# Cloud-Based Document Management System with AWS Elasticsearch

*A Project Based Learning Report Submitted in partial fulfilment of the requirements for the award of the degree*

*of*

**Bachelor of Technology**

**in the Department of Computer Science & Engineering**

**Cloud Based AI/ML Speciality (22SDCS07A)**

Submitted by
**2210030070: Deekshitha Bethireddy**

Under the guidance of

**Ms. P. Sree Lakshmi**


Department of Computer Science and Engineering

Koneru Lakshmaiah Education Foundation, Aziz Nagar

Aziz Nagar – 500075

March - 2025.

# Introduction

With the rapid increase in digital data, businesses and organizations require efficient document management solutions. A Cloud-Based Document Management System (DMS) leverages cloud infrastructure to store, manage, and retrieve documents efficiently. It provides scalability, cost-effectiveness, and security, making it a preferred choice for modern businesses. Traditional document storage methods often pose challenges such as limited accessibility, high maintenance costs, and security vulnerabilities. By adopting cloud-based solutions, organizations can overcome these limitations while ensuring seamless collaboration, workflow automation, and improved efficiency. Additionally, cloud-based DMS offers better data backup, version control, and integration with other business applications, enhancing overall productivity.

## AWS Services

1. **AWS S3:** Stores and manages documents securely with scalable cloud storage.
2. **AWS OpenSearch (Elasticsearch):** Enables fast document indexing and full-text search.
3. **Amazon Lambda:** Automates indexing by triggering OpenSearch updates when a document is uploaded.
4. **AWS API Gateway:** Provides a REST API for querying OpenSearch.
5. **Amazon Textract:** Extracts text from scanned documents for better searchability.
6. **Amazon CloudWatch** – Monitors system performance and Lambda execution.
7. **AWS IAM** – Ensures secure access control and authentication.

These services collectively enable a secure, scalable, and automated cloud-based document management system

## Project Purpose and Expected Outcome

The project aims to develop a Cloud-Based Document Management System using AWS OpenSearch, enabling efficient document storage, management, and retrieval. By integrating automation, security, and scalability, businesses can streamline document handling, enhance accessibility, and improve search efficiency.

- **Automated Document Indexing:** AWS Lambda triggers automatically index uploaded documents in OpenSearch.
- **Secure Storage & Access:** Amazon S3 stores documents securely, while AWS IAM controls permissions.
- **Efficient Search & Retrieval:** AWS OpenSearch enables fast and accurate document searches.
- **Scalable & Serverless Architecture:** API Gateway and AWS Lambda ensure seamless interaction with OpenSearch.

# Methodology

## Architecture and Workflow

The architecture consists of Amazon S3 for document storage, AWS OpenSearch for indexing and searching, AWS Lambda for automation, API Gateway for API exposure, IAM for authentication, and **CloudWatch** for monitoring.The workflow includes:

1. **Document Upload**: Users upload documents to an S3 bucket, ensuring secure cloud storage.
2. **Automatic Indexing**: S3 triggers AWS Lambda, which extracts text from documents and indexes them in AWS OpenSearch.
3. **Search API**: Users can search for documents via API Gateway, which connects to OpenSearch for fast retrieval.
4. **Authentication & Access Control**: AWS IAM ensures only authorized users can access and manage documents.
5. **Real-time Monitoring**: Amazon CloudWatch tracks API calls, Lambda executions, and search performance for optimization.

This architecture ensures scalable, automated, and secure document management, allowing businesses to efficiently store, index, and retrieve documents while improving workflow automation and accessibility.

## AWS Services Interaction

- **Amazon S3** stores documents securely and triggers AWS Lambda when a new file is uploaded.
- **AWS Lambda** extracts document content and indexes it into AWS OpenSearch for efficient searching.
- **AWS OpenSearch** enables fast retrieval of documents using metadata and full-text search.
- **AWS IAM** controls authentication and access permissions, ensuring secure document management.
- **Amazon API Gateway** provides a RESTful API for users to search and retrieve documents from OpenSearch.
- **Amazon CloudWatch** monitors system performance, tracking API calls, Lambda executions, and search analytics.

This interaction ensures a fully automated, secure, and scalable document management system, allowing seamless document indexing, retrieval, and access control.

## Justification for AWS Service Selection

The AWS services were selected for their scalability, security, and automation. Amazon S3 ensures secure document storage, AWS OpenSearch enables fast retrieval**,** Lambd**a** automates indexing, and API Gateway provides seamless access. IAM controls permissions, and CloudWatch monitors performance. This combination ensures a cost-effective, and high-performance document management system.

# Implementation

## AWS Infrastructure Setup

To set up the Cloud-Based Document Management System with AWS OpenSearch, the following AWS services are utilized:

- **Create Amazon S3 Bucket for Storage:** Stores uploaded documents securely, ensuring high availability. Public access is blocked to maintain security.
- **Set Up AWS IAM for Access Control:** Manages user authentication and enforces role-based permissions to protect sensitive data.
- **Deploy AWS OpenSearch for Indexing**: Enables fast and efficient full-text search, allowing users to retrieve documents quickly.
- **Develop AWS Lambda for Automation**: Automatically extracts document content upon upload and indexes it in OpenSearch, eliminating manual effort.
- **Integrate API Gateway for Search Access**: Provides a RESTful API that enables users to search and retrieve documents from OpenSearch seamlessly.
- **Enable CloudWatch for Monitoring**: Tracks system performance, logs API calls, and monitors search efficiency for optimization.

## Security Policies, IAM Roles, and Access Controls

- **IAM Roles & Policies:** Access is restricted based on roles such as Admin and User, ensuring the principle of least privilege. Only authorized users can manage or modify documents.
- **S3 Bucket Permissions:** Documents are stored in Amazon S3 with private access controls, preventing unauthorized access. Public access is blocked for security.
- **OpenSearch Access Control:** IAM policies restrict access to only approved users, ensuring that only authorized roles can query and index documents.
- **Lambda Execution Security**: IAM policies control which roles can execute AWS Lambda functions, preventing unauthorized modifications to backend logic.
- **API Gateway Security**: API access is restricted to authenticated users, preventing unauthorized queries.
- **Data Encryption**: Data in S3, OpenSearch, and API transmissions is encrypted at rest and in transit for security compliance.

These security measures ensure document integrity, controlled access, and protection against unauthorized modifications

## Automation

Automation in the Cloud-Based Document Management System ensures efficient operations with minimal manual intervention.

- **Automatic Lambda Triggering:** AWS Lambda automatically extracts content from documents uploaded to S3 and indexes them in OpenSearch.

- **Continuous Deployment:** AWS CodePipeline automates the deployment of Lambda functions, API Gateway, and OpenSearch updates.
- **Automated Testing**: AWS CodeBuild validates Lambda functions and API integrations before deployment.
- **CloudWatch Monitoring**: Amazon CloudWatch tracks system performance, logs errors, and sets alerts for failures.
- **CloudWatch Monitoring:** Amazon CloudWatch tracks performance, logs system errors, and sets up automated alerts for system anomalies.
- **Infrastructure as Code (IaC)**: AWS CloudFormation automates provisioning for S3, OpenSearch, Lambda, and IAM roles.