 A picture containing logo

Description automatically generated

Using Amazon Kendra for NLP Searching and Indexing

**About ICAI**

**About Challenge**

ICAI wanted a Search Engine functionality (Cognitive Searching) for contents of the icai.org on which Search is to be enabled (FAQs/HTML/Word/ppt/pdf/image-pdf). The main challenges faced by the customerin their current environment is that there is no NLP based Search based functionality available.

So, the customer desires highly robust and cost-effective solution capable of delivering a quality end user experience regardless of the demands on the platform using best in breed AWS services to achieve this goal.

To retrieve relevant content using semantic search based on the concept of Natural Language processing (NLP), the major challenges we faced are as below:

* NLP based Cognitive Searching of the contents
* User should be able to perform NLP based Searching (Cognitive Searching) based on the contents of Publications, Journals, Students FAQs, Members FAQs stored in Amazon S3
* Enabling user level restriction for each category

The Institute of Chartered Accountants of India (ICAI) is a statutory body established by an Act of Parliament, viz. The Chartered Accountants Act, 1949 (Act No.XXXVIII of 1949) for regulating the profession of Chartered Accountancy in the country. The Institute, functions under the administrative control of the Ministry of Corporate Affairs, Government of India. The ICAI is the second largest professional body of Chartered Accountants in the world, with a strong tradition of service to the Indian economy in public interest.

A picture containing text, clipart, building material

Description automatically generated Icon

Description automatically generated A picture containing text, clipart, table, stand

Description automatically generated

Icon

Description automatically generated Icon

Description automatically generated

**AWS Services used**

API Gateway

Cloudwatch

AWS Kendra

AWS S3

AWS Lambda

MIND team analyzed the problem and the approach for ICAI Kendra Searching and Indexing is described below.

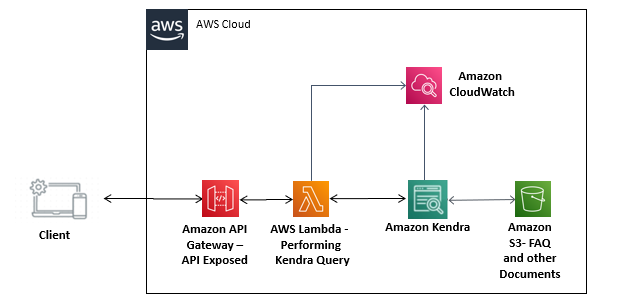
* ICAI account has an S3 bucket having category-wise documents in its subfolders and category-wise searching is to be performed. Every Kendra index comprises of configuring the data sources that connect to the S3 buckets having document store.
* We have controlled access to documents in an S3 data source using a configuration file. We specify the file in the console, a configuration file contains a JSON structure identifying S3 prefix and lists the access settings for the prefix.
* AWS Lambda to run integrated code serverless forbackend to call Kendra API’s. The [AWS Lambda](https://aws.amazon.com/lambda) function serves the API Gateway parses the path parameters and headers and issues an Amazon Kendra query call with AttributeFilters set to the username from the path parameter, the user access level, and category from the headers.
* Amazon Kendra returns the FAQs and documents for that particular category and filters them by the user access level and category. The Lambda function constructs a response with these search results and sends the FAQ and document search results back to the client application. Tagging is implemented for optimized results and improved relevancy.
* Deployment of Flask application to Elastic Beanstalk is done. Flask is basically an open-source web application framework for Python which is deployed to an AWS Elastic Beanstalk environment.

**Proposed Solution**

The user is now able to perform NLP based Searching (Cognitive Searching) based on the contents of Publications, Journals, Students FAQs, Members FAQs stored in Amazon S3. The search display pdf page number along with pdf link allowing the user to perform Restrictive Search within a selected topic / sub-topic. The optimization technique of tagging resulted in 66% improved results.

**Solution Outcome**

**Architecture Diagram**



**How AWS services helped in building ETL pipeline**

**AWS Lambda to run integrated code**

AWS Glue is used as serverless for backend to call Kendra API’s. managed ETL service for data transformation and data ingestion and in this solution all the major transformations are done using Glue ETL scripts.

**Amazon Kendra**

Amazon Kendra is an intelligent search service powered by machine learning (ML). This helped in creation of indexes and performing the NLP queries by sending requests.

**CloudWatch for monitoring**  
Amazon CloudWatch is a monitoring and observability service which keeps a tab on the set of services being used in the form of logs, metrics, and events.

**Amazon API Gateway**

Amazon API Gateway is a fully managed service that makes it easy for developers to create, publish, maintain, monitor, and secure APIs at any scale. Here it is used to get requests from the frontend and send it to lambda function.

**Amazon S3 for storage**

Amazon S3 is cloud object storage with industry-leading scalability, data availability, security, and performance. Here, the S3 Bucket is also used here to store the document set and FAQs of several categories.

.

**About the Partner**

**MothersonSumi INfotech &Designs Ltd.**

MothersonSumi INfotech & Designs Limited (MIND), a SEI CMMI Level 5 IT services company and the IT back bone of Motherson group. MIND is a trusted technology partner to over 200 clients globally. Our value proposition is in our strength in specific Industry segments and years of experience in the areas of intelligent warehousing, Supply chain enablement, software application development, smart ERP customization, infra managed services, cloud, IoT & Analytics. MIND is serving customers in 41+ countries with a strong team of 1500+ professionals.