DOI: doi.org/10.47363/JAICC/ICMLAIDS2026/2026(5)1

International Conference on Machine Learning, Artificial Intelligence and Data Science (ICMLAIDS 2026)

Conference Proceedings March 20 - 21, 2026 - Virtual

Optimizing Vector Embedding Storage and Indexing for AI at Scale

Suvendu Mohantyz

Senior ML Engineer at Amazon, Virginia, USA

Abstract

As large-scale AI systems continue to evolve, the demand for efficient storage and retrieval of dense vector embeddings has become a critical challenge for both training and inference. Vector databases such as FAISS and Milvus enable high-performance similarity search, but the underlying infrastructure costs—spanning GPU utilization, memory bandwidth, and SSD/NVMe storage—are escalating rapidly. This talk explores emerging strategies for optimizing embedding storage and indexing to balance cost efficiency with low-latency retrieval, a key requirement for production-scale AI applications.

We begin by highlighting advances in quantization, product quantization (PQ), and hybrid compression techniques, which significantly reduce embedding footprint without degrading model accuracy. We also discuss adaptive precision storage, where embeddings dynamically shift between low-precision and high-precision formats depending on workload criticality. Beyond compression, indexing innovations such as hierarchical navigable small-world graphs (HNSW), disk-aware indexing, and tiered memory hierarchies are pushing the boundaries of scalability by leveraging GPU-accelerated search and SSD-based caching.