Knowl-EDge

Serverless Learning Backend Service with Knowledge Base, Lambda and API Gateway

University of Toronto
School of Continuing Studies
SCS 3547 – Intelligent Agents

Student: Arjie Cristobal
Instructor: Larry Simon
March 2024

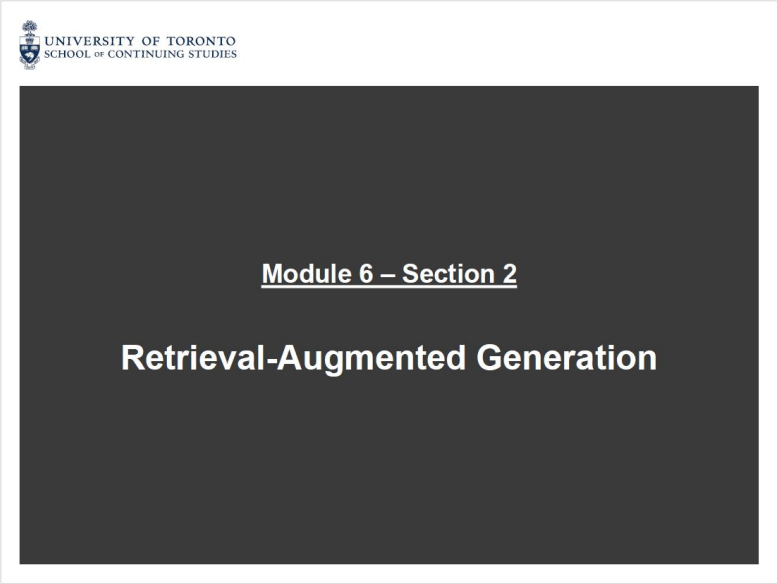
Objective

The primary objective of this study is to develop Knowl-EDge, a Serverless Backend Learning Service tailored for developers and technical support.

The study aims to take advantage of Generative AI with Large Language Models (LLMs), including RAG (Retrieve-Augmented Generation) and integrate Agents for Amazon Bedrock Knowledge Base to enhance the overall functionality and effectiveness of the service.

Consumers of this new API will be able to efficiently inquire about and access information related to the applications within our organizations.

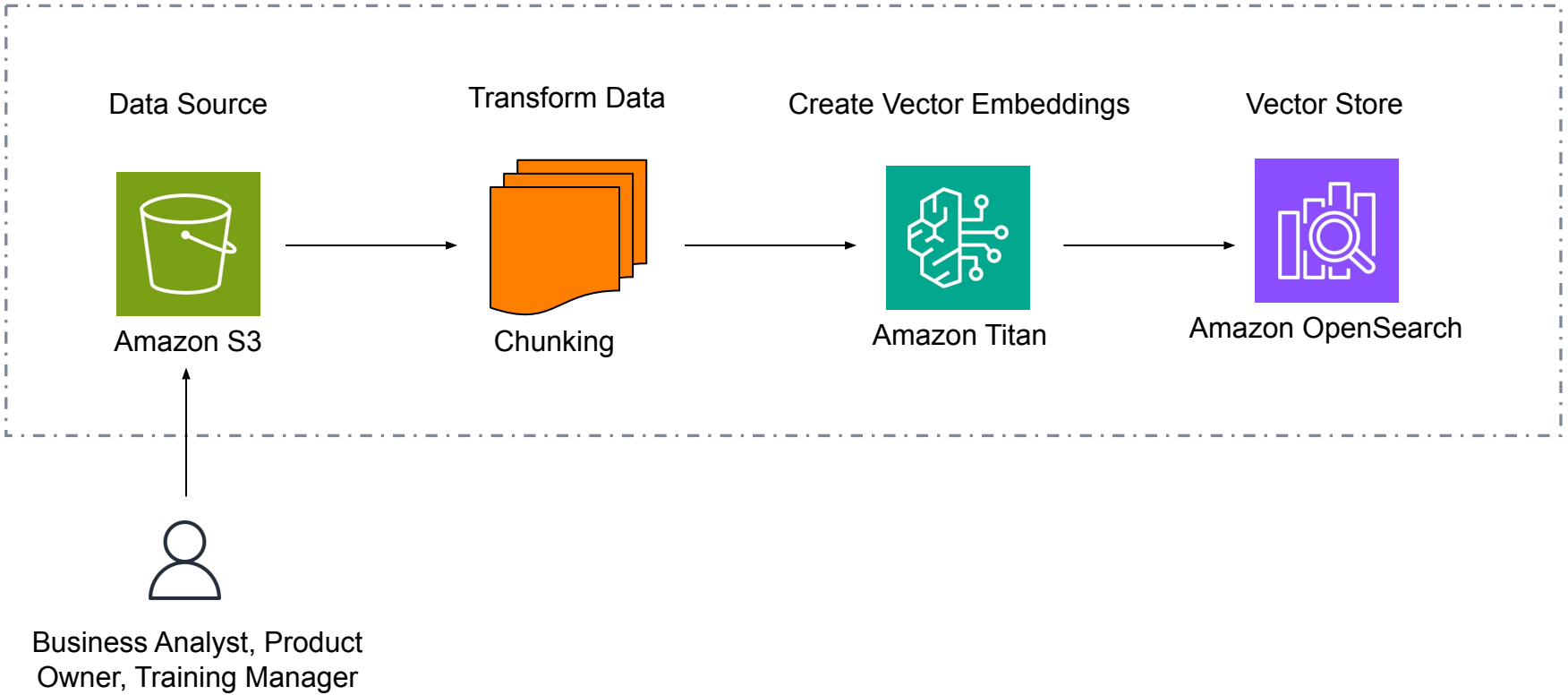
Concepts and Architecture



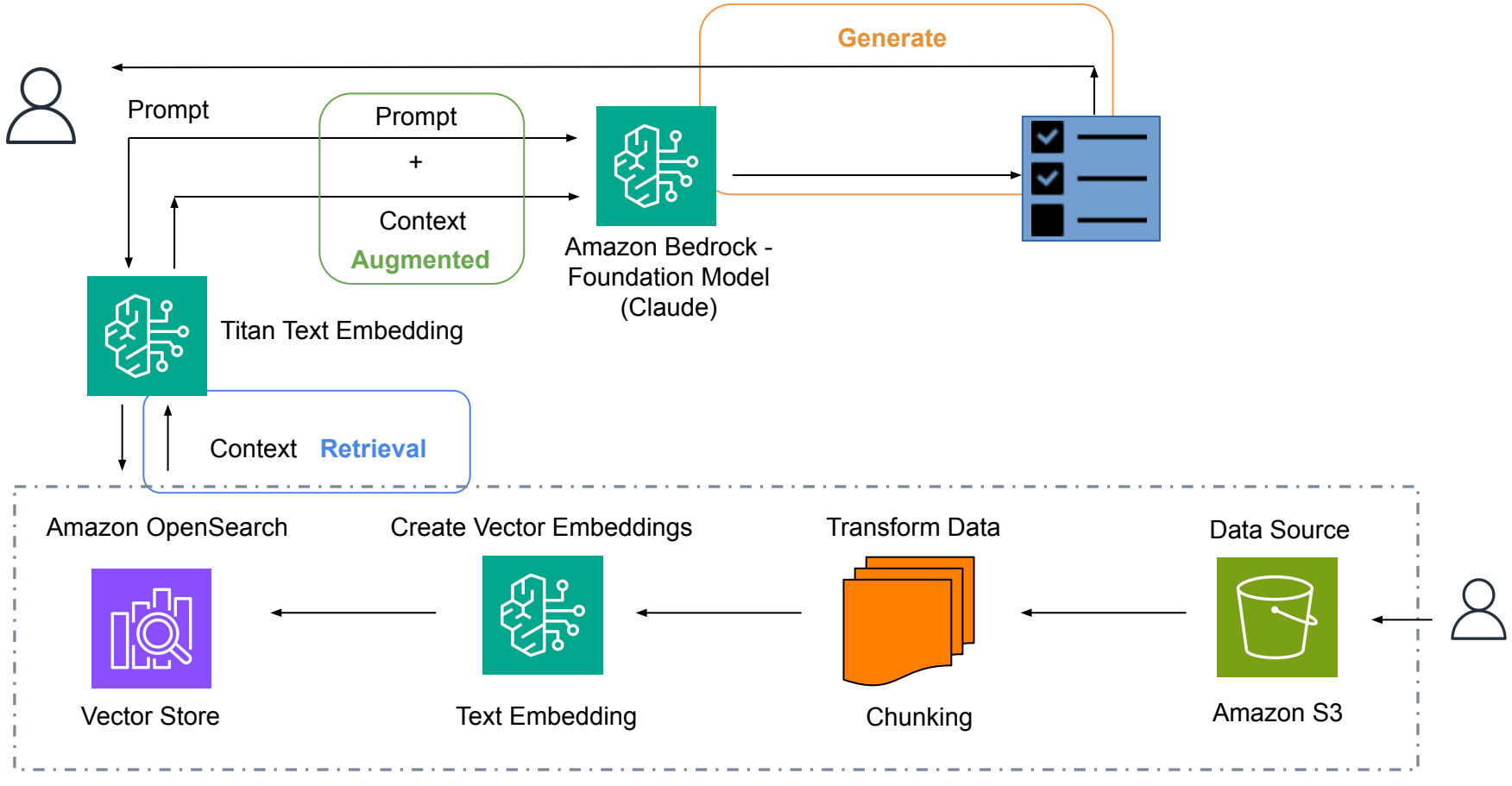
RAG:

- Allows to create a more customize version of LLM prompt that have an appearance of a fine-tuned document, but in reality we simply inserting text context into the prompt.
- The RAG Process can be a substitute for fine tuning a model, since it is far cheaper.

Data Ingestion Workflow



RAG Process


Generate
Prompt
Prompt
+
Context
Augmented
Amazon Bedrock -
Foundation Model
(Claude)
Titan Text Embedding
Context Retrieval
Data Ingestion Workflow
Amazon OpenSearch
Create Vector Embeddings
Transform Data
Data Source
Vector Store
Text Embedding
Chunking
Amazon S3

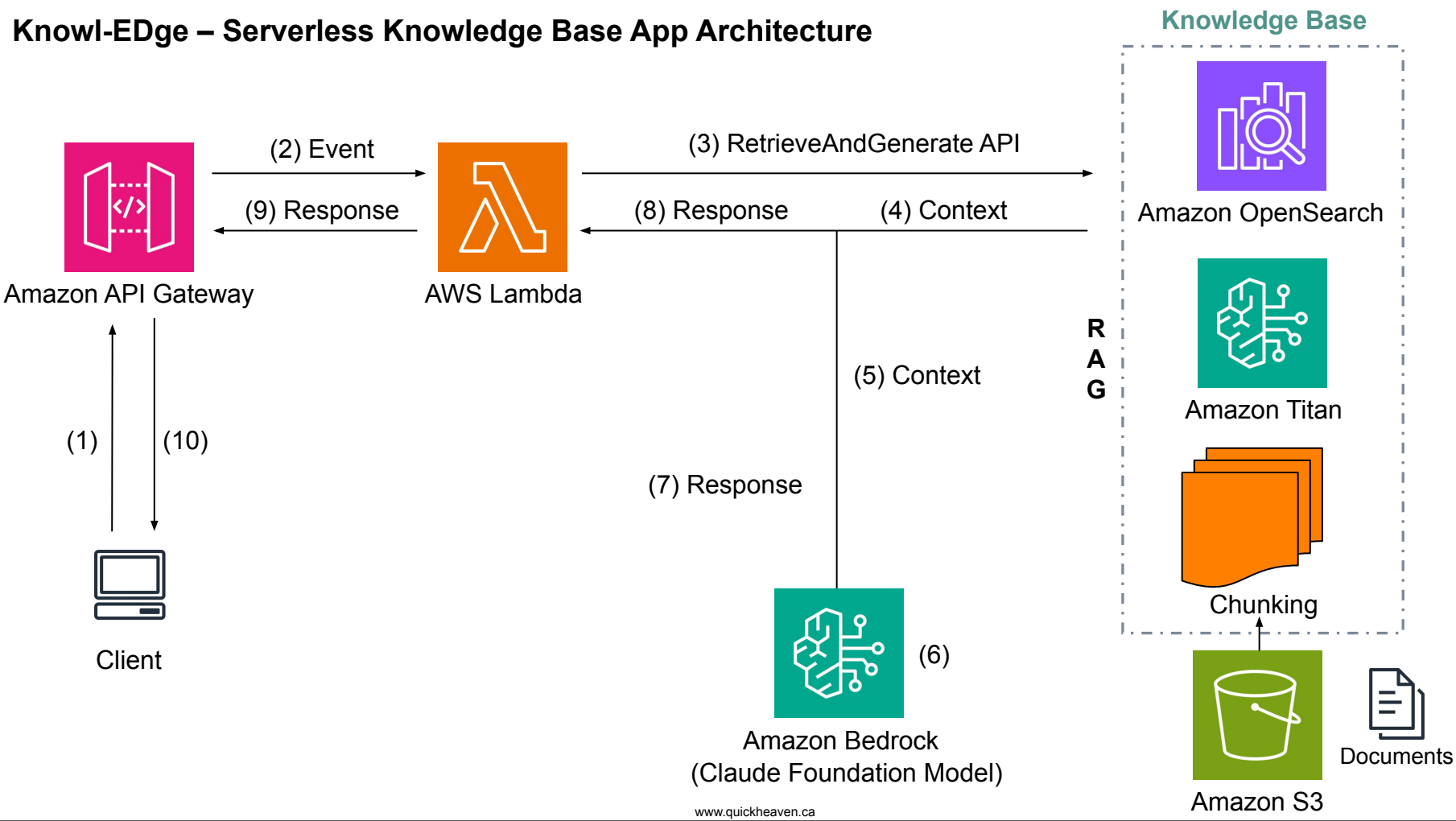
Data Ingestion Workflow

Knowledge Base for Amazon Bedrock

Business Analyst, Product Owner, Training Manager

AWS Free Tier, OpenSearch Service provides free usage of up to 750 hours per month of a t2.small.search or t3.small.search instance ($0.036 Price per hour)

Knowl-EDge – Serverless Knowledge Base App Architecture



Agents for Amazon Bedrock

Agents for Amazon Bedrock offers the ability to build and configure autonomous agents in the application. An agent helps end-users complete actions based on organization data and user input. Agents orchestrate interactions between foundation models (FMs), data sources, software applications, and user conversations.

AWS Lambda

```
boto3.client('bedrock-runtime')
```

```
boto3.client('bedrock-agent-runtime')
```

/ - GET method test results

Request	Latency	Status
/?prompt="What is Foundation Model?"	6208	200

Response body

```
{"statusCode": 200, "body": "A foundation model (FM) is an AI model with a large number of parameters and trained on a massive amount of diverse data. A foundation model can generate a variety of responses for a wide range of use cases."}
```

Response headers

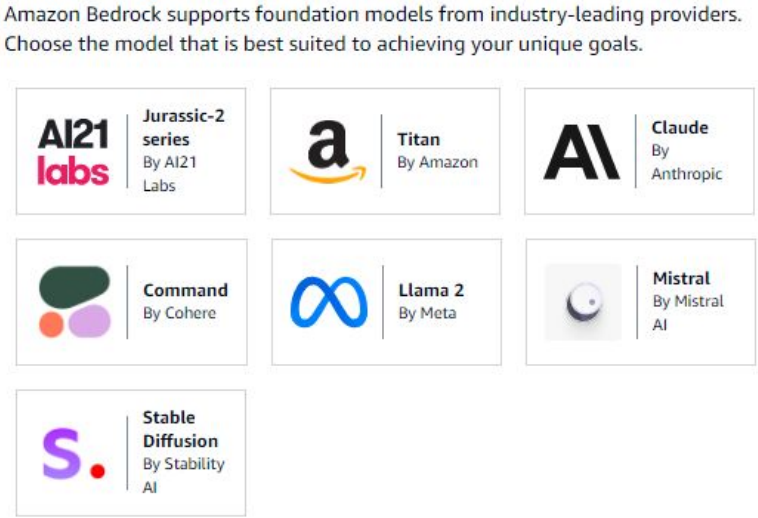
```
{
  "Content-Type": "application/json",
  "X-Amzn-Trace-Id": "Root=1-66021d75-4d0c490591d102a9092c7e2d;Parent=1fa338d8b25bb48f;Sampled=0;lineage=6fe4fe36:0"
}
```

Conclusion

In summary, our study successfully achieved its primary objective:the development of Knowl-EDge, a Serverless Backend Learning Service designed for developers and technical support. By using the power of Generative AI with LLM including RAG, we are able to improve the functionality and effectiveness of the service. The integration of Agents for Amazon Bedrock Knowledge Base ensures that consumers of this API can efficiently inquire about and access crucial information about our applications.

Recommendation

- Foundation Model Selection
- Prompt Engineering techniques
- Evaluate other Vector Databases



Thank you