iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** II **Month of publication:** February 2024

DOI: https://doi.org/10.22214/ijraset.2024.58394

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue II Feb 2024- Available at www.ijraset.com

Document Management System - AWS Comprehend, Generative AI and NoSQL DB

Niharika Tiwari[1], Prashansa Srivastava[2], Asst. Prof. Dr. Sadhana Rana[3]
[1, 2, 3]*CSE Department SRMCEM Lucknow, India*

Abstract: ***Classical document management systems like Microsoft word and Google Docs allow users to retrieve, make, improve and store documents using on device or cloud-based storage mechanisms with One-drive and Google Docs respectively. The growth in the amount of data that needs to be processed in organizations calls for a centralized system that allows users to store, retrieve, analyze, create and improve documents. To solve this, we have built a centralized system that allows users to create and improve documents using generative ai, store documents in a NoSQL based database MongoDB and analyze their documents using AWS Comprehend.***

Keywords: Generative AI, Intelligent Analysis, Natural Language Processing, AWS Comprehend, NoSQL, MongoDB, Intelligent Analysis

I. INTRODUCTION

Generative AI has recently taken the entire tech industry by storm. It has added new dimensions to how we create content. This has a direct impact on the content creator economy [1]. The impact of the technology is yet to be seen in its integration with existing technologies like simple document management systems. In this system we have integrates generative ai with AWS Comprehend to build one such document management system. Both generative ai and AWS Comprehend inherently use natural language process that provides improved user experiences [2].

Smart document analysis using AI/ML has for long been in use. This can be seen in its impact in syntax detection, sematic detection and PII detection systems [3]. This is an added layer on top of NLP Systems.

Also, we see that in the recent years, cloud computing technology has witnessed a huge boom in terms of its adoption. Everywhere around us we see organizations migrating to cloud platforms. Cloud makes traditional systems like storage and processing limitless when they are delivered as services via the internet [4].

This technology can be added to simple applications to give an impression of limitless storage to the user of the application. Addition of cloud-based storage systems allows for easy access of resources on a remote basis and manage large amounts of data [5].

Following the introduction of study in this section, Section 2 describes the literature review, Section 3 explains the methodology. Section 4 presents module description and their working. Section 5 discusses the results of the work followed by the conclusion and the future scope of this work.

II. LITERATURE REVIEW

AIGC is a field of computer science that focuses on the development of systems that can generate content, such as text, images, and music, autonomously. AIGC systems are trained on large datasets of existing content, and they use this training data to learn the patterns and rules that govern the generation of new content.

AIGC systems have become increasingly sophisticated in recent years, and they are now able to generate content that is indistinguishable from human-generated content in many cases. AIGC is being used in a variety of industries, including art, advertising, and education.

Here are some of the recent advances in the field of AIGC:

The development of new generative AI models, such as ChatGPT, which are able to generate more realistic and high-quality text than previous models.

The development of new applications for AIGC, such as the use of AIGC to generate personalized educational content for students.

Smart Document Classification

Smart document classification is a process of using AI and machine learning to automatically classify documents into different categories. This can be useful for a variety of tasks, such as organizing documents in a digital library or filtering out spam emails.

Here are some of the machine learning algorithms that are commonly used for smart document classification:

1) Support vector machines (SVMs)
2) Naive Bayes classifiers
3) Decision trees
4) Random forests
5) Gradient boosting machines

AIGC and smart document classification are two rapidly developing fields with a wide range of applications. AIGC can be used to generate personalized content for users, while smart document classification can be used to organize and filter large volumes of data. As these technologies continue to develop, we can expect to see them used in even more innovative and impactful ways in the future.

Cloud storage is a model of data storage in which the digital data is stored in logical pools, the physical storage of which is spread across multiple servers (potentially in different locations) and is typically managed by the cloud storage service provider.

III. PROPOSED METHODOLOGY

Our methodology to develop this document management system consists of a layered architecture. This pattern segregates the responsibilities of the different layers of code, ensuring security, scalability, and easy feature additions.

The layered architecture will consist of the following layers:

1) *Presentation layer:* This layer will be implemented using React, providing a user-friendly interface for users to manage their work on the cloud.
2) *Application layer:* This layer will contain the business logic of the application, such as processing user requests and generating responses. It will also use the AWS SDK for Java to establish a connection to the cloud and access the necessary services.
3) *Data layer:* This layer will be responsible for accessing and managing data, such as storing and retrieving data from a database. It will also use the AWS SDK for Java to interact with AWS services.

Each layer will be decoupled from the other layers, making the code more modular and reusable. This will also make it easier to scale the application and add new features in the future.

IV. MODULE DESCRIPTION

A. *Prompt Module*

The prompt module is responsible for handling all generative AI related content. It provides a number of features, including:

1) *Prompt generation:* The prompt module can generate prompts for a variety of generative AI tasks, such as text generation, image generation, and code generation.
2) *Prompt editing:* The prompt module allows users to edit prompts to improve their accuracy and specificity.
3) *Prompt evaluation:* The prompt module can evaluate prompts to identify language, tone and sentiments.

B. *Comprehend Module*

The comprehend module is a powerful tool for analysing and understanding large amounts of text data using AWS Comprehend. It makes it easy to extract entities, identify sentiment, model topics, and extract key phrases.

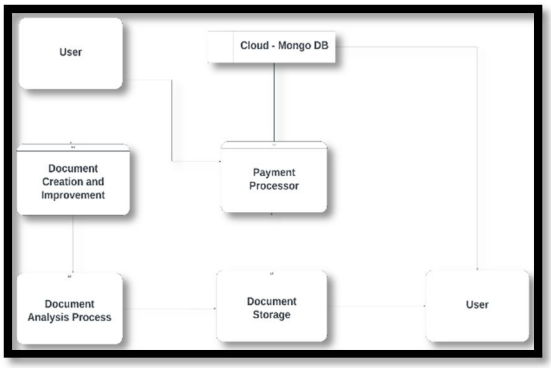
The comprehend module can be used for a variety of tasks, such as:

1) *Customer insights:* Analysing customer feedback to identify trends and patterns.
2) *Market research:* Analysing social media data to understand public opinion about a product or service.
3) *Risk analysis:* Analysing financial data to identify potential risks.

C. *Storage Module*

The storage module is responsible for providing cloud storage to users using MongoDB. It provides a number of features, including:

1) *Document storage:* The storage module can store documents in MongoDB, which is a NoSQL database that is well-suited for storing and managing large amounts of data.
2) *Document retrieval:* The storage module can retrieve documents from MongoDB quickly and efficiently.
3) *Document management:* The storage module allows users to manage their documents in MongoDB, such as creating, updating, and deleting documents.



V. RESULTS

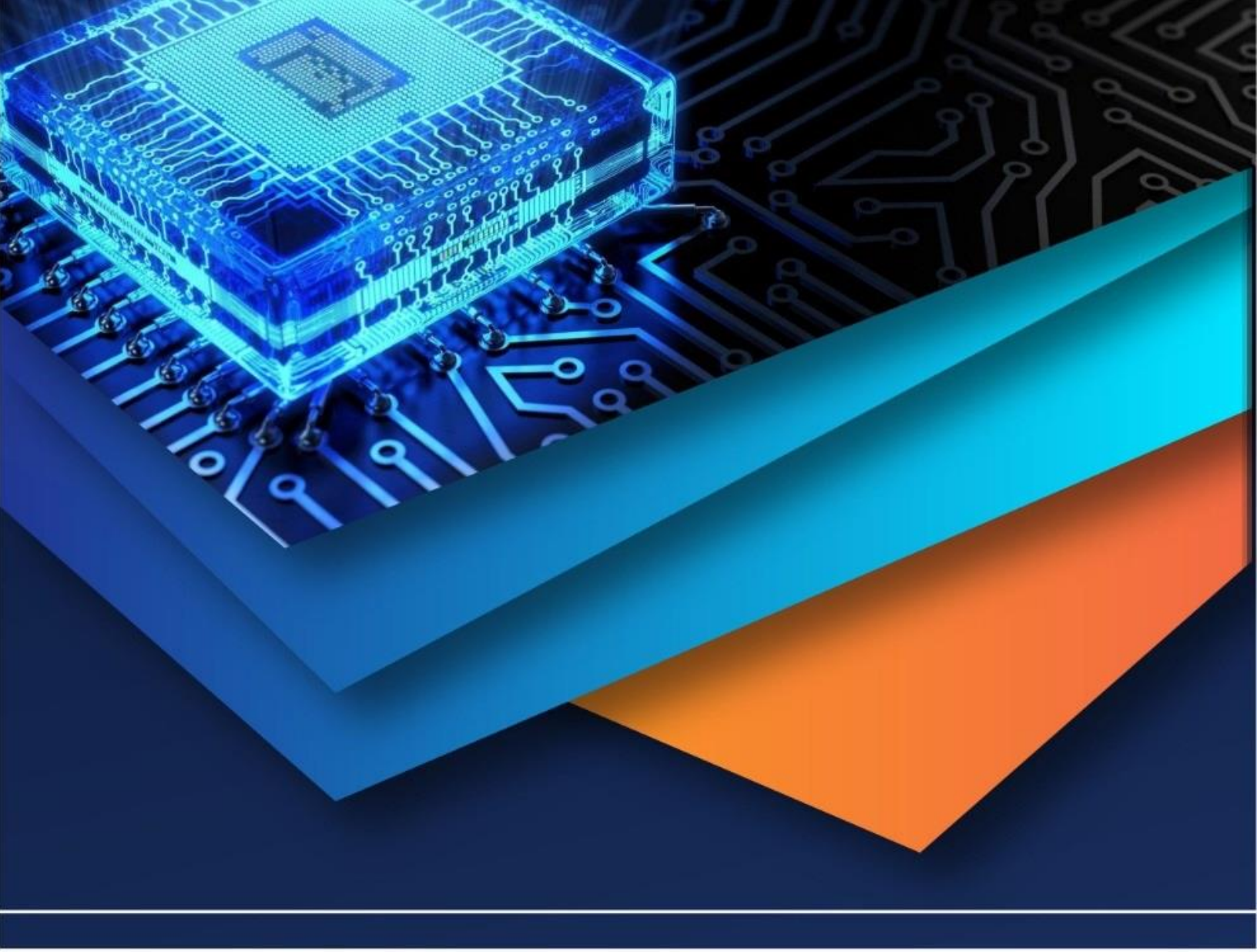
The performance of generative ai to generate documents and improve them depends largely on the api that they use in the background. Since these systems are in active development at organizations like Open AI and Google Gemini, we can expect a massive improvement from time to time. Similarly, the performance of analysis features depends upon the accuracy of the comprehend api by Amazon Web Services. The storage of documents depends upon our choice for MongoDB or PostgreSQL.

VI. CONCLUSION

The scope of expanding upon this field of work is huge. This involves adding technologies like AWS Quicksight for data visualization, DALL-E models for generative ai image processing, and other analysis models that we were limited to use by the free-tier of Amazon Web Services [7].

REFERENCES

[1] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4670714
[2] https://link.springer.com/chapter/10.1007/978-3-031-48057-7_14
[3] https://www.ijircst.org/DOC/6_irp681.pdf
[4] https://www.researchgate.net/publication/257723864_Research_on_Cloud_Data_Storage_Technology_and_Its_Architecture_Implementation
[5] https://typeset.io/papers/enterprise-web-based-file-management-a-system-architecture-o51krvvtkz
[6] https://aws.amazon.com/comprehend/
[7] https://aws.amazon.com/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY