JioCloud

AI Workbench

Overview

JioCloud AI Workbench is a GPU-backed compute environment on CloudXP, purpose-built for AI development and accelerated workloads. It delivers a fully configured virtual machine with your choice of GPU hardware - NVIDIA H200 NVL, NVIDIA H200 SXM, or AMD MI300X - with pre-installed GPU drivers and an up-to-date CUDA or ROCm stack. Teams can run model training, deep learning inference, RAG pipelines, and data-intensive experiments from day one, without configuring the underlying infrastructure. Select AI tools at provisioning time and receive a fully operational environment - JupyterLab, Ollama, MLflow, LangFlow, ChromaDB, and more - running the moment your workbench activates.

Key Features

- **Three GPU layouts - choose your compute power**
  Choose from NVIDIA H200 NVL, NVIDIA H200 SXM, or AMD MI300X GPU layouts, fully configured and AI-ready.
- **18+ pre-configured AI tools selectable at provisioning time**
  Provision 18+ AI tools-including Ollama, vLLM, JupyterLab, MLflow, LangFlow, ChromaDB, Weaviate, pgvector, Grafana, and Prometheus-pre-installed and ready to use.
- **Pre-validated software bundles**
  Deploy the complete Generative AI Stack as a pre-validated, dependency-free software bundle.
- **Zero manual setup - ready from day one**
  Start instantly with GPU drivers, CUDA/ROCm, and selected AI tools pre-installed, configured, and ready to use.
- **Real-time milestone tracking**
  Track every provisioning stage with live status updates and detailed error visibility.
- **Multi-instance provisioning for entire teams**
  Provision identical AI workbenches for entire teams in a single request with no environment drift.
- **CloudXP Marketplace guided wizard**
  Deploy AI workbenches through a guided workflow with no infrastructure expertise required.

JioCloud

Benefits

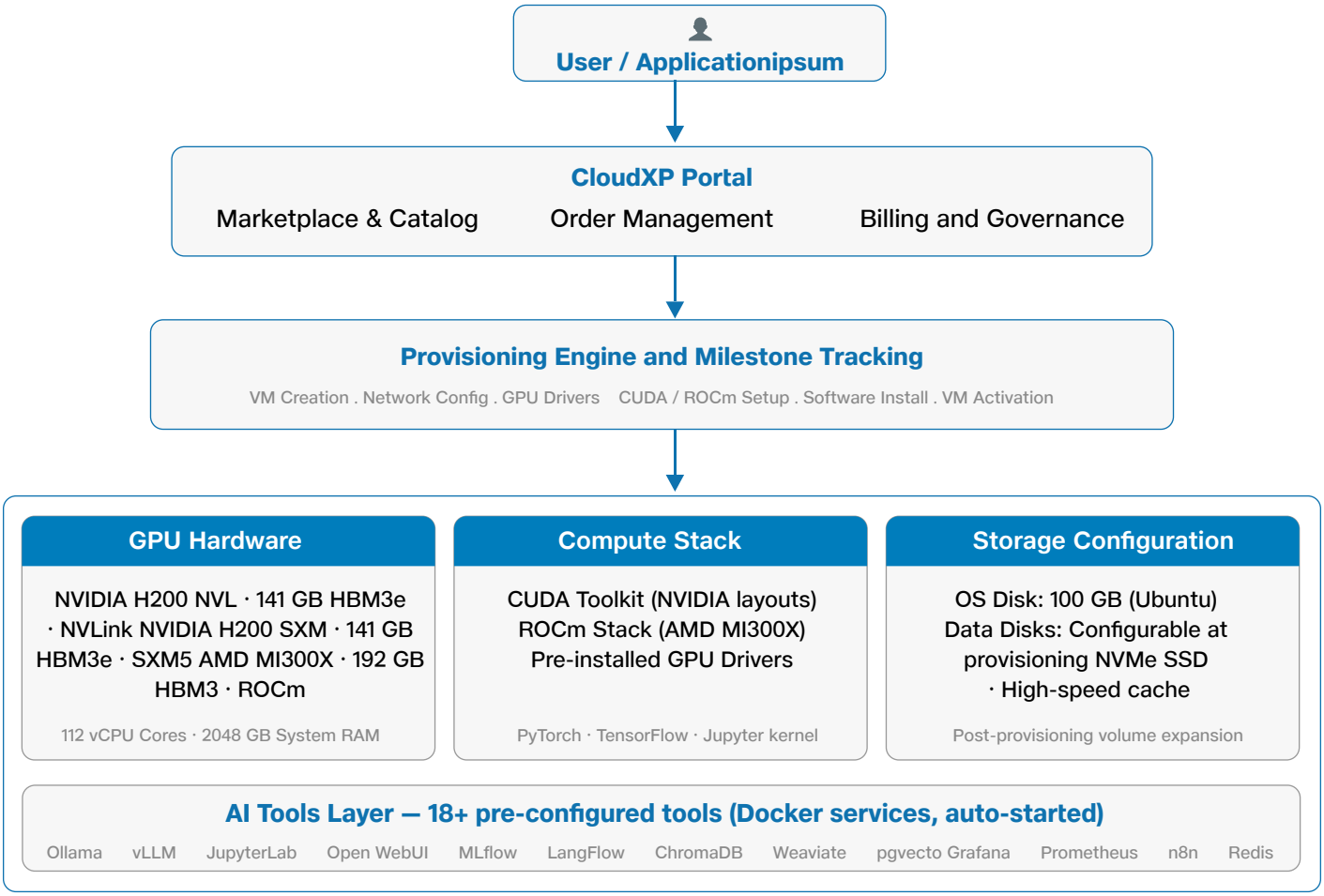
- **Ready in under 30 minutes**
  From order to running AI tools - Jupyter-Lab, Ollama, MLflow, and more - in under 30 minutes, with zero manual configuration after provisioning.
- **Reproducible team environments**
  Provision multiple identical instances in a single order. Every researcher gets the same tools, versions, and GPU configuration - eliminating onboarding delays.
- **Complete GenAI stack in one selection**
  The Generative AI Stack bundle deploys 8 validated tools - Ollama, Open WebUI, JupyterLab, ChromaDB, MLflow, LangFlow, Grafana, and Prometheus - in one click.
- **Faster time-to-experiment**
  Pre-validated bundles eliminate CUDA version hunting and dependency conflicts - your team starts building on day one, not day three.
- **GPU flexibility for any workload**
  Three hardware options spanning 141 GB to 192 GB GPU memory - NVLink for large multi-GPU inference, SXM5 for intensive training, HBM3 for memory-bound models.
- **Built-in governance and cost control**
  Every workbench is assigned to a hierarchy, project, and subscription at provisioning time - with billing allocation, quota boundaries, and full organisational visibility.

Technical Specifications

Category	Specifications
GPU Options	NVIDIA H200 NVL, NVIDIA H200 SXM, AMD MI300X
GPU Memory	1141 GB HBM3e (NVIDIA H200 NVL & SXM) / 192 GB HBM3 (AMD MI300X)
Compute Stack	CUDA Toolkit - NVIDIA layouts | ROCm - AMD MI300X (pre-installed)
AI Tools Available	18+ tools - LLM inference, notebooks, experiment tracking, vector DBs, monitoring, orchestration
Software Bundles	Generative AI Stack: 8 pre-validated tools in a single provisioning selection
Provisioning Time	Under 30 minutes (VM + GPU drivers + full tool stack)
Supported OS	Ubuntu
Access Method	SSH + browser-based tool UIs via port-mapped Docker services
Storage	Configurable OS disk + attachable data disks; post-provisioning expansion supported
Deployment	CloudXP Marketplace - single guided provisioning wizard

JioCloud

Architecture Diagram



Use Cases

- **Generative AI application development - end-to-end RAG pipelines**
  Build production-ready retrieval-augmented generation applications with Ollama for LLM inference, ChromaDB or Weaviate for vector storage, LangFlow for pipeline orchestration, and JupyterLab for development - all pre-installed and connected in one provisioned environment, ready in under 30 minutes.

- **LLM fine-tuning and model training at scale**
  Run full training and fine-tuning jobs on 141–192 GB GPU memory with PyTorch and TensorFlow-Jupyter pre-configured. MLflow automatically tracks every experiment run - metrics, parameters, and artefacts - with no additional setup required across H200 NVL, H200 SXM, or AMD MI300X layouts.

- **Multi-team AI research and experimentation environments**
  Provision 10, 20, or 50 identical workbenches in a single order. Every researcher receives the same tools, versions, and GPU configuration - eliminating onboarding delays, environment inconsistencies, and the "works on my machine" problem across your entire AI development organisation.