SUCCESS STORIES - SUMMARY

Project Name	Client	Brief Description	Key Analyses
Data Extraction from PDF Files	**Healthcare Provider**	Designed an automated process using **Optical Character Recognition ("OCR"),** to read and extract journal data from PDF files and ingest into SQL tables in a database	

Healthcare Provider

(Data Extraction from PDF Files)

Designed an automated process using **Optical Character Recognition ("OCR"),** to read and extract journal data from PDF files and ingest into SQL tables in a database

AUTOMATED PDF DATA EXTRACTION USING OCR FOR A HEALTHCARE PROVIDER

ABOUT THE CLIENT

Client is a U.S.-based leading specialty **women's health physician group** offering patient care in Obstetrics and Gynecology

SITUATION

- The client was manually updating journal adjustments from PDF files into its financial system for financial reporting purposes, and the process was manual and time consuming
- Merilytics partnered with the client to **build a semi-automated process using Optical Character Recognition ("OCR")** to convert data from scanned PDF files (> 1,000 pages per pdf) into excel and then ingest into a **central data warehouse** for further processing and financial reporting

VALUE ADDITION

- Converted the **source scanned pdf files** (i.e., text cannot be copied directly) **to computer vision images,** and **generated a data frame for extraction** by identifying the row and column dimensions of the data tables
- Developed **custom Python code using Tabula and Pandas libraries** to create an OCR script to read numbers/text from the data frame
- **Cleansed the data to handle spaces, blanks and similar characters (such as "zero", O, D etc.) and extracted the cleansed data** into Excel file before ingesting into Azure Synapse Data Warehouse

IMPACT

- Automation process was institutionalized and **led to accurate financial reporting for the client**, as the process was able to extract data from more than a 1,000 pages within a few minutes and significantly reduced the time spent and scope for human error
- Availability of journal adjustments **improved the data accuracy and completeness** by a **factor of 80%**

DATA EXTRACTION PROCESS DESIGN

Data Source	Data Processing	Data Cleaning	Data Storage
1,000+ pages of PDF documents available in shared network folder	Custom OCR Python code using Tabula & Pandas libraries was developed for data extraction	Data is extracted in Excel and cleansed for handling spaces, blanks, similar shaped characters (0, O, D, etc.)	Cleansed data is saved in the Dedicated SQL Pools of the Azure Synapse Analytics

A The Posted Adjustment journals were **scanned images published as pdf files**. Analyzed the dimensions of columns and rows to create lines around the text/numbers and created **data frame**

B Extracted data using Python code to **read and store the data into Excel** spreadsheet. Ran the semi-automated process for all the pdf files shared

C **Validated** the data against the subtotal and **cleansed data discrepancies** to prepare accurate data to be ingested into Data Warehouse

EXHIBIT 1: IMAGES OF PDF DOCUMENTS (SOURCE AND OUTPUT)

merilytics AN ACCORDION COMPANY

ILLUSTRATIVE

PDF DOCUMENTS

Dec 10 PALM VALLEY WOMENS CARE Page 1

Posted Adjustments Journal

From 10-08 Through 11-30

ID#	Guarantor Name	Acct Date	Adjustment Code	Amount	Unapplied Amount	Item Adjusted Date	Code	Description	DR	LC	Amount
153281	SI	10-08	COMWO	-205.52		08-24	99214	OV-LEVEL 4, EST	MCC	900	135.06
						08-24	58300	INSERT IUD	MCC	900	70.46
100590	KE	10-08	MEDWO	-939.72		09-03	58571	TLH W/T/O	HO	WVO	939.72
137022	BA	10-08	AHCWO	-316.42		09-22	76856	U/S, PELVIC	US2	200	161.62
						09-22	76830	U/S, TRANSVAG	US2	200	154.80
152802	GC	10-08	AHCWO	-140.28		09-20	99214	OV-LEVEL 4, EST	MCC	900	140.28
146501	LE	10-08	AHCWO	-95.85		04-08	99213	14-LEVEL 3, EST	YA	200	95.85
128762	MA	10-08	AHCWO	-131.68		08-10	99214	OV-LEVEL 4, EST	HO	900	131.68
146249	WI	10-08	AHCWO	-140.28		09-17	99214	OV-LEVEL 4, EST	MCC	900	140.28
139727	PE	10-08	CIGWO	-2317.52		09-29	59400	DEL, VAGINAL	HO	BEI	2317.52
126707	VI	10-08	UHCWO	-243.03		07-26	99395	18-39 YEARS	YA	200	133.00
						07-26	G0124	SCREENING PAP	YA	200	31.44
						07-26	Q0091	MEDICARE/PAP SCRE	YA	200	78.59
152131	BE	10-08	AHCWO	-160.66		08-18	76805	U/S, OB COMPLE	US2	200	160.66
135863	CC	10-08	AHCWO	-145.90		09-17	76856	U/S, PELVIC	US2	200	145.90
152533	HE	10-08	AHCWO	-152.03		09-17	76805	U/S, OB COMPLE	US2	200	152.03
153586	JA	10-08	AHCWO	-152.03		09-16	76805	U/S, OB COMPLE	US2	200	152.03
135161	JC	10-08	AHCWO	-152.03		09-07	76805	U/S, OB COMPLE	US2	200	152.03
107142	LE	10-08	AHCWO	-1.00		09-28	0500F	INITIAL PRENATAL	HO	900	1.00
152949	MC	10-08	AHCWO	-1.00		08-27	0500F	INITIAL PRENATAL	YA	200	1.00
153432	SM	10-08	AHCWO	-139.16		09-16	76856	U/S, PELVIC	US2	200	139.16
144384	TH	10-08	AHCWO	-1.00		09-28	0500F	INITIAL PRENATAL	MCC	900	1.00
153609	BE	10-08	AHCWO	-231.91		09-07	99385	18-39 YEARS	SV	200	153.32
						09-07	Q0091	MEDICARE/PAP SCRE	SV	200	78.59
139333	DI	10-08	AHCWO	-88.86		09-15	99213	14-LEVEL 3, EST	SV	900	88.86
105461	GI	10-08	AHCWO	-88.86		09-15	99213	14-LEVEL 3, EST	DA	900	88.86
150175	PE	10-08	AHCWO	-88.86		09-10	99213	14-LEVEL 3, EST	DA	900	88.86
134838	BC	10-08	AHCWO	-173.65		09-01	99204	OV-LEVEL 4, NEW	HO	200	170.15
						09-01	81002	URINALYSIS W/O MI	HO	200	3.50

DATA FRAME OUTPUT

152789	HU	10-11	AHCWO	-133.64		09-20	99214	OV-LEVEL 4, EST	HO	200	133.64
152980	LOP	10-11	AHCWO	-16.68		08-16	99213	14-LEVEL 3, EST	SV	900	16.68
152174	HUN	10-11	AHCWO	-22.08		05-07	99203	OV-LEVEL 3, NEW	DA	900	22.08
138654	PAL	10-11	AHCWO	-15.00		08-10	99214	OV-LEVEL 4, EST	HO	900	15.00
117209	SSR	10-11	AHCWO	-69.84		09-09	99395	18-39 YEARS	YA	200	69.84
153719	MIN	10-11	AHCWO	-281.22		09-21	99386	40-64 YEARS	SV	200	182.92
						09-21	G0124	SCREENING PAP	SV	200	19.71
						09-21	Q0091	MEDICARE/PAP SCRE	SV	200	78.59
116142	KH	10-11	AHCWO	-46.22		09-01	G0124	SCREENING PAP	YA	200	46.22
129534	GON	10-11	AHCWO	-46.22		09-08	G0124	SCREENING PAP	YA	200	46.22
135430	PAL	10-11	AHCWO	-46.22		09-01	G0124	SCREENING PAP	MCC	900	46.22
153601	SL	10-11	AHCWO	-46.22		09-13	G0124	SCREENING PAP	YA	200	46.22
150203	JOH	10-11	AHCWO	-124.81		09-22	G0124	SCREENING PAP	YA	200	46.22
						09-22	Q0091	MEDICARE/PAP SCRE	YA	200	78.59
153711	MC	10-11	BADWO	-70.00		10-04	76856	U/S, PELVIC	US2	200	35.00
	NEI					10-04	99213	14-LEVEL 3, EST	MCC	900	35.00
134397	EST	10-11	BADWO	-30.00		09-22	99214	OV-LEVEL 4, EST	MCC	900	30.00
	NEBE										
100681	KEL	10-11	BADWO	-36.60		09-22	99213	14-LEVEL 3, EST	MCC	900	34.56
	NEEDS TO PAY BAD DEBT					09-22	81002	URINALYSIS W/O MI	MCC	900	2.04
	Total Adjustments for 10-11			-28512.54*	0.00*						
153087	GRA	10-12	AHCWO	-295.90		09-30	76805	U/S, OB COMPLE	US2	200	295.90
	INCLUDED IN OB PACKAGE										
145149	MARQ	10-12	PVTWO	-233.89		10-12	99385	18-39 YEARS	DA	200	109.08
	PVT					10-12	G0124	SCREENING PAP	DA	200	46.22
						10-12	Q0091	MEDICARE/PAP SCRE	DA	200	78.59
	Total Adjustments for 10-12			-529.79*	0.00*						
153332	GAS	10-13	AHCWO	-18.31		07-15	G0124	SCREENING PAP	SV	900	18.31
150017	PET	10-13	AETWO	-91.65		09-02	99212	OV-LEVEL 2, EST	MCC	900	54.31
						09-02	96372	THERAPEUTIC INJEC	MCC	900	37.34
151801	BEN	10-13	AETWO	-123.56		09-02	99214	OV-LEVEL 4, EST	DA	900	123.56
134010	CHA	10-13	AETWO	-275.51		09-14	76856	U/S, PELVIC	US2	200	141.79
						09-14	76830	U/S, TRANSVAG	US2	200	133.72
136784	COT	10-13	AETWO	-154.94		09-14	76805	U/S, OB COMPLE	US2	200	154.94
134511	DELE	10-13	AETWO	-300.88		09-23	99396	40-64 YEARS	YA	200	176.07
						09-23	G0124	SCREENING PAP	YA	200	46.22
						09-23	Q0091	MEDICARE/PAP SCRE	YA	200	78.59
153133	GALI	10-13	AETWO	-278.12		09-27	99204	OV-LEVEL 4, NEW	SV	900	193.17
						09-27	99213	14-LEVEL 3, EST	SV	900	84.95
153189	GONI	10-13	AETWO	-154.94		09-22	76805	U/S, OB COMPLE	US2	200	154.94
149491	LOPE	10-13	AETWO	-154.94		09-22	76805	U/S, OB COMPLE	US2	200	154.94