FEDDATA

Generative Artificial Intelligence Solutions

Generative Artificial Intelligence (AI) comprises a set of AI models capable of generating new and original content, such as images, videos, music, or text. State-of-the-art generative AI models are based on deep neural networks designed with Generative Pre-trained Transformer (GPT); these mathematical constructs are the foundation for large language models (LLMs) as we know them today. LLM deep neural networks are trained on massive amounts of textual data, effectively developing "knowledge and comprehension" of human language. In the era of generative AI, LLMs have demonstrated intrinsic abilities to not just interact with the user using natural language, but also in problem solving, coding, pattern recognition, and other. Generative AI has the potential to revolutionize many industries, such as defense, intelligence, medical, financial, sports, incident response, education, and many others, by enabling the creation of new and innovative content that was previously very time consuming to produce.

FedData's On-Premises Generative Pre-trained Transformer AI Product Powered by Dell Technologies

OnPremGPT

A SCALABLE AND PRODUCTIZED RETRIEVAL AUGMENTED GENERATION SYSTEM FOR KNOWLEDGE EXPLORATION

The FedData AI Division has developed a proprietary Retrieval Augmented Generation (RAG) AI system with multi-modal capabilities that combines an information retrieval system with a powerful and efficient LLM. With a RAG it is possible to analyze and explore vast amounts of data, and extract actionable information, or solve problems. Our proprietary multi-user AI architecture enables concurrent serving of hundreds of users per compute node.

Examples of RAG use cases:

- Intelligent AI assistant for quick-response technical support
- Analysis of logfiles and ticketing systems
- Discover patterns in transactional data
- Instant translation from foreign languages, summarization, and context awareness
- Translate and transcribe audio conversations and use the RAG to find pertinent information in the conversations

ONPREMGPT* FEATURES:

- Enterprise-level leading edge RAG AI designed for multi-user environments, which can be scaled to thousands of users
- With on OnPrem computing, the customer has total control over the server hardware, the data content and configuration, and security, with no required access to third-party resources (i.e., LLM providers)
- Graphics Processing Unit (GPU) accelerated AI inferencing engine with the ability to be hosted on multi-node/multi-GPU architectures to serve hundreds of users concurrently
- Supports proven LLMs with the capability of adding customized LLMs based on customer need
- Data retrieval augmented with access to reference sources and document images to provide an additional layer of confidence in the AI responses
- Reference data used by the AI in generating the answer is filtered using a relevance score calculated from the semantic search engine
- Prompt engineering and guard-railing instructs the AI to generate accurate answers on the specific context
- State-of-the-art text embedding models and vector store databases to accurately transform and store data for AI consumption
- Capability to fine-tune AI models based on the specific use case to enhance response accuracy
- AI data management infrastructure to manage and catalogue large volumes of data:
  - Capability to work on separate data collections with ad-hoc user access policies
  - Support for a variety of data formats (i.e., PDF, Microsoft Office, raw text, web scraping)
  - Ability to integrate custom connectors for other data sources
- Audio RAG capable of transcribing and translating foreign languages and make the audio conversations available for exploration

FEDDATA

OnPremGPT

DELL Technologies
FEDERAL TITANIUM PARTNER

SOLUTION HARDWARE

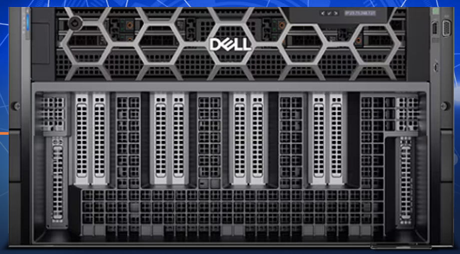
OnPremGPT* has been benchmarked and validated on industrial-grade architectures with GPU accelerators such as A100, L40S, and H100 on Dell and Nvidia platforms. In addition, our AI inferencing engine has been benchmarked on the latest hardware AI accelerators such as Intel Gaudi2. The system can be custom-sized based on customer requirements from small back-office deployments to large-scale enterprise solutions.

For very small deployments we also offer an Edge AI solution, providing a completely localized and on-premises RAG AI on laptop platforms; this solution is specifically tailored for denied areas or where network connection is weak or non-existent.

In our generative AI technical roadmap we have also considered the option to include enterprise storage systems and to be able to manage very large volumes of data sources. The storage solution will be engineered and sized appropriately based on the application (i.e., AI training or inference).

DEPLOYMENT SIZING

- **EDGE AI SCALE –** single user off-grid and on-premises AI on Alienware m18 laptop
- **SMALL SCALE –** 100s of users on single Dell or Nvidia node with at least 4X L40S GPUs
- **MEDIUM SCALE –** up to 500 users on 2-4 Dell or Nvidia node with at least 8X L40S or H100 GPUs
- **LARGE DATA CENTER –** greater than 1,000 users on multiple nodes, system will be custom sized



XE9680: Uncompromising 8 GPU performance for AI inferencing and training large language models:

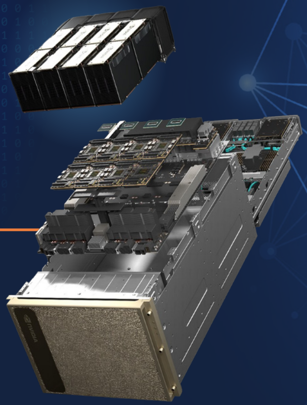
- Either 8 NVIDIA H100 SXM5, 8 NVIDIA A100 SXM4 or 8 AMD MI300X OAM GPUs
- Full interconnectivity between GPUs with NVIDIA NVLink or AMD Infinity Fabric
- Up to 10 front-facing full-height PCIe Gen 5 for expansion and AI network fabric connectivity

R760xa: Boost acceleration performance across the widest range of compute-intensive workflows:

- Optimized airflow for up to 4 double-width Gen5 PCIe or 12 single-width PCIe GPUs
- Broad array of NVIDIA, AMD, or Intel acceleration options
- Dual Intel Xeon Scalable CPUs with up to 56 cores and on-chip AI acceleration

Alienware Edge AI Laptop :

- Alienware m18 AMD
- 64 GB DDR5 RAM
- NVIDIA GeForce RTX 4090 – 16 GB GDDR5
- 8 TB RAID0 M.2 PCIe NVMe



NVIDIA DGX H100:

- 8x NVIDIA H100 GPUs With 640 Gigabytes of Total GPU Memory
- 4x NVIDIA NVSwitches™
- 10x NVIDIA ConnectX®-7 400 Gigabits-Per-Second Network Interface
- Dual Intel Xeon Platinum 8480C processors, 112 cores total, and 2 TB System Memory
- 30 Terabytes NVMe SSD

FEDDATA

FD002-10.30.24 FEDDATA.COM

FedData Technology Solutions
443.294.8290
sales@Feddata.com

FedData - MD Headquarters
9045 Junction Drive
Annapolis Junction, MD 20701

FedData - Florida Office
5338 W Crenshaw Street
Tampa, FL 33634