Contract digitization

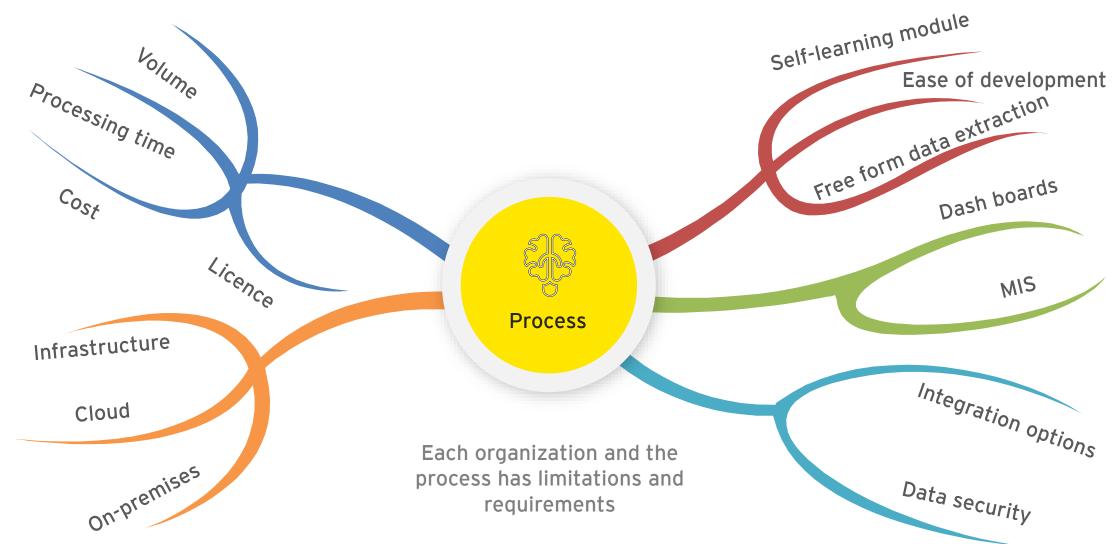# Data digitization solutions

Clauses extraction
Entity extraction
Tables extraction

Contract digitization

Invoice digitization
Purchase order automation

Accounts receivable/ payable

Intent and entities
Auto reply feature
BPM integration

Email processing

Digitization solutions

KYC

Proof of Identity

PAN
Passport
National ID

Bio-metrics

Fingerprint matching
Facial recognition

Digital Forms

New customer form digitized
Change application digitized

NLP   ML   RPA   BPM   OCR

Technologies used

● Case study available

Use-case specific considerations

Use case

Each use case for document digitization requires numerous feature consideration

Pdf/Jpeg
Hand written
Resolution
File feature
Tabular data
Contextual information
Variation in formats

Multiple page documents
Header less tables
Date type
Business rules
Languages
Data rules
User validation

Process specific considerations

Volume

Processing time

Cost

Licence

Infrastructure

Cloud

On-premises

Process

Each organization and the process has limitations and requirements

Self-learning module

Ease of development

Free form data extraction

Dash boards

MIS

Integration options

Data security

EY

Technology Stack

ICR pre-built platforms

IRIS — Document to Knowledge — A CANON COMPANY

ANTWORKS — Reimagine · Rethink · Recreate

Xtracta — your data engine

pendo systems — FUELING THE AI REVOLUTION

SIMPLEKYC

ess DOCS

KOFAX

docdigitizer

ABBYY

AUTOMATION ANYWHERE

Datacap — an IBM® Company

WorkFusion

LEADTOOLS — THE WORLD LEADER IN IMAGING SDKs

Custom ICR toolkit

Open source tools

Tensorflow

Python

Keras

Tesseract OCR

OpenCV

APIs (Computer vision/OCR/Language)

Azure

1. Computer vision API
2. Speech to text API
3. Language Understanding API

aws

1. Amazon Rekognition
2. Amazon Polly
3. Amazon Lex
4. Amazon Textract

Google Cloud

1. AutoML vision
2. Cloud vision API
3. Cloud text to speech API
4. Cloud Natural language API

InDEx Tech Stack

Open source tools
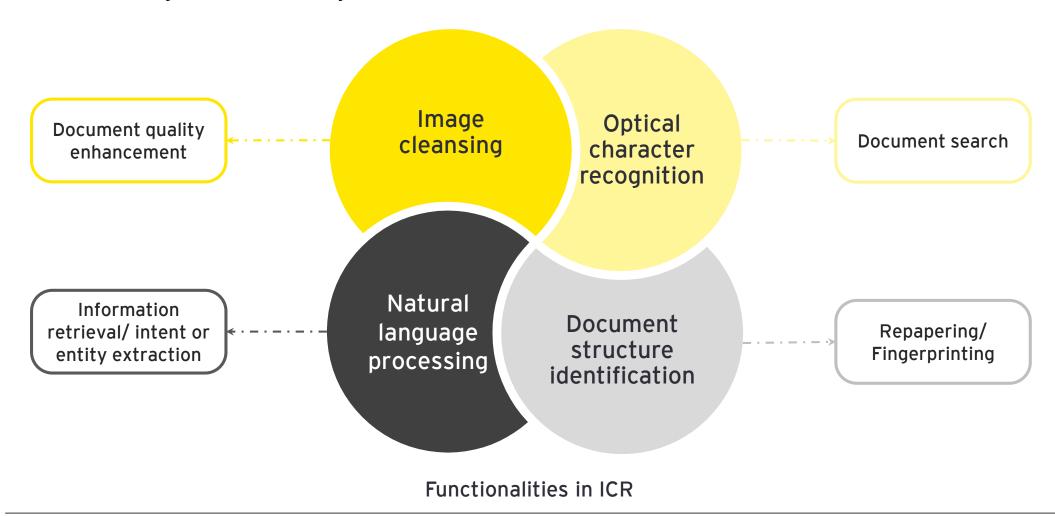
Python

OpenCV

APIs (Computer vision/OCR/Language)

Azure

1. Microsoft Azure Machine Learning services
2. Microsoft Azure Cognitive services
3. Azure Kubernetes services
4. Microsoft API management
5. Microsoft PowerApps

InDEx - Intelligent Data Extraction

To transform un-structured information to structured it is important that we are able to contextualize the information. Thus InDEx leverages machine learning to infer context based on historical information.

Document quality enhancement

Image cleansing

Optical character recognition

Document search

Information retrieval/ intent or entity extraction

Natural language processing

Document structure identification

Repapering/ Fingerprinting

Functionalities in ICR

InDEx in action

InDEx automatically extracts data fields from unstructured documents such as scanned contracts.

Input document (image)

The definitions and provisions contained in the 2006 ISDA Definitions, as published by the International Swaps and Derivatives Association, Inc., are incorporated into this Confirmation. In the event of any inconsistency between those definitions and provisions and this Confirmation, this Confirmation will govern.

This Confirmation supplements, forms part of, and is subject to the ISDA Master Agreement dated as of 29 December 2009 as amended and supplemented from time to time (the "Agreement") between Counterparty and ▮▮▮▮▮▮. All provisions contained in the Agreement govern this Confirmation except as expressly modified below.

The terms of the particular Rate Swap Transaction to which this Confirmation relates are as follows:

General Terms

Trade Date:	07 April 2011 (time of execution available upon request)
Effective Date:	11 April 2011
Termination Date:	11 April 2041, subject to adjustment in accordance with the Modified Following Business Day Convention.

Processed document

The definitions and provisions contained in the 2006 ISDA Definitions, as published by the International Swaps and Derivatives Association, Inc., are incorporated into this Confirmation. In the event of any inconsistency between those definitions and provisions and this Confirmation, this Confirmation will govern.

This Confirmation supplements, forms part of, and is subject to the ISDA Master Agreement dated as of 29 December 2009 as amended and supplemented from time to time (the "Agreement") between Counterparty and ▮▮▮▮▮▮. All provisions contained in the Agreement govern this Confirmation except as expressly modified below.

The terms of the particular Rate Swap Transaction to which this Confirmation relates are as follows:

General Terms

Trade Date:	07 April 2011 (time of execution available upon request)
Effective Date:	11 April 2011
Termination Date:	11 April 2041, subject to adjustment in accordance with the Modified Following Business Day Convention.

Required Data Fields

- ❖ **Master Agreement version**
- ❖ **Type of transaction**
- ❖ **Effective date**

Extracted Data Points

- ✓ **29 December 2009**
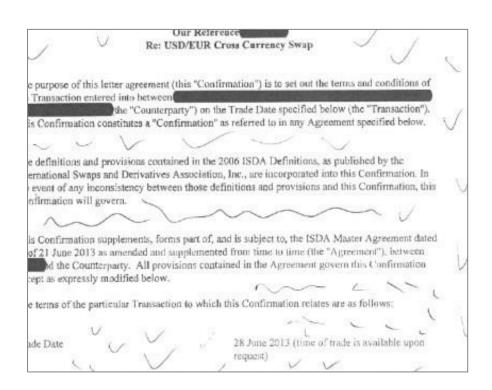- ✓ **Rate Swap Transaction**
- ✓ **11 April 2011**

Image cleansing

In order to **improve the success rate** of the OCR a set of **image processing and cleaning routines** are applied. Deep learning techniques such as convolutional neural networks are used to remove noise.

Raw data

Clean data

Information re-mapping/ fingerprinting

The layout of the document is analysed by algorithms able to locate headers, paragraphs, tables... **Dilating techniques** are used to identify **blocks of information** and implicit **table structures.**

1 Raw data

2 Use algorithms to learn format
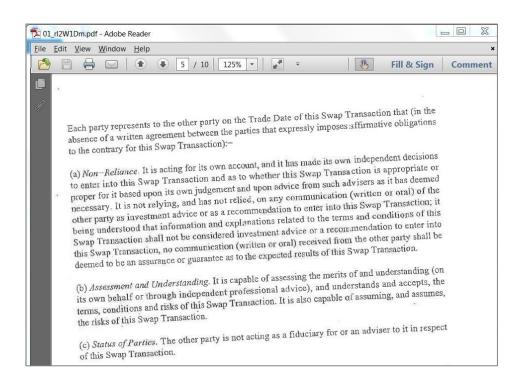
3 Locate and extract the relevant information
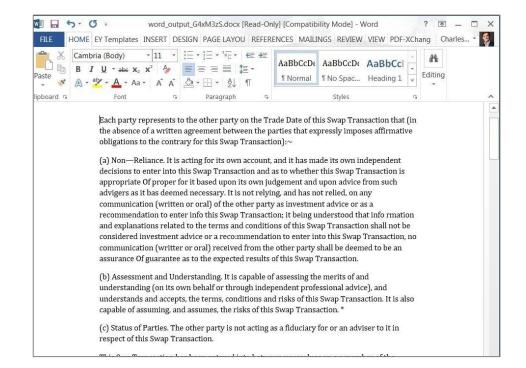
Optical character recognition

OCR engine is applied to each of the identified blocks of information to **transform unstructured data into text.**

Unstructured data

Structured text where further AI can be applied
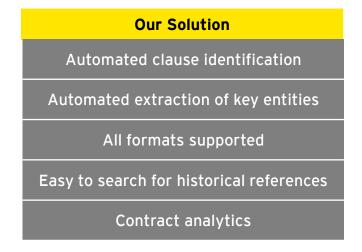
Contract Digitization
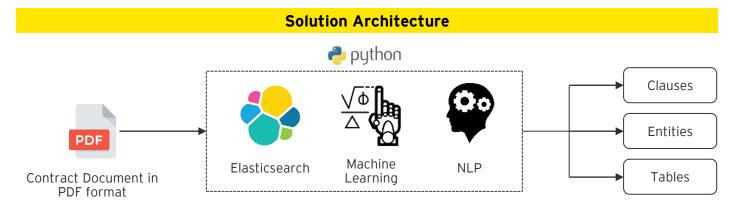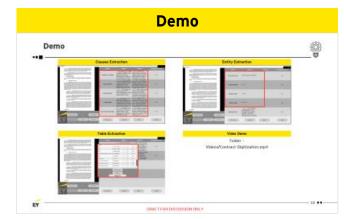
Every organization manages contracts such as service agreements, lease, SOW etc. qualified legal team is required to understand and manage these contracts. EY has developed solution to extract most commonly used information from the contracts automatically thereby significantly reducing the cost and improving efficiency.

Organizational Challenges	Our Solution	Application
Qualified lawyers required to extract basic information from contracts	Automated clause identification	Contract review
Difficult to search for data historically	Automated extraction of key entities	Contract management
High SLA for contract processing	All formats supported	Contract term comparison
Prone to human errors	Easy to search for historical references	Legal proceedings
	Contract analytics	

Solution Architecture

Demo

Demo

Clauses Extraction

Entity Extraction

Table Extraction
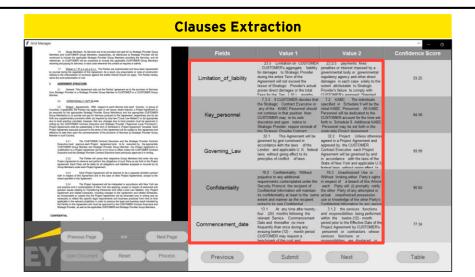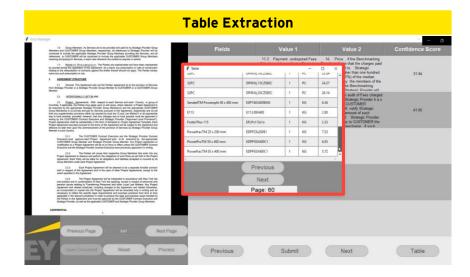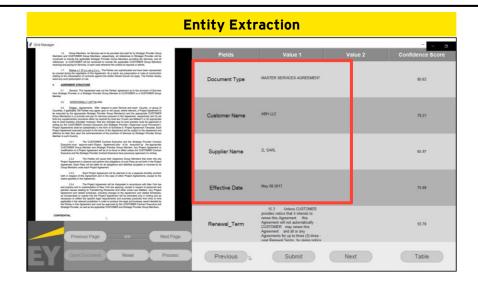
Video Demo

Folder –
Videos/Contract Digitization.mp4

Table digitization

Across organizations there are various documents containing tabular data in different unstructured formats coming from different sources which require human intervention for standardization and further processing. EY has built a customisable solution which can be leveraged to extract tabular information efficiently from different document formats automatically henceforth reducing processing time and saving manual effort.

Organizational Challenges

Multiple documents in different formats from multiple sources

Bottlenecks in end to end process automation plan

Data is mostly unstructured, unsearchable and requires huge processing time

Manual dependency and error prone

Our Solution

User friendly interface and validation supported

Customisable to any standard format
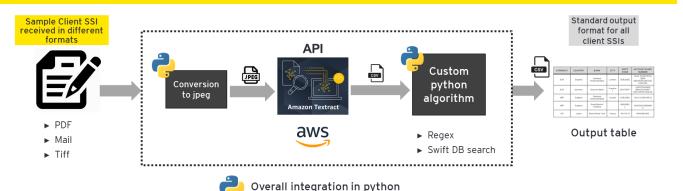
Converts any unstructured tabular data to structured tabular format

Lower handling time, high accuracy*

Application
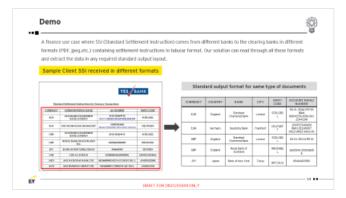
Any industry specific process which require digitizing tabular data from unstructured formats

Solution Architecture

Demo

Demo

A finance use case where SSI (Standard Settlement Instruction) comes from different banks to the clearing banks in different formats (PDF, jpeg,etc.) containing settlement instructions in tabular format. Our solution can read through all these formats and extract the data in any required standard output layout.

Sample Client SSI received in different formats

YES BANK
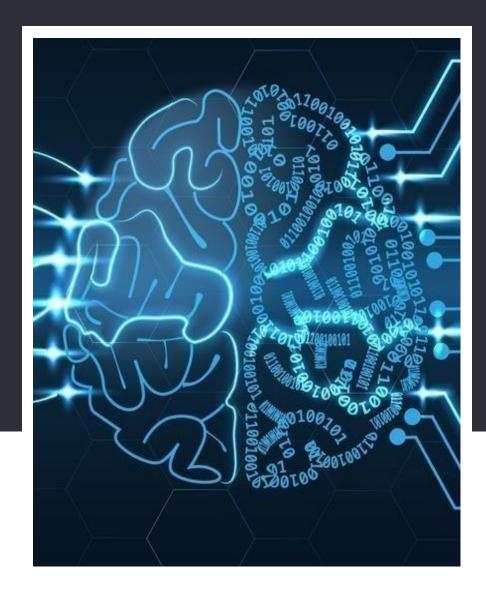
Standard Settlement Instructions for Treasury Transactions

CURRENCY	CORRESPONDING BANK	A/c NUMBER	SWIFT CODE
EUR	STANDARD CHARTERED BANK LONDON	56 01 2544199 96 IBAN GB64SCBL60910412544199	SCBLGB2L
EUR	DEUTSCHE BANK FRANKFURT	100953543600 IBAN DE245007 0010 0953 5436 00	DEUTDEFF
GBP	STANDARD CHARTERED BANK LONDON	00 01 2544199 01	SCBLGB2L
GBP	ROYAL BANK OF SCOTLAND PLC	160034/100006005	RBOSGB2L
JPY	BANK Of NEW YORK,TOKYO	8946483950	IRVTJPJX
CHF	UBS AG ZURICH	02300000036645050000K	UBSWCHZH80A
NZD	ANZ NATIONAL BANK LTD.	881169/00001NZD ACCOUNT NO. 1.	ANZBNZ22058
AUD	ANZ BANKING GROUP LTD.	81169/00001 CURRENT A/C NO.1.	ANZBAU3M

Standard output format for same type of documents

CURRENCY	COUNTRY	BANK	CITY	SWIFT CODE	ACCOUNT NAME/ NUMBER
EUR	England	Standard Chartered Bank	London	SCBLGB2L	56 01 2544199 96 IBAN GB64SCBL6091041 2544199
EUR	Germany	Deutsche Bank	Frankfurt	DEUTDEFF	100953543600 IBAN DE245007 0010 0953 5436 00
GBP	England	Standard Chartered Bank	London	SCBLGB2L	00 01 2544199 01
GBP	England	Royal Bank of Scotland		RBOSGB2L	160034/10000600 5
JPY	Japan	Bank of New York	Tokyo	IRVTJPJX	8946483950

Our Credentials

Business Case - ICR Order Entry

Background
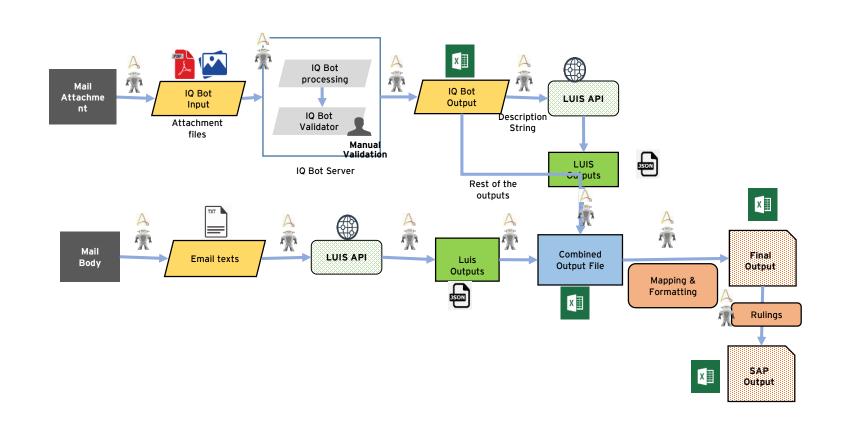
► A multinational steel manufacturing corporation performed the order entry process manually. The process involved manually recording and digitizing purchase orders and invoices with multiple vendors

EY Approach & Solution

► EY a study of the process and shortlisted OCR/ICR platforms to be integrated with the client's system

► An end to end ICR and NLP based solution to be designed for the vendor's PO and Invoice digitization process

► End to end automation integration to digitize PO and Invoice documents and combine the outputs to be fed into the ERP system in place
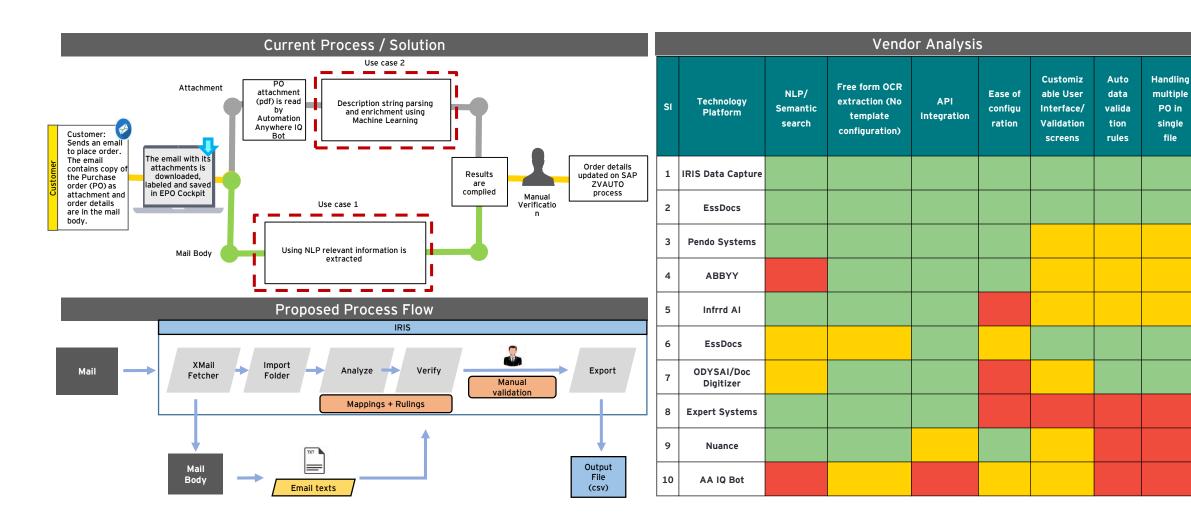
Outcome

► EY conducted extensive feasibility assessment exercises with multiple global vendors

► IRIS Xtract was finalized to conduct OCR/ICR digitization

► Microsoft LUIS was deployed to conduct NLP classification

► End to end automation by deploying Automation Anywhere was achieved

► The solution reduced total SLAs and increased overall efficiency of the process.

Business Case - ICR Vendor Selection

Current Process / Solution

Customer: Sends an email to place order. The email contains copy of the Purchase order (PO) as attachment and order details are in the mail body.

The email with its attachments is downloaded, labeled and saved in EPO Cockpit

Attachment — PO attachment (pdf) is read by Automation Anywhere IQ Bot

Use case 2 — Description string parsing and enrichment using Machine Learning

Mail Body — Use case 1 — Using NLP relevant information is extracted

Results are complied

Manual Verification

Order details updated on SAP ZVAUTO process

Proposed Process Flow

IRIS

Mail → XMail Fetcher → Import Folder → Analyze → Verify → Manual validation → Export

Mappings + Rulings

Mail Body → Email texts → Output File (csv)

Vendor Analysis

SI	Technology Platform	NLP/ Semantic search	Free form OCR extraction (No template configuration)	API Integration	Ease of configuration	Customizable User Interface/ Validation screens	Auto data validation rules	Handling multiple PO in single file
1	IRIS Data Capture							
2	EssDocs							
3	Pendo Systems							
4	ABBYY							
5	Infrrd AI							
6	EssDocs							
7	ODYSAI/Doc Digitizer							
8	Expert Systems							
9	Nuance							
10	AA IQ Bot							

Business Case – ICR Vendor Selection

Background

► A multinational investment bank's Accounts Payable process was handled by a vendor in-charge of data enrichment and validation services before further processing in ERP

► Key pain points: Multiple human touchpoints, regional variations, zero straight through processing, primitive reporting, no available audit trail

EY Approach & Solution

► EY conducts a study of the process, gathers data around invoice volumes and peak loads, conducts meetings with stakeholders to understand their key issues and challenges

► Best fit analysis to identify evaluation parameters and create an evaluation framework to assess solutions in the market

Outcome

► EY conducted demos, assessed functionalities & captured feedback from stakeholders & users. Built a model around those parameters to assess the solution based on technical , functional, financial and general vendor presence.

► Shortlisted solution based on agreed parameters and scoring was provided

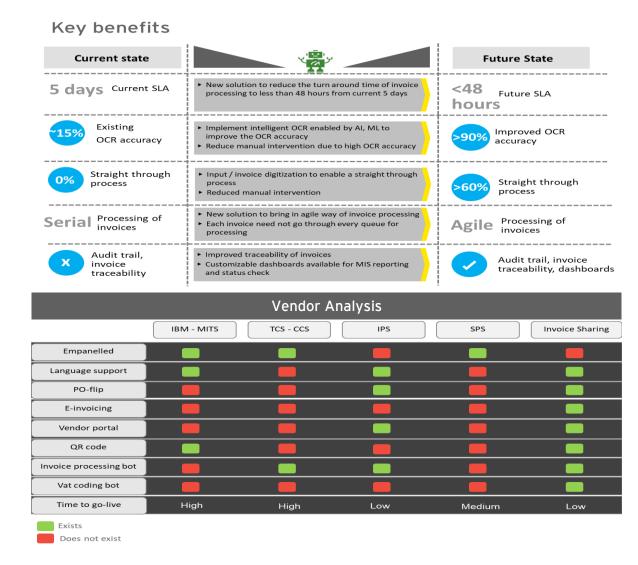► Suggest top 3 solutions : one primary solution and two alternate solutions & provide detailed implementation roadmap for the selected solution

Key benefits

Current state		Future State	
5 days	Current SLA	► New solution to reduce the turn around time of invoice processing to less than 48 hours from current 5 days	<48 hours	Future SLA
~15%	Existing OCR accuracy	► Implement intelligent OCR enabled by AI, ML to improve the OCR accuracy ► Reduce manual intervention due to high OCR accuracy	>90%	Improved OCR accuracy
0%	Straight through process	► Input / invoice digitization to enable a straight through process ► Reduced manual intervention	>60%	Straight through process
Serial	Processing of invoices	► New solution to bring in agile way of invoice processing ► Each invoice need not go through every queue for processing	Agile	Processing of invoices
X	Audit trail, invoice traceability	► Improved traceability of invoices ► Customizable dashboards available for MIS reporting and status check	✔	Audit trail, invoice traceability, dashboards

Vendor Analysis

	IBM – MITS	TCS – CCS	IPS	SPS	Invoice Sharing
Empanelled	🟩	🟩	🟥	🟩	🟥
Language support	🟩	🟥	🟩	🟥	🟩
PO-flip	🟥	🟥	🟩	🟥	🟩
E-invoicing	🟥	🟥	🟥	🟥	🟩
Vendor portal	🟥	🟥	🟩	🟥	🟩
QR code	🟩	🟥	🟥	🟥	🟩
Invoice processing bot	🟥	🟩	🟩	🟥	🟩
Vat coding bot	🟥	🟥	🟥	🟥	🟩
Time to go-live	High	High	Low	Medium	Low

🟩 Exists
🟥 Does not exist

Business Case - ICR Vendor Selection

Background

► A leading Global investment banking client performs KYC refresh for over 20,000 entities each year. The process involves manual tasks of document collection from various sources, data extraction and data validation.

► The process involves FTEs manually collecting publicly available document from different sources and extracting required data fields and conduct extensive validation through their downstream systems

EY Approach & Solution

► EY conducted a feasibility assessment of data extraction technologies that can be deployed on premises with the ability to extract data using NLP / semantic search.

► EY identified evaluation parameters and measured the technology platforms' ability to integrate with upstream and downstream systems

Outcome

► EY conducted extensive feasibility assessment exercises with multiple global vendors

► Data extraction was found feasible as per our study

► Suggested top 3 platforms: one primary platform and two alternate solutions & provided detailed implementation roadmap for the selected solution

► Due to limited time and training data, EY relied on the platform's robustness in a well trained scenario, first pass accuracy, credentials, references and assessment of the technology.

Vendor Analysis

Functionality Matrix

EY assessed the shortlisted technology platforms specifically for features relevant to GS. Below is a summary of top 5 Technloloy Platforms:

Technloloy Platform	On-Premise Deployment	NLP/ Semantic search	Rule engine	API Integration	Processing Speed	Self learning modules	Ease of on-boarding new document	Training & Support Services	Enterprise Class Management
Pendo Systems	Strong	Strong	Strong	Strong	Strong	Strong	Intermediate	Strong	Strong
Expert Systems	Strong	Strong	Strong	Strong	Strong	Strong	Intermediate	Weak	Strong
Ikarus	Strong	Strong	Strong	Strong	Intermediate	Strong	Intermediate	Intermediate	Strong
EY In-house	Strong	Intermediate	Intermediate	Intermediate	Intermediate	Intermediate	Intermediate	Strong	Strong
IBM Datacap	Strong	Intermediate	Weak	Weak	Intermediate	Strong	Weak	Intermediate	Strong

Strength definitions

Strong (green)
Denotes a capability that is; in a mature state of existence, easily demonstrable.

Intermediate (yellow)
Denotes a capability that is; approaching maturity but with gaps to fully satisfy the client requirement or business need.

Weak (red)
Denotes a capability that has; either not been developed or is in its infancy. The capability will quite often be in beta or not a focus for the Technloloy Platform