Edge LLM Distributed RAG

Multiple Edge LLM Synchronization of Distributed RAG

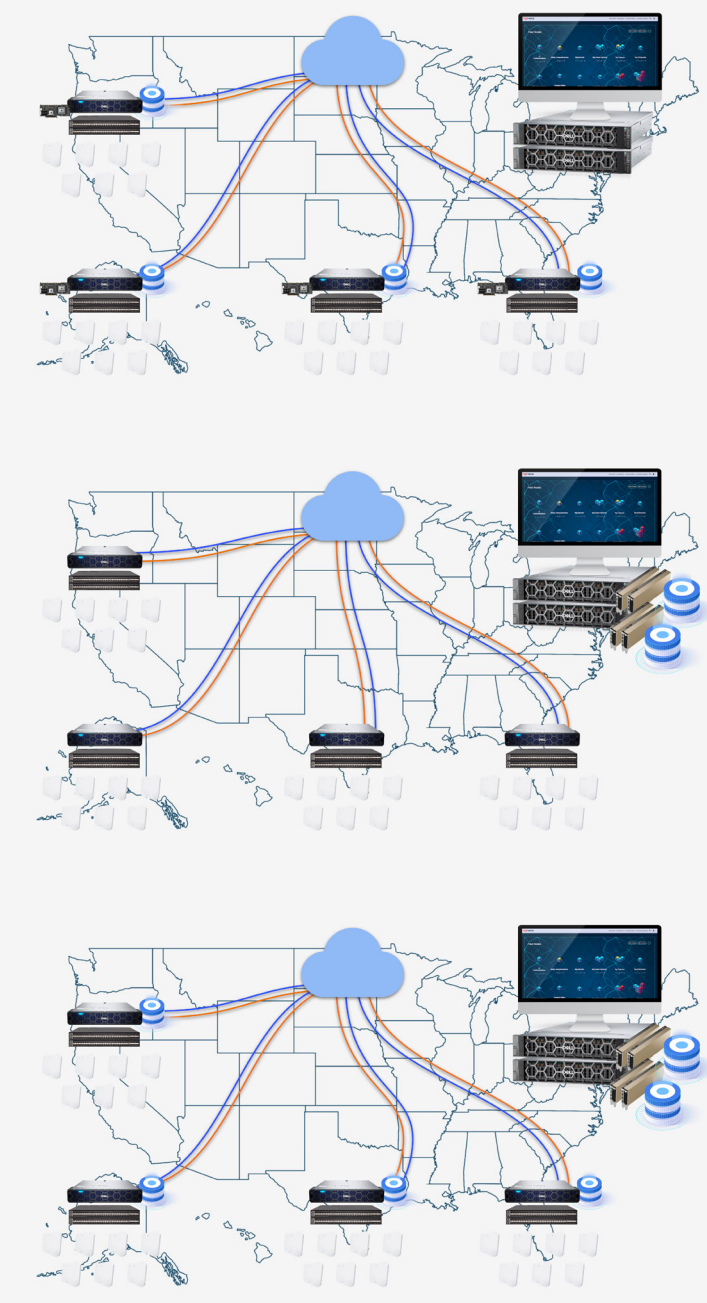
rgNets
The Edge. Redefined.

When we first experienced ChatGPT, we recognized it was not a question of “if” but “when” the world was about to change. At RG Nets, Inc., we took immediate action to address the challenges inherent in bringing about this new reality. Early on, we faced difficulties in using large language models (LLMs) in real-world scenarios. As soon as we attempted to provide proper context for our queries, we ran into problems that would spark the development of our distributed Retrieval Augmented Generation (RAG) system for LLMs.

The proliferation of large language models (LLMs) has revolutionized natural language processing (NLP) tasks, from text generation to machine translation, spurring our imaginations on and making the business world eager to harness their potential. However, the effectiveness of these models depends on their ability to access and leverage relevant information from vast datasets. Retraining and fine-tuning the models is extraordinarily expensive and time-consuming. RAG has emerged as the leading solution for dynamically augmenting responses with up-to-date information.

Most (RAG) approaches are extraordinarily effective at bridging the gap between fine-tuning models and the evolving datasets that a typical consumer would require as input context during prompt engineering. There is significant difference, however, in applying RAG in consumer applications and in enterprise environments.

Business datasets are proprietary, segmented, massive, and rapidly changing. It is very straightforward, albeit computationally and storage intensive, for an RAG system to index public datasets, like Wikipedia, for use as context, and to perform vector similarity search on this data. But this approach falls short when dealing with a proprietary business dataset.

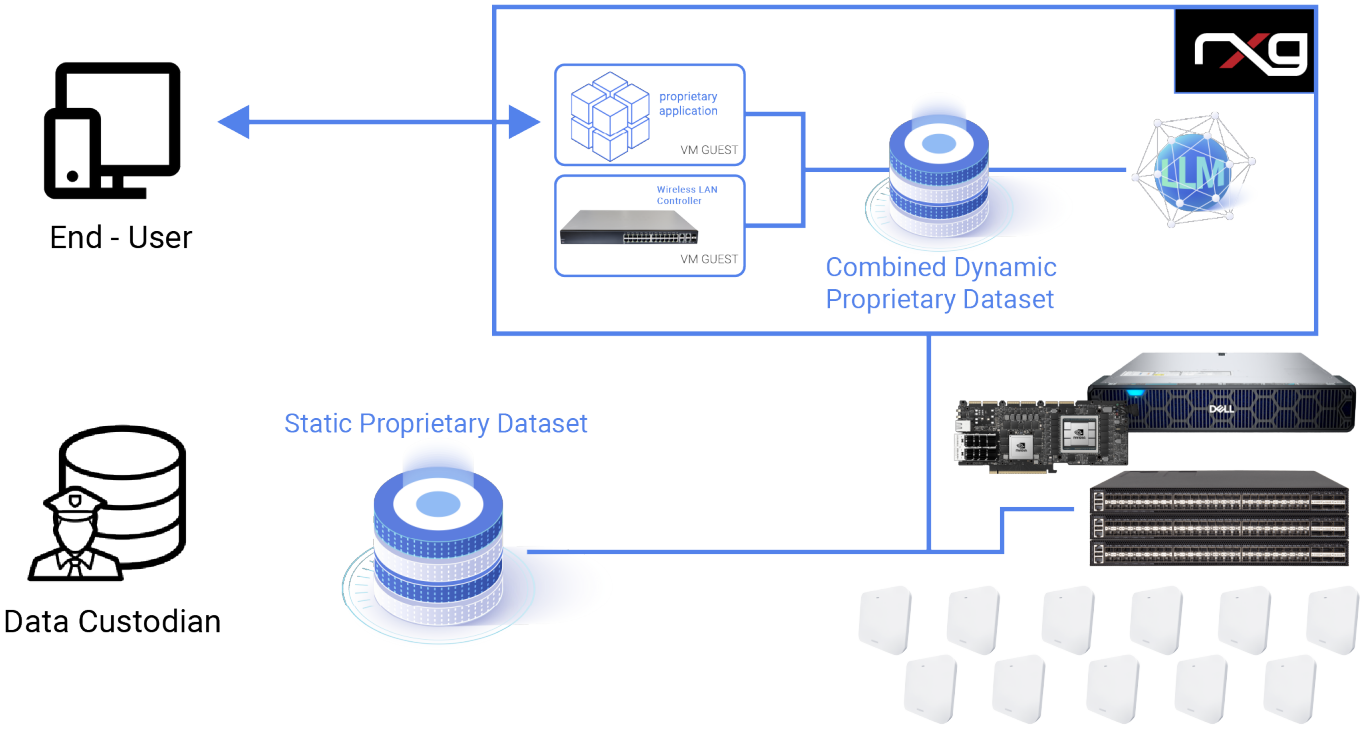


The complexities inherent in proprietary business datasets stem from several factors. Unlike public datasets, business datasets often include role-based access control (RBAC), limiting access to certain data to only specific users and teams. Furthermore, these datasets change frequently, and are often much larger, especially when largely populated by machine-to-machine communication.

The RG Nets, Inc., revenue eXtraction gateway (rXg) incorporates a distributed LLM architecture uniquely suited for overcoming the issues commonly associated with proprietary business datasets. The edge devices store the LLM sources and their related embeddings (indexes) used for vector similarity search, while The Fleet Manager synchronizes sources and propagates the security model. GPUs can be hosted on devices at the edge, on the Fleet Manager, or both, and operators may configure LLM workers to leverage whatever resources are available.

This decentralized dataset storage is particularly beneficial for large, dynamic business datasets, such as those generated by wireless network telemetry. Wireless controllers generate a constant stream of data, meaning that a regional aggregator could be presented with multiple terabytes of new data daily. The scale vastly eclipses even the most extensive public dataset. Managing these huge datasets while adhering to strict security and RBAC segmentation protocols is a daunting endeavor. Replicating these datasets to a centralized location is not only difficult and costly, but keeping them up to date poses additional challenges, making the task nearly impossible.

The RG Nets, Inc., rXg's LLM distributed RAG feature offers a groundbreaking solution for businesses grappling with the challenges posed by proprietary business datasets in LLM applications. By decentralizing the retrieval and generation processes, the rXg overcomes the limitations of traditional RAG approaches, and enables efficient, secure, and scalable access to vast, dynamic datasets, while significantly reducing costs, promoting efficiency and improving customer satisfaction. With RG Nets' distributed RAG solution, deploying, managing, and maintaining a distributed RAG system at scale has become incredibly easy. Contact us today for a demo!



rgNets

The Edge. Redefined.

www.rgnets.com
sales@rgnets.com
316 CALIFORNIA AVE
RENO, NV 89509