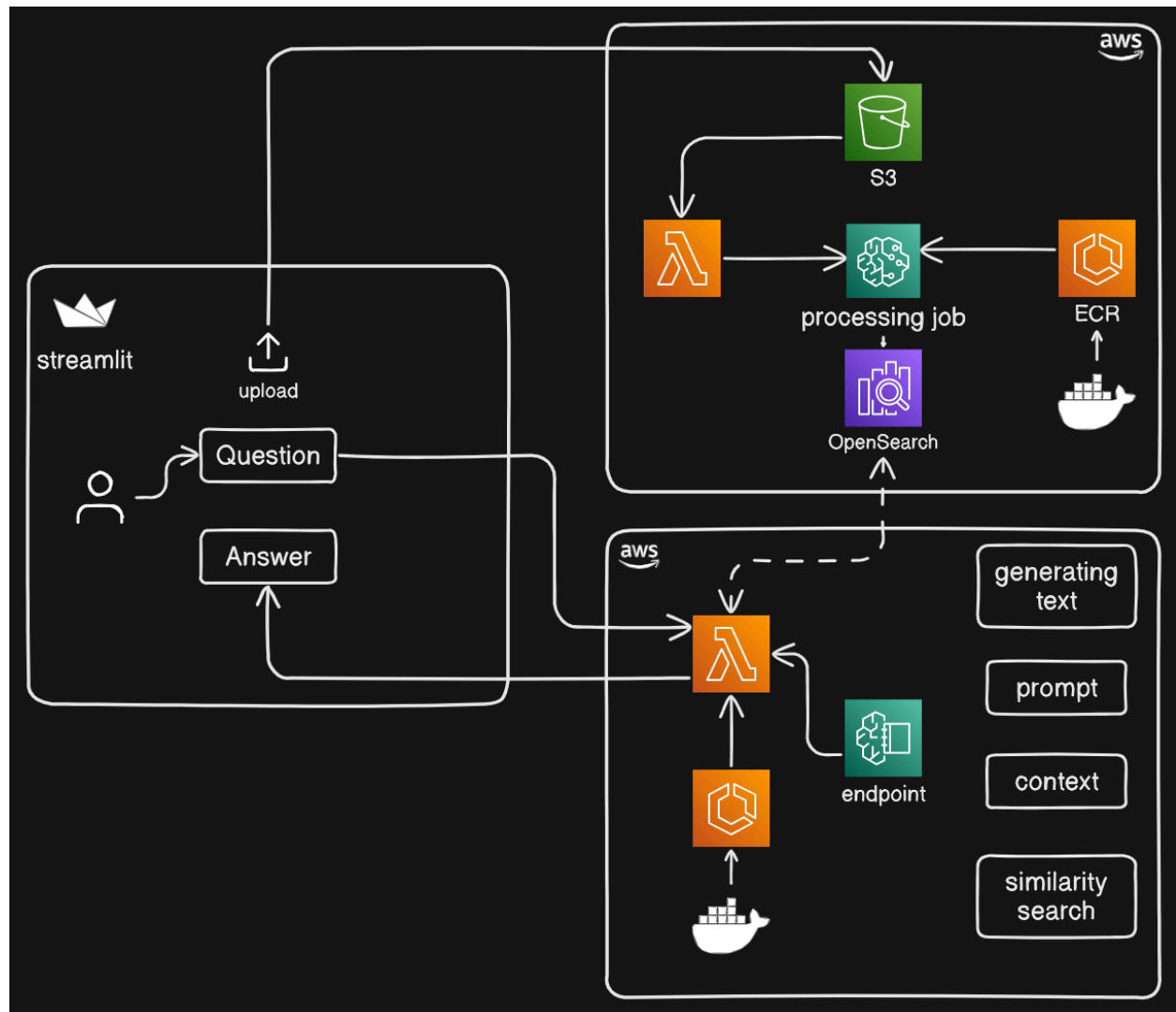


This is the v2 architecture of the bot inspired by aws's example

<https://aws.amazon.com/blogs/machine-learning/build-a-powerful-question-answering-bot-with-amazon-sagemaker-amazon-opensearch-service-streamlit-and-langchain/>



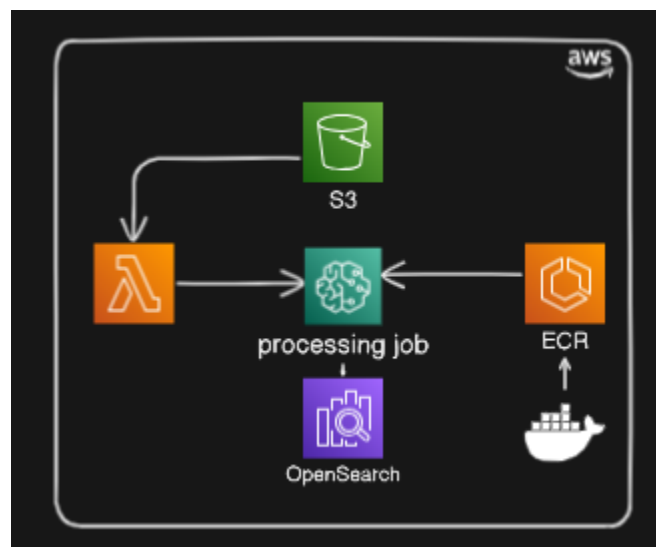
Required AWS Services:

1. S3 Bucket: For storing uploaded documents.
2. AWS Lambda: For triggering processing jobs and handling queries.
3. SageMaker Processing Job: For generating embeddings from documents.
4. SageMaker Endpoint: For generating text and processing queries.
5. OpenSearch Service Cluster: For storing embeddings and performing similarity searches.
6. ECR (Elastic Container Registry): For storing Docker images used by SageMaker Processing Jobs and endpoints.

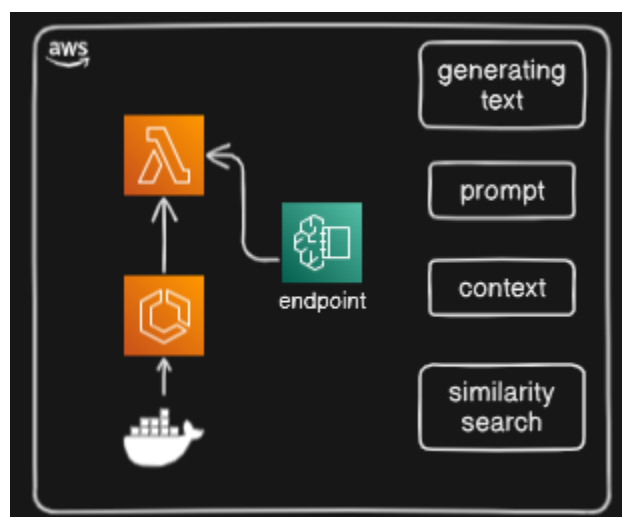
Workflow step-by-step:

The workflow can be

1. User will provide two inputs, a document and a question regarding the document. Each of these events will trigger a lambda.
 - a. When the user uploads a document from the UI, it gets stored in S3 bucket. This event triggers a lambda function.
 - i. The Lambda function starts a Sagemaker Processing Job
 1. The Sagemaker Processing Job processes the document to generate embeddings.
 2. The embeddings are stored in the OpenSearch index for future Similarity Search.
 - b. User will provide a query. This event triggers a Lambda function.
 - i. The Lambda function interacts with OpenSearch Service Cluster and does a similarity search.
 - ii. The similar documents are taken as context
 - c. These contexts are provided as prompts with the query in the prompt template to generate text using the Sagemaker LLM endpoint.
 - d. Generated texts are returned to the user as answers.
2. Lambda for Sagemaker Processing Job has a container image as entry point, this job is running on ml.c5.xlarge instance. The job contains a python script that uses LangChain framework and HuggingFace API and uses the open source embedding models to store embeddings to OpenSearch index.



3. Lambda for text generation is a Container Image which is dockerized to ECR. This wraps a python script that interacts with the Sagemaker LLM endpoint and performs a similarity search and the rest of the steps in 1b.



Testing and Validation

Document Page: 28-29

let's chat

Enter your query:

describe domain status values

Submit

Query processed successfully!

Answer: The domain status values in Amazon SageMaker are as follows:

1. Pending: Ongoing creation of domain.
2. InService: Successful creation of domain.
3. Updating: Ongoing update of domain.
4. Deleting: Ongoing deletion of domain.
5. Failed: Unsuccessful creation of domain. In this case, call the DescribeDomain API to see the failure reason for domain creation, delete the failed domain, and recreate the domain after fixing the error mentioned in FailureReason.
6. Update_Failed: Unsuccessful update of domain. Call the DescribeDomain API to see the failure reason for domain update, call the UpdateDomain API after fixing the error mentioned in FailureReason.

Document Page: 1-2

let's chat

Enter your query:

What are the key components involved in building AI solutions with SageMaker?

Submit

Query processed successfully!

Answer: The key components involved in building AI solutions with Amazon SageMaker include:

1. Machine learning (ML) lifecycle: Understanding the end-to-end process of building, training, and deploying ML models.
2. Solutions offered: Familiarizing yourself with the solutions provided by SageMaker, such as Amazon SageMaker Autopilot and SageMaker JumpStart.
3. Setting up Amazon SageMaker: Learning how to set up and use SageMaker based on your specific needs, including choosing the Amazon VPC ID, subnets, and security groups.
4. Storage: Managing your data storage using Amazon Elastic File System (Amazon EFS) and Amazon Elastic Block Store (Amazon EBS) file systems, and encrypting your data if necessary.

Document Page: 17

let's chat

Enter your query:

does sagemaker automatically assign timestamp while creating domain?

Submit

Query processed successfully!

Answer: Yes, when creating a domain using the "Set up for single user (Quick setup)" option in Amazon SageMaker, the domain is automatically assigned a name with a timestamp in the format "QuickSetupDomain-YYYYMMDDTHHMMSS".

Document Page: 50

let's chat

Enter your query:

How to update a domain from aws cli? command only

Submit

Query processed successfully!

Answer: To update a domain from AWS CLI, you can use the `update-domain` command with the following parameters:

```
aws sagemaker update-domain \
--domain-id domain-id \
--default-user-settings default-user-settings \
--default-space-settings default-space-settings \
--domain-settings-for-update settings-for-update \
--region region
```

Replace `domain-id` with the ID of the domain you want to update, `default-user-settings` and `default-space-settings` with the JSON-formatted settings you want to apply, and `settings-for-update` with a JSON-formatted object describing the settings you want to update.

For example, to update the execution role for the default user settings, you can use the following command:

```
aws sagemaker update
```

Document pg 3-4

let's chat

Enter your query:

what are the steps of a typical machine learning workflow?

Submit

Query processed successfully!

Answer: A typical machine learning workflow involves the following three stages:

1. Generate example data: This is the first stage where you collect and prepare the data for training the model. This may involve cleaning, preprocessing, and labeling the data.
2. Train a model: In this stage, you use an algorithm and the prepared data to train the machine learning model. The model learns patterns and relationships from the data to make predictions or inferences.
3. Deploy the model: Once the model is trained, it is integrated into an application to generate inferences in real-time and at scale. This involves deploying the model into a secure and scalable environment, such as Amazon SageMaker.

Document Page: 9

let's chat

Enter your query:

What is the purpose of Amazon SageMaker Feature Store?

Submit

Query processed successfully!

Answer: Amazon SageMaker Feature Store is a fully managed service that makes it easy to build, store, and serve machine learning features for machine learning models. It allows users to manage and share features across different teams, projects, and models, and provides features in real-time for online prediction and offline for model training and batch inference. It also supports various data sources, processing frameworks, and feature processing pipelines to help users prepare and manage their data for machine learning models.

let's chat

Enter your query:

what are the ChannelType of AutoMLChannel?

Submit

Query processed successfully!

Answer: The ChannelType of an AutoMLChannel in Amazon SageMaker can be set to either "training" or "validation" values. The "training" ChannelType is the default value and is used for building a machine learning model. The "validation" ChannelType is used for validation data, which is used to evaluate the performance of the model during training. A maximum of two data sources is allowed, one for training data and one for validation data. The format and schema of both data sources must be the same. Setting the ValidationFraction value in AutoMLDataSplitConfig causes an error when using two data sources.

let's chat

Enter your query:

Which planet do we live in?

Submit

Query processed successfully!

Answer: I'm unable to answer that question based on the provided context. The context only discusses Amazon SageMaker, an AWS machine learning service, and instructions for setting up an AWS account. There is no mention of planets in the text.

