AI-DRIVEN PDF DOCUMENT COMPARISON

Author & Presenter : Sanduni Perera

Problem & Objective

Problem Statement

- Manual comparison of product specifications is time consuming and error prone
- Differences in text, numeric tables, and images need to be identified accurately

Objective

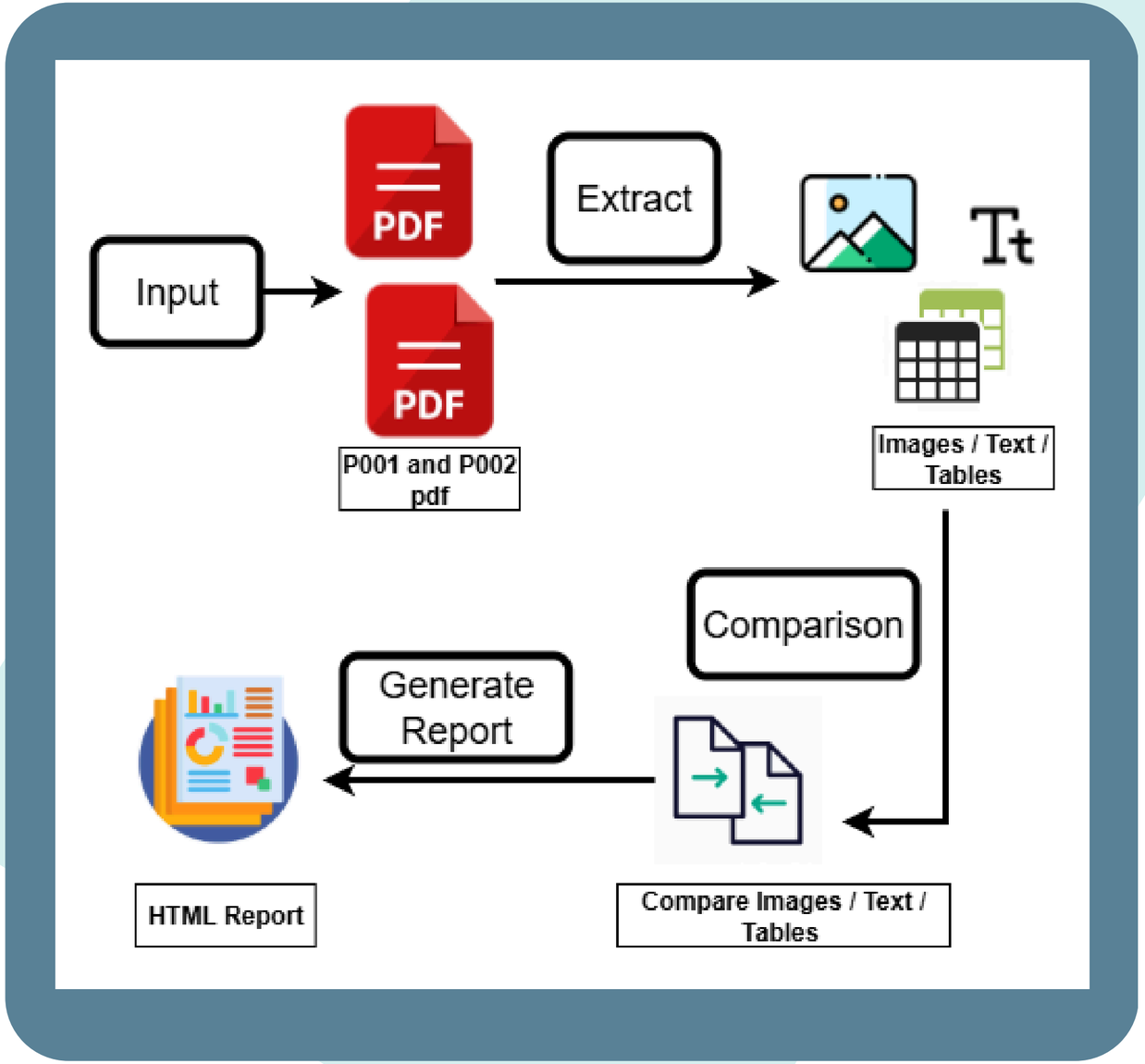
- Automatically compare two PDF documents (text, tables, images) and highlight key differences.

Goal

- Demonstrate AI and Python-based automation for real-world use cases such as product specification validation and version control.

SOLUTION APPROACH

- **Extract Data -** Extract text, tables, and images from both PDFs using pdfplumber and camelot.
- **Compare Text -** Perform section-wise, line-by-line comparison, highlighting differences in red (P001) and green (P002).
- **Compare Tables -** Normalize rows/columns and compare numeric and textual values to find differences .
- **Compare Images -** Use AI-based similarity techniques (imagehash and SSIM) to assess image differences.
- **Generate Report -** Summarize all differences in an HTML report with visual highlights for text, tables, and images.



Results – Text Comparison

PDF Comparison Report

19	28	1
Text Differences	Table Differences	Image Differences

Text Comparison

Section	Line No.	P001 Value	P002 Value
Product Weights and Measures	1	Product & primary	Product & primary
Product Weights and Measures	2	Product only Secondary packaging Transit packaging	Product only Secondary packaging Transit packaging
Product Weights and Measures	3	packaging	packaging
Product Weights and Measures	4	Item Quantity 1 1 10 150	Item Quantity 1 1 3 90
Product Weights and Measures	5	Width mm 37 37 80 300	Width mm 25 27 100 450
Product Weights and Measures	6	Depth mm 15 15 240 350	Depth mm 15 15 18 350
Product Weights and Measures	7	Height mm 210 210 60 300	Height mm 240 260 290 350
Product Weights and Measures	8	Weight 126 126 1294 20.1	Weight 50 54 168 5.8
Packaging Information	1	Card use 0 Box Case	Card use 0 0 Case
Packaging Information	2	Card gms 0 34 675	Card gms 0 0 787.5
Packaging Information	3	Plastic use 0 0 -	Plastic use Wallet Bag -
Packaging Information	4	Plastic gms 0 0 0	Plastic gms 4 6 0
Packaging Information	5	Metal use 0 0 -	Metal use 0 0 -
Packaging Information	6	Metal gms 0 0 0	Metal gms 0 0 0
Packaging Information	7	Timber use 0 0 -	Timber use 0 0 -
Packaging Information	8	Timber gms 0 0 0	Timber gms 0 0 0
Packaging Information	9	14 Aug 2019 12:55PM Created via cki.skoocloud.com 1 of 1	05 May 2020 03:15PM Created via cki.skoocloud.com 1 of 1
Header	1	Product Specification PDF	Product Specification PDF
Header	2	Date Correct as of 14th Aug 2019	Date Correct as of 5th May 2020
Header	3	Brand C.K	Brand C.K
Header	4	Product T4343M 17	Product T0083 6
Header	5	Description Combination Spanner 17mm	Description C.K Engineers File Round 150mm 2nd Cut
Header	6	Barcode 5013969550603	Barcode 5013969390704
Header	7	Commodity code 8204110000	Commodity code 8203100000
Header	8	Country of origin TW	Country of origin PL
Header	9	Features & ~ Drop forged chrome vanadium steel, hardened & tempered for strength & durability ~ Heavy duty chrome plated for	Features & ~ Manufactured from special file steel, precision cut and hardened for outstanding performance and durability ~
Header	10	Benefits corrosion resistance ~ 12° offset bi-hex ring ~ According to DIN 3113 ~ Size: 17 mm ~ Overall length: 210 mm	Benefits Ergonomic soft grip handle for superior comfort and control ~ Specially designed solid inner core for a safe and secure
Header	11		handle to tang bond ~ Profile: Round ~ Second cut ~ Tapered towards the point ~ Length: 150 mm

Results – Table Comparison

Table 1 — Product Information

Field	P001	P002
Date	Correct as of 14th Aug 2019	Correct as of 5th May 2020
Brand	C.K	C.K
Product	T4343M 17	T0083 6
Description	Combination Spanner 17mm	C.K Engineers File Round 150mm 2nd Cut
Barcode	5013969550603	5013969390704
Commodity code	8204110000	8203100000
Country of origin	TW	PL

Table 2 — Product Weights and Packaging Information

Item / Measurement	Product only P001	Product only P002	Product & primary packaging P001	Product & primary packaging P002	Secondary packaging P001	Secondary packaging P002	Transit packaging P001	Transit packaging P002
	Product only	Product only	Product & primary\npackaging	Product & primary\npackaging	Secondary packaging	Secondary packaging	Transit packaging	Transit packaging
Item Quantity	1	1	1	1	10	3	150	90
Width mm	37	25	37	27	80	100	300	450
Depth mm	15	15	15	15	240	18	350	350
Height mm	210	240	210	260	60	290	300	350
Weight	126	50	126	54	1294	168	20.1	5.8
	—	—	—	—	—	—	—	—
Packaging Information	Card use	Card use	0	0	Box	0	Case	Case
	Card gms	Card gms	0	0	34	0	675	787.5
	Plastic use	Plastic use	0	Wallet	0	Bag	-	-
	Plastic gms	Plastic gms	0	4	0	6	0	0
	Metal use	Metal use	0	0	0	0	-	-
	Metal gms	Metal gms	0	0	0	0	0	0
	Timber use	Timber use	0	0	0	0	-	-
	Timber gms	Timber gms	0	0	0	0	0	0

Results – Image Comparison

Image Comparison

P001

Carl Kammerling International | Brand Innovation

P002

Carl Kammerling International | Brand Innovation

Hash Diff: 0
SSIM: 1.0
Similarity: High

P001

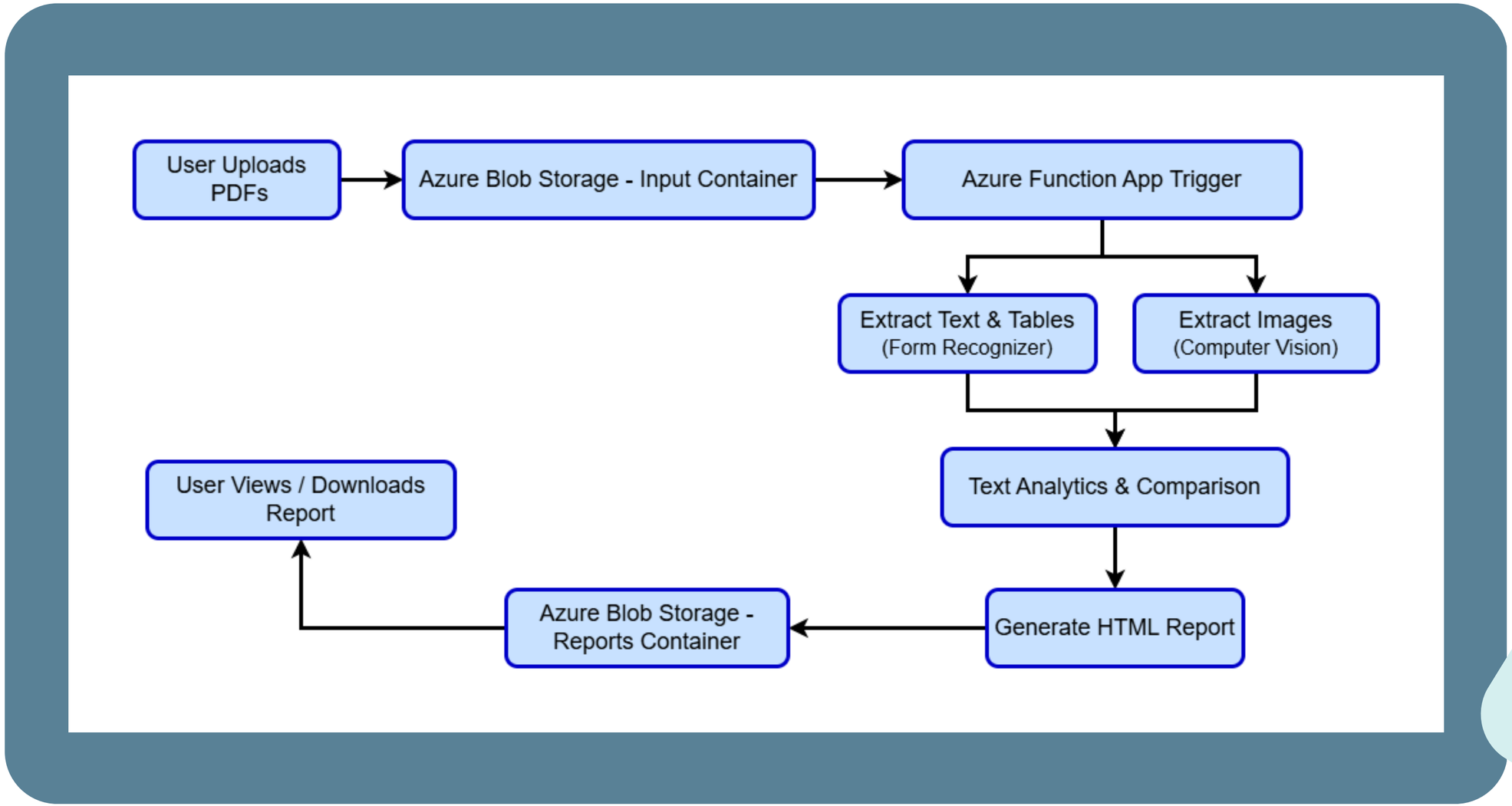
P002

Hash Diff: 32
SSIM: 0.89
Similarity: Moderate

How Azure AI Services Fit Into the Solution

- **Form Recognizer -** Automatically extracts text, tables, and layout from PDFs, improving accuracy and reducing manual parsing errors.
- **Text Analytics (Cognitive Services) -** Detects key phrases, entities, and semantic differences, enabling smarter comparisons beyond exact string matches.
- **Computer Vision -** Analyzes images, detects differences, and extracts text from scanned PDFs.
- **Azure Functions & Logic Apps -** Serverless, on-demand processing, triggers automated workflows like email notifications and parallel processing.

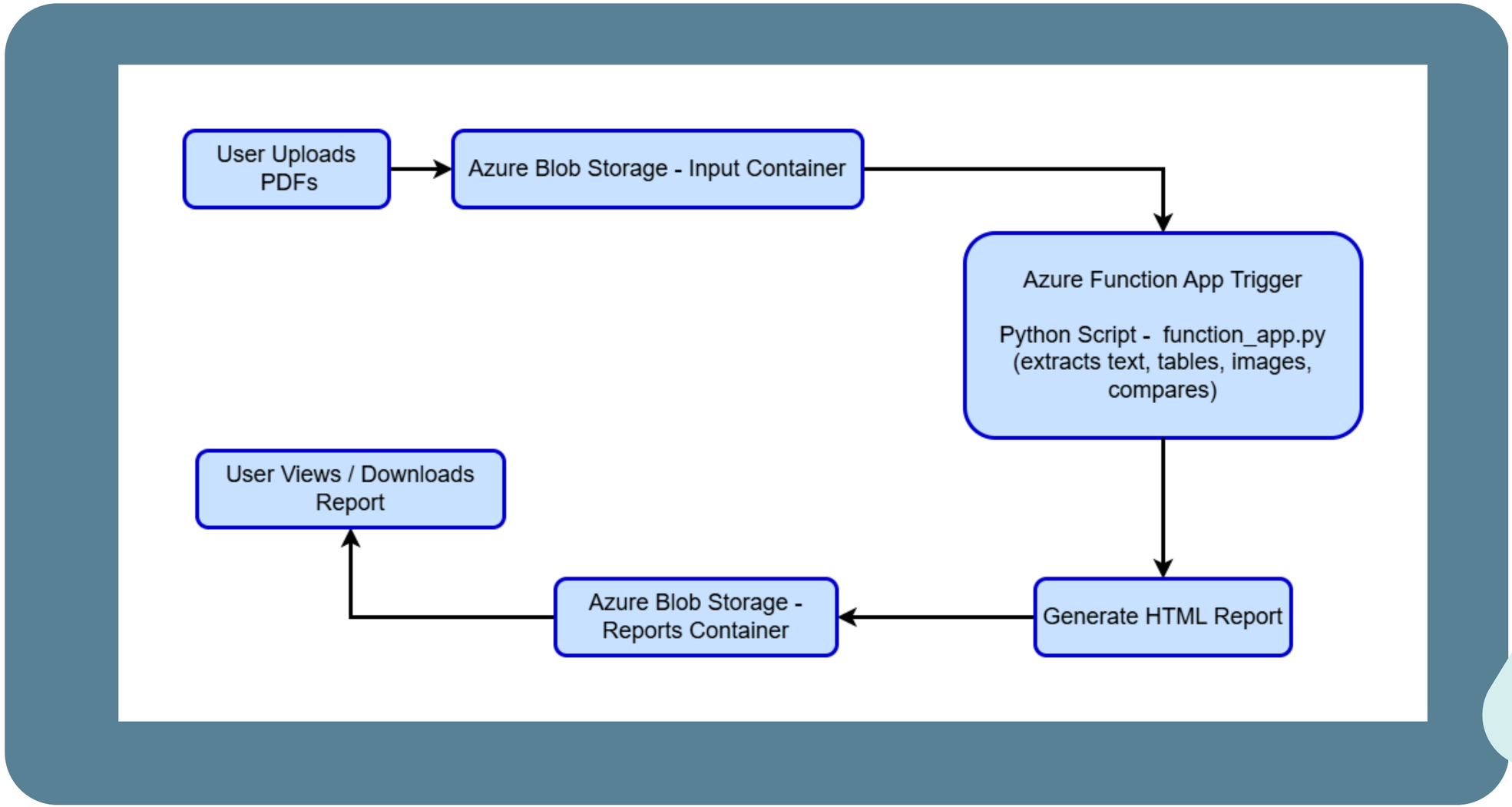
AZURE WORK FLOW



RECOMMENDATIONS FOR IMPROVEMENTS & SCALABILITY

- **Automation -** Integrate Logic Apps or Power Automate to notify users when reports are ready.
- **Parallel Processing -** Use Azure Functions to process multiple PDF uploads simultaneously for faster throughput.
- **Semantic Comparison -** Leverage Text Analytics for deeper understanding of content differences.
- **Monitoring & Logging -** Use Azure Monitor and Application Insights to track processing and handle errors efficiently.
- **Future Enhancements -** Consider integrating AI models to summarize PDF differences or classify document changes automatically.

AZURE IMPLEMENTED WORK FLOW



REFERENCES

1. Microsoft. Azure AI Services Overview. Microsoft Learn. - Available at: https://learn.microsoft.com/en-us/azure/ai-services/what-are-ai-services
2. Microsoft. Azure AI Document Intelligence (Form Recognizer). Microsoft Learn. - Available at: https://learn.microsoft.com/en-us/azure/ai-services/document-intelligence
3. Microsoft. Azure Blob Storage. Microsoft Learn. - Available at: https://learn.microsoft.com/en-us/azure/storage/blobs
4. Microsoft. Azure Functions. Microsoft Learn. - Available at: https://learn.microsoft.com/en-us/azure/azure-functions
5. Microsoft. Azure Text Analytics. Microsoft Learn. - Available at: https://learn.microsoft.com/en-us/azure/ai-services/text-analytics
6. Microsoft. Azure Computer Vision. Microsoft Learn. - Available at: Automation: Integrate Logic Apps or Power Automate to notify users when reports are ready.

QUESTIONS?

THANK YOU FOR YOUR TIME AND ATTENTION.