Project Name: A Distributed Key-Value Store for Retrieval-Augmented Generation (RAG) in the Cloud
Project proposer: Qi Chen (cheqi@microsoft.com)
Open Source: Yes
Mentors: (Qi Chen, cheqi@microsoft.com)

Preferred Experience
Required (for most or all team members) : C++ coding skill, basic knowledge about RAG
Required (at least one team member) : knowledge on distributed computing, storage and communication systems
Valuable: team work experience
Nice to have: familiar with vector retrieval algorithms

Project Background
Retrieval-Augmented Generation (RAG) has emerged as a key paradigm for large language model (LLM) applications, where model outputs are enhanced by retrieving relevant external knowledge from large-scale data stores. In practical deployments, RAG systems impose stringent requirements on the underlying key-value (KV) storage, including extremely high read throughput, low tail latency, elastic scalability, and strong availability in cloud environments.

Existing distributed KV stores are often designed for general-purpose workloads and may not be optimized for read-heavy, latency-sensitive RAG workloads, nor for emerging requirements such as near-data computation and dynamic value growth. This project aims to explore the design and implementation of a new distributed KV storage system optimized specifically for RAG applications in the cloud.

Project Description
The objective of this project is to design and prototype a read-optimized distributed KV store that:

- Prioritizes high read throughput and low end-to-end latency
- Supports efficient load balancing under skewed access patterns
- Provides fault tolerance and high availability in a cloud setting
- Enables near-data computing to reduce data movement
- Efficiently supports dynamic and incremental value growth, which is common in RAG pipelines (e.g., document chunk expansion, metadata enrichment)

Through this project, we aim to study the trade-offs involved in tailoring computation-storage co-design systems for modern AI-driven workloads.

Learning Outcomes
By the end of the project, we expect to deliver:

- A clear design and prototype of a RAG-oriented distributed KV store
- An experimental evaluation demonstrating strengths and limitations of the design
- Insights into storage system optimizations specifically tailored for AI and LLM applications in the cloud

This project provides hands-on experience with:

- Distributed storage system design
- Cloud scalability and fault tolerance
- Performance evaluation of large-scale systems
- Systems challenges arising from modern AI workloads