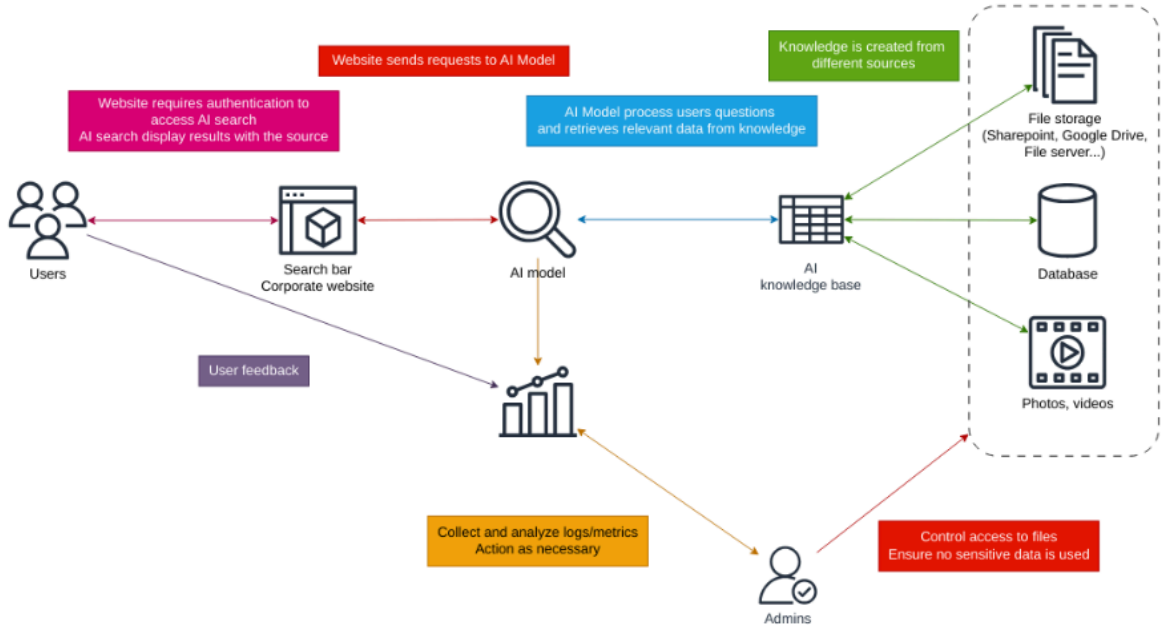
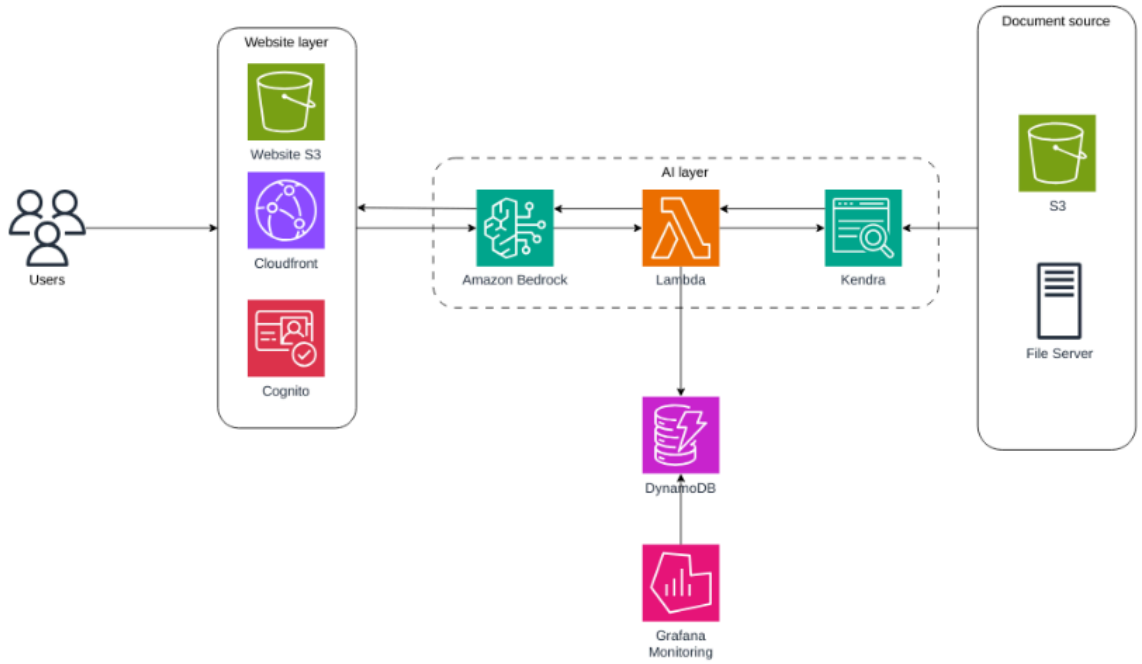
# Capstone Project: Part 1

## 1. Conceptual design




## 2. High level diagram



# 3. Project risks

## a. Access to sensitive data

As we are trying to retrieve data from different locations, we will need to ensure that the relevant corporate processes are in place. Restricted access to the location RAG is reading from, continuous monitoring and file reviews.

## b. Users mis-use

Malicious users could get access to the search engine, we need to make sure that the system won't divulge sensitive data, or allow mis-use of the system (i.e. ask irrelevant questions). Guardrail will be configured and regularly tested/monitored to be sure this kind of behaviour won't happen.

## c. Cost and up-time

System will be monitored and use auto-scaling rules to make best use of serverless capabilities of AWS. Rules will be configured to ensure users won't face any slowness of the service.

# 4. Tools

After some research, AWS Kendra could be a really good tool to use for this use case. Simple to implement and avoids a lot of development/deployment work. This would speed up the release timeframe of the solution. Plus, if based on the questions below the customer is only looking for an "AI search box" we could remove the Bedrock entry point out of the solution and point directly to Kendra.
However, if the customer is looking for something fully customizable, then implementing an embedding bedrock model will be the way to go.

# 5. Questions

- Budget approved for the solution? Deadline? POC?
- How much usage are we expecting? Fluctuation in the load? Geolocation of the users?
- Will some data get obsolete? Will we need to regularly clean up the knowledge base? Where are the files stored? Do we need to sync them with some 3rd party storage?
- Do we want an AI search box or AI chat assistant?