



# Neural Search Simplified

## The Easiest Way to Add Production-Grade Neural Search

Neural search is quickly surpassing traditional keyword search in many NLP applications because it can better understand user intent—and this means a better user search experience. But how do you add neural search to your search stack without having to invest the time and resources needed to learn a new software platform?

### We have the solution

GSI's OpenSearch and Elasticsearch neural search plugins allow you to extend your current Elasticsearch or OpenSearch software investment with a production-grade neural search solution that searches billions of items in milliseconds, with high recall.

That means you don't have to learn a new software platform to add neural search to your existing Elasticsearch or OpenSearch workflow—saving you valuable time and resources.

So now you can focus on what you do best—creating great applications.

### Easy Integration

You don't have to learn new APIs—just connect the plugin to your Elasticsearch or OpenSearch installation to add neural search and a vector database to your application. This saves valuable time and resources and enables rapid deployment and use.

### Flexible

GSI's neural search is offered in both on-premises and SaaS deployments.

We offer an on-premises solution if physical control of your vector database is critical. Just add our plugin and neural search hardware to your OpenSearch/Elasticsearch installation, and it turns your server into a vector database.

If you do not want to manage the infrastructure, our SaaS offering makes it easy to add neural search to your existing Elasticsearch or OpenSearch cloud deployment with a simple plugin.

The SaaS solution offers flexible usage-based hourly pricing based on the amount of APU resources needed to support your application.

*“GSI’s OpenSearch plugin made building a multilingual and multimodal search application super easy. If you’re already using Elasticsearch / OpenSearch, the plugin allows you to implement neural search without having to move to a completely new vector search database. This is a big boost to quickly launching neural search in production at scale. Also, it allowed us to achieve low latency similar to keyword search but in our vector search application, which is quite cool.”*

**Dmitry Kan – Host of Vector Podcast, and CEO & Co-founder, Muves Ltd**

## Metadata Filtering

The plugins support pre-filtering using standard metadata filters from Elasticsearch or OpenSearch. Filter on product metadata such as item description, item color, category, or brand for a tailored search that produces more relevant results.

## Hybrid Search

When users know exactly what they're searching for, keyword search is still a good option. Our plugins allow you to build hybrid search using native OpenSearch or Elasticsearch keyword search along with neural search to leverage the strengths of each and provide a complementary set of relevant results.

## Batch Query Search

The plugins support both single query and batch query searches. Batch query processes multiple queries in parallel.

For example, if a user has multiple interests, such as news, finance, etc., batch query can search each document against their different interests in parallel to serve recommendations with a lower per-query latency than single-queries.

## Conserve Server Resources

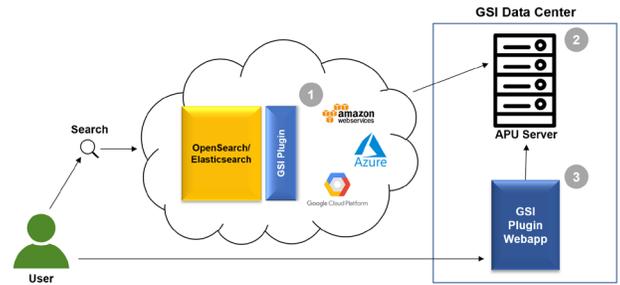
The plugins address CPU/memory scalability concerns by managing vector search and its associated memory allocation in the APU backend. That means the vector search workload doesn't burden your Elasticsearch/OpenSearch server resources, freeing them for other important tasks.

### Built Specifically for High-Performance, Scalable Search

Unlike other solutions that rely on general purpose CPUs, GSI's neural search solutions are based on in-memory computing technology built specifically for vector search at scale.

The result—you get a cost-effective solution that searches billions of items in milliseconds, leading to a better user experience and lower cost per query.

## GSI Neural Search SaaS System Architecture



**Figure 1: Architecture for GSI's SaaS-based Neural Search Option**

As seen in **Figure 1**, the key components for the SaaS plugin option are:

1. **GSI Plugin**—connects an Elasticsearch or OpenSearch index to the GSI APU (Associative Processing Unit) backend. It indexes vectors and accesses the APU hardware backend, which is in-memory computing technology that is built specifically for vector search at scale.
2. **GSI APU Server**—contains APUs to search billion-scale vector databases in milliseconds, with high recall.
3. **GSI OpenSearch Webapp**—a simple web user interface for index admin tasks like uploading indexed vectors to the APU backend along with their metadata for filtering.

**Ready to improve your customer's search experience with a high-performance, scalable neural search solution?**

Sign up for a free trial at [www.searchium.ai](http://www.searchium.ai).



Searchium.ai is a SaaS Platform owned by GSI Technology.