IBM GEN AI REPORT

Project Proposal: AI-Powered Document Digitization and Analysis using IBM Watsonx

VIT Bhopal University

Name : Asmita Sarkar
Registration Number : 22BSA10100
Email : asmitasarkar2022@vitbhopal.ac.in

AI-Powered Document Digitization and Analysis using IBM Watsonx

---

1. 1. Problem Statement

In today's digital-first enterprises, organizations handle massive volumes of documents—PDFs, scanned files, forms, reports, and contracts—that often remain locked in unstructured formats. These documents hold valuable information but are not easily searchable or analyzable, impeding automation, compliance, and decision-making.

Manual extraction of data is labor-intensive, error-prone, and inefficient. There is a pressing need for an automated solution that enables enterprises to extract structured data from unstructured documents quickly and accurately.

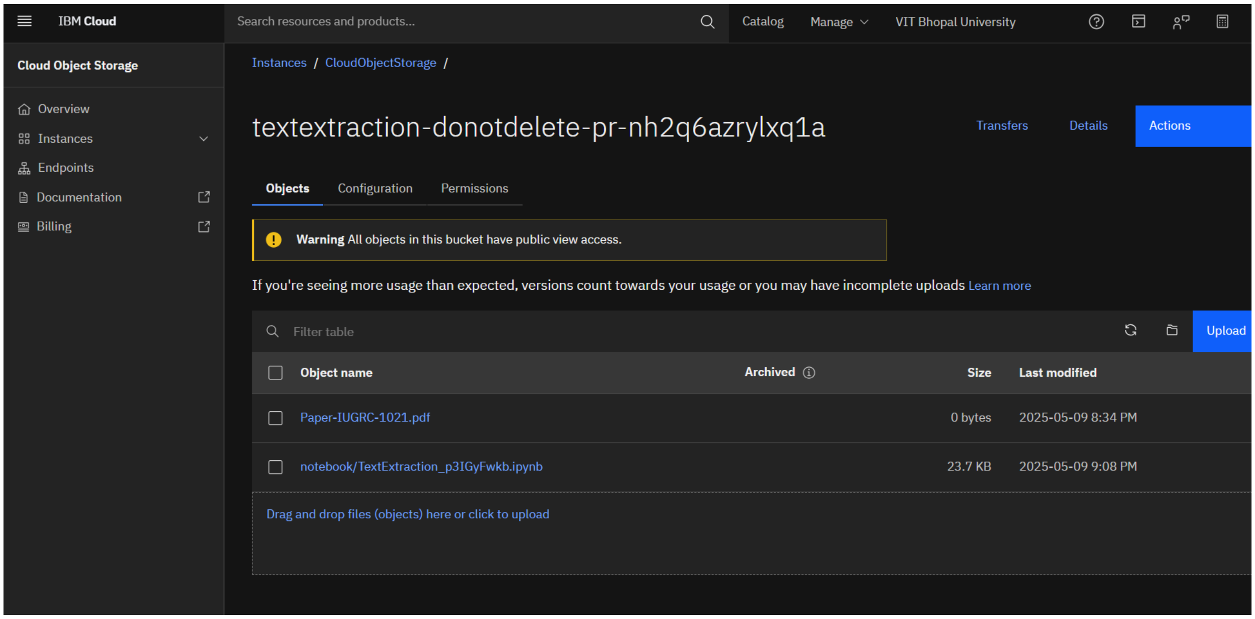
---

2. Project Objective

To develop an AI-powered intelligent document analysis system that utilizes the IBM watsonx.ai Text Extraction API to automate the extraction of relevant text and data from PDFs and scanned documents. The system will:

- Upload documents to IBM Cloud Object Storage
- Extract structured data using Watsonx
- Export results in machine-readable formats (e.g., JSON/Markdown)

- Enable integration into enterprise workflows (e.g., analytics, automation)



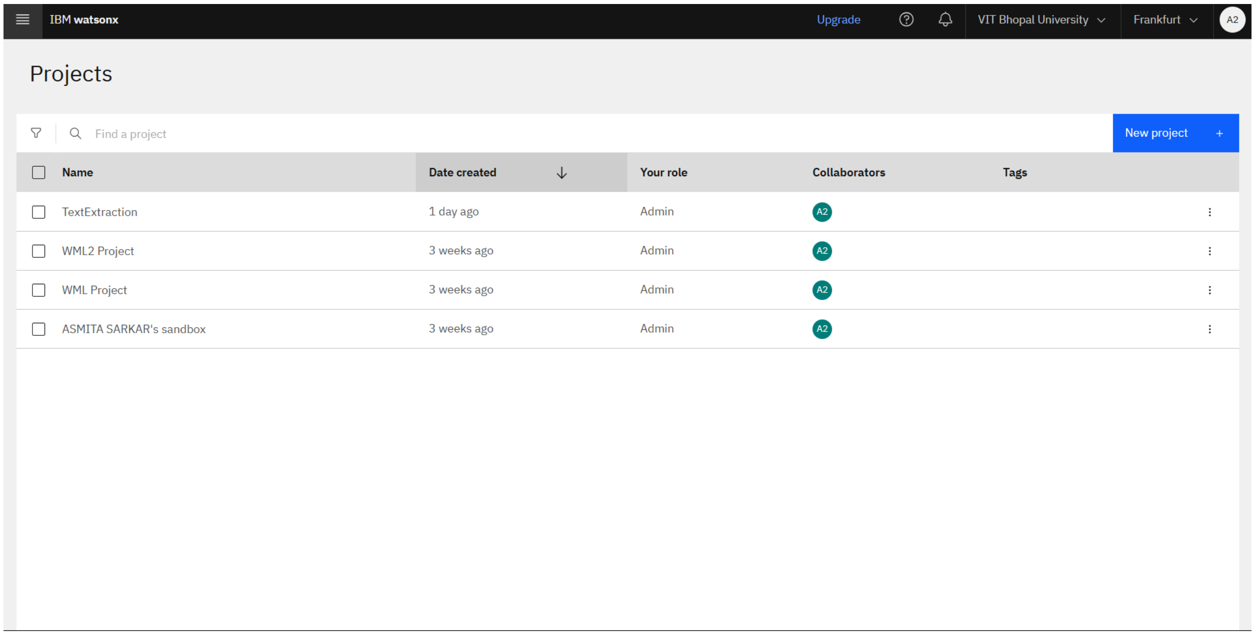
textextraction-donotdelete-pr-nh2q6azrylxq1a

Warning All objects in this bucket have public view access.

If you're seeing more usage than expected, versions count towards your usage or you may have incomplete uploads Learn more

Object name	Archived	Size	Last modified
Paper-IUGRC-1021.pdf		0 bytes	2025-05-09 8:34 PM
notebook/TextExtraction_p3IGyFwkb.ipynb		23.7 KB	2025-05-09 9:08 PM

Drag and drop files (objects) here or click to upload



Projects

Name	Date created	Your role	Collaborators	Tags
TextExtraction	1 day ago	Admin	A2	
WML2 Project	3 weeks ago	Admin	A2	
WML Project	3 weeks ago	Admin	A2	
ASMITA SARKAR's sandbox	3 weeks ago	Admin	A2	

3. Proposed Solution Overview

- We propose a cloud-based system that employs IBM Cloud Object Storage and watsonx.ai Text Extraction API to automate document intelligence workflows.
- **Key Functionalities:**

Feature	Description
Upload Interface	**Drag-and-drop or programmatic upload of documents to a secure COS bucket**
Text Extraction Engine	**Uses watsonx foundation model to extract plain text, tables, and layout**
JSON Conversion	**Output structured data in JSON for easy integration with downstream tools**
Analytics-Ready	**Searchable, filterable, and analyzable text format from raw PDFs**
Python SDK	**A backend script that connects, uploads, extracts, and downloads data**

4. Innovation and Technical Relevance

Innovation:

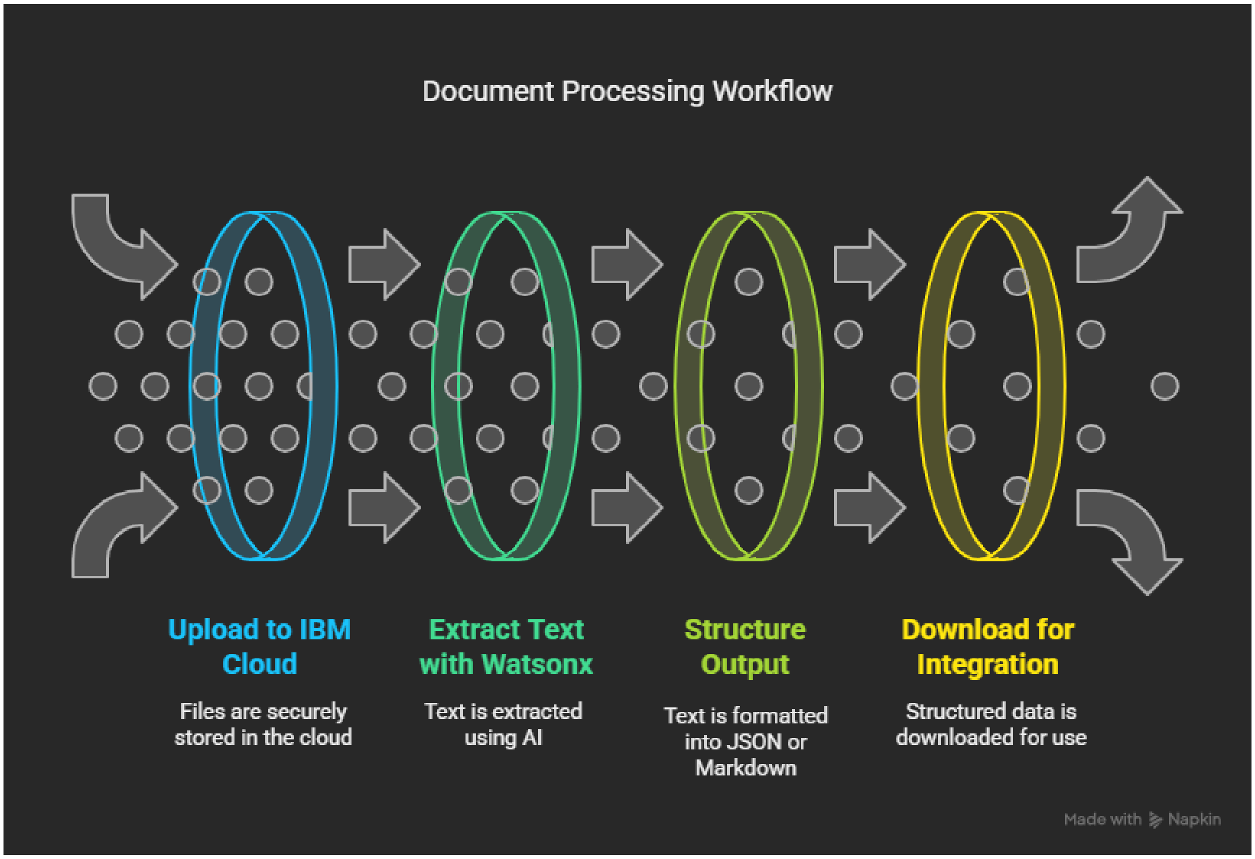
- Uses cutting-edge foundation models on IBM watsonx.ai for intelligent OCR and NLP.
- Converts high-volume unstructured documents into structured datasets.
- Supports multiple input formats, including scanned PDFs, with multi-language capability.

Technologies:

- IBM Cloud Object Storage (COS): Stores uploaded documents and extracted outputs.
- watsonx.ai Text Extraction API: Extracts textual and tabular information from documents.
- Python SDK: Interfaces the extraction process using programmatic control.

5. System Architecture

Architecture Diagram:


Document Processing Workflow
Upload to IBM Cloud
Files are securely stored in the cloud
Extract Text with Watsonx
Text is extracted using AI
Structure Output
Text is formatted into JSON or Markdown
Download for Integration
Structured data is downloaded for use
Made with Napkin

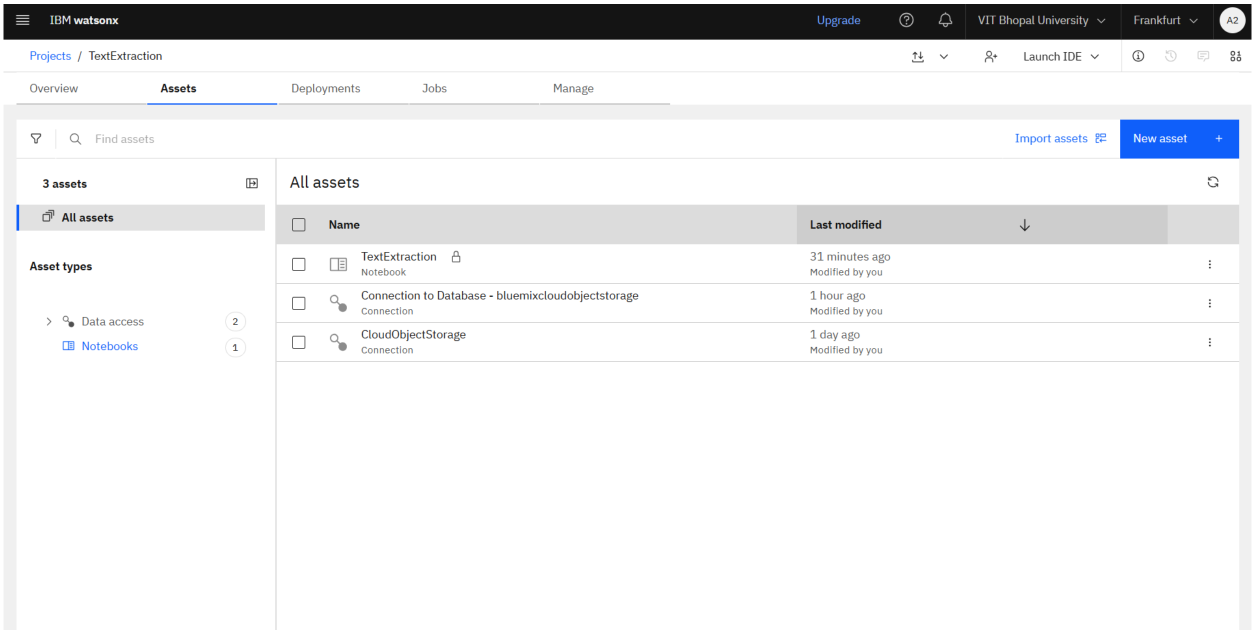
6. Implementation Steps

Step 1: Setup Cloud Infrastructure

- Create an IBM Cloud Object Storage bucket.
- Generate HMAC credentials and connect watsonx.ai with the bucket.

Step 2: Upload Document(s)

- Use the IBM COS web interface or Python `ibm_boto3` library to upload files.

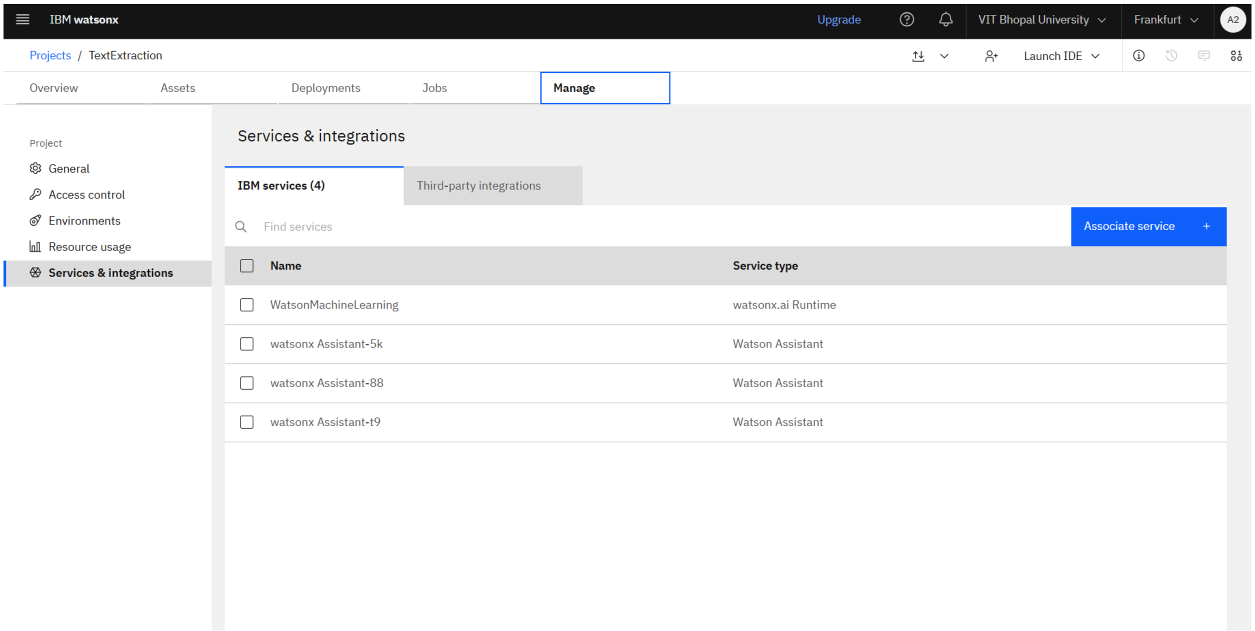


All assets

Name	Last modified
TextExtraction Notebook	31 minutes ago Modified by you
Connection to Database - bluemixcloudobjectstorage Connection	1 hour ago Modified by you
CloudObjectStorage Connection	1 day ago Modified by you

Step 3: Text Extraction Process

- Authenticate using API Key and Project ID.
- Initialize the Watsonx client and connect the COS bucket.
- Use the `TextExtractions.run_job()` method with custom parameters (language, table extraction, OCR).
- Store results back in the COS bucket.



Services & integrations

Name	Service type
WatsonMachineLearning	watsonx.ai Runtime
watsonx Assistant-5k	Watson Assistant
watsonx Assistant-88	Watson Assistant
watsonx Assistant-t9	Watson Assistant

Step 4: Download and Process Results

- Parse JSON output locally.
- Enable integration with analytics dashboards or reporting tools.

7. Sample Code Snippet (Python)

python

```python
from ibm_watsonx_ai import APIClient, Credentials

from ibm_watsonx_ai.foundation_models.extractions import TextExtractions

from ibm_watsonx_ai.helpers import DataConnection, S3Location

from ibm_watsonx_ai.metanames import TextExtractionsMetaNames

import ibm_boto3

credentials = Credentials(url=″https://us-south.ml.cloud.ibm.com″, api_key=″your_api_key″)

client = APIClient(credentials, project_id=″your_project_id″)



cos_client = ibm_boto3.client(

   service_name=′s3′,

   aws_access_key_id=′access_key′,

   aws_secret_access_key=′secret_key′,

  endpoint_url=′https://s3.us-south.cloud-object-storage.appdomain.cloud′

)

document_reference = DataConnection(connection_asset_id=″conn_id″, location=S3Location(bucket=″bucket″, path=″file.pdf″))

results_reference = DataConnection(connection_asset_id=″conn_id″, location=S3Location(bucket=″bucket″, path=″file.json″))

extraction = TextExtractions(api_client=client, project_id=″your_project_id″)

steps = {TextExtractionsMetaNames.OCR: {′language_list′: [′en′]}, TextExtractionsMetaNames.TABLE_PROCESSING: {′enabled′: True}}

extraction.run_job(document_reference=document_reference, results_reference=results_reference, steps=steps)
```

8. Expected Outcomes

- Efficient Data Processing: Significant reduction in manual labor and processing time.
- Structured Intelligence: Conversion of unstructured document data to structured formats.
- Seamless Integration: API-driven approach enables integration with enterprise workflows (RPA, analytics, BI dashboards).
- Scalability: Cloud-native solution capable of handling large datasets.

9. Use Cases

1. Healthcare – Extract patient data from prescriptions or medical reports.
2. Finance – Digitize and analyze contracts, invoices, and statements.
3. Legal – Structure legal documents and case files.
4. Education – Extract data from scanned academic records and certificates.

10. Conclusion

This project presents a robust, scalable, and intelligent solution to one of the most persistent problems in digital transformation: extracting meaning from documents. Using IBM watsonx.ai, we automate the transition from raw unstructured content to structured, usable data—improving efficiency, accuracy, and enabling next-gen enterprise workflows.