Building Your First RAG AI Agent

A complete guide to creating intelligent chatbots with N8N, Pinecone, and Google AI

Created by Chinmay Kaitade

MERN Stack Developer | AI Enthusiast

What You'll Learn Today

01

Prerequisites & Setup

Essential API keys and account configurations needed to get started

02

Data Ingestion Pipeline

Transform your documents into searchable vector embeddings

03

Retrieval System

Build the query-to-answer pipeline that powers your AI agent

04

Integration & Testing

Connect everything together and validate your RAG implementation

Created by Chinmay Kaitade

MERN Stack Developer | AI Enthusiast

Essential Prerequisites

Google AI Studio

Free API key for text embedding and generation. Get yours at Google AI Studio

Pinecone Account

Vector database for semantic search. Sign up free at Pinecone for API key and environment

N8N Workflow

Automation platform to orchestrate your RAG pipeline and handle integrations

---

Created by Chinmay Kaitade

MERN Stack Developer | AI Enthusiast

Critical Pinecone Index Configuration

Required Settings

1

Dimension: 768

Matches Gemini's text-embedding-004 model output exactly

2

Metric: Cosine

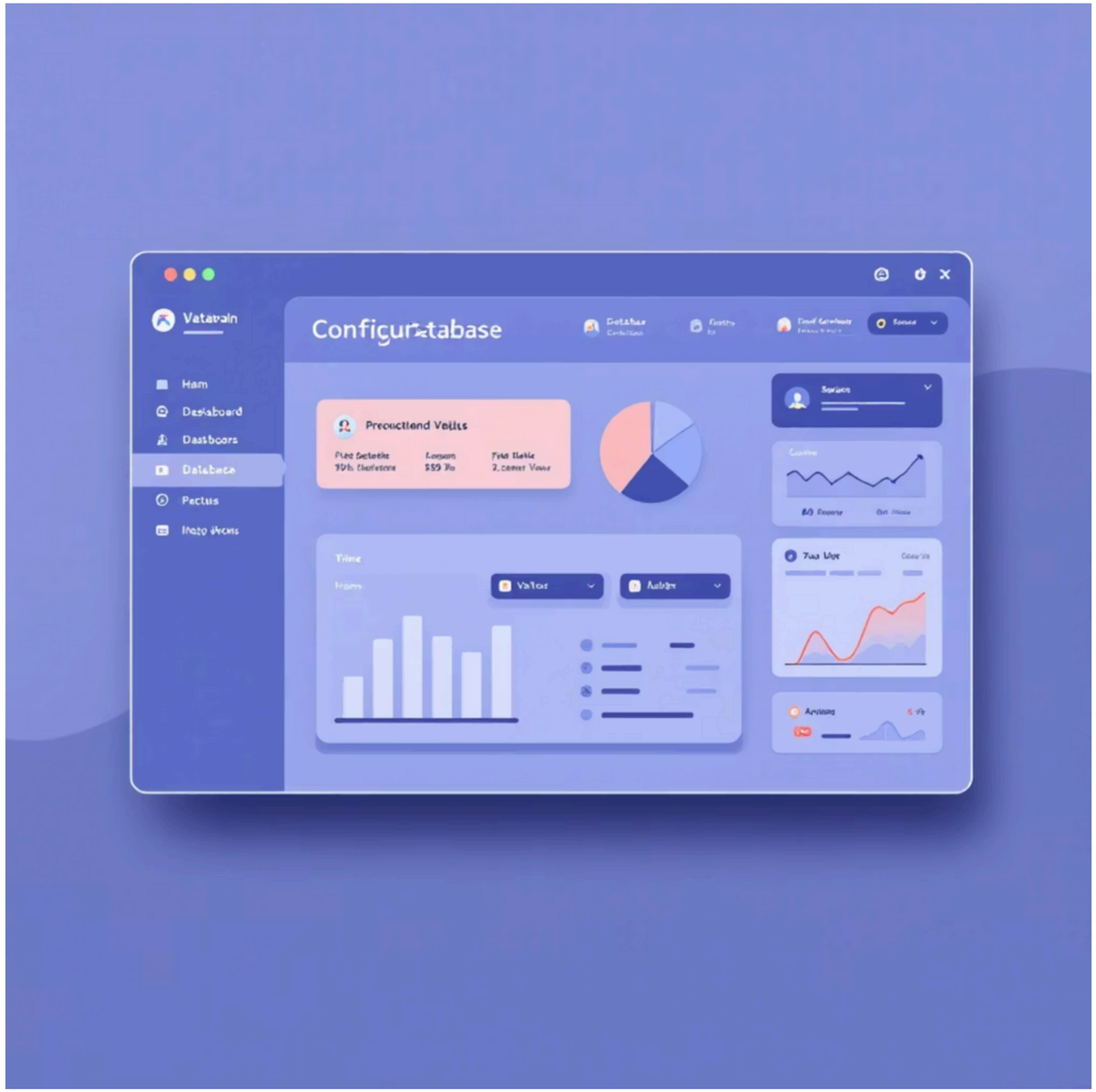
Optimal for measuring semantic similarity between text vectors

3

Vector Type: Dense

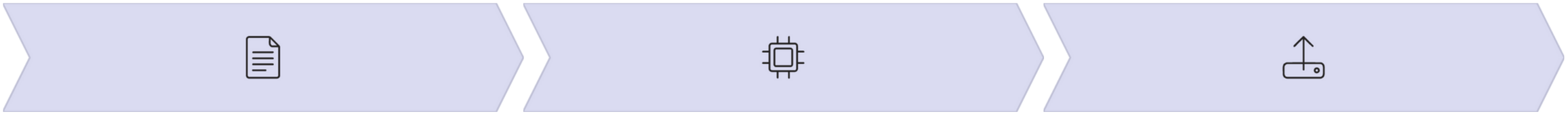
Fixed-length arrays of floating-point numbers for precise embeddings

Pro Tip: These settings are non-negotiable. Any mismatch will cause integration failures between Gemini and Pinecone.



Data Ingestion Pipeline

Transform your knowledge base into searchable vectors



Collect Documents

Gather PDFs, text files, and notes into a centralized knowledge base

Generate Embeddings

Use Gemini text-embedding-004 to convert text into 768-dimension vectors

Upsert to Pinecone

Upload vectors with unique IDs and metadata for efficient retrieval

---

Created by Chinmay Kaitade

MERN Stack Developer | AI Enthusiast

The Magic Happens Here

Vector Similarity Search

When users ask questions, your RAG agent performs semantic search across thousands of documents in milliseconds, finding the most relevant context to generate accurate, personalized responses.

Created by Chinmay Kaitade

MERN Stack Developer | AI Enthusiast

Retrieval Pipeline Architecture

Query Embedding

Convert user question into vector using same Gemini embedding model

1

Vector Search

Pinecone finds top-k most semantically similar documents

2

Context Augmentation

Combine retrieved text with original user query

3

LLM Generation

Gemini-pro generates accurate answer using retrieved context

4

> The key insight: Your AI agent doesn't rely on general training data—it uses your specific knowledge base to provide truly personalized responses.

Created by Chinmay Kaitade

MERN Stack Developer | AI Enthusiast

N8N Integration Benefits

Visual Workflow Design

Build and debug your RAG pipeline with drag-and-drop interface, making complex AI workflows accessible to developers

Native Integrations

Connect seamlessly to Airtable, Google Sheets, and 400+ other services without writing custom API code

Automated Triggers

Set up automatic data ingestion when new documents are added to your knowledge base

---

Created by Chinmay Kaitade

MERN Stack Developer | AI Enthusiast

Key Success Metrics

<100ms

Query Response Time

Vector search delivers lightning-fast semantic matching

95%

Context Relevance

Properly configured embeddings ensure high-quality retrieval

768

Vector Dimensions

Gemini's embedding model captures rich semantic meaning

∞

Scalability

Pinecone handles millions of vectors with consistent performance

With proper implementation, your RAG agent will provide accurate, contextual responses that feel genuinely helpful to users.

Created by Chinmay Kaitade

MERN Stack Developer | AI Enthusiast

Connect & Continue Learning

Ready to build your own RAG AI agent? Let's stay connected and share the journey!

📸 Instagram

@chinmaykaitade_hunter

Daily AI learning updates

👨‍💻 GitHub

ChinmayKaitade

Code examples and projects

🌐 LinkedIn

Professional Network

Technical discussions

🐦‍🔥 X (Twitter)

@chinmaydotcom

Quick tips and insights

Questions? Feedback? Just say hello! 👋

Created by Chinmay Kaitade

MERN Stack Developer | AI Enthusiast