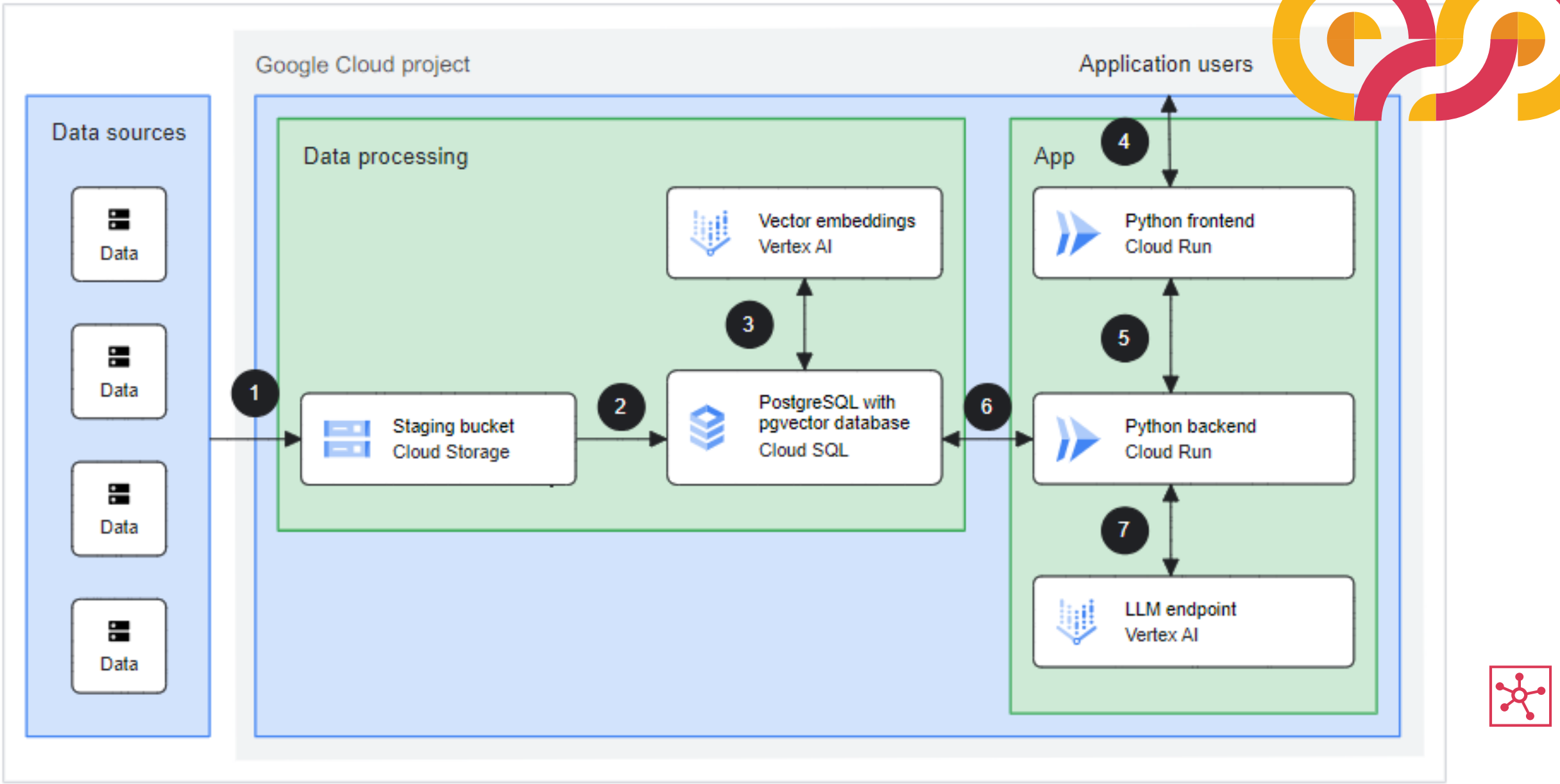
Generative AI RAG with Cloud SQL



Generative AI RAG with Cloud SQL

1. Data is uploaded to a Cloud Storage bucket.
2. Data is loaded to a PostgreSQL database in Cloud SQL.
3. Embeddings of text fields are created by using Vertex AI and stored as vectors.
4. You open the application in a browser.
5. The frontend service communicates with the backend service for a generative AI call.
6. The backend service converts the request to an embedding and searches existing embeddings.
7. Natural language results from the embeddings search, along with the original prompt, are sent to Vertex AI to create a response.

source:
https://cloud.google.com/architecture/ai-ml/generative-ai-rag