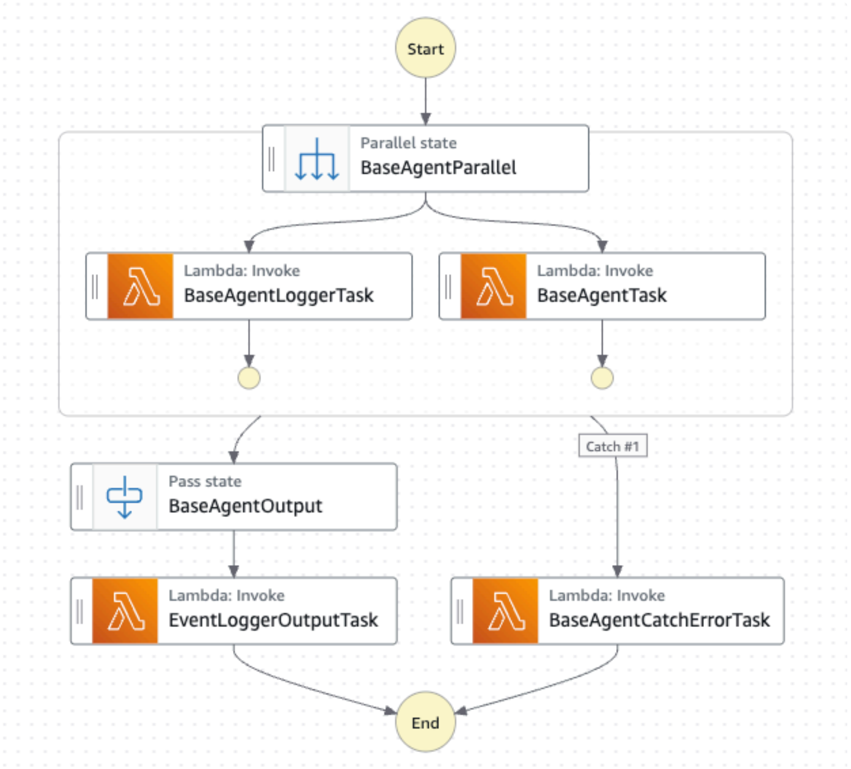
Knowledge Base RAG Fusion Agent

The current query architecture is implemented using a combination of Step Functions, Lambdas, and Bedrock Agents invoked from Lambdas. Logging is performed using Lambdas and information is stored in JSON formats on a dedicated S3 bucket.

Knowledge Base Agent

This implements the basic functionality of Knowledge Base agent invoking workflow.

Note: The component is glued to the Knowledge base RAG Fusion workflow and used to run the alternative queries against the Knowledge Base.



Knowledge Base workflow

The main component of the workflow is **BaseAgentTask**, a Lambda function that invoke the **Knowledge Base Amazon Bedrock Agent**.

The Agent is invoked with:

- agentId - the Id of the Amazon Bedrock Agent invoked
- agentAliasId - the AliasId for the Amazon Bedrock Agent invoked

- inputText - this is the question for the Knowledge Base
- sessionId - if not changed between calls, one can keep the context of the chat
- sessionAttributes - these are the session attributes - we include here the metadata for filtering the search in the Knowledge Base. Note: for sessionAttributes to apply, we need to include im the Amazon Bedrock Agent instructions the relevant directive $prompt_session_attributes$ is included in the Agent instructions.
- knowledgeBaseConfigurations - these are parameters that are used to control the functionality of Knowledge Base (we set here the maximum number of results expected).

Additional components are for logging the input, execution and output data (**BaseAgentLoggerTask**, **EventLoggerOutputTask**), and for catching error (**BaseAgentCatchErrorTask**) in the Step Function execution.

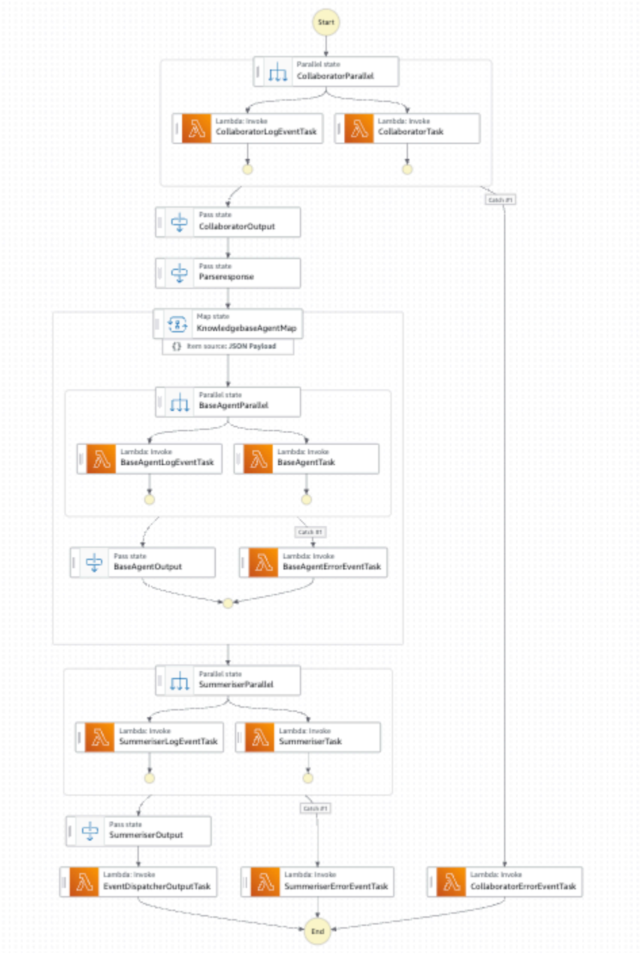
Knowledge Base Amazon Bedrock Agent

The full list of Amazon Bedrock Agents in operation is given in. Knowledge Base and Amazon Bedrock components.

Agent Input/Output contract

See the description in Agent Input/Output contract.

Knowledge Base RAG Fusion



RAG Fusion workflow

The current implementation for the RAG Fusion workflow contains three main steps, implemented as Lambda functions:

- **Query generator (Collaborator Task)** - The initial question is passed to an Anthropic Claude 3 Sonnet model that is tasked to generate variations of the questions, to ensure a broader context coverage in the Knowledge Base. A total of 6 (original and 5 variations) queries are run in parallel.
- **Query execution (Base Agent Task)** - With each question variation, (a total of 6) Knowledge Base Bedrock Agents are invoked in parallel, to retrieve and summarise relevant context from the Knowledge Base.
- **Response aggregator (Summariser Task)** - A specialized Bedrock Agent is invoked to process the dictionary of questions/responses from the previous step and to produce a summary of the alternative answers (RAG Fusion).

Aggregated workflow

The initial query is analysed and distributed between the (RAG Fusion) Knowledge Base workflow (described in previous section) and the SQL Query workflow. (Vitic Agent V1).

Structure and implementation are presented in Combined RAG Fusion V2.