Vertex AI Search POC Guide

This guide explains how to set up and deploy a Vertex AI Search proof-of-concept (POC) locally and on Render, with two integration options: an embedded Google widget (JWT secured with auto-refresh) and a direct API mode.

1. Google Cloud Setup - Create a Vertex AI Search Data Store (Custom / Unstructured). - Ingest test documents (HTML/PDF). - Create App/Widget config and copy configId. - Add authorized domains: localhost, 127.0.0.1, and your Render URL. - Create a Service Account with "Discovery Engine Searcher" role and download JSON key.

2. Project Structure my-search-app/ ■■■ app.py ■■■ requirements.txt ■■■ Procfile ■■■ .gitignore ■■■ templates/index.html ■■■ static/style.css

3. Python Dependencies flask, gunicorn, google-cloud-discoveryengine, PyJWT

4. App Features - /token: issues a JWT from service account (valid 60 min). - Auto-refresh JWT every 55 minutes in frontend JS. - /api_search: queries Vertex AI Search directly. - /: serves a page embedding both widget and API search UI.

5. Local Development - Create venv and install requirements. - Set environment variables: GOOGLE_APPLICATION_CREDENTIALS=./secrets.json GCP_PROJECT_ID, GCP_LOCATION, GCP_DATA_STORE_ID, WIDGET_CONFIG_ID - Run: python app.py - Visit http://127.0.0.1:5000

6. Deployment on Render - Push repo to GitHub (exclude secrets.json, venv). - Render → New Web Service → Python → connect repo. - Build: pip install -r requirements.txt - Start: gunicorn app:app - Secrets: upload JSON key as Secret File at /etc/secrets/secrets.json - Env Vars: set same as local (GOOGLE_APPLICATION_CREDENTIALS, PROJECT_ID, etc.). - Add Render domain to authorized domains.

7. Testing - GET /token should return a JWT. - POST /api_search {"query":"..."} should return results. - Widget search should display results after domain authorization.

8. Troubleshooting - ModuleNotFoundError: use python -m pip install flask - Configuration not authorized: add domain in authorized domains. - DefaultCredentialsError: check GOOGLE_APPLICATION_CREDENTIALS path. - Empty results: wait for document indexing to finish.

This completes the setup for a working POC with Vertex AI Search.