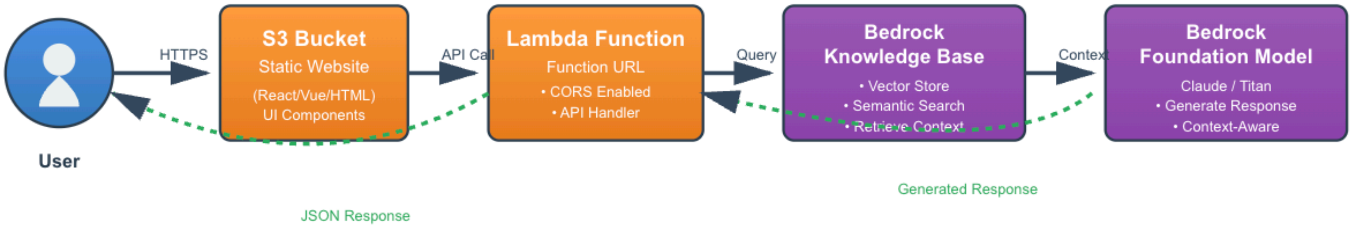
AWS Serverless RAG Architecture



S3 Bucket
Static Website
(React/Vue/HTML)
UI Components

Lambda Function
Function URL
- CORS Enabled
- API Handler

Bedrock Knowledge Base
- Vector Store
- Semantic Search
- Retrieve Context

Bedrock Foundation Model
Claude / Titan
- Generate Response
- Context-Aware

Request/Response Flow

1 **User Query**
- User enters question in UI
- Frontend sends POST request

2 **Lambda Processing**
- Receives request via Function URL
- Validates input & handles CORS

3 **Knowledge Retrieval**
- Query embeddings generated
- Vector similarity search
- Relevant context retrieved

4 **Response Generation**
- FM receives query + context
- Generates contextual answer
- Returns to Lambda

5 **Response Delivery**
- Lambda formats response
- UI displays answer to user

Key Features

- **Serverless Architecture:** No servers to manage
- **Auto-scaling:** Handles traffic spikes automatically
- **CORS Enabled:** Secure cross-origin requests
- **RAG Pattern:** Retrieval-Augmented Generation
- **Low Latency:** Fast response times
- **Cost-Effective:** Pay per request
- **Secure:** IAM-based access control
- **Semantic Search:** Context-aware retrieval
- **State-of-the-art AI:** Latest foundation models

AWS Serverless RAG Application | S3 + Lambda + Bedrock Knowledge Base + Foundation Models