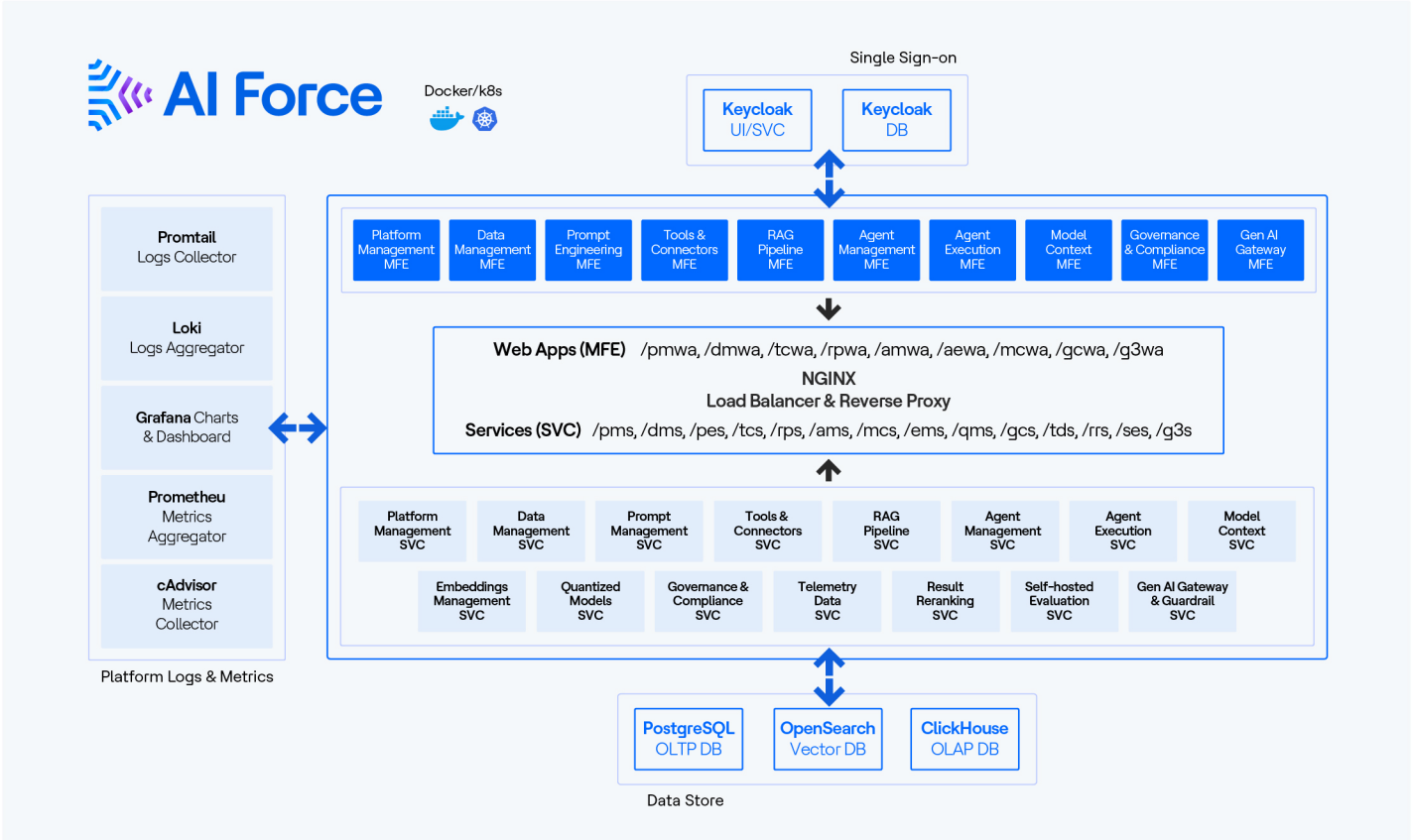
HCLTech | Supercharging Progress™

AI Force

Unlocking enterprise efficiency and productivity through GenAI and Agentic AI-led service transformation

AI Force 2.0 is a GenAI and Agentic AI platform that automates and augments workflows across software and data engineering, IT operations and enterprise business processes to significantly improve and accelerate business outcomes.

AI Force 2.0 technical architecture



The HCLTech AI Force 2.0 platform has become more modular, adopting a microservices architecture. It has 25 native AI Force containers (15 Micro Services and 10 Micro Front ends) and 11 external prebuilt containers for database, authentication and platform observability.

These web applications are the primary components of the AI Force platform that users interact with. Each application has a specific job, such as managing data, building prompts, or checking system safety. Together, they make it easy for people to use and control the platform. Some of the Micro Frontend web applications and what it does are listed below:

Web application name	What it does
Platform Management	Control and set up the platform, users and projects
GenAI Gateway and Guardrail	Manage AI models and set safety rules
Data Management	Upload, organize and find files and data
Prompt Engineering	Create, test and improve AI instructions (prompts)
Tools and Connector	Add and use tools, connect with other systems
RAG Pipeline	Build and manage how information is found and used
Agent Execution	Run and control AI agents
Agent Management	Track rules, safety and compliance
Model Context Protocol	Manage how information is shared between AI models
Governance, Risk and Compliance	Manage enterprise policies and regulatory standards

The backend microservices are the hidden engines that power the platform. They handle important tasks such as managing data, running AI agents and ensuring everything runs smoothly and securely. These services operate in the background to ensure the platform is fast, secure and reliable. Below are the backend microservices:

Microservices	What it does
Platform Management Service (PMS)	– Exposes REST APIs for authentication, user management, project configuration and RBAC setup – Provides support for managing user roles, project access and activity audit logging
GenAI Gateway and Guardrail (G3S)	– Offers APIs for managing LLMs, embeddings and speech model interactions – Supports model-specific configuration and generation of chat completions and embeddings – Enforces AI safety with keyword monitoring and access policy-based guardrails
Data Management Service (DMS)	– Deliver APIs for data ingestion, indexing and file management operations – Supports workflows including file upload, download, deletion and indexing status retrieval
Prompt Engineering Service (PES)	– Introduces APIs to manage prompt creation, evaluation, updates and trace log retrieval – Enables evaluation of prompt effectiveness and version control for prompt assets
Tools and Connector Service (TCS)	– Enables APIs to manage custom tools and connectors in the platform's Tools Library – Supports publishing, updating, executing tools, running scripts and generating API function wrappers
RAG Pipeline Service (RPS)	– Provides APIs to create, update, duplicate and manage RAG pipelines – Supports pipeline testing, OpenSearch indexing, vector storage and knowledge graph querying – Includes metadata retrieval for dropdowns and access to detailed project configurations
Embedding Model	– Provides REST APIs for generating embedding from user inputs – Enables model selection, download and metadata retrieval for embedding operations
Telemetry Data Service	– Implements APIs to collect and query telemetry and performance metrics – Supports batch metric storage, health checks and data querying for analytics

Microservices	What it does
Model Context Protocol Service	– Enables APIs to manage offline quantified LLM models – Supports downloading of models, listing available models, generating completions locally and checking service health
Agent Management Service (AMS)	– Enables APIs to create, configure and manage AI agents Supports agent lifecycle operations, including create, update, delete, list and configuration dropdowns
Agent Execution Service (AES)	– Enables APIs to execute AI agents
Re-ranking Services	– API provides endpoints for retrieving and reranking documents based on user queries. Key functionalities include retrieving relevant documents from an index using various retrieval techniques (keyword, vector, hybrid) and reranking documents based on similarity scores
Governance Compliance Services	– Ensures organizations operate ethically, manage risks and meet regulatory requirements

AI Force 2.0: Turning ideas into intellectual assets

AI Force 2.0 has made significant strides in advancing the frontiers of AI and generative AI, with a robust portfolio of over 35 patents that reflect deep innovation and strategic foresight. These patents span both traditional AI, automation and generative AI, showcasing the team's ability to address complex enterprise challenges through intelligent, scalable solutions. A total of 16 patents has already been granted, including recognitions from both the Indian Patent Office (IPO) and the United States Patent and Trademark Office (USPTO), underscoring the global impact and credibility of HCLTech's intellectual property. The remaining patents are in various stages of filing, with seven focused on cutting-edge GenAI technologies.

The patents cover areas like:

- How can a system understand requirements and generate them
- Optimization of the resource allocation plan during software product sustenance
- Process to generate test scripts
- System-driven output being contextual, etc.

Technical requirements

AI Force can be deployed in a containerized environment with

Docker based deployment
Available for Linux and macOS environments.

Kubernetes based Deployment Suitable for cloud native environments

The following are the hardware requirements for docker based deployment

- **RAM**: Minimum 64 GB (depending on the number of users)
- **Disk**: Minimum 512 GB HDD **Architecture**: X64 or x86-64 CPU

Supported OS

- Linux: Ubuntu 24.04 LTS (recommended)
- Windows: Windows 11 and Docker Desktop (check for compatibility)

Core connectivity requirement

→ Firewall ports:

- 443 (HTTPS): Mandatory for secure user access via Nginx Load Balancer
- 80 (HTTP): Redirects to HTTPS

→ User access:

- Root access to the virtual machine or physical machine

→ SSL/TLS certificates:

- Trusted CA certs mounted in Nginx container for secure access
- Container registry access:o Access to DockerHub (docker.io/hcltechaiforce) for pulling AI Force images

The following are the cluster requirements for Kubernetes based deployment:

- **Kubernetes version:** v1.28 or later
- **Nodes:** Minimum of 8 worker nodes
  RAM per node: 64 GB (minimum)
  CPU per node: 16 vCPUs (minimum)
- **Persistent volume:** Minimum 1 TB
- **Local node storage:** Minimum 2 TB (approximately 100 GB per VM)
- **Architecture:** X64 or x86-64 CPU

Supported OS

- Linux: Ubuntu 24.04 LTS (recommended)

Persitent volume

core connectivity requirements

→ Firewall ports:

- 443 (HTTPS): Mandatory for secure user access with ingress/nginx
- 0 (HTTP): Redirects to HTTPS

→ SSL/TLS certificates:

- Trusted CA certs mounted for secure access.

→ Container registry access:

- Access to DockerHub (docker.io/hcltechaiforce), including access to the Helm chart stored in the DockerHub repository

Load and scalability benchmarks

The objective of load testing was to evaluate system performance under both normal and peak concurrent user conditions.

API	Data	Results (60 sec ramp-up)
Login / Authentication (Keycloak)	NA	480 concurrent users → 11 sec average response time
Save Prompt	2K Prompts related to various tasks (Code Gen in various programming languages - .py, .cpp, .java, .c and Q&A)	3000 concurrent users → 0.13 sec average response time
Test Prompt		
Save RAG	Various file formats in different sizes (100 MB .txt, 53 pages.docx, 53 pages .pdf)	1400 concurrent users → 21 sec average response time
Test RAG		2100 concurrent users → 4.3 sec average response time
		200 concurrent users → 2.8 min average response time

UI performance testing aimed to ensure smooth and responsive user interactions, with a target page load time of under three seconds. Measurements across various screens confirmed that most pages met this expectation. For example, the sign-in screen loaded in 1.68 seconds and the home screen in 1.44 seconds. The objective of volume testing was to assess the system's ability to handle large data volumes across different modules. The testing was designed to validate whether the application could support file uploads and interactions involving code, documents and artifacts of varying sizes.

High volume	Medium volume
Modules - Data upload, RAG	Modules - Prompts, Agents, BYOU
Handles up to 100MB data	Handles up to 10MB data

AI Force 2.0 connectors

AI Force makes it easy to connect with both Large Language Models (LLMs) and Small Language Models (SLMs) from different providers. These connectors enable you to select the optimal AI model for your needs—whether you require powerful, general-purpose models or smaller, faster ones tailored to specific tasks. This flexibility helps you get the right results while saving time and costs.

Commercial LLMs supported are:
- Azure Open AI
- Google Gemini
- Anthropic Claude on AWS Bedrock
- IBM Granite
- NVIDIA Nemotron

Open source LLMs/SLMs supported are:
- Phi 3
- Meta Llama
- Mistral AI

Prebuilt tool connectors are:
- JIRA
- GitHub
- Confluence

HCLTech | Supercharging Progress™

hcltech.com