Informatica® Cloud Application Integration
July 2024

Simple RAG Consumption with Pinecone

Informatica Cloud Application Integration Simple RAG Consumption with Pinecone
July 2024

Publication Date: 2024-07-31

Table of Contents

Preface

Use *Simple RAG Consumption with Pinecone* to learn how to receive a query from the user, convert it into vectors, form the context, and return a comprehensive response using a Large Language Model (LLM). The recipe is based on REST and SOAP APIs and you use an HTTP request to call the process.

CHAPTER 1

Introduction to Simple RAG Consumption with Pinecone recipe

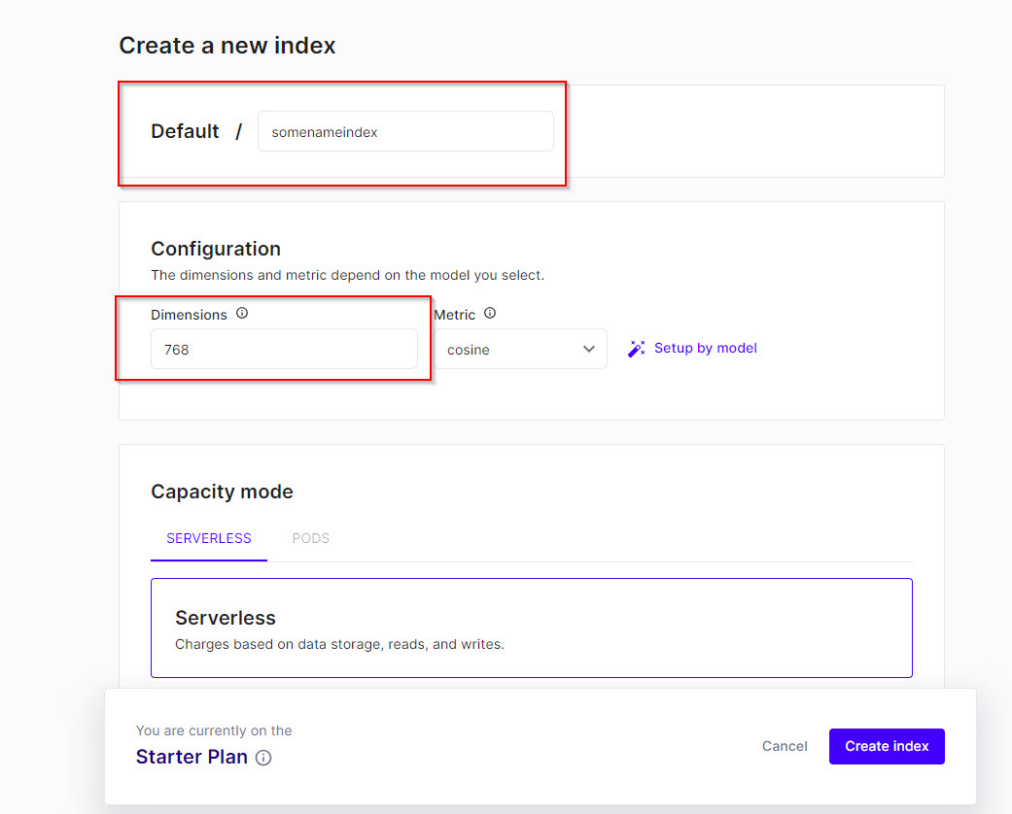
The Simple Retrieval Augmented Generation (RAG) Consumption with Pinecone recipe is based on REST and SOAP APIs.

The process submits a query that is received from the user, converts it into a vector, and uses it to search for similar vectors in a database. The top K matches are retrieved, filtered by a cutoff score, and used to form a context. This context including the original query is passed to a Large Language Model (LLM) to generate and return a comprehensive response.

Prerequisites for creating an index in Pinecone

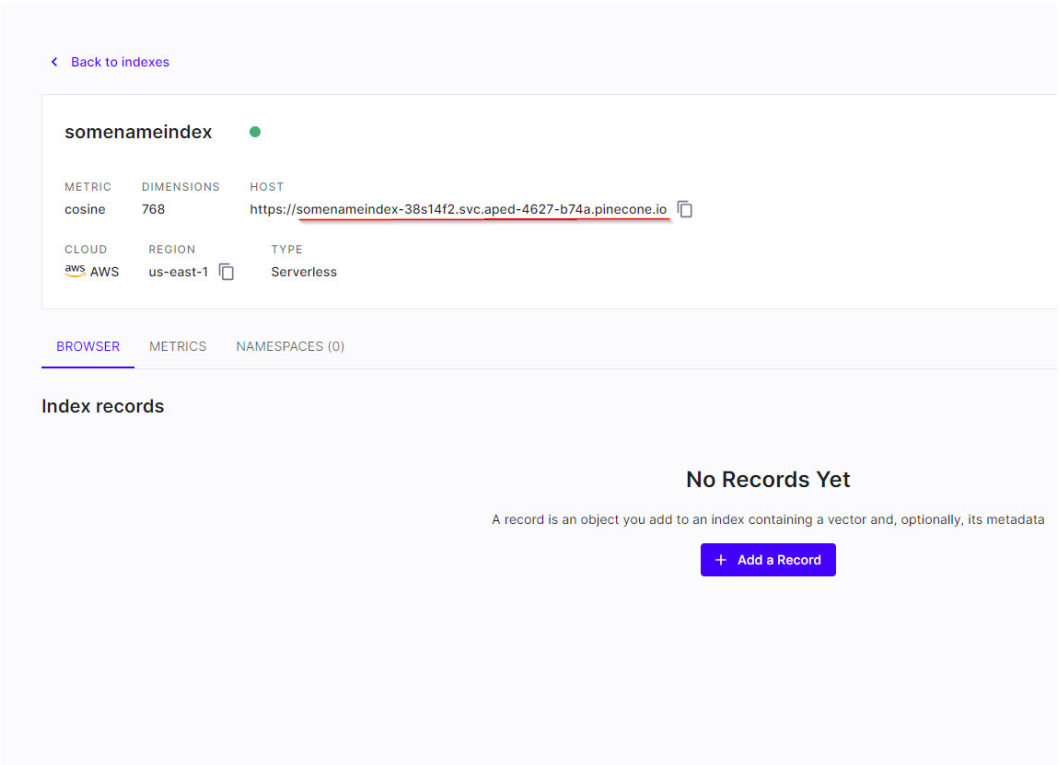
To ensure that the process works correctly, you must create an index in Pinecone and add text for the context. The context is then converted into vectors using the Gemini Embeddings process.

1. Open the **Create a new index** page in Pinecone.
2. In the **Default** field, enter an index name and in the **Dimensions** field, enter the value **768** as shown in the following image:



3. Click **Create index**.

After creating the index, you can use the **HOST** value without `https://` in the **Index_Host** input parameter as shown in the following image:

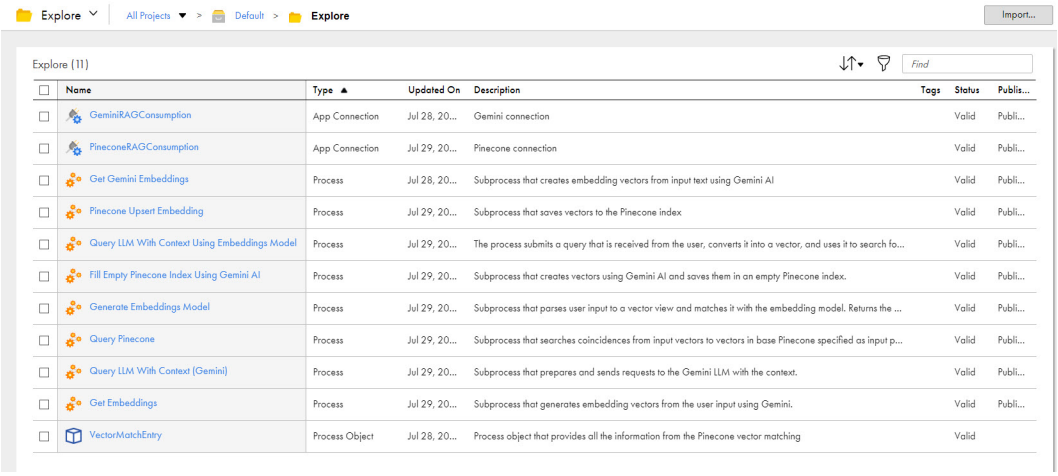

Back to indexes
somenameindex
METRIC
cosine
DIMENSIONS
768
HOST
https://somenameindex-38s14f2.svc.aped-4627-b74a.pinecone.io
CLOUD
AWS
REGION
us-east-1
TYPE
Serverless
BROWSER
METRICS
NAMESPACES (0)
Index records
No Records Yet
A record is an object you add to an index containing a vector and, optionally, its metadata
Add a Record

CHAPTER 2

Recipe contents

The Simple RAG Consumption with Pinecone recipe contains app connections, processes, and a process object.

The following image shows the assets that the Simple RAG Consumption with Pinecone recipe package contains:



Name	Type	Updated On	Description	Tags	Status	Publis...
GeminiRAGConsumption	App Connection	Jul 28, 20...	Gemini connection		Valid	Publi...
PineconeRAGConsumption	App Connection	Jul 29, 20...	Pinecone connection		Valid	Publi...
Get Gemini Embeddings	Process	Jul 28, 20...	Subprocess that creates embedding vectors from input text using Gemini AI		Valid	Publi...
Pinecone Upsert Embedding	Process	Jul 29, 20...	Subprocess that saves vectors to the Pinecone index		Valid	Publi...
Query LLM With Context Using Embeddings Model	Process	Jul 29, 20...	The process submits a query that is received from the user, converts it into a vector, and uses it to search fo...		Valid	Publi...
Fill Empty Pinecone Index Using Gemini AI	Process	Jul 29, 20...	Subprocess that creates vectors using Gemini AI and saves them in an empty Pinecone index.		Valid	Publi...
Generate Embeddings Model	Process	Jul 29, 20...	Subprocess that parses user input to a vector view and matches it with the embedding model. Returns the ...		Valid	Publi...
Query Pinecone	Process	Jul 29, 20...	Subprocess that searches coincidences from input vectors to vectors in base Pinecone specified as input p...		Valid	Publi...
Query LLM With Context (Gemini)	Process	Jul 29, 20...	Subprocess that prepares and sends requests to the Gemini LLM with the context.		Valid	Publi...
Get Embeddings	Process	Jul 29, 20...	Subprocess that generates embedding vectors from the user input using Gemini.		Valid	Publi...
VectorMatchEntry	Process Object	Jul 28, 20...	Process object that provides all the information from the Pinecone vector matching		Valid	

Simple RAG Consumption with Pinecone recipe assets

The following table lists the assets that the Simple RAG Consumption with Pinecone recipe package contains:

Asset Name	Asset Type	Description
PineconeRAGConsumption	App connection	Pinecone connection.
GeminiRAGConsumption	App connection	Gemini connection.
VectorMatchEntry	Process object	Provides all the information from the Pinecone vector matching.

Asset Name	Asset Type	Description
Get Gemini Embeddings	Process	Subprocess that creates embedding vectors from input text using Gemini AI.
Pinecone Upsert Embedding	Process	Subprocess that saves vectors to the Pinecone index.
Fill Empty Pinecone Index Using Gemini AI	Process	Subprocess that creates vectors using Gemini AI and saves them in an empty Pinecone index.
Generate Embeddings Model	Process	Subprocess that parses user input into a vector view and matches it with the embedding model. Returns the matching context score and metadata.
Query Pinecone	Process	Subprocess that searches coincidences from input vectors to vectors in the base Pinecone index specified as input parameter Index_Host_Pinecone. Returns a result with metadata.
Query LLM With Context (Gemini)	Process	Subprocess that prepares and sends requests to the Gemini LLM with the context.
Get Embeddings	Process	Subprocess that generates embedding vectors from the user input using Gemini.
Query LLM with Context using Embeddings Model	Process	Submits a query that is received from the user, converts it into a vector, and uses it to search for similar vectors in a database. The top K matches are retrieved, filtered by a cutoff score, and used to form a context. This context including the original query is passed to a Large Language Model (LLM) to generate and return a comprehensive response.

CHAPTER 3

Using the Simple RAG Consumption with Pinecone recipe

To use the Simple RAG Consumption with Pinecone recipe, you must perform the following steps manually:

Step 1: Copy and access the recipe

Step 2: Configure and publish the GeminiRAGConsumption connection

Step 3: Configure and publish the PineconeRAGConsumption connection

Step 4: Configure and publish the processes

Step 5: Invoke the process

Copying and accessing the recipe

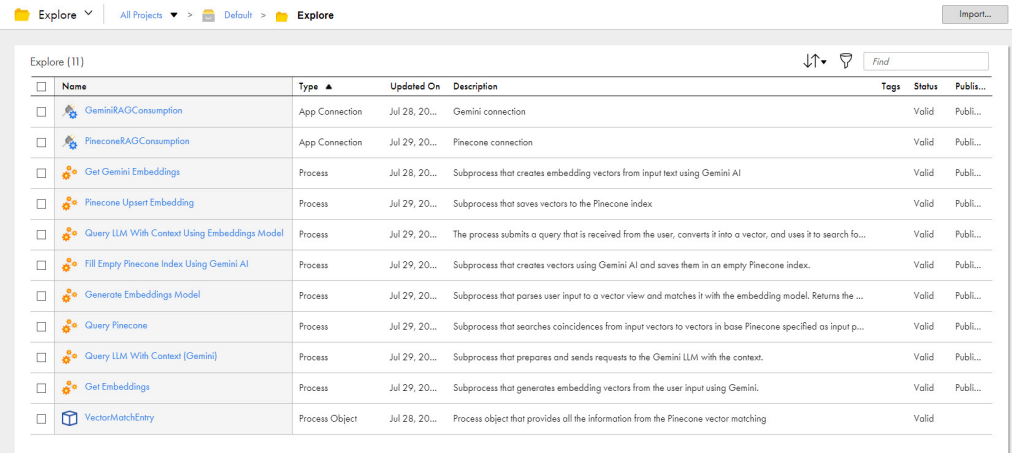
To copy and access the recipe content, perform the following steps:

1. Open the **Simple RAG Consumption with Pinecone** recipe and click **Use**.
2. Select the location where you want to copy the recipe, and then click **Continue**.
3. In the **Copying the recipe** dialog box, click **OK**.

   It might take some time for the recipe to get copied. You will receive a notification when the recipe is ready for use.
4. After the recipe is copied, click **Explore** to access the recipe content.

5. Navigate to the project or folder where you copied the recipe or enter the recipe name in the **Find** box.

   All the assets in the recipe are displayed as shown in the following image:



Name	Type	Updated On	Description	Tags	Status	Publis...
GeminiRAGConsumption	App Connection	Jul 28, 20...	Gemini connection		Valid	Publi...
PineconeRAGConsumption	App Connection	Jul 29, 20...	Pinecone connection		Valid	Publi...
Get Gemini Embeddings	Process	Jul 28, 20...	Subprocess that creates embedding vectors from input text using Gemini AI		Valid	Publi...
Pinecone Upsert Embedding	Process	Jul 29, 20...	Subprocess that saves vectors to the Pinecone index		Valid	Publi...
Query LLM With Context Using Embeddings Model	Process	Jul 29, 20...	The process submits a query that is received from the user, converts it into a vector, and uses it to search fo...		Valid	Publi...
Fill Empty Pinecone Index Using Gemini AI	Process	Jul 29, 20...	Subprocess that creates vectors using Gemini AI and saves them in an empty Pinecone index.		Valid	Publi...
Generate Embeddings Model	Process	Jul 29, 20...	Subprocess that parses user input to a vector view and matches it with the embedding model. Returns the ...		Valid	Publi...
Query Pinecone	Process	Jul 29, 20...	Subprocess that searches coincidences from input vectors to vectors in base Pinecone specified as input p...		Valid	Publi...
Query LLM With Context (Gemini)	Process	Jul 29, 20...	Subprocess that prepares and sends requests to the Gemini LLM with the context.		Valid	Publi...
Get Embeddings	Process	Jul 29, 20...	Subprocess that generates embedding vectors from the user input using Gemini.		Valid	Publi...
VectorMatchEntry	Process Object	Jul 28, 20...	Process object that provides all the information from the Pinecone vector matching		Valid	

Configuring and publishing the GeminiRAGConsumption connection

To configure and publish the GeminiRAGConsumption connection, perform the following steps:

1. Open the **GeminiRAGConsumption** connection.
2. In the **Type** field, select **Gemini**.
3. In the **Run On** field, select **Cloud Server or any Secure Agent**.
4. In the **Connection Properties** section, enter the API key in the **API_Key** property. The **API_Key** property authenticates Gemini connection requests.
5. Save, test, and publish the connection.

Configuring and publishing the PineconeRAGConsumption connection

To configure and publish the PineconeRAGConsumption connection, perform the following steps:

1. Open the **PineconeRAGConsumption** connection.
2. In the **Type** field, select **Pinecone**.
3. In the **Run On** field, select **Cloud Server or any Secure Agent**.
4. In the **Connection Properties** section, enter the API key in the **API_Key** property. The **API_Key** property authenticates Pinecone connection requests.
5. Save, test, and publish the connection.

Configuring and publishing the processes

The Get Gemini Embeddings process, Pinecone Upsert Embedding process, and Fill Empty Pinecone Index Using Gemini AI process consume the user's input text, modify it to a vector representation, and save it to the Pinecone index.

To configure and publish the processes, perform the following steps:

1. Open the following processes in the order specified below:
    1. Get Gemini Embeddings
    2. Pinecone Upsert Embedding
    3. Fill Empty Pinecone Index Using Gemini AI
    4. Get Embeddings
    5. Query Pinecone
    6. Generate Embeddings Model
    7. Query LLM With Context (Gemini)
    8. Query LLM with Context using Embeddings Model
2. For each process, on the **Start** tab of the Start step, select **Cloud Server** from the **Run On** list.
3. Optionally, for the **Query LLM with Context using Embeddings Model** process, in the **Set LLM Models** step, in the **Assignments** tab, you can update the values in the **Set_Context_Model** and **Set_Embedding_Model** fields.
4. Save and publish all the processes.

Invoking the process

When you invoke the Query LLM With Context Using Embeddings Model process, the user sees the answer matching the context that was used in the LLM request.

You can run the process using one of the following options:

- REST or SOAP API endpoints in any API client such as cURL, Postman, SOAP UI, or any programming language
- Web browser by passing the input parameters