FEDDATA

DELL Technologies
FEDERAL TITANIUM PARTNER

OnPremGPT

Actionable Intelligence at the Speed of Thought...

With OnPremGPT, it is possible to literally "talk to the data" and to very quickly and effectively extract answers and solutions to complex problems.

- A Fully Integrated Solution for Multi-Int Data Ingestion and Interrogation using Generative AI
- An OnPrem AI tool in the arsenal for the next generation warfighter
- Let the AI draw connections, evaluate options, and guide you through critical aspects of a mission with unprecedented speed, accuracy, and efficiency

Benefits of OnPremGPT

- The customer has total control over the hardware infrastructure, data content and localization, configuration, and security, with no required access to third-party resources (e.g., Large Language Model [LLM] providers or third party services requiring an API key).

FEDDATA's OnPrem-GPT Features

- Enterprise-level RAG AI for multi-user environments, on-prem solution with no external resources needed, supporting thousands of concurrent users on multiple nodes
- Proven and customizable LLMs (e.g., Mistral, Mixtral), with a bring-your-own LLM option available
- Data retrieval augmented with reference sources and document images, with AI responses filtered using relevance scores from semantic search
- Advanced text embedding models and vector store databases with fine-tuning capabilities for specific use cases, and AI data management for large volumes of data
- Separate data collections with ad-hoc user access policies, supporting various data formats (PDF, Microsoft Suite, raw text, web scraping)
- Integration capabilities with other data sources and prompt engineering and guard-railing for accurate context-specific answers

OnPrem AI Planning & Analytics

- Integrates with diverse data sources and destinations
- Supports fine-grained access control and data encryption
- Real-time data ingestion, transformation, and processing
- Scales horizontally to handle large data volumes
- Language translation (i.e., Korean and Japanese)

Data Source

UNSTRUCTURED

- Multimedia (Audio/Video)
- Sensor Data
- Communications
- Word / PowerPoint

SEMI-STRUCTURED

- XLM / JSON
- NoSQL
- Email
- Reports

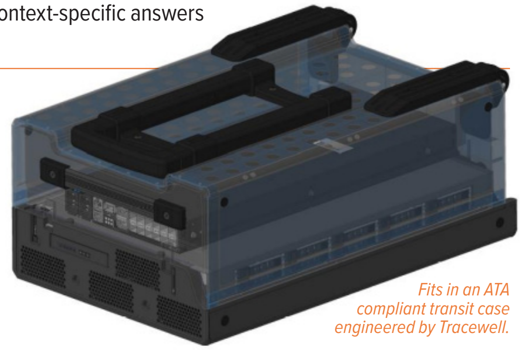
STRUCTURED

- Relational Databases
- Proprietary Databases
- Spreadsheets, emails, documents, reports, presentations

OnPremGPT

TRACEWELL & DELL 6615 BOX
Form Factor Highlights

Ground-breaking compute capability (64 cores and 2TB of RAM) in a highly compact platform footprint, making it the ideal platform for high-performance computing at the edge. Available in an Air Transportation Association (ATA) carry-on compliant case. Can be scaled to accommodate additional components and technology.



Fits in an ATA compliant transit case engineered by Tracewell.

Maximum Configuration

- Processors: 64 cores
- Memory: 2TB
- Storage: 10, 2.5" drives (8 SATA/SAS or 10 NVMe)
- PCIe Slots: 2 (capable of holding Nvidia L4 GPU or equivalent form factor GPUs)
- SYSTEM: 18"D, 13" W, 3.47"H

Note: AMD EPYCTM processor

Performance Highlights

- Support for 20 concurrent users with latencies under 10 seconds for large context queries (average of 20 tokens)
- Throughput of 300 queries per minute (QPMs) for 10 concurrent users
- Average ingest time for PDF documents: 170 pages/minute

Solution Hardware

- We have Subject Matter Experts that can assist with full spectrum hardware & software solutions; for tactical & strategic solutions.
- Benchmarking and validation conducted on industrial-grade architectures with Nvidia GPU accelerators (H100 and L40S)
- Capable of supporting customer requirements from small Tactical-Edge deployments to large-scale enterprises

Tested on x86_64 architectures such as:
Dell PowerEdge XE9680
Dell PowerEdge R760xa

FEDDATA
443.294.8290 | alliances@feddata.com