electra Link, Inc.	REFET_TO_NAME: Rumit To: TradeSTAR, Inc.	0
Results: DATE: 87/22/001 SWPLIE: PWOLE WARKE: TOTAL_INVICE_AM: \$15,355.43 TABLER DATA RD4 1 RD4 1 RD4 1 RD5201107368: Re11. Clubs Les	RELLTRANNE: Electra Line, Inc.	RENIT_TO_NAME: Resit To: TradeSTAR, Inc.	0
Results: DATE: 87/23/2025 SWALER_WARKE: TATA_SWALER_WARKE: TATA_SWALER_AN: 515,255-43 TABLAR DATA TABLAR DATA TABLAR DATA Result Parts	BELL_TO_ANNE: Electra Link, Inc.	RERET_TO_NAME: Rumit To: TradeSTAR, Inc.	0
Results: DATE: 07/22/020 Severing Noviet Nemeric: TATE, Noviet Nemeric: TATELAR DATA Res 1 Computer New Sector Sector Sector Despiter New Sector Despiter New Sector Despiter New Sector	REA_TO_ANNE: Electra Line, Inc.	RERIT_TO_NAME: Resit To: TradeSTAR, Inc.	0
Results: OATE: 87/23/0015 0076.25/0016.944608: 1076.25/0016.945 515,055-03 12804.84 DATA PSOLUTION: Following Comparison PSOLUTION: Following Comparison PSOLU	BELL_TO_ANNE: Flactra Line, Inc.	RENIT_TO_MANG: Remain To: TradesTAN, Inc.	0
Results: DATE: 07/22/0021 SUPPLIER_DATE: TOTAL_NUMBER: TABLER DATA RNM 1 COSCEPTOR: Suff. Clyok Lee Paylor Pay NUCL_PRU_MATE_20 PAULE_PRU	REA_TO_NAME: Electra Line, Inc.	RERIT_TO_NAME: Remit To: TradeSTAR, Inc.	0

Screenshot6: Processed data