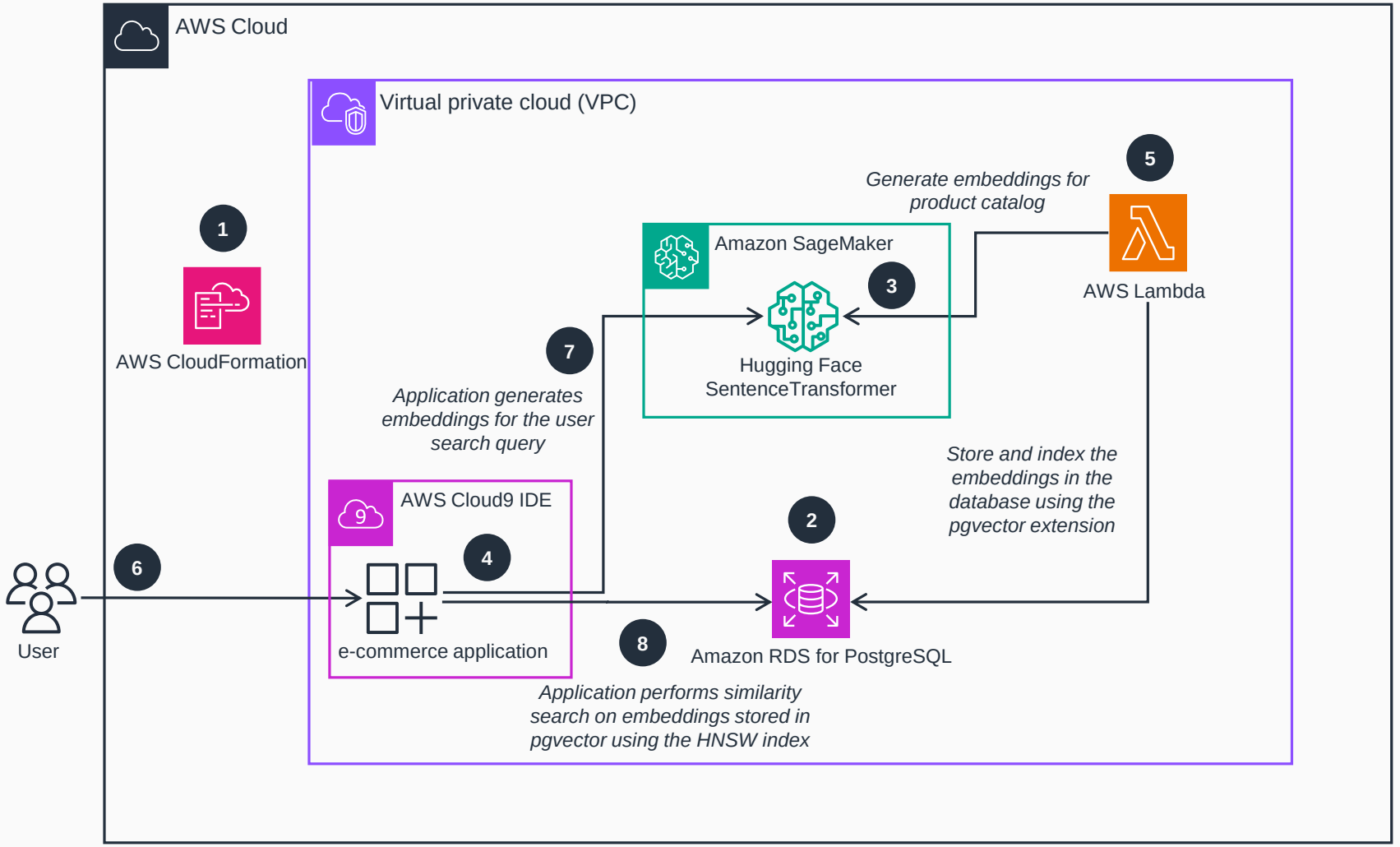
Guidance for E-Commerce Products Similarity Search on AWS

This architecture diagram shows how to build a product catalog with a similarity search capability. It uses artificial intelligence (AI), Amazon SageMaker, Amazon RDS for PostgreSQL, and the pgvector extension. Steps 1-4 are shown here, for steps 5-8, refer to the next slide.



1. Deploy resources in your AWS account using **AWS CloudFormation**. This includes deploying instances of **Amazon Relational Database Service** (**Amazon RDS) for PostgreSQL**, **Amazon SageMaker**, **AWS Cloud9**, **AWS Lambda**, and a Custom Resource.

2. **RDS for PostgreSQL** stores both the product catalog and embeddings for the products using the PostgreSQL open-source extension pgvector to store and index these high-dimensional vector embeddings.

3. **SageMaker** runs the pre-trained HuggingFace large language model (LLM) (SentenceTransformer) for real-time inference. You also have the flexibility to select other models that best fit your needs. **Amazon Bedrock** offers an alternative way for running diverse foundation models and conducting inference to produce text embeddings.

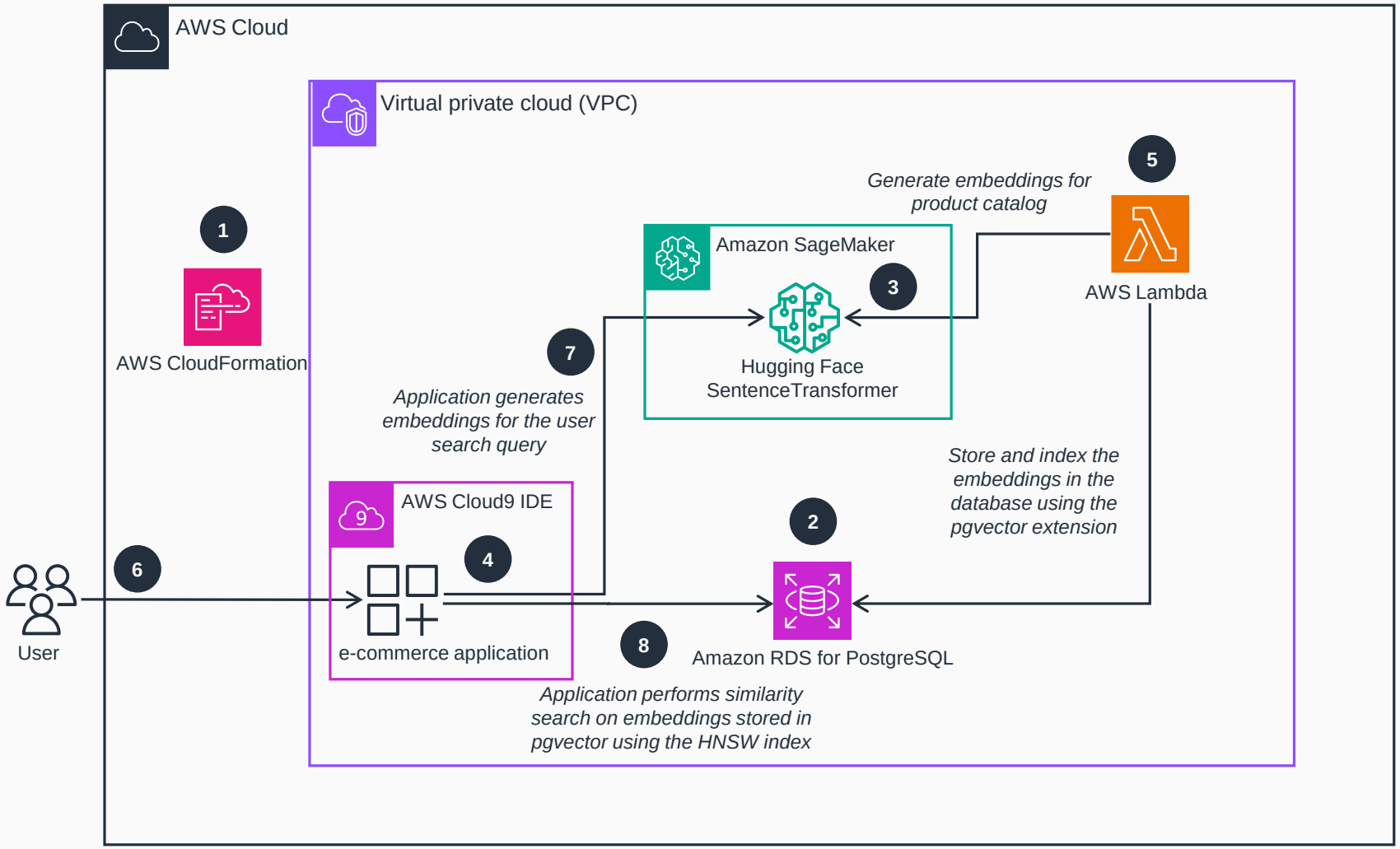
4. The **AWS Cloud9** integrated development environment (IDE) hosts a sample Streamlit application that provides product information retrieved from **RDS for PostgreSQL**. You have the option to replace this sample application with one of your choice. For alternative compute choices to run the application, consider deploying it on **Amazon Elastic Container Service (Amazon ECS)**, **Amazon Elastic Kubernetes Service (Amazon EKS)**, **AWS Fargate**, or **Amazon Elastic Compute Cloud (Amazon EC2)**.

Reviewed for technical accuracy June 21, 2024

AWS Reference Architecture

Guidance for E-Commerce Products Similarity Search on AWS

Steps 5-8



5 Custom Resources in **CloudFormation** run user-defined provisioning logic during stack creation, update (if the Custom Resource is modified), or deletion. When combined with a **Lambda** function, **CloudFormation** invokes and runs the function accordingly.

The Custom Resource from the **CloudFormation** stack invokes **Lambda** to bootstrap **RDS for PostgreSQL**. During this process, the system creates the pgvector extension, initializes the Product Catalog schema, and generates embeddings using **SageMaker** near real-time inference. The system stores these embeddings along with product catalog metadata in **RDS for PostgreSQL** and indexes the embeddings using the pgvector index type hierarchical navigable small world (HNSW).

6 Run the e-commerce product catalog application on **AWS Cloud9** and preview the application while it's running.

7 Perform a search on the product catalog in the application, which generates embeddings for the search query using the **SageMaker** near real-time inference endpoint.

8 The application connects to **RDS for PostgreSQL,** runs the similarity search query on embeddings using the pgvector *HNSW* index, and then displays the product catalog similarity search results on the application screen.

Reviewed for technical accuracy June 21, 2024

AWS Reference Architecture