

POC Minimum GPT System Requirements

Minimum System Specifications for On-Premises Proof of Concept Deployments

Overview

This document outlines the basic small-size hardware requirements (typically supporting around 150 users) and other relevant system requirements for all sizes. Please refer to the Hardware Sizing guide to determine the size needed for your deployment.

Hardware Requirements

Component	Specification
GPU	2 GPUs required: <ul style="list-style-type: none">• 1 x 12 GB VRAM or larger• 1 x 24 GB VRAM or larger CUDA 12.x-compatible GPU Compute Capability \geq 7.0 (Volta / Ampere / Ada / Hopper) Unsupported: Maxwell (SM 5.x) or Pascal (SM 6.x) GPUs Note: Additional GPU VRAM may be required for data analysis tasks.
CPU	12 cores total
Memory	48 GB RAM total
Storage	256 GB SSD

Server and Software Requirements

Two virtual or physical servers are required — one Windows server and one Linux server to host the Docker containers. The Docker containers may be collocated on an existing Linux Docker host.

Server	Specification
Windows VM	OS: Windows Server 2019 or newer, including IIS Web Server features configured by installer. Allocated Resources: <ul style="list-style-type: none">• 8 GB RAM• 4 CPU cores Services Hosted: <ul style="list-style-type: none">• Dashboard• Ingestor• API
Linux Docker Host	OS: Ubuntu Server 22.04 LTS or newer. Other distributions supported.

Server	Specification
	<p>Allocated Resources:</p> <ul style="list-style-type: none"> • Uses the remaining hardware resources from the total. <p>Containers Deployed:</p> <ul style="list-style-type: none"> • Gateway • Embedder • LLM • MS SQL Database • Postgres Vector Database <p>Additional Software Dependencies:</p> <ul style="list-style-type: none"> • CUDA 12.8 • Docker CE version 27.3.1 or later

Network

Source	Destination	Ports	Notes
Windows Dashboard / Ingestor VM	Linux Docker Host	8080, 443, 1433	Gateway API + MS SQL DB
File Uploading Agent installation location	Windows Dashboard / Ingestor VM	8080, 443	Required if using file uploading agent
End user workstations	Windows Dashboard / Ingestor VM	443	—

Internet Access

Source	Hosts	Notes
Docker Container Host	*.docker.io, *.docker.com, *.auth0.com production.cloudflare.docker.com docker-images-prod.6aa30f8b08e16409b46e0173d6de2f56.r2.cloudflarestorage.com	Authenticating to Docker Hub and downloading container images.
Docker Container Host	*.hf.co, huggingface.co, cdn-lfs.huggingface.co	Downloading LLM models and relevant files.
Installer workstation	customers.agatsoftware.com downloads.agatsoftware.com	Download installer files and upload logs for AGAT review if required.

Permissions

SQL Server — Data Analysis / Ingestion

A user with full read-only permissions to the existing company database to be used as the source for data analysis or ingestion.

Windows Dashboard / Ingestor VM

Should be domain joined to enable user authentication. Alternatively, Azure SSO may be used.

File Uploading Agent

Needs to run from a location where it has access to files to be uploaded (locally or via network share) and run as Local System user.

SharePoint

Service user needs to be Site Collection Admin as well as have read-only access to all relevant sites.

Miscellaneous

- Customer may provide an SSL certificate to secure HTTPS access by end users to the dashboard. The certificate will be installed on the Dashboard VM certificate store and accessed by IIS.
- Remote installation and support sessions using Microsoft Teams or Webex with remote screen control.

For technical enquiries and deployment support, visit www.defenix.ai