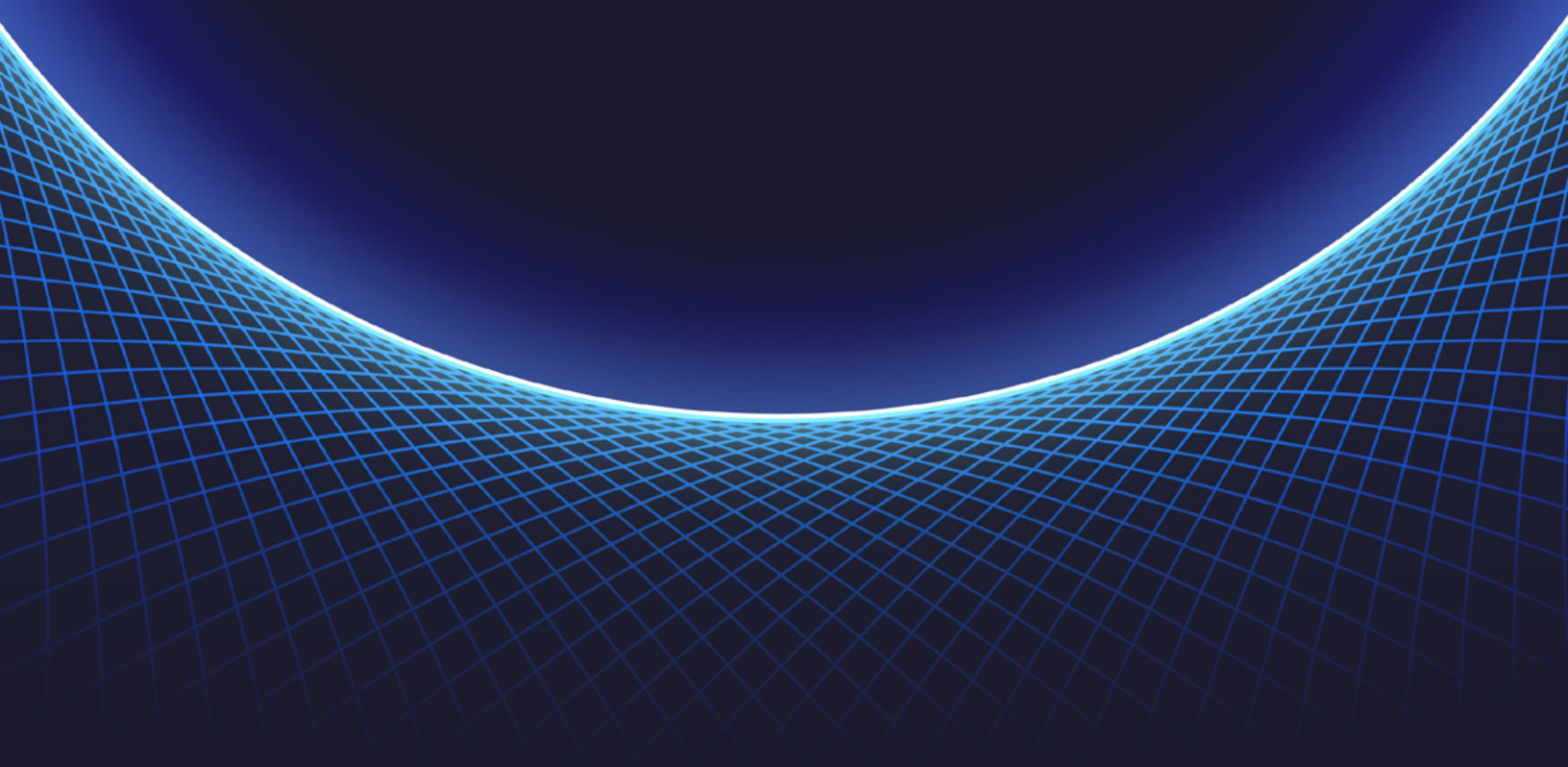
Pinecone

Optimizing and accelerating data classification with Pinecone and AWS



aws Available in AWS Marketplace

Pinecone

Table of contents

Why use a vector database for classification?

In machine learning (ML) and artificial intelligence (AI), classification is a fundamental pillar, enabling algorithms to categorize and predict outcomes based on input data.

What is classification?

Classification is the process of categorizing data into different classes based on different variables. Classification enables better data management, improves predictions, and supports more informed decision-making across a wide range of applications. Its use cases include email spam detection, medical diagnostic testing, fraud detection, image classification, speech recognition, document classification, product classification, data classification, and generative AI.

Optimizing classification with Pinecone

While useful and effective, classification does have some challenges—especially when traditional databases, such as MySQL, and traditional classification models, such as logistic regression, are involved. For example, classification tasks often require finding similar data points in large datasets that include images, documents, or user profiles. Traditional databases typically use lexical or keyword search, which can lead to slow and inefficient queries and miss critical context.

A vector database isn't constrained by tables and keywords. Instead, it represents each data item as a vector (called a vector embedding), which is a list of numbers capturing its key characteristics, enabling efficient similarity search.

Pinecone is a fully managed, serverless vector database that runs on Amazon Web Services (AWS) and integrates and works with many other AWS services. Pinecone excels at finding semantically similar meanings, relevance, and context in large datasets based on vector proximity.

Because Pinecone is serverless, it significantly speeds up the process of finding the most similar items to a given query, yielding accurate classification based on nearest neighbors. These fast similarity search capabilities, combined with metadata filtering and live index update support, make Pinecone a powerful tool for building and deploying classification applications that deliver high performance at scale.

Search isn't the only classification challenge for traditional databases. Training a classification model with traditional databases is costly and time-consuming. The process involves teaching a machine learning algorithm to categorize data into predefined labels. It starts with data preprocessing, such as normalization and feature extraction, and then feeds the data into a neural network with input, hidden, and output layers. The model learns by adjusting weights and biases through backpropagation. When new data arrives, the model must be retrained to maintain prediction accuracy.

Pinecone addresses this challenge by enabling the identification of relevant vs. irrelevant vectors in the proximity of the model's learned weights. Through this identification, it homes in on the critical area of vectors near the target vector space, saving time, resources, and compute. As a result, you only train on the data that matters, and fine-tuning models is faster. The result is lower costs without affecting performance.

> **Gong's transition to Pinecone serverless resulted in 10x cost reduction while maintaining peak performance.[1]**

[1] https://www.pinecone.io/blog/serverless/

What classification use cases can you build with Pinecone?

With Pinecone, you can build systems and applications that automatically classify high-dimensional data into predefined categories, perform similarity searches, and much more. The common use cases that customers are using Pinecone to address are:

Model training	Active learning systems	Reducing fraud and security alerts
Sentiment analysis	Genomic research	Object recognition

The following chapters cover these use cases in detail, starting with model training.

Pinecone in a nutshell

Pinecone's serverless architecture on AWS delivers many benefits and capabilities. These include cost reduction, scalability, multiple integrations, and improved performance. There are also helpful resources, an active community, and a plan that enables you to start for free.

Pinecone is specifically designed to store, index, and retrieve high-dimensional data as vector embeddings, which reduces latency down to milliseconds. Thanks to this low latency paired with high throughput for large datasets, Pinecone can support applications with up to billions of vectors.

Multitenancy: Namespaces, metadata filtering, and more enables you to partition your data, ensuring you're getting the right answers all the time at the lowest possible cost.

Pinecone automatically scales based on usage, eliminating the need for manual capacity planning and management.

It also uses vector clustering—via proprietary algorithms—on top of AWS object storage. And with distributed storage, vector data is clustered on AWS object storage, providing virtually limitless data scalability and high availability.

Cost savings: Pinecone separates read and write paths, which allows the independent scaling of compute resources. Combined with its consumption-based pricing model, where you only pay for what you use, it delivers cost efficiency at any scale (up to 50x less than pod-based indexes).[2]

[2] https://www.pinecone.io/blog/serverless/

The magic behind model training

Traditional model training involves feature engineering, vector embeddings generation, and the recursive computation of weights and fine-tuning. As the data in the model increases, so does the likelihood of errors, inefficiency, and longer training time. By contrast, Pinecone automates vector embeddings and indexing, allowing for quick retrieval and management of complex data like genomic sequences. As a result, Pinecone can significantly enhance the data annotation process due to semantic search capabilities, making it more precise and relevant for training ML models.

This is particularly useful in scenarios where identifying the most relevant samples is key to improving model performance. For example, you can use Pinecone's vector search to fine-tune classifiers. Traditional classification ML models have a series of neural network layers. These layers convert data (text and images) into vector embeddings and use learning functions to perform classification. Typically, the model consists of many layers, with the final being classification. Pinecone uses these vector embeddings and their corresponding labels to classify data into categories by focusing on the essential data samples that are near the group boundary to train the classifier. Pinecone then uses these predetermined groups to classify new data samples.

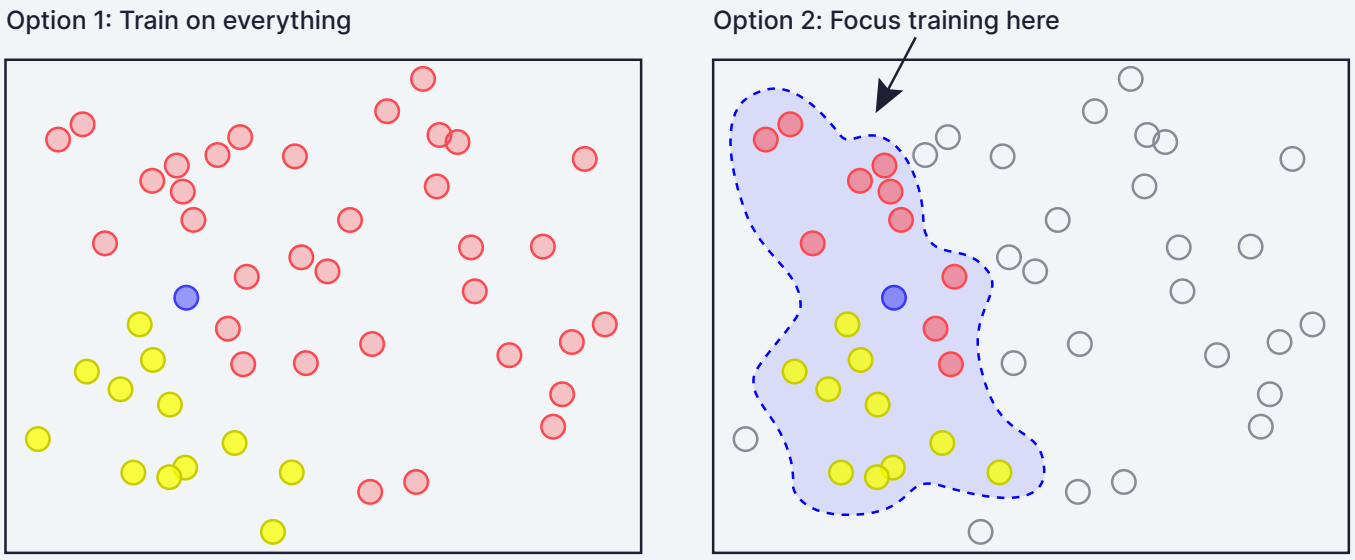


Figure 1. Model training options: Training on everything vs. focused training

All about active learning systems

While traditional model training lays the groundwork for ML, as mentioned, it often struggles with inefficiency as data volumes grow. That's where active learning comes in.

Active learning focuses on choosing the most useful data to learn from, enhancing the model's performance. Pinecone enables active learning use cases by capturing the nuances and contextual variations present in user conversations, facilitating the accurate labeling of each concept. In addition, Pinecone serves up the most relevant results and then lets users vote yea or nay on them. This exercise informs the model on what's relevant and what's not. This helps provide precise examples for concept tracking in user conversations.

Customer success

Gong quickly finds and automatically labels data

Gong is a revenue intelligence platform that collects conversation data from everywhere a company interacts with customers to provide full understanding of the customer lifecycle. To help customers act on this data, Gong developed an innovative active learning system that uses AI to detect and track complex concepts in the conversations it collects. It does this by using Pinecone for vector searches to identify sentences that are like the provided examples. Here is the workflow:

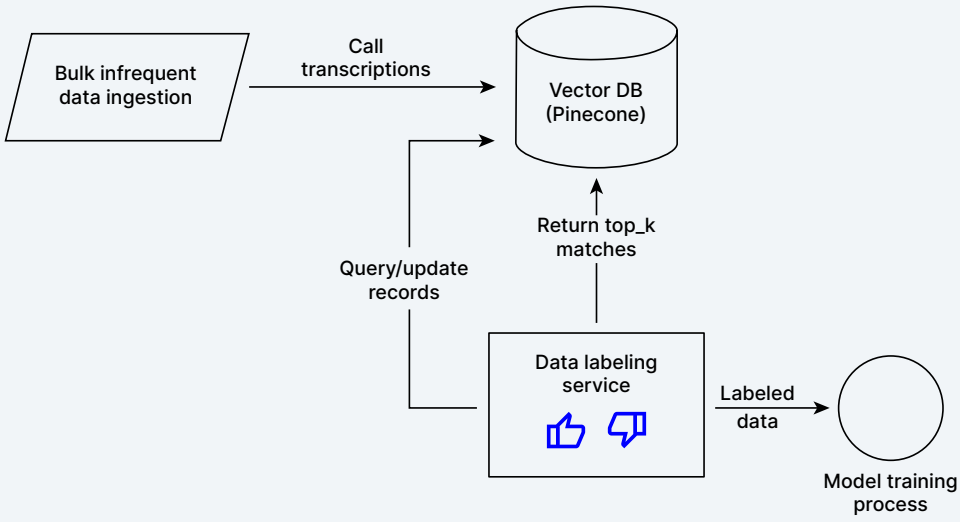


Figure 2. High-level diagram: Classification use case

It is Pinecone's serverless architecture on AWS that allows Gong to seamlessly use vector storage at any scale while achieving substantial cost reductions.

For more details, read the full Gong story.

Reducing fraud and security alerts

Vector databases can significantly enhance security and reduce fraud in financial systems by efficiently detecting patterns and anomalies in high-dimensional data. The following examples show how Pinecone supports multiple security use cases with classification.

Fraud and anomaly detection

Companies can use Pinecone to detect anomalies or outliers in high-dimensional data, such as network traffic patterns, financial transactions, or sensor readings. By representing data as vectors and using Pinecone's query capabilities, organizations can quickly identify and respond to unusual or suspicious activities. This efficiency lends itself to faster results. For example, in financial systems, Pinecone can compare vectors with known patterns of fraud, swiftly identifying and flagging suspicious activities for further investigation.

Detecting similar security alerts

The number of alerts that security operations centers (SOCs) receive in a day can reach into the thousands—and security analysts must review and process all of them. Many of those alerts are genuine, but others are duplicates. Regardless, using the same alert "noise" throughout can make it difficult to determine which alerts are genuine. With Pinecone's similarity search capabilities, you can implement a classification system that relies on the nearest neighbors of a query vector. This helps identify security alerts that are expressed differently but are similar in nature.

Employing sentiment analysis

Sentiment analysis is a technique used in natural language processing (NLP) to determine the emotional undertone of text—positive, negative, or neutral. Organizations use sentiment analysis to identify and group opinions about their products, services, and ideas. A vector database allows for precise sentiment detection by efficiently processing complex textual patterns, resulting in quick and accurate insights.

This is evident in how Pinecone conducts sentiment analysis. Pinecone can store and manage vector embeddings, making sentiment analysis faster and more scalable. When you want to analyze the sentiment of a new piece of text, its vector is compared against existing vectors in the database. It can determine its sentiment based on similarity. The vector database quickly retrieves and analyzes similar text vectors, providing an immediate sentiment analysis result. Pinecone's live index updates allow you to do this in real time and on the freshest dataset.

Notebook

Sentiment analysis with Pinecone

Pinecone offers comprehensive documentation and technical artifacts to help users get hands-on experience with their vector databases. These cover several AI use cases and are hosted on Pinecone's GitHub repository, which includes Jupyter Notebooks optimized for learning and exploring AI techniques. One such notebook applies sentiment analysis to the hotel industry to understand customer perception and potential areas that need improvement. After indexing reviews in Pinecone, hotel staff can search for what customers consider important when they stay at a hotel and then analyze their opinions.

Pinecone makes it easy to search a topic and get customer reviews relevant to the search query along with sentiment labels as metadata. For example, it can return the top 500 reviews related to specific London hotel room sizes. From those reviews, you can determine that the general opinion of the room size is positive. By expanding the search to include cleanliness, staff, air conditioning, and food during a specific time range, travelers can learn which hotel has the most positive sentiment in all five areas. These are the sample results.

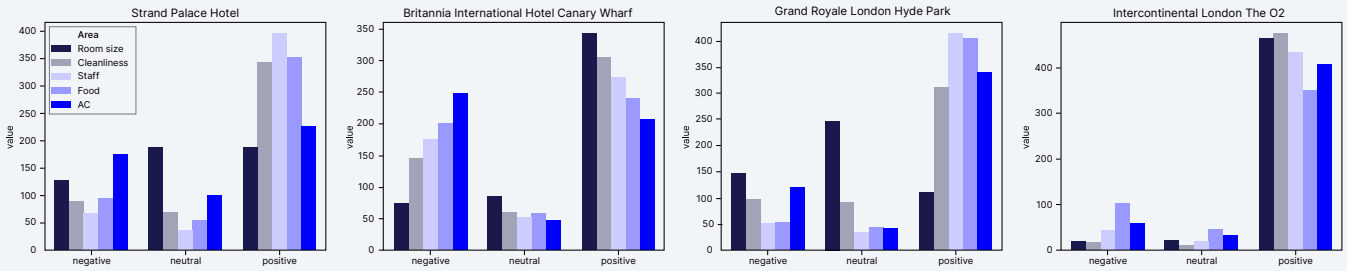


Figure 3. Sample results from the hotel sentiment analysis

To experiment with hotel sentiment analysis, visit the notebook on GitHub.

Classification in genomic research

Genomic researchers are turning to vector databases for their ability to swiftly process and analyze intricate genomic data (the information related to the structure and function of an organism's genome). The advanced indexing and similarity search capabilities drive fast, precise identification of genetic patterns.

Pinecone supports genomic data classification in several ways. First, it can manage and query high-dimensional genomic data vectors, making it easier to classify complex genetic sequences accurately—even in real time. Pinecone's optimized and adaptive indexing allows for rapid similarity searches among large genomic datasets for quick identification of similarities and patterns. Finally, Pinecone's serverless architecture can accommodate the growing volume of genomic data without compromising performance, an essential capability in the field of genetic research.

Object recognition in images and videos

Object recognition is the technique of identifying the object present in images and videos. Pinecone's fast similarity search and flexible indexing make it an ideal choice for real-world object recognition and classification applications. Pre-trained deep learning models generate vector embeddings that capture essential image features. You can then use the labels of the most similar embeddings to identify and classify the objects in new images.

When presented with a new image, the labels associated with the nearest neighbors can be used to classify the object therein. Using Pinecone's capabilities, developers can build robust and scalable object recognition systems that can classify objects accurately and handle large-scale datasets.

Classification use cases with Pinecone: A quick summary

Model training

Pinecone allows for the quick retrieval of similar vectors, enhancing the data annotation process and making it more precise and relevant for training ML models.

Active learning systems

Pinecone enables active learning use cases that capture conversational nuances and contextual variations, facilitating accurate learning.

Reducing fraud

Pinecone can enhance security and reduce fraud in financial systems by efficiently detecting patterns and anomalies in high-dimensional data.

Sentiment analysis

Pinecone can store semantically similar data points that are closer together, while dissimilar points are farther apart in the latent space, making sentiment analysis faster and more scalable.

Genomic research

Pinecone can search by similarity or relevance to other genomic data vectors, making it easier to classify complex genetic sequences accurately—even in real time.

Object recognition

Pinecone's positioning of semantically similar points allows the given vector of an object to be matched with similar vectors in the database, helping identify which group it belongs to and to determine its label.

Now that you understand how Pinecone addresses common classification use cases, let's look at a reference architecture example.

Example

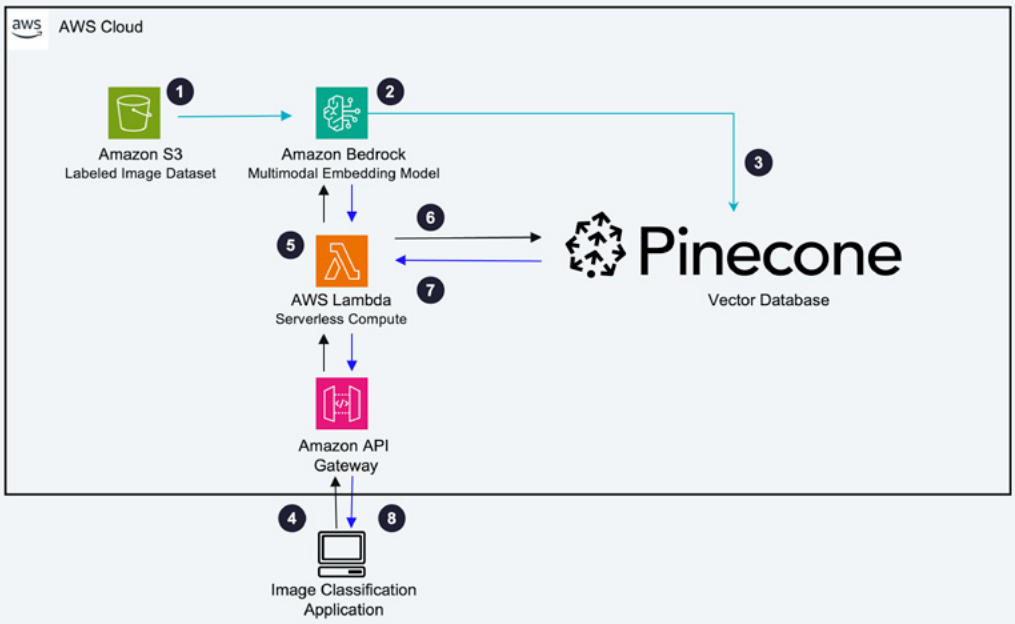
Reference architecture for image classification

This chapter guides you through a reference architecture for building an image classification using Pinecone, Amazon Bedrock, and AWS services. This architecture allows you to classify new images based on similarities to pre-labeled images without any machine learning training. It consists of two main components: the backend labeled image dataset pipeline and the front-end classification pipeline.

The backend uses Amazon Titan, a multi-modal embedding model from Amazon Bedrock, to generate vector embeddings for labeled images. These embeddings, along with their labels as metadata, are stored in Pinecone. The front-end pipeline classifies new images by converting them into vector embeddings and using the embeddings as a payload to query Pinecone. Pinecone then returns the matching embeddings along with their corresponding labels.

In this reference architecture (Figure 4), the Amazon Titan multi-modal computer vision embedding model is used to generate vector embeddings. The embeddings and their corresponding labels can classify data into categories. These predetermined groups can be used to classify new data samples. Implementing the proposed solution involves the following two pipelines.

Figure 4. Image classification reference architecture using Pinecone, Bedrock, and AWS services



Labeled dataset pipeline

The labeled dataset pipeline is shown as numbers 1 through 3 in Figure 4:

1. Store the labeled dataset, consisting of images and their classification labels, in an Amazon Simple Storage Service (S3) bucket.
2. For each image, generate vector embeddings using the Titan multi-modal embedding model in Amazon Bedrock.
3. Store the vector embeddings, along with their respective classification labels, as metadata in the Pinecone vector database.

Query pipeline

The query pipeline is represented as numbers 4 through 8 in Figure 4.

4. When a new image is received, send it to the backend pipeline via an API Gateway.
5. The new image is converted to vector embeddings using Titan. An Amazon Lambda function orchestrates the process of converting the image to a vector, performing the search, and relaying the results.
6. Use the vector embeddings of the image as a payload to query the Pinecone vector database.
7. Fetch the nearest neighbors for the query vector embeddings and retrieve the corresponding labels.
8. Return the classification label to the front-end application.

And that's how Pinecone, Amazon Bedrock, and AWS services can enable image classification based on similarities to pre-labeled images without any machine learning training.

To learn more, visit the image classification notebook on GitHub.

Pinecone

Supercharge your apps with Pinecone on AWS

Pinecone on AWS provides organizations with a highly scalable cloud infrastructure and automated vector embedding and indexing for efficient, performant handling of large-scale, high-dimensional data. Together, Pinecone and AWS offer more flexibility and capabilities—plus seamless integration with other AWS services—for optimal, accurate application performance.

Start building today with Pinecone or find Pinecone in AWS Marketplace.

©2024 Pinecone – All rights reserved.

aws Available in AWS Marketplace