SOBHA

REALTY

Business Requirement Document - Document Classification and Key Data Extraction

Document details:

Author: Samji Varghese Alex	**Business Unit:** Sobha Realty	**Project ID:**
Reviewer: Deepu Nair	**Business Function:** Information & Technology	**Release Date:**
Approver:	**Published:**	**File Version:** 1.1

Version Details

Version: 1.1	**Creation Date:** 27-Nov-2025
Status: Draft	**Last Revision Date:** 11-Dec-2025
Owner: Samji Varghese Alex	**Pages:** 11

Revision History:

Version	Description	Amended By	Change Date
1.0	Draft Version	Samji Varghese Alex	27 Nov 2025
1.1	Updated Sections	Deepu Nair	11 Dec 2025

Table of Contents

Introduction:

This BRD defines the objectives and requirements for the AI-based Document Classification & Key Data Extraction solution for Sobha Realty. The system will automatically process customer and sales documents stored in SharePoint, classify them using OCR/LLM models, extract the required metadata, and migrate the documents to ADLS along with their corresponding JSON outputs.

The processed documents and JSON output will be made accessible through secure API endpoints for use by Salesforce. This initiative aims to improve compliance, eliminate manual document migration to Salesforce.

Objectives:

- The project will be executed in two phases: Phase 1 will cover Solution Build & Run the Documents for Project Greens and One Park Avenue, while Phase 2 will include all remaining projects documents to classify & move it to ADLS. Phase 2 will follow the same functional & technical scope as Phase 1, and the list of projects will be shared by the business during or after Phase 1 execution.
- Migrate documents from SharePoint to ADLS maintaining original folder structure
- Automate document classification and key metadata extraction using OCR/LLMs, generating a corresponding JSON file for each document.
- Improving accuracy through business-provided Key Identifiers.
- Validate signature and seal presence (presence) to determine whether customer/seller documents are Signed / Not Signed.
- Support exception handling through categorization, email notifications, and Power BI-based exception reporting.
- Provide reconciliation reports to track processed, failed, and pending documents for audit transparency.
- Support exception reprocessing through manual classification.
- Provide secure API endpoints to retrieve documents and metadata for Salesforce.

Scope:

In Scope

- Migration of customer-related documents from SharePoint to ADLS, maintaining the SharePoint folder hierarchy
- Automated document classification and metadata (key data) extraction for supported formats using OCR (Azure DI) and OpenAI O4 Mini.
- Migration of PDF, JPG, PNG and JPEG with classification and metadata extraction.
- Migration of non-supportive OCR document like Outlook emails, Excel files, and password-protected documents, without classification or metadata extraction.
- Generation of JSON metadata files capturing classification output and extracted key fields stored alongside each migrated document.
- Validation of signature and seal presence for signed/not signed documents.
- Exception categorization: Low Confidence Classification, Corrupted Files
- Automated email notifications, and Power BI exception dashboards for operational tracking.
- Reconciliation reporting to track successfully processed, failed, and pending documents.
- Phase 2 will follow the same functional & technical scope as Phase 1. The list of remaining projects will be provided by business during execution.

Limitations or Constraints:

1- **OCR & LLM Accuracy Depends on Document Quality** -Low-resolution scans, handwritten notes, blurred images, cropped pages, or tilted/rotated documents may reduce classification or metadata extraction accuracy.
2- **Password-Protected and Encrypted Files -** Such files will be migrated to ADLS but will not undergo classification or metadata extraction, as OCR cannot process locked documents unless passwords are provided.
3- **Unsupported File Formats -** Files other than PDF, JPG, PNG, JPEG will be migrated without document classification or key data extraction.
4- **Manual Data Cleanup or Editing -** The system does not alter or correct document content; it only extracts and classifies based on what is present.
5- **Performance Dependent on Document Volume & Network Bandwidth -** Migration speed may vary based on network throughput, SharePoint response time, and pipeline resource allocation.
6- **Signature/Seal Detection –** The system will detect only the presence of signatures/seals and will not validate authenticity, matching, or legal correctness.

Functional Requirements:

The system must perform the following functional capabilities as part of the Document Migration, Classification & Key Data Extraction solution:

FR01. SharePoint to ADLS Document Migration

- The system must migrate documents to ADLS while preserving the folder structure.
- The system must migrate all supported file formats (PDF, JPG, PNG, JPEG) with full processing.
- The system must migrate unsupported file types (Excel, Outlook files, password-protected files) without classification or key data.
- The system must retain original file names during migration unless exceptions are raised and mapped.

FR02. Document Classification

- The system must automatically classify documents using OCR/LLM algorithms.
- The business will provide a Key Identifier Document containing keywords or reference tags, which must be used as part of the classification logic to accurately determine document types.
- The system must verify the presence of a signature and/or seal to classify applicable documents as Signed and Not Signed.
- The system must generate a confidence score (%) for each classification result.
- Documents that cannot be confidently classified must be routed to the Exception folder.

FR03. Key Data (Metadata) Extraction

- The system must extract key metadata fields from eligible document types based on business-defined requirements.
- The system must generate average confidence score (%) for key data extraction.
- The system must generate a Completion Score, indicating the percentage of required metadata fields successfully extracted out of the total expected fields.
- The system must identify and classify document into below document types and extract Key data from the document

Sno	Document classification	Attributes to Extract
1	Booking form -Signed	Primary Applicant Name
		Project
		Unit Number
2	Booking form -Not Signed	Primary Applicant Name
		Project
		Unit Number
3	SPA- Signed	Purchaser Name

		Community
		DLD Unit no
4	SPA- Not Signed	Purchaser Name
		Community
		DLD Unit no
5	Receipt	Project
		Unit No
		Amount
6	Emirates ID	ID Number
		Name
		Nationality
		Expiry Date
7	Passport	Passport No
		Name
		Date of expiry
8	Visa Copy	U.I.D No
		Name
		Expiry Date
9	Quotation	
10	File Note Document	
11	DEWA Bill	Reference Number
		Premise Number
12	Feedback Form	
13	Handover of Possession	Unit No
14	Initial Contract of sale	Contract No
		Participant Name
15	Agreement of Property Sale	
16	Checklists	Unit No
		Name
17	Undertaking Letter	Reference
18	Mortgage Receipt	Name
		Receipt Number
		Unit Number
19	Mortgage Contract	
20	Resale NOC	Project Name
		Unit Number
		Buyer Details - Name
21	Transfer Letter	Unit No
22	TAA	unit number
23	Resale Contract	Contract Number
		Buyer Name
		Contract Number
		Buyer Details
24	Handover Confirmation	
25	Authorization for Handover	Unit Number

FR04. Confidence-Based Routing

- The system must route documents based on Classification confidence thresholds:
    - High Confidence (≥ defined threshold) → Success folder
    - Low Confidence (< defined threshold) → Exception folder
- Confidence thresholds will be finalized and configurable based on business approval.

FR05 Exception Handling

- The system must route the following exceptions to the designated exception folder:
    - Unclassified Document
    - Documents with low classification confidence
    - Corrupted Files
- The system must generate notification emails for exceptions via SMTP.
- The system must generate Power BI-based exception reports using classification logs.
- If a document is placed in the Exception folder, the system must support the following two resolution options:
    - **Manual Classification by Business Users** - The business can manually assign the document type by providing a CSV file containing the list of exception documents and their corrected document types. Once the CSV is uploaded, the system must reprocess the exception documents using the updated rules.
    - **Miscellaneous Files** – If the business is unable to classify a document and the document type remains blank in the provided CSV; the system must classify the file as Miscellaneous and move it to the respective Unit's Miscellaneous folder in ADLS.
- The system must allow reprocessing of failed or exception files without duplicating files.

FR06. JSON Metadata Generation

- For every processed file, the system must generate a JSON file containing:
    - Document Classification result.
    - Document Classification Confidence score.
    - Extracted key metadata fields.
    - Metadata Extraction Avg Confidence score
    - Metadata Extraction Completion Score
    - Source and destination folder path.
    - Timestamp of processing

- JSON must be stored in the same destination folder as the original document.

FR07. API-Based Consumption

- The system must expose processed documents and their JSON metadata through secured API endpoints.
- The system must allow Salesforce to retrieve documents and metadata using these APIs.
- The system must support asynchronous processing using Task IDs for document retrieval.
- When a user requests documents for a Unit and selected document types, the system must generate a Task ID representing the retrieval request.
- Unit Mapping will be used at the API layer to translate Salesforce Unit Names to their matching ADLS Unit folders. No folder renaming will occur in ADLS.
- A separate API must allow the user to check the status of a Task ID (In Progress, Completed, Failed).
- Once processing is completed, the user must be able to retrieve the document payload and its JSON metadata using the Task ID.
- The API must support acknowledgment of completed tasks to maintain audit traceability.

FR08. Reconciliation & Reporting

- The system must provide reconciliation reports capturing:
    - Processing Date
    - Total files processed
    - Successful migrations
    - Exception files
    - Skipped Files
    - JSON generation count
- Report output must be made available via Power BI Dashboard

Non-Functional Requirements

- Security: Authentication must be managed through Azure Active Directory (AAD), and all secrets/credentials must be securely stored in Azure Key Vault.
- Monitoring: System failures, exceptions, and processing statistics must be monitored through Power BI reports, and critical alerts must be communicated via Outlook email notifications.
- Performance: Processing throughput and SLA (migration speed, classification accuracy, batch size) must be mutually agreed and configurable based on document volume.

- Compliance: Extracted metadata must support compliance checks such as signature/seal detection (presence only) where applicable.
- Audit: All processed files must include a JSON metadata file stored in ADLS to maintain full audit history, traceability, and document lineage.

Risk

- Delays in Business Inputs – Delays in providing mapping files, Key Identifier rules, metadata extraction rules, sample documents, or required system access may impact project timelines and downstream activities.
- Exception Handling Delays – Delays from the business in reviewing and resolving exception documents may impact the overall processing cycle and completion timelines.
- BRD Sign-off Delays – Any delay in BRD sign-off may postpone implementation activities, impacting the overall project schedule.

Key Assumptions & Dependencies

- Sobha will provide required SharePoint access (read + download) for VM, PAM access, Unit Mapping file, and sample documents for each supported document type before implementation begins.
- The metadata fields to be extracted and the classification confidence thresholds will be finalized by the business before development begins.
- Sobha's IT/Identity team will provide azure AD App Registration, SMTP-enabled email account, Power BI Pro license, and Key Vault access.
- The Salesforce team will support API consumption of processed documents and metadata stored in ADLS.
- Any changes in document types, metadata rules, exception logic, or folder mapping after BRD sign-off will follow a formal Change Request (CR) process.
- Business stakeholders will validate extracted metadata, approve exception categories, and sign off the final outputs for UAT and production release.

Key Stakeholder- Sign Off:

Name	Role, Organization	Reviewed	Approval
Mohan Shanmugam	Senior Manager - IT - Digital Products & CX, Group IT, Sobha Realty		
Akash Lall	Assistant Vice President - IT - Digital Products & CX, Group IT, Sobha Realty		
Sunny Pradhan	Manager – CRM, CRM, Sobha Realty		
Kapil Shah	General Manager – CRM, CRM, Sobha Realty		
Gautam Sawhney	Chief Customer Experience Officer, CRM, Sobha Realty		