Server Prerequisite

1. Hardware Requirements:
   - Bare metal access to an HPE ProLiant XL645d Gen10 server
   - At least 1x or more NVIDIA A100 80GB SXM4 GPUs with MIG support
   - At least 512 GB RAM (recommended for multi-GPU ML workloads)
   - 1TB+ primary SSD/NVMe for OS and containers

2. BIOS & Firmware:
   - UEFI Boot Mode
   - Virtualization Enabled (VT-x/AMD-V)
   - Resizable BAR support (if required by GPUs)
   - GPU ECC mode enabled
   - Update to latest BIOS & BMC firmware

3. Network and Access:
   - **Reliable internet access for downloading packages, drivers, containers**
   - **SSH access enabled with sudo/root permission**
   - Firewall rules allow:
     - Ports 22 (SSH), 443 (HTTPS), 80 (HTTP)
     - Docker or app-specific ports (e.g., 6006 for TensorBoard)

4. Software & System Requirements:
   - Bootable media or ISO for Ubuntu Server 22.04 LTS
   - Basic tools post-install:
     - curl, wget, build-essential, git, vim, htop
     -python3, python3-pip (in case of pre-Anaconda use)

5. Credentials & Configs:
   - Sudo-enabled user for all commands
   - Access to Docker Hub (or a private registry if used)
   - Proxy info (if behind a corporate firewall)
   - SSH keys for Git or remote deployment (if applicable)

Scope of Work (Included)

1. Containerization:
    - Installation and configuration of Docker
    - Creating custom Dockerfiles for CUDA + ML/AI stack
    - Building images with compatible versions of:
        - PyTorch + CUDA
        - TensorFlow + CUDA
        - JAX, Transformers, OpenCV, etc.
    - Running containers with GPU/MIG access

2. Local Python Environment (Anaconda):
    - Installing Anaconda on the host
    - Setting up conda environments with:
        - CV2 (OpenCV)
        - PyTorch
        - TensorFlow
        - Transformers, JAX, DL4J, XGBoost, Theano, etc.
    - CUDA toolkit and cuDNN setup in conda for GPU support

3. System Services and Auto-Start:
    - Creating startup scripts for auto-MIG allocation
    - Service enablement for persistent GPU configs

4. Monitoring and Visualization:
    - Installing DCGM Exporter for GPU telemetry

5. Deep Learning Tools:
    - Host or container-based setup for
        - DeepStream
        - Deep Cognition Studio (if required via external binary)
        - W&B client
        - TorchVision, NLTK, CHROMA

Out of Scope (Not Included)

- Kubernetes setup
- Model training/ML experiments
- Long-term storage/network design
- Enterprise license management (e.g., for DeepStream or TensorRT Pro features)
- Hardware procurement or replacement
- Custom web UI development
- Remote user onboarding/training
- Production CI/CD pipelines for model deployment